SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 1, pp. 1–6, January 1991

LARGEST SINGULAR VALUE SUBMULTIPLICATIVITY*

CHARLES R. JOHNSON† AND PETER NYLEN‡

Abstract. A combinatorial technique by which several different products on matrices may each be represented as a conventional product of transformed matrices is described. When the matrix transformation does not increase the largest singular value, a submultiplicativity inequality for the product may be deduced. An example is given of a product that is submultiplicative, but for which there is no such representation in terms of the ordinary product. The Hadamard product on infinite matrices and a mapping defined on triples of matrices X, Y, $B \to (XY) \circ B$ are also considered.

Key words. Hadamard product, submultiplicativity inequality

AMS(MOS) subject classifications. 15A18, 15A42, 15A45, 15A60

1. Introduction. The primary purpose of this note is to exhibit a simple technique for deducing some submultiplicativity inequalities for nonstandard products (such as the Hadamard product) from the corresponding inequality for the usual product. We illustrate the technique with some examples, including a new inequality (spectral norm submultiplicativity for the "box" product). Although we do not attempt to determine all implications of the technique, it is likely that the idea will be useful in analogous settings.

Let $\mathbf{M_{n,m}}$ denote the set of all n-by-m complex matrices. By a *product* (on matrices) we mean a function

$$\bullet : \mathbf{M_{n,m}} \times \mathbf{M_{p,q}} \to \mathbf{M_{r,s}}, \tag{1.1}$$

whose value at A, $B \in \mathbf{M_{n,m}} \times \mathbf{M_{p,q}}$ is denoted by $A \cdot B$. Note that we are making no assumption of linearity, associativity, or any other properties usually associated with the term product. We consider several products that can be defined on matrices.

Given $A \in \mathbf{M_{n,p}}$ and $B \in \mathbf{M_{p,m}}$, the *usual product* of $A = [a_{ij}]$ and $B = [b_{ij}]$ is denoted by AB.

Let $A \in \mathbf{M_{n,m}}$ and $B \in \mathbf{M_{p,q}}$. The *Kronecker product* of $A = [a_{ij}]$ and $B = [b_{ij}]$, denoted by $A \otimes B$, is the member of $\mathbf{M_{np,mq}}$ defined blockwise by $[a_{ij}B]$.

Let $A = [a_{ij}]$ and $B = [b_{ij}] \in \mathbf{M_{m,n}}$. The *Hadamard product* of A and B is the matrix $[a_{ij}b_{ij}]$, which we denote by $A \circ B$.

Let n, m, p, q and r be positive integers. Let $A \in \mathbf{M_{np,mq}}$ and $B \in \mathbf{M_{nq,mr}}$. Partition A into an n-by-m matrix of blocks whose i,jth block is a p-by-q matrix, $A = [A_{ij}]$ in block notation. Partition B as an n-by-m block matrix whose i,jth block is q-by-r, so that $B = [B_{ij}]$. The *box product* of A and B is defined by $A \blacklozenge B = [A_{ij}B_{ij}]$. When $n = m = 1$, the box product reduces to the usual product; when $p = q = r = 1$, the box product becomes the Hadamard product. The box product has also been considered in [HMN].

2. The largest singular value. The largest singular value of a matrix A is defined to be the nonnegative square root of the largest eigenvalue of the matrix A^*A, which we

* Received by the editors August 15, 1988; accepted for publication (in revised form) October 23, 1989. The work of both authors was supported in part by a North Atlantic Treaty Organization (NATO) travel grant.

† Department of Mathematics, College of William and Mary, Williamsburg, Virginia 23185 (#CRJOH2@WMMVS.BITNET). The work of this author was supported in part by National Science Foundation grant DMS 87 13762 and by Office of Naval Research contract N00014-87-K-0661.

‡ Department of Mathematics ACA, Auburn University, Auburn, Alabama 36849 (PMNYLEN@AUDUCVAX.BITNET). The work of this author was supported in part by Office of Naval Research contract N00014-87-K-0012.

denote by $\sigma_1(A)$. Of course, $\sigma_1(\cdot)$ restricted to the set of matrices of a particular size is the spectral norm.

We list here several well-known properties of σ_1 that we will need.

Permutation Invariance. Let $A \in \mathbf{M}_{n,m}$. Let P and Q be n-by-n and m-by-m permutation matrices, respectively. Then $\sigma_1(PAQ) = \sigma_1(A)$.

Direct Sum. Let $A \in \mathbf{M}_{n,m}$ and $B \in \mathbf{M}_{p,q}$. Then

$$\sigma_1(A \oplus B) = \sigma_1\left(\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}\right) = \max\{\sigma_1(A), \sigma_1(B)\}.$$

Submatrix. Let $A \in \mathbf{M}_{n,m}$ and let B be a matrix obtained from A by deleting some rows and/or columns. Then $\sigma_1(B) \leqq \sigma_1(A)$.

Submultiplicativity. Let $A \in \mathbf{M}_{n,p}$ and $B \in \mathbf{M}_{p,m}$. Then $\sigma_1(AB) \leqq \sigma_1(A)\sigma_1(B)$.

3. Submultiplicativity. Let $\cdot$ be a product as defined in (1.1). We say that $\cdot$ is submultiplicative with respect to $\sigma_1(\cdot)$ if for all $A \in \mathbf{M}_{n,m}$ and all $B \in \mathbf{M}_{p,q}$,

$$\sigma_1(A \cdot B) \leqq \sigma_1(A)\sigma_1(B).$$

All four products we have described have this property. For the usual product, see [HJ1, p. 296]. For a treatment of inequalities involving the singular values of the Hadamard product, including this one, see [AHJ]. For the Kronecker product the inequality is an equality (see [HJ2, Chap. 4]). Submultiplicativity of the box product has also been independently discovered by others [HMN]. Our purpose here is simply to illustrate a combinatorial embedding technique that facilitates a unified proof of submultiplicativity for the latter three products (and others) based on the submultiplicativity of the usual product. We suspect this approach will be useful elsewhere.

OBSERVATION 3.1. Let $\cdot$ be a product on matrices. Suppose there exist mappings F and G such that for all A and B for which $A \cdot B$ is defined,

(a) $A \cdot B = F(A)G(B)$,

(b) $\sigma_1(F(A)) \leqq \sigma_1(A)$ and $\sigma_1(G(B)) \leqq \sigma_1(B)$.

Then $\cdot$ is submultiplicative with respect to σ_1, since

$$\sigma_1(A \cdot B) = \sigma_1(F(A)G(B)) \leqq \sigma_1(F(A))\sigma_1(G(B)) \leqq \sigma_1(A)\sigma_1(B).$$

Of course, this observation is equally valid for any nonnegative-valued function (in place of $\sigma_1(\cdot)$) provided it satisfies the submultiplicativity inequality with respect to the usual product and is defined on the different sizes of matrices appearing in (a) and (b). There are many candidates for product/function pairs that we might try.

Now we construct mappings F and G that satisfy the hypotheses of Observation 3.1 for the latter three products.

For the Kronecker product, let the sizes n-by-m and p-by-q be given. We consider the Kronecker product mapping $\mathbf{M}_{n,m} \times \mathbf{M}_{p,q}$ into $\mathbf{M}_{np,mq}$. We start with the formula [HJ2, Chap. 4]

$$A \otimes B = (A \otimes I_p)(I_m \otimes B),$$

where I_p denotes the p-by-p identity matrix and I_m the m-by-m identity matrix. Define the mappings F and G, respectively, by

$$F(A) = (A \otimes I_p) \quad \text{and} \quad G(B) = (I_m \otimes B).$$

By the direct sum property of σ_1,

$$\sigma_1(G(B)) = \sigma_1(B).$$

There exist permutation matrices P and Q such that

$$P(A \otimes I_p)Q = I_p \otimes A$$

(see [HJ2, Chap. 4]). Then

$$\sigma_1(F(A)) = \sigma_1(I_p \otimes A) = \sigma_1(A).$$

For the Hadamard product, let the size n-by-m be given. F will be a mapping from $\mathbf{M}_{n,m}$ into $\mathbf{M}_{n,nm}$ and G a mapping from $\mathbf{M}_{n,m}$ into $\mathbf{M}_{nm,m}$. For $x \in \mathbf{C}^n$, define $D(x)$ to be the n-by-n diagonal matrix with the entries of x placed in order on the main diagonal. Let A and $B \in \mathbf{M}_{n,m}$ be given. Denote the columns of A and B by $a_1, \cdots, a_m$ and $b_1, \cdots, b_m$, respectively. Define F by

$$F(A) = [D(a_1), D(a_2), \cdots, D(a_m)]$$

and define G by

$$G(B) = b_1 \oplus b_2 \oplus \cdots \oplus b_m.$$

The usual product of $F(A)$ and $G(B)$ is the Hadamard product $A \circ B$ since

$$\begin{aligned} F(A)G(B) &= [D(a_1)b_1, D(a_2)b_2, \cdots, D(a_m)b_m] \\ &= [a_1 \circ b_1, a_2 \circ b_2, \cdots, a_m \circ b_m] = A \circ B. \end{aligned}$$

By applying the direct sum and submatrix properties of σ_1, we have

$$\sigma_1(G(B)) \leqq \sigma_1(B).$$

There exists an nm-by-nm permutation matrix P such that

$$F(A)P = a_1 \oplus a_2 \oplus \cdots \oplus a_m.$$

Thus, similarly,

$$\sigma_1(F(A)) \leqq \sigma_1(A).$$

Submultiplicativity of $\sigma_1(\cdot)$ with respect to the Hadamard product was first noted in [S]. In this case our technique actually exhibits the stronger inequality

$$\sigma_1(A \circ B) \leqq r_1(A)c_1(B),$$

in which $r_1(A)$ denotes the largest row length of A and $c_1(B)$ the largest column length of B. This inequality was noted in [AHJ].

For the box product, let the integers n, m, p, q, and r in the definition of the box product be given. Let $A \in \mathbf{M}_{np,mr}$ and $B \in \mathbf{M}_{nr,mq}$. Utilize the same block partitioning for $A = [A_{ij}]$ and $B = [B_{ij}]$.

We define the mappings $F : \mathbf{M}_{np,mr} \to \mathbf{M}_{np,nmr}$ and $G : \mathbf{M}_{nr,mq} \to \mathbf{M}_{nmr,mq}$ by the same procedure as that used with the Hadamard product, except instead of placing entries of A and B in specified locations, place the p-by-r blocks of A in those locations occupied by entries of A and the r-by-q blocks of B in those locations occupied by entries of B.

An extension of the argument used in the Hadamard product case based on block multiplication of matrices gives

$$A \blacklozenge B = F(A)G(B).$$

Now, we obtain the bounds on σ_1. For $j \in \{1, \cdots, m\}$, let B_j denote the nr-by-q submatrix of B in the columns of B indexed by $(j-1)m+1$ through $(j-1)m+q$. Then

$$G(B) = B_1 \oplus B_2 \oplus \cdots \oplus B_n,$$

from which the inequality

$$\sigma_1(G(B)) \leqq \sigma_1(B)$$

follows. For $i \in \{1, \cdots, n\}$ let A_i denote the p-by-mr submatrix of A consisting of the rows indexed by $(i-1)n+1$ through $(i-1)n+p$. Let P denote the permutation matrix used at this stage in the Hadamard product argument. The permutation matrix $P \otimes I_r$ accomplishes the equality

$$F(A)(P \otimes I_r) = A_1 \oplus A_2 \oplus \cdots \oplus A_n,$$

from which

$$\sigma_1(F(A)) \leqq \sigma_1(A)$$

follows.

We may note that these proofs of submultiplicativity hold for any function defined on all sizes of matrices that has the permutation invariance, direct sum, submatrix, and submultiplicativity properties. Besides σ_1, the norms induced by the l_p norm, $1 \leqq p \leqq \infty$, have these properties. Of the unitarily invariant norms, only multiples of $\sigma_1(\cdot)$, $c\sigma_1(\cdot)$, with $c \geqq 1$, have both the submultiplicativity and direct sum property.

4. A nonexample. In this section, we show that a product having the submultiplicativity property need not satisfy the hypotheses of Observation 3.1. Define the product

$$\bullet : \mathbf{M}_2 \times \mathbf{M}_2 \to \mathbf{M}_2$$

by the following:

$$A \cdot W = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \cdot \begin{pmatrix} w & y \\ x & z \end{pmatrix} = \begin{pmatrix} aw & ay \\ bw & 0 \end{pmatrix}.$$

We will use the following upper and lower bounds on σ_1. Let $B \in \mathbf{M}_{n,m}$. A lower bound on $\sigma_1(B)$ is the maximum of the Euclidean length of the rows and columns of B. This is a special case of the submatrix property. An upper bound on $\sigma_1(B)$ is trace $(B^*B)^{1/2}$. We then may obtain the submultiplicativity inequality

$$\begin{aligned} \sigma_1(A \cdot W) &\leqq (|aw|^2 + |ay|^2 + |bw|^2)^{1/2} \\ &\leqq (|aw|^2 + |ay|^2 + |bw|^2 + |by|^2)^{1/2} \\ &= (|a|^2 + |b|^2)^{1/2}(|w|^2 + |y|^2)^{1/2} \\ &\leqq \sigma_1(A)\sigma_1(W). \end{aligned}$$

To show that $\cdot$ does not satisfy the hypotheses of Observation 3.1, we first need a lemma.

LEMMA 4.1. *Let $x, y \in \mathbf{C}^n$ with Euclidean length at most one and suppose that $x^t y = 1$. Then $x = y^c$, where the superscript c denotes the complex conjugate.*

Proof. The proof follows from the well-known characterization of cases of equality for the Cauchy–Schwarz inequality. □

Now we suppose $\cdot$ has a representation as in the hypotheses of Observation 3.1 and we derive a contradiction. Denote the rows of $F(A)$ and the columns of $G(B)$, respectively, by

$$F(A) = \begin{pmatrix} f_1(A)^t \\ f_2(A)^t \end{pmatrix}$$

and

$$G(B)=[g_1(B),g_2(B)].$$

Denoting the member of $\mathbf{M_2}$ with a one in the i,j position and zero in all other positions by E_{ij}, we have

$$E_{11}\cdot E_{11}=E_{11},\qquad E_{11}\cdot E_{12}=E_{12},\qquad E_{21}\cdot E_{11}=E_{21}.$$

This implies

$$f_1(E_{11})^t g_1(E_{11})=1,\qquad f_1(E_{11})^t g_2(E_{12})=1,\qquad f_2(E_{21})^t g_1(E_{11})=1.$$

By applying the lower bound for $\sigma_1(F(E_{ij}))$ and $\sigma_1(G(E_{ij}))$, all these vectors have at most unit length. By applying Lemma 4.1, we have

$$f_1(E_{11})^c=g_1(E_{11}),\qquad f_1(E_{11})^c=g_2(E_{12}),\qquad f_2(E_{21})^c=g_1(E_{11}).$$

Thus,

$$f_2(E_{21})^c=g_2(E_{12}).$$

However, this is a contradiction since it implies that the 2,2 element of $E_{21}\cdot E_{12}$ is one, whereas this product is the zero matrix. We conclude that $\cdot$ does not satisfy the hypotheses of Observation 3.1.

5. Extensions. In this section we show that this approach can be applied to demonstrate the submultiplicativity of the Hadamard product of infinite matrices that represent operators on the Hilbert sequence space l_2 and that other inequalities appearing in the literature can be simply deduced with this method.

Let $A=[a_{ij}]$ and $B=[b_{ij}]$ be semi-infinite matrices representing operators in the Hilbert sequence space l_2 with respect to the standard orthogonal basis. We may deduce Hadamard submultiplicativity in this setting by noting that the previous construction of G and F yield mappings from l_2 into a countable direct sum of l_2, and back again, respectively. Here we give an explicit construction of infinite matrices $F(A)$ and $G(B)$ representing these mappings, such that $F(A)G(B)=A\circ B$.

Define the function $p\colon \mathbf{N}\to\mathbf{N}$ ($\mathbf{N}=\{1,2,3,\cdots\}$) by setting $p(i)$ to be the ith largest prime number. Define functions F and G mapping the set of semi-infinite matrices into itself by the following: $F(A)=[f_{ik}]$,

$$f_{ik}=\begin{cases} a_{ir} & \text{if } k=p(r)^i \text{ for some } r\in\mathbf{N},\\ 0 & \text{otherwise,}\end{cases}$$

and $G(B)=[g_{kj}]$,

$$g_{kj}=\begin{cases} b_{qj} & \text{if } k=p(q)^j \text{ for some } q\in\mathbf{N},\\ 0 & \text{otherwise.}\end{cases}$$

With these definitions of F and G, we continue to have $A\circ B=F(A)G(B)$, $\sigma_1(A)\geqq\sigma_1(F(A))$, and $\sigma_1(B)\geqq\sigma_1(G(B))$. Thus, we again have $\sigma_1(A\circ B)\geqq\sigma_1(A)\sigma_1(B)$.

Let A and $B\in\mathbf{M_{n,m}}$. Let $r_i(A)$ denote the ith largest Euclidean length among the rows of A and $c_i(A)$ the ith largest Euclidean column length of A. Now, let X and Y be matrices such that $A=XY$. In the recent work [AHJ], the family of inequalities

$$\sum_{i=1}^{k}\sigma_i(A\circ B)\leqq\sum_{i=1}^{k}r_i(X)c_i(Y)\sigma_i(B),\qquad k=1,\cdots,n \tag{5.1}$$

was proven. We note that the $k = 1$ case of (5.1) may be demonstrated using the present methodology.

It is easily verified that for $x \in \mathbf{C}^n$, $y \in \mathbf{C}^m$, and $B \in \mathbf{M}_{n,m}$, $(xy^t) \circ B = D(x)BD(y)$. Let the matrices $X \in \mathbf{M}_{n,p}$, $Y \in \mathbf{M}_{p,m}$, and $B \in \mathbf{M}_{n,m}$ be given. Denote the columns of X by $x_1, \cdots, x_p$ and the rows of Y by $y_1^t, \cdots, y_p^t$. Then

$$\begin{aligned}(XY) \circ B &= (x_1 y_1^t) \circ B + \cdots + (x_p y_p^t) \circ B \\ &= D(x_1)BD(y_1) + \cdots + D(x_p)BD(y_p) \\ &= [D(x_1), \cdots, D(x_p)](I_p \otimes B) \begin{pmatrix} D(y_1) \\ \cdots \\ D(y_p) \end{pmatrix}.\end{aligned}$$

This is of the form $F(X)G(B)H(Y)$. It is readily seen, using the permutation invariance and direct sum properties of σ_1, that $\sigma_1(F(X)) = r_1(X)$, $\sigma_1(H(Y)) = c_1(Y)$, and $\sigma_1(G(B)) = \sigma_1(B)$, and thus the inequality is proved by two applications of the submultiplicativity property of σ_1.

This representation may be further exploited to carry out the first step in the proof of (5.1), namely, to show that

$$\sum_{i=1}^{k} \sigma_i(A \circ B) \leqq \sigma_1(B) \sum_{i=1}^{k} r_i(X) c_i(Y), \qquad k = 1, \cdots, n.$$

At present we do not know if our method can be used to prove (5.1) directly, or even the weaker family of inequalities from [HJ3],

$$\sum_{i=1}^{k} \sigma_i(A \circ B) \leqq \sum_{i=1}^{k} \sigma_i(A)\sigma_i(B), \qquad k = 1, \cdots, n.$$

However, in sequel to this paper we will show how Observation 3.1 may be modified to obtain the inequalities

$$\Phi(A \circ B) \leqq \Phi(A)\Phi(B) \tag{5.2}$$

and

$$\Phi(A \blacklozenge B) \leqq \Phi(A)\Phi(B) \tag{5.3}$$

for unitarily invariant norms $\Phi(\cdot)$ that dominate the spectral norm. Inequalities (5.2) and (5.3) appear in [HJ3] and [HMN], respectively.

REFERENCES

[AHJ] T. ANDO, R. A. HORN, AND C. R. JOHNSON, *The singular values of a Hadamard product: A basic inequality*, Linear and Multilinear Algebra, 21 (1987), pp. 345–365.

[HJ1] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[HJ2] ———, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1989.

[HJ3] ———, *Hadamard and conventional submultiplicativity for unitarily invariant norms on matrices*, Linear and Multilinear Algebra, 20 (1987), pp. 91–106.

[HMN] R. A. HORN, R. MATHIAS, AND Y. NAKAMURA, *Inequalities for unitarily invariant norms and bilinear matrix products*, Tech. Report 516, Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD, 1989.

[S] I. SCHUR, *Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen*, J. Reine Angew. Math., 140 (1911), pp. 1–28.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 1, pp. 7–15, January 1991

DETERMINANTS OF HESSENBERG *L*-MATRICES*

JEFFREY L. STUART†

Abstract. A determinantal formula for Hessenberg matrices is presented. The formula uses paths in an associated directed graph. The qualitative properties of Hessenberg matrices are investigated. Necessary and sufficient conditions are given for when the matrix is an L-matrix, and for when the determinant is sign positive or sign negative.

Key words. Hessenberg matrix, L-matrix, determinant, qualitative determinant

AMS(MOS) subject classifications. 15A09, 15A15, 15A57

1. Introduction. The recent literature contains many papers which relate the sign patterns of matrices or their inverses to other properties of the matrix including invertibility [2], [7], [8], stability [4], [5], solvability [2], [6], [8], and determinantal formulae [1], [9]. We investigate the relationship between the sign pattern of a Hessenberg matrix and the positivity or negativity of its determinant. For this purpose, we employ a combinatorial formula for the determinant of a Hessenberg matrix in terms of products along certain paths in an associated, directed graph. This formula is used to characterize which sign patterns for Hessenberg matrices yield sign-positive or sign-negative determinants, and hence which Hessenberg matrices are L-matrices.

2. Hessenberg matrices. Throughout this paper, $\mathcal{M}_n(F)$ will denote the set of all $n \times n$ matrices over the set F, where F is $\mathbb{R}$, $\mathbb{C}$ or $\{-1, 0, 1\}$. If A is in $\mathcal{M}_n(\{-1, 0, 1\})$, A will be called a *pattern*, and the entries of A will be represented by the characters "+," "−," and "0."

Let A be in $\mathcal{M}_n(\mathbb{C})$. The matrix $A = [a_{ij}]$ is called an upper Hessenberg matrix if $a_{ij} = 0$ whenever $i > j + 1$. An upper Hessenberg matrix is called unreduced if $a_{ij} \neq 0$ whenever $i = j + 1$.

Lower Hessenberg matrices and unreduced lower Hessenberg matrices are defined analogously. Since such matrices are the transposes of upper Hessenberg matrices, since the determinant is transpose-invariant, and since inversion and transposition are commuting operations, we will consider only upper Hessenberg matrices in this paper.

Suppose that A is an $n \times n$ upper Hessenberg matrix which is *not* unreduced. That is, $a_{i+1,i} = 0$ for some i. Then A partitions into a block upper triangular matrix of the form

$$A = \left[\begin{array}{c|c} A_1 & A_{12} \\ \hline 0 & A_2 \end{array}\right]$$

where A_1 is an $i \times i$ upper Hessenberg matrix, and where A_2 is an $(n - i) \times (n - i)$ upper Hessenberg matrix. Consequently, an arbitrary upper Hessenberg matrix can be represented as a block upper triangular matrix each of whose diagonal blocks is an *unreduced* upper Hessenberg matrix. It follows that many formulae which require unreduced Hessenberg matrices can be extended to arbitrary upper Hessenberg matrices by applying the formulae to each of the unreduced, diagonal blocks.

* Received by the editors December 27, 1988; accepted for publication (in revised form) January 4, 1990.
† Department of Mathematics, University of Southern Mississippi, Hattiesburg, Mississippi 39406 (stuart@usmcp6.bitnet).

For every real number r, the weak sign function wsgn (r) is defined by

$$\text{wsgn}(r) = \begin{cases} 1 & \text{if } r \geqq 0, \\ -1 & \text{if } r < 0. \end{cases}$$

Observe that for every real number r, wsgn $(r) \cdot r = |r|$.

PROPOSITION 1. *Let A be in $\mathcal{M}_n(\mathbb{R})$. Suppose that A is an upper Hessenberg matrix. Then there exists a diagonal matrix D with diagonal entries ± 1 such that DAD^{-1} is an upper triangular Hessenberg matrix with the same zero pattern as A, and such that $[\mathrm{DAD}^{-1}]_{i,i-1} = |a_{i,i-1}|$ for $2 \leqq i \leqq n-1$. Furthermore, D can be chosen to be $D = \text{diag}(d_1, d_2, \cdots, d_n)$ where $d_1 = 1$, and $d_i = d_{i-1} \cdot \text{wsgn}(a_{i,i-1})$ for $2 \leqq i \leqq n$. In particular, if A is unreduced, then DAD^{-1} is unreduced with a positive subdiagonal.*

An unreduced, upper Hessenberg matrix with a positive subdiagonal will be called a *Hessenberg matrix in standard form*.

3. Sign patterns, sign-positive determinants, and L-matrices. For every real number r, the sign function sgn (r) is defined by

$$\text{sgn}(r) = \begin{cases} 1 & \text{if } r > 0, \\ 0 & \text{if } r = 0, \\ -1 & \text{if } r < 0. \end{cases}$$

Let A be in $\mathcal{M}_n(\mathbb{R})$. Define sgn (A) to be the matrix in $\mathcal{M}_n(\{-1, 0, 1\})$ such that for each i and j, $[\text{sgn}(A)]_{ij} = \text{sgn}(A_{ij})$. Let $Q(A)$ be the subset of $\mathcal{M}_n(\mathbb{R})$ given by

$$Q(A) = \{B\colon \text{sgn}(A) = \text{sgn}(B)\}.$$

Thus for each A, $Q(A)$ has a canonical representative: sgn (A). The matrix A is called an *L-matrix* if every matrix in $Q(A)$ is invertible.

If A is in $\mathcal{M}_n(\mathbb{R})$, then A is said to have *sign-positive determinant* if $\det(B) > 0$ for every matrix B in $Q(A)$. *Sign-negative*, *sign-nonnegative*, and *sign-nonpositive determinants* are similarly defined. Clearly, A is an L-matrix if and only if $\det(B) \neq 0$ for every B in $Q(A)$.

PROPOSITION 2. *Let A be in $\mathcal{M}_n(\mathbb{R})$. A is an L-matrix if and only if A has either sign-positive determinant or sign-negative determinant.*

Proof. Assume A is an L-matrix. Suppose that there exist matrices B and B' in $Q(A)$ such that $\det(B) > 0$ and $\det(B') < 0$. Since B and B' have the same sign pattern, there is a continuous path in $\mathbb{R}^{n^2}$ from B to B' that remains in $Q(A)$. Since the map $\det(\cdot)\colon \mathbb{R}^{n^2} \to \mathbb{R}$ is continuous, the intermediate value theorem for continuous, real-valued functions implies there must be a point C on the path at which $\det(C) = 0$, a contradiction. The converse is clear. □

4. Triangular embeddings and $\mathcal{DG}(A)$. If B is in $\mathcal{M}_n(\mathbb{C})$ and $1 \leqq i, j \leqq n$, then $B(i|j)$ will denote the submatrix of A obtained from B by deleting row i and column j.

Let $A = [a_{ij}]$ be an upper Hessenberg matrix in $\mathcal{M}_n(\mathbb{C})$. Then A embeds in an $(n+1) \times (n+1)$ upper triangular matrix T_A, called the *triangular embedding of A*, as follows:

$$T_A = \left[\begin{array}{c|c} 1 & \\ 0 & \\ \vdots & A \\ 0 & \end{array}\right] = \left[\begin{array}{cccc} 1 & a_{11} & a_{12} & \cdots a_{1n} \\ 0 & a_{21} & a_{22} & \\ \vdots & & a_{32} \ddots & \vdots \\ & & & \ddots a_{nn} \\ \hline 0 & \cdots & 0 & 1 \end{array}\right].$$

Observe that $T_A(n+1|1) = A$. Let $\hat{A}$ denote A with the indexing inherited from T_A. Thus the rows of $\hat{A}$ are indexed by $\{1, 2, \cdots, n\}$ and the columns by $\{2, 3, \cdots, (n+1)\}$. That is, for $1 \leqq i \leqq n$ and $1 < j \leqq n+1$, $[T_A]_{ij} = \hat{a}_{ij}$.

If A is an upper Hessenberg matrix in $\mathscr{M}_n(\mathbb{R})$, let $\mathscr{DG}(A)$ denote the edge-weighted, loop-free, directed graph on $(n+1)$ vertices for which sgn (T_A) is the adjacency matrix. That is, $\mathscr{DG}(A)$ has vertex set $\{1, 2, 3, \cdots, n+1\}$, and there is an edge from i to j if and only if $\hat{a}_{ij} \neq 0$ and $i < j$. If there is an edge from i to j, then it is assigned weight sgn $(\hat{a}_{ij})$. (In diagrams, this weight will be denoted with either a "+" or a "−.")

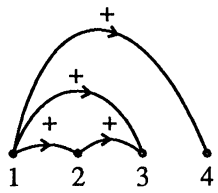
Example 1. Let A be any matrix for which

$$\text{sgn}\,(A) = \begin{bmatrix} + & + & + \\ + & + & 0 \\ 0 & + & 0 \end{bmatrix}.$$

Then

$$\text{sgn}\,(T_A) = \begin{bmatrix} + & + & + & + \\ 0 & + & + & 0 \\ 0 & 0 & + & 0 \\ 0 & 0 & 0 & + \end{bmatrix},$$

and $\mathscr{DG}(A)$ is the graph:



Since the graph $\mathscr{DG}(A)$ will play a crucial role in our results, it is appropriate to characterize the irreducibility of A in terms of this graph rather than in terms of the standard associated directed graph $\mathscr{G}(A)$.

LEMMA 3. *Let A be an unreduced, upper Hessenberg matrix. A is irreducible if and only if for each i with $1 \leqq i \leqq n-1$, there is an edge in $\mathscr{DG}(A)$ from $\{1, 2, \cdots, i\}$ to $\{i+2, i+3, \cdots, n+1\}$.*

Proof. Since A is unreduced, A is irreducible if and only if there is a directed path in $\mathscr{G}(A)$ from 1 to n. Let i be the largest vertex such that there is a directed path in $\mathscr{G}(A)$ from 1 to i. Since A is unreduced, the vertex set $\mathscr{S} = \{1, 2, \cdots, i\}$ is strongly connected, and for all $j > i$, there can be no edge from a vertex in $\mathscr{S}$ to vertex j. Thus, $\mathscr{G}(A)$ is strongly connected if and only if for each i with $\leqq i \leqq n-1$, there is an edge from $\{1, 2, \cdots, i\}$ to $\{i+1, i+2, \cdots, n\}$. Since an edge from α to β in $\mathscr{G}(A)$ becomes an edge from α to $\beta+1$ in $\mathscr{DG}(A)$, the result holds. □

5. Paths and path products. Suppose i and j are positive integers with $i < j$. Let $\mathscr{P}_{ij}$ denote the set of all increasing sequences of integers starting with i and ending with j. If $P \in \mathscr{P}_{ij}$, P is called a *path from i to j*. If $P \in \mathscr{P}_{ij}$, let $|P|$ denote the number of elements in P considered as a set, and let $P^c = \{i, i+1, i+2, \cdots, j\} \setminus P$.

Let $B = [b_{rs}]$ be in $\mathscr{M}_n(\mathbb{C})$ such that $b_{rr} \neq 0$ for each r. Let i and j be positive integers with $1 \leqq i < j \leqq n$. Let P in $\mathscr{P}_{ij}$ be the sequence $\{i = i_1, i_2, \cdots, i_{|P|} = j\}$. Let $\prod_P b_{\gamma\gamma}^{-1}$ denote the product of all of the terms $(b_{\gamma\gamma})^{-1}$ such that γ is in P. If P^c is nonempty, let $\prod_{P^c} b_{\gamma,\gamma}$ denote the product of all of the terms $b_{\gamma,\gamma}$ such that γ is in P^c. If P^c is empty,

let $\prod_{P^c} b_{\gamma,\gamma} = 1$. Finally, the *path product for P*, denoted by $\prod_P b_{\alpha\beta}$, is the product $b_{i_1,i_2}b_{i_2,i_3}\cdots b_{i_{|P|-1},i_{|P|}}$.

This notation facilitates the following formula for the entries of the inverse of an upper triangular matrix (see [3, p. 264]).

THEOREM 4. *Let B be in $\mathcal{M}_n(\mathbb{C})$. Suppose that B is an invertible, upper triangular matrix. Then*

$$[B^{-1}]_{ij} = \begin{cases} (b_{ii})^{-1} & \text{if } i=j, \\ \sum_{P\in\mathcal{P}_{ij}} (-1)^{|P|+1}\left[\prod_P b_{\alpha\beta}\right]\left[\prod_P b_{\gamma\gamma}^{-1}\right] & \text{if } i<j, \\ 0 & \text{if } i>j. \end{cases}$$

6. Parity, consistency, and full patterns. Let $B = [b_{rs}]$ be in $\mathcal{M}_n(\mathbb{R})$. Let i and j be positive integers with $1 \leqq i < j \leqq n$. Let P be in $\mathcal{P}_{ij}$. The *sign of the path P* is defined to be sgn $(\prod_P b_{\alpha\beta})$. The path P is called a *nonzero path* if its sign is nonzero. Then the path P has *parity* if sgn $(\prod_P b_{\alpha\beta}) = (-1)^{|P|}$, and it has *antiparity* if sgn $(\prod_P b_{\alpha\beta}) = (-1)^{|P|+1}$. If all nonzero paths in $\mathcal{P}_{ij}$ have parity, or if all nonzero paths in $\mathcal{P}_{ij}$ have antiparity, we say all paths from i to j have *consistent parity*. Let $\mathcal{G}$ be the edge-weighted, loop-free directed graph for which sgn (B) is both the adjacency matrix and the weighting matrix. The graph $\mathcal{G}$ has *consistent parity* if all paths in $\mathcal{G}$ have parity or if all paths in $\mathcal{G}$ have antiparity.

If A is in $\mathcal{M}_n(\{-1, 0, 1\})$ such that $a_{ij} = 0$ implies $i - j > 1$, then A is called a *full pattern*.

LEMMA 5. *Let A in $\mathcal{M}_n(\{-1, 0, 1\})$ be a Hessenberg matrix in standard form. If A is a full pattern, and if all paths from vertex 1 to vertex $(n + 1)$ in $\mathcal{DG}(A)$ have consistent parity, then $\mathcal{DG}(A)$ has consistent parity.*

Proof. Choose $i < j$. Since A is a full pattern, then every path

$$P = \{i_1 = i, \cdots, i_h = j\}$$

in $\mathcal{DG}(A)$ extends to a path $P^* = \{1, i_1, \cdots, i_h, n + 1\}$ in $\mathcal{DG}(A)$. Now apply the consistent parity for all paths of the type P^*. □

Note that the requirement that A be a full pattern cannot be removed. For the matrix of Example 1, all paths from 1 to 4 in $\mathcal{DG}(A)$ have consistent parity, but the paths from 1 to 3 do not.

7. A determinantal formula for Hessenberg matrices. We present a combinatorial formula for the determinant of a Hessenberg form. While this formula does not provide an efficient means of computing the determinant since it involves 2^{n-1} summands, it is useful for studying the relationship between the sign pattern of a matrix and the sign of its determinant.

THEOREM 6. *Let A in $\mathcal{M}_n(\mathbb{C})$ be an upper Hessenberg matrix. Let $\hat{A} = [\hat{a}_{ij}]$ be the matrix obtained from A by indexing the columns of A by the integers 2, 3, $\cdots$, $(n + 1)$. Then*

$$\det(A) = \det(\hat{A}) = (-1)^{n+1} \sum_{P\in\mathcal{P}_{1,n+1}} (-1)^{|P|}\left[\prod_P \hat{a}_{\alpha,\beta}\right]\left[\prod_{P^c} \hat{a}_{\gamma,\gamma}\right].$$

Proof. Since $\det(A)$ is continuous in each of the entries of A, the case when at least one subdiagonal entry $a_{j+1,j}$ is zero follows by a continuity argument from the case when all entries $a_{j+1,j}$ are nonzero. That is, it suffices to prove the result in the case where A is unreduced.

Let A be unreduced. Embed A in the upper triangular matrix $B = T_A$. Since A is unreduced, T_A is nonsingular. As is well known,

$$[B^{-1}]_{1,n+1} = [\det(B)]^{-1}(-1)^{(n+1)+1} \det[B(n+1\,|\,1)]$$
$$= \left[\prod_{i=1}^{n+1} b_{ii}\right]^{-1} (-1)^n \det(A).$$

Now use Theorem 4 to obtain $[B^{-1}]_{1,n+1}$, and note that for each P in $\mathscr{P}_{1,n+1}$,

$$\left[\prod_{P} b_{\gamma\gamma}^{-1}\right]\left[\prod_{i=1}^{n+1} b_{ii}\right] = \left[\prod_{P^c} b_{\gamma,\gamma}\right].$$

Finally, 1 and $n+1$ are both in P for all P in $\mathscr{P}_{1,n+1}$, so

$$\prod_{P^c} b_{\gamma,\gamma} = \prod_{P^c} \hat{a}_{\gamma,\gamma}. \qquad \square$$

It should be noted that if A is not unreduced, then $\det(A)$ can also be expressed as the product of the determinants for each of the unreduced Hessenberg diagonal blocks as discussed in § 2.

8. Sign positivity and sign negativity for $\det(A)$. Theorem 6 has, as direct consequences, the following two theorems relating necessary and sufficient conditions for a qualitatively signed determinant to parity or antiparity in $\mathscr{DG}(A)$. The first result is immediate.

THEOREM 7. *Let A be in $\mathscr{M}_n(\mathbb{R})$ be an upper Hessenberg matrix. If there is no path from 1 to $n+1$ in $\mathscr{DG}(A)$, then $\det(B) = 0$ for all B in $Q(A)$.*

THEOREM 8. *Let A be in $\mathscr{M}_n(\mathbb{R})$ and be a Hessenberg matrix in standard form. Suppose that there is at least one path from 1 to $n+1$ in $\mathscr{DG}(A)$. The conclusion depends on whether n is even or odd.*

Suppose that n is odd. Then $\det(A)$ is sign positive if and only if every path from 1 to $n+1$ in $\mathscr{DG}(A)$ has parity; $\det(A)$ is sign negative if and only if every path from 1 to $n+1$ in $\mathscr{DG}(A)$ has antiparity.

Suppose that n is even. Then $\det(A)$ is sign positive if and only if every path from 1 to $n+1$ in $\mathscr{DG}(A)$ has antiparity; $\det(A)$ is sign negative if and only if every path from 1 to $n+1$ in $\mathscr{DG}(A)$ has parity.

COROLLARY 9. *Let A be in $\mathscr{M}_n(\mathbb{R})$ and be a Hessenberg matrix in standard form. The matrix A is an L-matrix if and only if there is at least one path from 1 to $n+1$ in $\mathscr{DG}(A)$, and either every path from 1 to $n+1$ in $\mathscr{DG}(A)$ has parity or every path from 1 to $n+1$ in $\mathscr{DG}(A)$ has antiparity.*

Proof of Theorem 8. First we prove that the parity/antiparity conditions are sufficient to determine the sign of the determinant. In the formula for the determinant given by Theorem 6, each summand has sign $(-1)^{n+1+|P|}\cdot\operatorname{sgn}[\prod_P \hat{a}_{\alpha,\beta}]$. If every path P from 1 to $n+1$ in $\mathscr{DG}(A)$ has parity, then the signs of the nonzero summands are $(-1)^{n+1+2|P|} = (-1)^{n+1}$. If every path P from 1 to $n+1$ has antiparity, the signs of the nonzero summands are $(-1)^{n+2+2|P|} = (-1)^n$. Finally, since there is a path from 1 to $n+1$ in $\mathscr{DG}(A)$, there is at least one nonzero summand.

Next we prove that the parity/antiparity conditions are necessary for the sign of the determinant to be implied by the sign pattern of the matrix. The case for n odd and all paths from 1 to $n+1$ is proven. The proofs for the remaining cases are analogous.

Suppose that n is odd and that $\det(A)$ is sign positive, but that some path P' from 1 to $n+1$ in $\mathscr{DG}(A)$ has antiparity. Then $\operatorname{sgn}\prod_{P'} \hat{a}_{\alpha,\beta} = (-1)^{1+|P'|}$. Let r be a real

number with $r \geqq 1$. Let $\hat{B}_r$ be the matrix in $\mathcal{M}_n(\mathbb{R})$ whose entries are defined as follows: $\hat{b}_{ij} = 0$ if $\hat{a}_{ij} = 0$; $\hat{b}_{ii} = 1$ for all i; if $\hat{a}_{ij} \neq 0$ and $\hat{a}_{ij}$ is not on P', let $\hat{b}_{ij} = \hat{a}_{ij}(r^{-n})$; and if $\hat{a}_{ij}$ is on P', then let $\hat{b}_{ij} = r\hat{a}_{ij}$. Then $B_1 = A$, and B_r has the same sign pattern as A for all $r \geqq 1$. If P is in $P_{1,n+1}$ and $P \neq P'$, then $\prod_P \hat{b}_{\alpha,\beta}$ either is zero or contains at least one factor of r^{-n}. Thus as r becomes arbitrarily large, $\det(\hat{B})$ is dominated by the term

$$(-1)^{n+1+|P'|}\left[\prod_{P'} \hat{b}_{\alpha,\beta}\right] = (-1)^{n+2|P'|+2} \cdot r^{|P'|}\left|\prod_{P'} \hat{a}_{\alpha,\beta}\right|,$$

which is clearly negative, contradicting the sign positivity of $\det(A)$. □

Remark. It can occur that A is a Hessenberg matrix in standard form with $\det(A) > 0$, but $\det(A)$ is not sign positive. For example, let

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 1 \\ 0 & 1 & -1 \end{bmatrix},$$

and let

$$B = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 6 \\ 0 & 1 & -1 \end{bmatrix}.$$

Then $\det(A) = 1 > 0$, and B is in $Q(A)$, but $\det(B) = -4 < 0$.

9. Sign patterns for Hessenberg *L*-matrices. The following result is immediate.

COROLLARY 10. *Let A be a Hessenberg matrix in standard form. Let B be obtained from A by arbitrarily choosing nonzero entries of A occurring on or above the diagonal, and replacing those entries with zeros. If* $\det(A)$ *is sign nonnegative (sign nonpositive), then so is* $\det(B)$.

By direct observation, it is impossible to choose a nonzero value for the 3,3-entry of the matrix A given in Example 1 in order that the filled in Hessenberg matrix is still an L-matrix. A partial converse to the preceding corollary is, however, still possible. That is, there are conditions under which complete fillin preserves sign nonnegativity or sign nonpositivity of the determinant.

THEOREM 11. *Let A be a Hessenberg L-matrix. Suppose that $\mathcal{DG}(A)$ has consistent parity. Then there is a Hessenberg L-matrix B that is a full pattern such that when $a_{ij} \neq 0$,* $\operatorname{sgn}(a_{ij}) = \operatorname{sgn}(b_{ij})$.

Proof. Filling in entries of A corresponds to adding weighted edges to $\mathcal{DG}(A)$. Starting with $\mathcal{DG}(A)$, create a sequence of graphs by adding one edge at a time. When an edge is added, there are two possible cases: If there is no path in the graph between the vertices for the edge to be added, assign the weight of the edge arbitrarily; if there is already a path between the two vertices, weight the edge so as to preserve the consistency of parity. This can be done since $\mathcal{DG}(A)$ has consistent parity, and the weighting rule for added edges guarantees that each subsequent graph has consistent parity. Label the final graph as $\mathcal{G}^*$. Then $\mathcal{G}^* = \mathcal{DG}(B)$ where B is a full pattern such that the nonzero entries of A have the same sign as the corresponding entries of B. Since $\mathcal{G}^*$ has consistent parity, B is an L-matrix. □

THEOREM 12. *There are exactly 2^n matrices in $\mathcal{M}_n(\{-1, 0, 1\})$ that are Hessenberg L-matrices in standard form and also full patterns. Exactly half of these have sign-positive determinant, and the other half have sign-negative determinant.*

For n even (n odd), these full patterns are determined from the pattern given below by first arbitrarily and independently assigning ± 1 to each of the $n - 1$ border entries

denoted by "□," and then uniquely determining the entries denoted by "*" by imposing the requirement that all paths from 1 to $n+1$ in $\mathscr{DG}(A)$ have consistent parity. When all paths from 1 to $n+1$ have parity (antiparity), the determinant is sign negative. When all paths from 1 to $n+1$ have antiparity (parity), the determinant is sign positive:

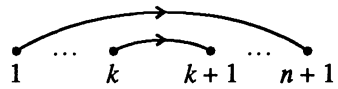
n even $(n = 2k)$
n odd $(n = 2k-1)$

$$\begin{array}{c} \begin{array}{cccccccccc} 2 & 3 & & k & k+1 & & & n & n+1 & \\ * & * & \cdots & * & \square & \square & \cdots & \square & * & 1 \\ + & * & \ddots & * & * & * & \cdots & * & \square & 2 \\ 0 & + & & & & & & \vdots & \vdots & \\ & 0 & \ddots & & & & * & & & \\ \vdots & \vdots & \ddots & & & & & * & \square & k \\ & & & & & & \ddots & \vdots & * & k+1 \\ & & & & & & \ddots & * & \vdots & \\ & & & & & & \ddots & + & * & \\ 0 & 0 & \cdots & & & & & 0 & + & * \end{array} \\ \begin{array}{cccccccccc} & & & & & & & & & n \end{array} \end{array}$$

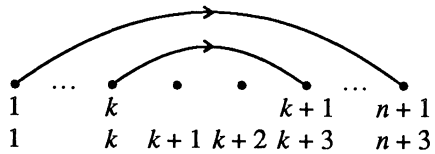
Proof. By Theorem 8, it is necessary and sufficient that all paths from 1 to $n+1$ in $\mathscr{DG}(A)$ have consistent parity in order for the matrix to be an L-matrix. By Lemma 5, $\mathscr{DG}(A)$ must have consistent parity if A is an L-matrix and a full pattern. By choosing n even or odd, and by choosing consistent parity to parity or antiparity, all possible cases are covered. We prove the theorem for n even. The case for n odd is similar.

Choose k so that $2k = n$. The proof is by induction on k. The case for $k = 1$ is easily checked.

Assume that the result holds for $n = 2k$. Then $n+2 = 2(k+1)$. Choose an $n \times n$ full pattern A_n that is a Hessenberg L-matrix in standard form. The graph $\mathscr{DG}(A_n)$ can be transformed into the graph for an $(n+2) \times (n+2)$ full pattern A_{n+2} by adding two additional vertices and all of the edges arising at or terminating at those two vertices. Consider the added vertices as being positioned between vertices k and $k+1$ of $\mathscr{DG}(A_n)$, and relabel the vertices as indicated:



$\mathscr{DG}(A_n)$



By the induction hypothesis, $\mathscr{DG}(A_n)$ has consistent parity. (Parity or antiparity is determined by the sign of the edge from 1 to $n+1$ in $\mathscr{DG}(A_n)$.) We now add the edges so as to preserve consistent parity. Add the edge $1 \to (k+2)$. Since it does not lie on a path from 1 to $((n+2)+1)$, arbitrarily assign it a weight of ± 1. Similarly, add the edge $(k+1) \to ((n+2)+1)$, and arbitrarily assign it a weight of ± 1. (Thus there are 2^2 assignments for these two edges.) Now for $j \leqq k$, every edge of the form $j \to (k+1)$ lies on a path from 1 to $((n+2)+1)$, and thus must be assigned the weight required for consistent parity. Add these edges to the graph and weight them as required. Similarly, consistent parity implies that every edge of the form $(k+2) \to j$ with $j \geqq k+3$ has its weight uniquely determined. Add these edges and assign their weights as required. The

one remaining edge to be added is the edge $(k+1) \to (k+2)$, which together with the edges $1 \to (k+1)$ and $(k+2) \to ((n+2)+1)$, forms a path from 1 to $((n+2)+1)$. Hence the weight of $(k+1) \to (k+2)$ is determined by the parity consistency condition. Since there were 2^n choices for the full pattern A_n, and since two further arbitrary weight choices were made, there are $2^{n+2}(n+2) \times (n+2)$ full patterns A_{n+2} that satisfy parity consistency, and hence are Hessenberg L-matrices in standard form. Since exactly half of these have parity and exactly half have antiparity, exactly half of these have sign positive determinant. The result holds by induction. □

10. Selected patterns for sign positivity of det (A). In light of Theorem 12, it is impossible to list all of the full patterns that correspond to Hessenberg L-matrices. Here we list a few interesting patterns that guarantee sign-positive determinants:

Banded for all $n \geqq 2$:

$$\begin{bmatrix} + & - & + & - & + & \cdots & \\ + & + & - & + & - & \cdots & \\ & + & + & - & + & \cdots & \\ & & \ddots & \ddots & \ddots & \ddots & \\ & & & + & + & - & + \\ & 0 & & & + & + & - \\ & & & & & + & + \end{bmatrix};$$

Half-bordered for odd $n \geqq 3$:

$$\begin{bmatrix} + & + & + & + & \cdots & + \\ + & - & - & - & \cdots & - \\ & + & - & - & \cdots & - \\ & & \ddots & \ddots & \ddots & \ddots & \\ & & & + & - & - & - \\ & 0 & & & + & - & - \\ & & & & & + & - \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} - & - & - & \cdots & - & + \\ + & - & - & \cdots & - & + \\ & + & - & \cdots & - & + \\ & & \ddots & \ddots & \ddots & \\ & & & + & - & - & + \\ & 0 & & & + & - & + \\ & & & & & + & + \end{bmatrix};$$

Fully bordered for even $n \geqq 2$:

$$\begin{bmatrix} + & + & + & + & \cdots & + & - \\ + & - & - & - & \cdots & - & + \\ & + & - & - & \cdots & - & + \\ & & \ddots & \ddots & \ddots & \ddots & \\ & & & + & - & - & - & + \\ & 0 & & & + & - & - & + \\ & & & & & + & - & + \\ & & & & & & + & + \end{bmatrix}.$$

11. Inverse patterns. In [7], Lady and Maybee prove the following theorem on sign patterns for the inverses of L-matrices.

THEOREM 13. *Let A be an irreducible L-matrix with $a_{ii} \neq 0$ for $1 \leqq i \leqq n$. Let a_{ij}^{-1} denote the i,j-entry of A^{-1}. Then:*

(1) *If $a_{ij} \neq 0$, $\operatorname{sgn}(a_{ij}) = \operatorname{sgn}(a_{ji}^{-1})$; and*

(2) *If $a_{ij} \neq 0$, then a_{ji}^{-1} has a determined sign if and only if every directed path from j to i in $\mathcal{G}(A)$ has the same sign.*

Two difficulties arise in applying this theorem to Hessenberg L-matrices in standard form. First, standard form does not require diagonal entries to be nonzero. Second, there

appears to be no simple relationship between $\mathcal{G}(A)$ and $\mathcal{DG}(A)$, and consequently no simple relationship between existence of signed cycles in $\mathcal{G}(A)$ and parity consistency in $\mathcal{DG}(A)$.

The first difficulty can be partially addressed as follows. Suppose that A is an irreducible, Hessenberg L-matrix in standard form and that $a_{ii} = 0$. For small positive ε, adding an ε multiple of row $i+1$ (column $i-1$) to row i (column i) yields an invertible matrix that is an irreducible, Hessenberg matrix in standard form whose sign pattern differs from A only where A has zero entries in row i (column i). This new matrix need not be an L-matrix, however. Alternatively, an L-matrix with a nonzero i,i entry can be obtained from A via a row or column permutation. In this case, however, the resultant matrix will be neither irreducible nor a Hessenberg matrix in standard form since it has a zero on its subdiagonal.

The second issue, that of relating signs of cycles to consistent parity, appears to be rather difficult. From numerical experiments with randomly generated Hessenberg L-matrices A in standard form ($n \leqq 8$), it appears that the presence of even a few paths from 1 to $n+1$ in $\mathcal{DG}(A)$ is sufficient to control the signs of the *lower* Hessenberg part of A^{-1} and to permit all possible signs for the entries of A^{-1} for which $j - i > 1$. In closing, we offer the following conjecture.

CONJECTURE. Let A be a Hessenberg L-matrix in standard form. If A is a full pattern, and if B is in $Q(A)$, then the lower Hessenberg portion of sgn (B^{-1}) is the upper Hessenberg portion of A, and the remaining entries of B^{-1}, which correspond to the zero entries in A, can occur with any sign.

REFERENCES

[1] W. BARRETT AND C. R. JOHNSON, *Determinantal formulae for matrices with sparse inverses*, II: *Asymmetric zero patterns*, Linear Algebra Appl., 56 (1984), pp. 73–88.

[2] L. BASSETT, J. MAYBEE, AND J. QUIRK, *The correspondence principle in a qualitative environment*, Econometrica, 36 (1968), pp. 544–563.

[3] M. FIEDLER, *Special Matrices and Their Applications in Numerical Mathematics*, Martinus Nijhoff, Boston, 1986.

[4] C. JEFFRIES AND C. R. JOHNSON, *Some sign patterns that preclude matrix stability*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 19–25.

[5] C. JEFFRIES, V. KLEE, AND P. VAN DEN DRIESSCHE, *When is a matrix sign stable?*, Canad. J. Math., 29 (1977), pp. 315–326.

[6] V. KLEE, R. LADNER, AND R. MANBER, *Sign solvability revisited*, Linear Algebra Appl., 59 (1984), pp. 131–157.

[7] G. M. LADY AND J. MAYBEE, *Qualitatively invertible matrices*, Math. Social Sci., 6 (1983), pp. 397–407.

[8] G. M. LADY, *The structure of qualitatively determinate relationships*, Econometrica, 51 (1983), pp. 197–218.

[9] J. MAYBEE, D. OLESKY, P. VAN DEN DRIESSCHE, AND G. WIENER, *Matrices, determinants and digraphs*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 500–519.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 1, pp. 16–23, January 1991

003

POLE ASSIGNMENT AND ADDITIVE PERTURBATIONS OF FIXED RANK*

ION ZABALLA†

Abstract. This paper is devoted to solving the general problem of pole assignment, as stated by Rosenbrock and Hayton [*Internat. J. Control*, 27 (1978), pp. 837–852], under certain restrictions for uncontrollable systems. The solution is used to give some results about the changes of the Jordan structure of a matrix subjected to additive perturbations of fixed rank.

Key words. invariant factors, interlacing inequalities, majorization, additive perturbations

AMS(MOS) subject classifications. 15A21, 15A36, 93B55

1. Introduction. The general problem of pole assignment as given by Rosenbrock and Hayton [8] can be stated in the following general way. Let $\mathbf{R}$ be the field of real numbers, and let $\mathbf{R}[s]$ denote the ring of polynomials with coefficients in $\mathbf{R}$. Let $A(s) \in \mathbf{R}[s]^{n\times n}$ and $B(s) \in \mathbf{R}[s]^{n\times m}$ be $n \times n$ and $n \times m$ polynomial matrices, respectively, such that $|A(s)| \not\equiv 0$ ($|\cdot|$ means determinant). Assume that the rational matrix $A(s)^{-1}B(s)$ is strictly proper (i.e., $A(s)^{-1}B(s) \to 0$ when $s \to \infty$). When do there exist matrices $C(s) \in \mathbf{R}[s]^{m\times n}$ and $D(s) \in \mathbf{R}[s]^{m\times n}$ such that $|D(s)| \not\equiv 0$, $D(s)^{-1}C(s)$ is proper (i.e., $D(s)^{-1}C(s) \to H \in \mathbf{R}^{m\times n}$ when $s \to \infty$), and

$$G(s)=\begin{bmatrix} A(s) & B(s) \\ C(s) & D(s) \end{bmatrix}$$

has prescribed invariant factors?

The symbol $:>$ will be used to mean "divides" and $\prec$ is the symbol of majorization in the Hardy, Littlewood, and Pólya sense [6]. That is to say, if $a = (a_1, a_2, \cdots, a_n)$ and $b = (b_1, b_2, \cdots, b_n)$ are two n-tuples of real numbers, we will write $a \prec b$ if and only if

$$\sum_{i=1}^{k} a_{[i]} \leqq \sum_{i=1}^{k} b_{[i]}, \qquad 1 \leqq k \leqq n-1$$

with equality holding for $k = n$. $a_{[1]} \geqq \cdots \geqq a_{[n]}$ and $b_{[1]} \geqq \cdots \geqq b_{[n]}$ are the components of a and b in nonincreasing order.

If $A(s)$ and $B(s)$ are assumed to be relatively left prime (i.e., the invariant factors of $[A(s), B(s)]$ are all equal to one), then Rosenbrock and Hayton gave a sufficient condition for the problem to have a solution as follows.

THE ROSENBROCK–HAYTON THEOREM. *A sufficient condition for the existence of a proper $m \times n$ rational matrix $D(s)^{-1}C(s)$ such that $\tau_1 :> \cdots :> \tau_{n+m}$ are the invariant factors of $G(s)$ is*

$$\tau_i = 1, \qquad 1 \leqq i \leqq n, \tag{1}$$

* Received by the editors July 1, 1988; accepted for publication (in revised form) September 19, 1989. This work was carried out in part within the Program "Ayudas a la Investigación para Profesores Universitarios, 1987," and was finished during the author's visit to The College of William and Mary in Virginia. It was presented at the Mathematical Sciences Lecture Series on Matrix Spectral Inequalities at The Johns Hopkins University, June 20–24, 1988. The principal lecturer for the 1988 series was Professor Robert C. Thompson.

† Departamento de Matemáticas, Universidad del País Vasco, Apdo 450, 01080 Vitoria-Gasteiz, Spain. Present address, Department of Mathematics, University of California, Santa Barbara, California 93106 (zaballa@henri.UCSB.edu).

$$(2) \qquad (k_1+b-1, \cdots, k_m+b-1) \prec (d(\tau_{n+m}), \cdots, d(\tau_{n+1}))$$

where $k_1 \geqq \cdots \geqq k_m \geqq 0$ *are the controllability indices of* $A(s)^{-1}B(s)$, b *is its biggest observability index, and* $d(\cdot)$ *denotes degree.*

There are several ways of defining the controllability indices of a rational matrix (see [3]), but quickly speaking we can say that the controllability indices of $A(s)^{-1}B(s)$ ($A(s)$, $B(s)$ are not required to be relatively left prime) are those of any pair (X, Y) such that

$$A(s)^{-1}B(s) = H(sI-X)^{-1}Y$$

for some H, (X, H) being a completely observable pair. The observability indices of $A(s)^{-1}B(s)$ can be defined in a similar way.

If $A(s)$ and $B(s)$ are not relatively left prime and we are allowed to construct $C(s)$ and $D(s)$ without specific requirements, then Sà [5] and Thompson [11] have given a necessary and sufficient condition for a more general problem to have a solution as follows. (From now on $\mathbf{F}$ will be an arbitrary field.)

THE SÀ–THOMPSON THEOREM. *If* $A(s) \in \mathbf{F}[s]^{n\times p}$ *and* $\alpha_1 : > \cdots : > \alpha_n$ *are its invariant factors* ($\alpha_i = 0$ *for* $i >$ rank $A(s)$), *then there exist matrices* $B(s) \in \mathbf{F}[s]^{n\times q}$, $C(s) \in \mathbf{F}[s]^{m\times p}$, *and* $D(s) \in \mathbf{F}[s]^{m\times q}$ *such that*

$$\begin{bmatrix} A(s) & B(s) \\ C(s) & D(s) \end{bmatrix}$$

has $\tau_1 : > \cdots : > \tau_{n+m}$ (*if* $p+q < n+m$, *then* $\tau_i : = 0$ *for* $i > p+q$) *as invariant factors if and only if*

$$(3) \qquad \tau_i : > \alpha_i : > \tau_{i+m+q}, \qquad 1 \leqq i \leqq n.$$

The Sà–Thompson result for the case when the prescribed submatrix is $[A(s), B(s)]$ and the Rosenbrock–Hayton theorem are particular cases of the general problem of pole assignment, and this paper is devoted to giving a new result under the following assumptions:

(i) $A(s)$ is a p-characteristic matrix and $A(s)^{-1}B(s)$ is strictly proper, i.e.,

$$A(s) = s^p I_n + \sum_{j=0}^{p-1} A_j s^j, \qquad B(s) = \sum_{j=0}^{p-1} B_j s^j.$$

(ii) $D(s)$ is prescribed to be q-characteristic, $D(s)^{-1}C(s)$ is proper, and $q \geqq p-1$.

Since $\mathbf{F}$ is an arbitrary field we should define what is meant by proper and strictly proper rational matrices. $A(s)^{-1}B(s)$ will be said to be *proper* (*strictly proper*) if the degrees of the polynomials in the ith row of $[A(s), B(s)]$ are not bigger (are less, respectively) than the degree of the polynomial in the position (i, i).

Before continuing, let us say something about the above restrictions. If we do not constrain ourselves to the case in which $A(s)$ is p-characteristic we can still solve the problem [15] using a similar methodology and under the assumption that q is greater than the largest degree appearing among the polynomials of $A(s)$. This degree turns out to be the largest observability index of $A(s)^{-1}B(s)$ when $A(s)$ and $B(s)$ are relatively left prime. That is to say, we can generalize the result by Rosenbrock and Hayton to the case when $A(s)$ and $B(s)$ are not relatively left prime. We make the assumption that $A(s)$ is p-characteristic because we gain clarity in the proofs and just this case is enough to obtain the results we wish concerning the change of the Jordan structure of a matrix under perturbations of fixed rank. We will deal with this problem in § 3. The result we

have obtained generalizes an earlier result by Thompson [12] to perturbations of arbitrary rank.

Finally, we say a few words about the assumption $q \geqq p-1$. As far as we know, nothing has been said about the case $q < p-1$, and this seems to be a hard case (see [1]). For example, if $a(s)$ and $b(s)$ are monic polynomials of degrees 3 and 2, respectively, the $a(s)^{-1}b(s)$ is a one-by-one strictly proper rational matrix. If we are looking for polynomials $d(s)$ and $c(s)$, $d(s)$ monic and $c(s)$ with no greater degree than $d(s)$, such that

$$\begin{bmatrix} a(s) & b(s) \\ c(s) & d(s) \end{bmatrix}$$

has a prescribed determinant, it is easily seen that we can always find such polynomials if $d(s)$ is allowed to have degree greater than one, but this is not always possible if $d(s)$ is prescribed to be linear (see [1]).

2. Invariant factor assignment. We begin with our most general result.

THEOREM 1. *Let* $A(s) = s^p I_n + \sum_{j=0}^{p-1} A_j s^j$ *and* $B(s) = \sum_{j=0}^{p-1} B_j s^j$, *where* B_j *may be a zero matrix for some* $j = 0, 1, \cdots, p-1$, $A(s) \in \mathbf{F}[s]^{n\times n}$ *and* $B(s) \in \mathbf{F}[s]^{n\times m}$. *Then there exist matrices* $C(s) \in \mathbf{F}[s]^{m\times n}$ *and* $D(s) \in \mathbf{F}[s]^{m\times m}$ *such that* $D(s)$ *is* q*-characteristic*, $q \geqq p-1$, $D(s)^{-1}C(s)$ *is proper and*

$$\begin{bmatrix} A(s) & B(s) \\ C(s) & D(s) \end{bmatrix}$$

has $\tau_1 :> \cdots :> \tau_{n+m}$ *as invariant factors if and only if*

$$\tau_i :> \alpha_i :> \tau_{i+m}, \qquad 1 \leqq i \leqq n, \tag{4}$$

$$(k_1+q, \cdots, k_m+q) \prec (d(\sigma_m), \cdots, d(\sigma_1)) \tag{5}$$

where $\alpha_1 :> \cdots :> \alpha_n$ *are the invariant factors of* $[A(s), B(s)]$,

$$\sigma_j = \frac{\beta^j}{\beta^{j-1}}, \quad \beta^j = \beta_1^j \times \cdots \times \beta_{n+j}^j, \quad \beta_i^j = \text{l.c.m.}\,(\alpha_{i-j}, \tau_i),$$

$1 \leqq i \leqq n+j$, $0 \leqq j \leqq m$ *and* $k_1 \geqq \cdots \geqq k_m$ *are the controllability indices of* $A(s)^{-1}B(s)$.

Proof. First, if $D(s)$ is q-characteristic and $D(s)^{-1}C(s)$ is proper, then

$$C(s) = \sum_{j=0}^{q} C_j s^j \tag{6}$$

where some of the matrices C_j may be zero.

Let X be the first companion matrix of $A(s)$ [4], i.e.,

$$X = \begin{bmatrix} 0 & I_n & 0 & \cdots & 0 & 0 \\ 0 & 0 & I_n & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & I_n \\ -A_0 & -A_1 & -A_2 & \cdots & -A_{p-2} & -A_{p-1} \end{bmatrix}. \tag{7}$$

Let $C(s)$ be a matrix as in (6) and define the following matrices:

$$\begin{aligned} &L_j(s) = sL_{j-1}(s) + A_{p-j}, \qquad 1 \leqq j \leqq p-1, \qquad L_0(s) = I_n, \\ &C_j(s) = \sum_{k=j}^{q} C_k s^{k-j}, \qquad 1 \leqq j \leqq p-1, \qquad C_0(s) = C_0. \end{aligned} \tag{8}$$

A simple computation shows that there exist matrices $Y_0, Y_1, \cdots, Y_{p-1}$ such that

$$\sum_{j=0}^{p-1} L_j(s)Y_j = B(s).$$

Assume that $D(s)$ is a q-characteristic matrix and write

$$Y = \begin{bmatrix} Y_{p-1} \\ \vdots \\ Y_1 \\ Y_0 \end{bmatrix}, \qquad Z(s) = [C_0 C_1 \cdots C_{p-2} C_{p-1}(s)], \tag{9}$$

$$T(s) = D(s) - \sum_{j=1}^{p-1} C_j(s)Y_{p-j}. \tag{10}$$

We claim that

$$\begin{bmatrix} I_{n(p-1)} & 0 & 0 \\ 0 & A(s) & B(s) \\ 0 & C(s) & D(s) \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} sI_{np} - X & Y \\ Z(s) & T(s) \end{bmatrix}$$

are equivalent polynomial matrices. In fact, let

$$P(s) = \begin{bmatrix} I_n & 0 & \cdots & 0 & 0 & 0 \\ 0 & I_n & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & I_n & 0 & 0 \\ L_{p-1}(s) & L_{p-2}(s) & \cdots & L_1(s) & L_0(s) & 0 \\ C_1(s) & C_2(s) & \cdots & C_{p-1}(s) & 0 & I_m \end{bmatrix}$$

and

$$\begin{bmatrix} sI_n & -I_n & 0 & \cdots & 0 & 0 & Y_{p-1} \\ 0 & sI_n & -I_n & \cdots & 0 & 0 & Y_{p-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & sI_n & -I_n & Y_1 \\ I_n & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & I_m \end{bmatrix}.$$

Then $P(s)$ and $Q(s)$ are unimodular and

$$P(s)\begin{bmatrix} sI_{np} - X & Y \\ Z(s) & T(s) \end{bmatrix} = \begin{bmatrix} I_{n(p-1)} & 0 & 0 \\ 0 & A(s) & B(s) \\ 0 & C(s) & D(s) \end{bmatrix} Q(s). \tag{11}$$

On the other hand, from (8) and (10) it is easily seen that $T(s)$ is q-characteristic.

Next, assume that $\tau_1 : > \cdots : > \tau_{n+m}$ are the invariant factors of

$$\begin{bmatrix} A(s) & B(s) \\ C(s) & D(s) \end{bmatrix};$$

then, by defining $\delta_i = \tau_{i-(p-1)n}$, $1 \leqq i \leqq np + m$, where we agree that $\tau_i = 1$ for $i < 1$, we have that $\delta_1 : > \cdots : > \delta_{np+m}$ are the invariant factors of

$$\begin{bmatrix} I_{n(p-1)} & 0 & 0 \\ 0 & A(s) & B(s) \\ 0 & C(s) & D(s) \end{bmatrix}$$

and then of

$$\begin{bmatrix} sI_{np}-X & Y \\ Z(s) & T(s) \end{bmatrix}.$$

As $sI_{np}-X$ is regular, there exist (unique) matrices $Z \in \mathbf{F}^{m\times np}$ and $S(s) \in \mathbf{F}[s]^{m\times np}$ such that

$$Z(s)=S(s)(sI_{np}-X)+Z.$$

As $sI_{np}-X$ is regular and $d(Z(s)) \leqq q$, it turns out that $d(S(s)) \leqq q-1$. (The degree of a polynomial matrix is that of its entry with highest degree.) Now

$$\begin{bmatrix} I_{np} & 0 \\ -S(s) & I_m \end{bmatrix}\begin{bmatrix} sI_{np}-X & Y \\ Z(s) & T(s) \end{bmatrix}=\begin{bmatrix} sI_{np}-X & Y \\ Z & R(s) \end{bmatrix}$$

has $\delta_1, \cdots, \delta_{np+m}$ as invariant factors and $R(s) = T(s) - S(s)Y$ is a q-characteristic matrix.

From the preceding discussion we can conclude that if there exist matrices $C(s) \in \mathbf{F}[s]^{m\times n}$ and $D(s) \in \mathbf{F}[s]^{m\times m}$ such that $D(s)$ is q-characteristic (q an arbitrary nonnegative integer), $D(s)^{-1}C(s)$ is proper, and $\tau_1 :> \cdots :> \tau_{n+m}$ are the invariant factors of

$$\begin{bmatrix} A(s) & B(s) \\ C(s) & D(s) \end{bmatrix}, \tag{12}$$

then there exist matrices $Z \in \mathbf{F}^{m\times np}$ and $R(s) \in \mathbf{F}[s]^{m\times m}$ such that $R(s)$ is q-characteristic and $\delta_1, \cdots, \delta_{np+m}$ are the invariant factors of

$$\begin{bmatrix} sI_{np}-X & Y \\ Z & R(s) \end{bmatrix}, \tag{13}$$

X and Y being the matrices defined by (7) and (9) and determined uniquely by $A(s)$ and $B(s)$.

Conversely, if $q \geqq p-1$ and Z, $R(s)$ are matrices such that $R(s)$ is q-characteristic and $\delta_1, \cdots, \delta_{np+m}$ ($\delta_i = \tau_{i-n(p-1)}$) are the invariant factors of the matrix in (13), then there exist matrices $C(s)$, $D(s)$ satisfying the requirements of the theorem such that the matrix in (12) has $\tau_1, \cdots, \tau_{n+m}$ as invariant factors. (It should be noted that if $q < p-1$, then some additional restrictions on Z are needed in order to obtain from the above process matrices $C(s)$ and $D(s)$ with degree q.)

According to Theorem 2.5 of [14], a necessary and sufficient condition for the existence of matrices $Z \in \mathbf{F}^{m\times np}$ and $R(s) \in \mathbf{F}[s]^{m\times m}$ such that $R(s)$ is q-characteristic and

$$\begin{bmatrix} sI_{np}-X & Y \\ Z & R(s) \end{bmatrix}$$

has $\delta_1, \cdots, \delta_{np+m}$ as invariant factors is

$$\delta_i :> \varepsilon_i :> \delta_{i+m}, \qquad 1 \leqq i \leqq np, \tag{14}$$

$$(k_1+q, \cdots, k_m+q) \prec (d(\theta_m), \cdots, d(\theta_1)) \tag{15}$$

where $\varepsilon_1 :> \cdots :> \varepsilon_{np}$ and $k_1 \geqq \cdots \geqq k_m$ are the invariant factors and the controllability indices of (X, Y),

$$\theta_j=\frac{\mu^j}{\mu^{j-1}}, \qquad \mu^j=\mu_1^j\times\cdots\times\mu_{np+j}^j, \qquad \mu_i^j=\text{l.c.m.}\,(\varepsilon_{i-j}, \delta_i),$$

$$1 \leqq i \leqq np+j, \qquad 0 \leqq j \leqq m.$$

From (11) we get that $[sI_{np} - X \quad Y]$ and

$$\begin{bmatrix} I_{n(p-1)} & 0 & 0 \\ 0 & A(s) & B(s) \end{bmatrix}$$

are equivalent polynomial matrices. So, if $\alpha_1 : > \cdots : > \alpha_n$ are the invariant factors of $[A(s)B(s)]$, we have $\varepsilon_i = \alpha_{i-(p-1)n}$ where $\alpha_i = 1$ for $i < 1$.

On the other hand, if

$$H = [I_n \quad 0 \cdots 0] \in \mathbf{F}^{n \times np},$$

then it turns out that $H(sI_{np} - X)^{-1}Y$ is a completely observable state-space realization of $A(s)^{-1}B(s)$ (see [16]). Thus, bearing in mind that the controllability indices of $A(s)^{-1}B(s)$ are those of (X, Y) and the characterization of δ_i and ε_i, it is easily seen that (4) and (5) are equivalent to (14) and (15), respectively, and the theorem follows. □

Remarks. (1) From the proof of the theorem we can conclude that conditions (4) and (5) are necessary even if $0 \leqq q < p - 1$.

(2) If q is prescribed to be greater than $p - 1$, then $C(s)$ can be constructed to have degree at most $p - 1$ and then $D(s)^{-1}C(s)$ would be strictly proper.

As a consequence of Theorem 1, we have the following corollary.

COROLLARY 1. *Under the same conditions as in Theorem* 1, *if* $A(s)$ *and* $B(s)$ *are relatively left prime, then there exist matrices* $C(s) \in \mathbf{F}[s]^{m \times n}$ *and* $D(s) \in \mathbf{F}[s]^{m \times m}$ *such that* $D(s)^{-1}C(s)$ *is proper,* $D(s)$ *is q-characteristic, and* $\tau_1, \cdots, \tau_{n+m}$ *are the invariant factors of*

$$\begin{bmatrix} A(s) & B(s) \\ C(s) & D(s) \end{bmatrix}$$

if and only if

$$\tau_i = 1, \qquad 1 \leqq i \leqq n \tag{16}$$

and

$$(k_1 + q, \cdots, k_m + q) \prec (d(\tau_{n+m}), \cdots, d(\tau_{n+1})). \tag{17}$$

Proof. If $A(s)$ and $B(s)$ are relatively left prime, then the invariant factors of $[A(s), B(s)]$ are all equal to one, and in this case (16) and (17) are equivalent to (4) and (5), respectively. □

As noted in the Introduction, if $A(s)$ and $B(s)$ are relatively left prime, $A(s)$ is p-characteristic and $A(s)^{-1}B(s)$ is strictly proper, then by [7, p. 103], the observability indices of $A(s)^{-1}B(s)$ are all equal to p, and from Corollary 1, we get the Rosenbrock–Hayton theorem by prescribing $q = p - 1$.

Our next result is a slight generalization of a result of Sà [5].

THEOREM 2. *Let* $A(s) \in \mathbf{F}[s]^{n \times n}$ *be a p-characteristic matrix, and let* $\alpha_i : > \cdots : > \alpha_n$ *be its invariant factors. Let* $\tau_1 : > \cdots : > \tau_{n+m}$ *be monic polynomials such that* $\sum_{j=1}^{n+m} d(\tau_j) = np + mq$ *for some nonnegative integer* $q \geqq p - 1$. *Then there exist matrices* $B(s) \in \mathbf{F}[s]^{n \times m}$, $C(s) \in \mathbf{F}[s]^{m \times n}$, *and* $D(s) \in \mathbf{F}[s]^{m \times m}$ *such that* $A(s)^{-1}B(s)$ *is strictly proper,* $D(s)^{-1}C(s)$ *is proper,* $D(s)$ *is q-*characteristic, *and*

$$\begin{bmatrix} A(s) & B(s) \\ C(s) & D(s) \end{bmatrix}$$

has $\tau_1 : > \cdots : > \tau_{n+m}$ *as invariant factors if and only if*

$$\tau_i : > \alpha_i : > \tau_{i+2m}, \qquad 1 \leqq i \leqq n. \tag{18}$$

Sà's result is the case $q = p$.

Proof. The necessity of (18) for any $q \geqq 0$ is a consequence of the Sà–Thompson theorem. We will show its sufficiency. Let X be the first companion matrix of $A(s)$ and let $\varepsilon_i = \alpha_{i-(p-1)n}$, $1 \leqq i \leqq pn$, be its invariant factors. Define $\delta_i = \tau_{i-(p-1)n}$, $1 \leqq i \leqq pn + m$. Now use (18) and the proof of Theorem 3.2 of [13] (using Theorem 2.5 of [14] instead of Lemma 2.11 of [13]) to show that there exist matrices $Y \in \mathbf{F}^{np \times m}$, $Z \in \mathbf{F}^{m \times np}$, and $D_1(s) \in \mathbf{F}[s]^{m \times m}$ such that $D_1(s)$ is q-characteristic and

$$\begin{bmatrix} sI_{np} - X & Y \\ Z & D_1(s) \end{bmatrix}$$

has $\delta_1, \cdots, \delta_{np+m}$ as invariant factors. Now, as in the proof of Theorem 1, this matrix is equivalent to

$$\begin{bmatrix} I_{n(p-1)} & 0 & 0 \\ 0 & A(s) & B(s) \\ 0 & C(s) & D(s) \end{bmatrix}$$

for some matrices $B(s) \in \mathbf{F}[s]^{n \times m}$, $C(s) \in \mathbf{F}[s]^{m \times n}$, and $D(s) \in \mathbf{F}[s]^{m \times m}$, obtained from Y, Z, and $D_1(s)$.

So, $\tau_1 : > \cdots : > \tau_{n+m}$ are the invariant factors of

$$\begin{bmatrix} A(s) & B(s) \\ C(s) & D(s) \end{bmatrix}.$$

□

3. Changes of the Jordan structure. As a consequence of the previous section, we can give some results related to the possible invariant factors that can be attained when a given matrix is subjected to an additive perturbation of fixed rank. Since the eigenvalues and the Jordan structure of a complex matrix are determined by its elementary divisors, and hence by its invariant factors, our next results apply in an obvious way to the study of the changes of the Jordan structure of a matrix under additive perturbations.

If we take $p = 1$ and $q = 0$ in Theorem 2, we obtain the following result.

THEOREM 3. *Let $A \in \mathbf{F}^{n \times n}$, and let $\alpha_1 : > \cdots : > \alpha_n$ be its invariant factors. If $\tau_1 : > \cdots : > \tau_n$ are monic polynomials such that $\sum_{j=1}^{n} d(\tau_j) = n$, then there exists a matrix $P \in \mathbf{F}^{n \times n}$ with* rank $P \leqq m$ *such that $A + P$ has $\tau_1, \cdots, \tau_n$ as invariant factors if and only if*

$$\tau_{i-m} : > \alpha_i : > \tau_{i+m}, \qquad 1 \leqq i \leqq n \tag{19}$$

where $\tau_i = 1$ for $i < 1$ and $\tau_i = 0$ for $i > n$.

Proof. Define $\mu_i = \tau_{i-m}$, $1 \leqq i \leqq n + m$. From Theorem 2, there exist matrices $B \in \mathbf{F}^{n \times m}$ and $C \in \mathbf{F}^{m \times n}$ such that

$$\begin{bmatrix} sI_n - A & B \\ C & I_m \end{bmatrix} \tag{20}$$

has $\mu_1, \cdots, \mu_{n+m}$ as invariant factors if and only if

$$\mu_i : > \alpha_i : > \mu_{i+2m}, \qquad 1 \leqq i \leqq n. \tag{21}$$

Now, the matrix in (20) is equivalent to

$$\begin{bmatrix} sI_n - (A + BC) & 0 \\ 0 & I_m \end{bmatrix}.$$

So, the invariant factors of the matrix in (20) are those of $A + BC$ and m-invariant factors equal to one. Put $P = BC$; then rank $P \leqq m$ and $\tau_1, \cdots, \tau_n$ are the invariant

factors of $A + P$ if and only if $\tau_i = \mu_{i+m}$ and (21) holds; that is, if and only if (19) is satisfied. □

Remarks. (i) It is easily seen that

$$r = \min \{ t: \tau_{i-t} :> \alpha_i :> \tau_{i+t} \}, \qquad 1 \leqq i \leqq n$$

is the minimum rank of the matrices $P \in \mathbf{F}^{n \times n}$ such that $A + P$ has $\tau_1, \cdots, \tau_n$ as invariant factors.

(ii) If we take $m = 1$, we get an earlier result due to Thompson [12]. The general case has also been solved by Silva [9], [10], but his approach is completely different.

(iii) Theorem 3 is not a complete characterization of the possible Jordan structures of a matrix subjected to perturbations of fixed rank. For instance, there could be matrices P of rank $r \geqq 1$ such that $A + P$ and A are similar. A complete answer to this problem for the case when $\mathbf{F}$ is an algebraically closed field is given by Silva in [9]. His proof is large and complicated and does not apply to the case of general fields.

Acknowledgment. The results of this paper were presented at Robert Thompson's lectures held at The Johns Hopkins University, Baltimore, MD, June 20–24, 1988. Thanks are due to the organizers for inviting the author to such an interesting conference.

REFERENCES

[1] T. E. DJAFERIS AND S. K. MITTER, *Some generic invariant factor assignment results using dynamic output feedback*, Linear Algebra Appl., 50 (1983), pp. 103–131.

[2] J. M. GRACIA, I. DE HOYOS, AND I. ZABALLA, *Perturbation of linear control systems*, Linear Algebra Appl., to appear.

[3] G. D. FORNEY, *Minimal bases of rational vector spaces, with applications to multivariable linear systems*, SIAM J. Control, 13 (1975), pp. 493–520.

[4] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Matrix Polynomials*, Academic Press, New York, 1982.

[5] E. MARQUES DE SÀ, *Imbedding conditions for λ-matrices*, Linear Algebra Appl., 24 (1979), pp. 33–50.

[6] A. W. MARSHALL AND I. OLKIN, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, London, 1979.

[7] H. H. ROSENBROCK, *State-Space and Multivariable Theory*, Thomas Nelson, London, 1970.

[8] H. H. ROSENBROCK AND G. E. HAYTON, *The general problem of pole assignment*, Internat. J. Control., 27 (1978), pp. 837–852.

[9] F. C. SILVA, *Somas de matrizes com factores invariantes prescritos*, Ph.D. thesis, University of Lisbon, Lisbon, Portugal, 1986.

[10] ———, *The rank of the difference of matrices with prescribed similarity classes*, Linear and Multilinear Algebra, 24 (1988), pp. 51–58.

[11] R. C. THOMPSON, *Interlacing inequalities for invariant factors*, Linear Algebra Appl., 24 (1972), pp. 1–32.

[12] ———, *Invariant factors under rank one perturbations*, Canad. J. Math., 22 (1980), pp. 240–245.

[13] I. ZABALLA, *Interlacing inequalities and control theory*, Linear Algebra Appl., 101 (1988), pp. 9–31.

[14] ———, *Interlacing and majorization in invariant factor assignment problems*, Linear Algebra Appl., 121 (1989), pp. 409–421.

[15] ———, *The general problem of pole assignment without complete control*, manuscript.

[16] ———, *Invariant factor assignment on higher-order systems using state-feedback*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 147–154.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 1, pp. 24–40, January 1991

004

ON THE RECONSTRUCTION OF LAYERED MEDIA FROM REFLECTION DATA*

ALFRED BRUCKSTEIN†, THOMAS KAILATH‡, ISRAEL KOLTRACHT§, AND PETER LANCASTER¶

Abstract. The problem of reconstructing an elastic layered medium from a discrete reflection response is considered. Using matrix methods, a family of models is defined that is parametrized by the surface reflection coefficient. The relationship between a general response and that with a perfect reflector at the surface is established and is used to provide a new proof of a recently established representation for the reflection coefficients. A (known) thresholding strategy for the prediction of reflection coefficients is presented and is shown to be a "maximum a posteriori" estimation process. Numerical examples are given.

Key words. Levinson algorithm, Toeplitz matrices, reflection coefficients, layered media

AMS(MOS) subject classifications. 86-08, 65F05, 15A90

Introduction. In this paper we consider the problem of reconstructing a layered medium from noisy reflection response data. It is assumed that the medium is made up of a sequence of horizontal homogeneous layers (the Goupillaud model), and that the measurement noise is bounded in magnitude by ε. We also admit some a priori knowledge of the reflection coefficient sequence; namely, that most of the reflection coefficients are zero and if different from zero, that they are uniformly distributed between $[-1, 1]$.

Both the standard reconstruction procedures, known as dynamic deconvolution (Claerbout [3], Aki and Richards [1], Robinson and Treitel [13]) and the layer peeling procedure (Bruckstein, Koltracht, and Kailath [2]), are unstable in the presence of noise. A thresholding strategy for stabilizing this procedure has recently been introduced in Bruckstein, Koltracht, and Kailath [2] and Koltracht and Lancaster [8] (see also Ferber [5]). This strategy consists of careful estimation of error magnification in the recursive reconstruction procedure and of the use of recursive estimates for setting to zero small computed reflection coefficients. The estimation of errors is based on a new representation of reflection coefficients for a general surface condition first obtained in Koltracht and Lancaster [8].

Section 1 contains a new derivation of this formula that is both simpler than the original one and also has more physical intuition behind it. It contains some relevant results of error analysis from Koltracht and Lancaster [9] as well.

In § 2 the thresholding strategy is described and we show that it can be viewed as an approximate *maximum* a posteriori estimation process for the reflection coefficients. The strategy is also compared with the minimum entropy deconvolution method proposed by Wiggins [17] for geophysical reflection seismology.

The stabilizing effects of the thresholding strategy in the presence of noise are illustrated in the numerical experiments of § 3, with synthetic as well as field data.

* Received by the editors August 29, 1988; accepted for publication (in revised form) October 3, 1989.

† Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 3200, Israel.

‡ Faculty of Electrical Engineering, Stanford University, Stanford, California 94305 (barb@isl.stanford.edu). This work was supported in part by U.S. Army Research Office contract DAAL03-86-K-0045 and Air Force Office of Scientific Research, Air Force Systems Command contract AF83-0228.

§ Department of Mathematics, University of Connecticut, Storrs, Connecticut 06268 (koltrach@uconnvm.bitnet). The work of this author was supported in part by National Science Foundation grant DMS-8801961.

¶ Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada T2N 1N4 (lancaster@uncamult.bitnet). The work of this author was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada.

A similar idea of setting to zero small reflection coefficients appeared simultaneously in Ferber [5] (both this paper and the paper of Bruckstein, Koltracht, and Kailath [2] were submitted in 1984). The error estimates in Ferber [5], however, are less accurate and apply to the perfect surface reflector only.

1. Representations of reflection coefficients. The one-dimensional inverse scattering problem amounts to the reconstruction of an acoustic medium from its response to a known input pressure wave. Discretizing the medium into a large number of thin layers, we can assume that each layer has a constant impedance, and that changes of impedance occur only at layer interfaces. Such interfaces are characterized by their reflection coefficients. To define a reflection coefficient, consider a vertically incident unit impulse on the interface from above (and measured in terms of units that represent the square root of energy). The part of the impulse that is reflected upward gives the value of the corresponding reflection coefficient c; hence $|c| \leqq 1$. The transmitted part t can be calculated from the energy conservation law as $t = \sqrt{1-c^2}$. If a unit impulse is incident on the same interface from below, the reflected amplitude is equal to $-c$ (see Robinson [14, p. 48], for example).

Let the controlled input signal, which is sent downward, be measured just above the surface at uniform intervals of time τ, giving the input sequence $\{d_0, d_1, \cdots, d_N\}$. Starting with time τ, after each τ units of time, some reflected upcoming signal (possibly zero) will reach the surface from below. Denote this upcoming sequence of signals just below the surface by $\{0, v_1, v_2, \cdots, v_N\}$. Each v_j represents a superposition of a primary reflection from the jth interface with multiple reflections from previous layers. (Note that the width of each layer is determined by the half travel time $\tau/2$ of the pressure wave; thus, the physical width depends on the velocity of propagation in the medium of this particular layer.)

Assuming that the surface reflection coefficient c_0 is known, the sequence of downgoing signals just below the surface can then be seen to be $\{t_0 d_0, t_0 d_1 - c_0 v_1, \cdots, t_0 d_N - c_0 v_N\}$.

Let $u_j = t_0 d_j - c_0 v_j$, $j = 1, \cdots, N$, and define the following nested sequence of matrices:

$$R_k = L(\mathbf{u}_k)L^T(\mathbf{u}_k) - L(\mathbf{v}_k)L^T(\mathbf{v}_k) \tag{1}$$

for $k = 0, 1, \cdots, N$ where T denotes transpose, $\mathbf{u}_k = [1, u_1, \cdots, u_k]^T$, $\mathbf{v}_k = [0, v_1, \cdots, v_k]^T$, and for any vector $\mathbf{a} = [a_0, \cdots, a_k]^T$ of any length $k+1$, $L(\mathbf{a})$ denotes a lower triangular Toeplitz matrix whose first column is $\mathbf{a}$. Thus,

$$L(\mathbf{a}) = \begin{bmatrix} a_0 & 0 \cdots 0 \\ a_1 & a_0 \cdots 0 \\ \cdot & \cdot \quad \cdot \\ a_k & a_{k-1} \cdots a_0 \end{bmatrix}.$$

Conservation of energy arguments (Kailath, Bruckstein, and Morgan [6]; see also Lev-Ari and Kailath [12]) show that R_N is a positive-definite matrix.

THEOREM 1. *Let $\{d_0, d_1, \cdots, d_N\}$ be the controlled input sequence and let $\{0, v_1, \cdots, v_N\}$ be the upcoming sequence measured just below the surface of a layered medium defined by the sequence of reflection coefficients $\{c_0, c_1, \cdots, c_N\}$. Then for $k = 0, \cdots, N-1$*

$$c_{k+1} = \sum_{j=0}^{k} v_{j+1}\gamma_k(j) \tag{2}$$

where $\gamma_k = [\gamma_k(0), \cdots, \gamma_k(k)]^T$ *is the solution of the equation*

$$R_k\gamma_k = [0, \cdots, 0, 1]^T$$

and R_k *are defined via* (1) *with* $\mathbf{u} = t_0\mathbf{d} - c_0 v$.

The representation formula (2) was established in Koltracht and Lancaster [8]. The new derivation of Theorem 1 is based on reduction of the general case when $\mathbf{d} = \{d_0, d_1, \cdots, d_N\}$ and $c_0 \in [-1, 1]$ to the special case when $d_j = 0$, $j = 1, \cdots, N$, and $c_0 = -1$ (or perfect reflection of upcoming waves at the surface). In reflection seismology this case is called "marine" and the representation of the reflection coefficients given by (2) of Theorem 1 is well known in this particular situation (see Kunetz [10], Robinson [14]).

Let us first show how the transformations from $\mathbf{d}$ to $\mathbf{v}$ and from $\mathbf{u}$ to $\mathbf{v}$ in the Goupillaud model can be interpreted as discrete, causal, linear systems (see Robinson [14], for example). We also show how, using a limiting process, the "marine case" can be included in a family of systems parametrized by c_0, the reflection coefficient at the surface.

Assuming that $|c_0| < 1$, it is not difficult to see that $\mathbf{v}$ (in the first layer) is related to the input vector $\mathbf{d}$ by $\mathbf{v} = t_0 B\mathbf{d}$, where B is a strictly lower triangular Toeplitz matrix:

$$B = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 \\ b_1 & 0 & & & \vdots \\ b_2 & b_1 & 0 & \ddots & \vdots \\ \vdots & & & 0 & 0 \\ b_N & & \cdots & b_1 & 0 \end{bmatrix},$$

$b_1 = c_1$, $b_2 = c_2 t_1^2 - c_1^2 c_0$, and for $j = 2, \cdots, N$, b_j is a polynomial in $c_0, c_1, \cdots, c_j$. This relation implies that the transformation $\mathbf{d} \to \mathbf{v}$ is a discrete causal linear system.

For $|c_0| < 1$ we have

$$\mathbf{u} = t_0\mathbf{d} - c_0\mathbf{v}, \tag{3}$$

and it follows that

$$B\mathbf{u} = (I - c_0 B)\mathbf{v},$$

or $A\mathbf{u} = \mathbf{v}$ where

$$A = (I - c_0 B)^{-1} B. \tag{4}$$

As A is also lower triangular and Toeplitz it is seen that, as claimed above, the transformation $\mathbf{u} \to \mathbf{v}$ is also a discrete causal linear system. Furthermore, the system (i.e., A and B) both depend continuously on c_0.

Next we show how to include the cases when $|c_0| = 1$ in our discussion. Observe that either case $c_0 = \pm 1$ means that no finite signal above ground can produce a signal below ground. However, it we consider the limiting process $c_0 \to -1$, and simultaneously let $d_0 \to \infty$ in such a way that $t_0 d_0 \to 1$ (while $\{d_j\}_{j=1}^{\infty}$ remains bounded), then it follows from the equation $\mathbf{v} = t_0 B\mathbf{d}$ that

$$\hat{\mathbf{v}} \overset{\text{def}}{=} \lim_{c_0 \to -1} \mathbf{v} = B_{-1}\mathbf{e}_0$$

where B_{-1} denotes B evaluated at $c_0 = -1$ and $\mathbf{e}_0$ is the first unit vector, i.e., $\mathbf{e}_0^T = [1, 0, \cdots, 0]$. Furthermore, it follows from (3) that, in this case,

$$\hat{\mathbf{u}} = \mathbf{e}_0 + \hat{\mathbf{v}},$$

as physical reasoning also requires. Equation (3) also applies in the sense that $A_{-1}\hat{\mathbf{u}} = \hat{\mathbf{v}}$ where

$$A_{-1} = (I + B_{-1})^{-1} B_{-1}.$$

Thus, the equation $\mathbf{v} = t_0 B\mathbf{d}$ still makes sense in the limit as $c_0 \to -1$ and represents the physical situation when $c_0 = -1$ and the disturbing signal $\mathbf{e}_0$ is applied just *below* the surface. A similar argument applies when $c_0 \to 1$. Thus, *the matrix* A: $\mathbf{u} \to \mathbf{v}$ *defines a causal linear system for any* $c_0 \in [-1, 1]$ *and* A *depends continuously on* c_0. The limiting case when $c_0 = -1$, $\hat{\mathbf{u}} = \mathbf{e}_0 + \hat{\mathbf{v}}$ is known as the *marine case* and the vector $\hat{\mathbf{v}}$ is called the *marine response*.

As the transformation B from input $\mathbf{d}$ to output $\mathbf{v}$ is a time-invariant causal linear system (i.e., a filter) and depends only on $c_0, c_1, \cdots, c_N$, we may write, absorbing t_0 into B,

$$\mathbf{v} = B(c_0, c_1, \cdots, c_N)\mathbf{d}.$$

In this notation the marine response of the model is

$$\hat{\mathbf{v}} = B(-1, c_1, \cdots, c_N)\mathbf{e}_0.$$

As $\mathbf{u} = t_0\mathbf{d} - c_0\mathbf{v}$, the marine response is also characterized by the property that when $\mathbf{v} = \hat{\mathbf{v}}$ we have $\mathbf{u} = \mathbf{e}_0 + \hat{\mathbf{v}}$. We use this to prove the following reduction theorem.

THEOREM 2. *For any* $c_0 \in [-1, 1]$, *let*

$$\mathbf{u} = \begin{bmatrix} 1 \\ u_1 \\ \vdots \\ u_N \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} 0 \\ v_1 \\ \vdots \\ v_N \end{bmatrix}$$

represent the downgoing and upcoming signals in the first layer, (*respectively*), *and write* $U = L(\mathbf{u})$, $V = L(\mathbf{v})$. *Then the marine response of the model is given by*

$$\hat{\mathbf{v}} = (U - V)^{-1}\mathbf{v}. \tag{5}$$

Proof. The model associated with surface reflection coefficient c_0 is a filter. Let $\mathbf{a}$ be its impulse response and $A = L(\mathbf{a})$ (so that $\mathbf{a} = A\mathbf{e}_0$). The marine response of the model is the vector $\hat{\mathbf{v}}$ for which $A(\hat{\mathbf{v}} + \mathbf{e}_0) = \hat{\mathbf{v}}$, i.e.,

$$\hat{\mathbf{v}} = (I - A)^{-1}A\mathbf{e}_0.$$

We have $A\mathbf{u} = \mathbf{v}$, or $AU\mathbf{e}_0 = UA\mathbf{e}_0 = V\mathbf{e}_0$ so that $\mathbf{a} = U^{-1}V\mathbf{e}_0$ and it follows that $A = U^{-1}V$. Substitute in the equation for $\hat{\mathbf{v}}$ and use the fact that lower triangular Toeplitz matrices commute to obtain

$$\begin{aligned} \hat{\mathbf{v}} &= (I - U^{-1}V)^{-1}U^{-1}V\mathbf{e}_0 \\ &= (U - V)^{-1}V\mathbf{e}_0 = (U - V)^{-1}\mathbf{v}. \end{aligned}$$

□

Now let us complete the proof of Theorem 1. This depends on the reduction to the "marine case" as described in Theorem 2. We use a subscript k (as in $\mathbf{u}_k$, $\mathbf{v}_k$) to denote vectors of length $k + 1$.

For the "marine case" it is well known (see Kunetz [10], Robinson [14]) that, if

$$T_k \overset{\text{def}}{=} L(\hat{\mathbf{v}}_k + \mathbf{e}_0)L(\hat{\mathbf{v}}_k + \mathbf{e}_0)^T - L(\hat{\mathbf{v}}_k)L(\hat{\mathbf{v}}_k)^T \tag{6}$$

(a positive-definite Toeplitz matrix) and $\mathbf{w}_k$ is defined by $T_k\mathbf{w}_k = \mathbf{e}_k$, then the subsurface reflection coefficients are given by the (Levinson–Durbin) formula

$$c_{k+1} = \hat{\mathbf{v}}_k^T\mathbf{w}_k, \qquad k = 0, 1, \cdots, N. \tag{7}$$

From (5) we have for the general case

$$L(\mathbf{v}_k) = L(\mathbf{u}_k - \mathbf{v}_k)L(\hat{\mathbf{v}}_k).$$

But also

$$\begin{aligned} L(\mathbf{u}_k) &= L(\mathbf{u}_k - \mathbf{v}_k) + L(\mathbf{v}_k) \\ &= L(\mathbf{u}_k - \mathbf{v}_k) + L(\mathbf{u}_k - \mathbf{v}_k)L(\hat{\mathbf{v}}_k) \\ &= L(\mathbf{u}_k - \mathbf{v}_k)(L(\hat{\mathbf{v}}_k) + I) \\ &= L(\mathbf{u}_k - \mathbf{v}_k)L(\hat{\mathbf{v}}_k + \mathbf{e}_0). \end{aligned}$$

Consequently, using (1) and (3) we obtain

$$L(\mathbf{u}_k - \mathbf{v}_k)T_k L(\mathbf{u}_k - \mathbf{v}_k)^T = R_k.$$

As $L(\mathbf{u}_k - \mathbf{v}_k)$ is nonsingular and T_k is positive definite, it follows that R_k is positive definite. Furthermore, as $L(\mathbf{u}_k - \mathbf{v}_k)\mathbf{e}_k = \mathbf{e}_k$, $T_k\mathbf{w}_k = \mathbf{e}_k$ implies

$$L(\mathbf{u}_k - \mathbf{v}_k)T_k L(\mathbf{u}_k - \mathbf{v}_k)^T (L(\mathbf{u}_k - \mathbf{v}_k)^T)^{-1}\mathbf{w}_k = e_k,$$

or $R_k(L(\mathbf{u}_k - \mathbf{v}_k)^T)^{-1}\mathbf{w}_k = \mathbf{e}_k$. Thus, if $\boldsymbol{\gamma}_k$ is defined by $R_k\boldsymbol{\gamma}_k = \mathbf{e}_k$, then $\mathbf{w}_k = L(\mathbf{u}_k - \mathbf{v}_k)^T\boldsymbol{\gamma}_k$ and (7) and (5) give

$$\begin{aligned} c_{k+1} &= (L(\mathbf{u}_k - \mathbf{v}_k)^{-1}\mathbf{v}_k)^T L(\mathbf{u}_k - \mathbf{v}_k)^T\boldsymbol{\gamma}_k \\ &= \mathbf{v}_k^T\boldsymbol{\gamma}_k \end{aligned}$$

as required. □

2. The effects of noisy data. In practice the measured response of a layered medium is contaminated with noise arising from measurement errors, spatial effects, and the discretization of the continuous medium. Thus we can write $\mathbf{v} = \hat{\mathbf{v}} + \varepsilon$ where $\hat{\mathbf{v}}$ is the vector of measured noisy response. In what follows we assume that the errors ε_j are uniformly distributed

$$|\varepsilon_j| < \varepsilon, \qquad j = 1, \cdots, N,$$

ε being a known estimate. Under this assumption it is possible to show (Koltracht and Lancaster [9]) that the matrix R_N defined in (1) will be perturbed by a certain matrix $F = \{f_{ij}\}_{i,j=0}^{N}$

$$R_N = \hat{R}_N + F,$$

where, with a high probability (of 99.8 percent), elements of F satisfy the inequality

$$|f_{ij}| < \sqrt{3}\varepsilon(|c_0| + 2(1 - c_0^2)^{1/2}), \qquad i,j = 0, \cdots, N. \tag{8}$$

(Note that when $c_0 = \pm 1$, the right-hand side of (8) is simply equal to $\sqrt{3}\varepsilon$.) Similar estimates can be obtained for measurement noise with other statistical properties. Given the representation (2) of the reflection coefficients and the estimate (8) of the size of the perturbation matrix, we can estimate the error in the reflection coefficients as follows (Koltracht and Lancaster [9]).

THEOREM 3. *Let $\hat{\mathbf{v}}$ be the recorded response of a layered medium with a known surface reflection coefficient c_0. Let ε denote the noise level, so that*

$$|v_j - \hat{v}_j| < \varepsilon, \qquad j = 1, \cdots, N,$$

and let $\hat{c}_k$, $k = 1, \cdots, N$ *denote the reflection coefficients corresponding to the recorded response and the known input vector* $\mathbf{d}$. *Then for sufficiently small* ε, *with a probability of* 0.998, *and for* $k = 0, \cdots, N-1$

$$(9)\quad |c_{k+1}-\hat{c}_{k+1}| < \sqrt{3}\varepsilon\left(\sum_{j=0}^{k}\hat{\gamma}_k^2(j)\right)^{1/2}\times\left[1+(|c_0|+2(1-c_0^2)^{1/2})\sum_{j=0}^{k}|\hat{x}_k(j)|\right]$$

where c_k, $k = 1, \cdots, N$, *are the true reflection coefficients, and for* $k = 0, \cdots, N-1$, $\hat{\gamma}_k$ *and* $\hat{\mathbf{x}}_k$ *are defined by* $\hat{R}_k\hat{\gamma}_k = \mathbf{e}_k$ *and* $\hat{R}_k\hat{\mathbf{x}}_k = [\hat{v}_1, \cdots, \hat{v}_{k+1}]^T$.

Efficient algorithms for computing the bounds of Theorem 3 can be found in Koltracht and Lancaster [8] (see also Lev-Ari and Kailath [11]). Note that in the case of a Toeplitz matrix $\hat{R}_N$ the right-hand side of (9) simplifies to

$$(10)\qquad |c_{k+1}-\hat{c}_{k+1}| < \sqrt{3}\varepsilon\left(\sum_{j=0}^{k}\hat{\gamma}_k^2(j)\right)^{1/2}\sum_{j=0}^{k+1}|\hat{\gamma}_{k+1}(j)|/\hat{\gamma}_{k+1}(k+1)$$

where $\hat{\gamma}_k$ can be computed via the usual Levinson algorithm (see Koltracht and Lancaster [8] for more details). We remark that the estimate (10) is more accurate than the one suggested in Ferber [5] for the marine case only.

3. Inverse scattering with thresholding: An approximate "maximum a posteriori" estimation process. The discretization of the pressure wave and the elastic medium, and the presence of noise, imply that most of the observed reflecting interfaces are an artificial byproduct of the chosen discretization interval, and do not correspond to real reflectors. Moreover, because of these facts the reflection coefficients are computed approximately with the precision of the bound of (9) at best. This means in particular, that the zero reflection coefficients, which correspond to artificial layers, can become nonzero values within this bound. It is, of course, our objective to reconstruct the real layered structure of the medium, and the first priority is therefore to distinguish the real reflecting interfaces from the artificial ones.

In order to use our prior information, which says that most of the reflection coefficients are zero, the following thresholding strategy is useful (see Ferber [5], Bruckstein, Koltracht, and Kailath [2], Koltracht and Lancaster [8]).

(i) Start with the known data c_0, $\{d_1, \cdots, d_N\}$, $\{\hat{v}_1, \cdots, \hat{v}_N\}$ and $k = 0$.

(ii) Compute $\hat{\gamma}_k$, $\hat{\mathbf{x}}_k$, and $\hat{c}_{k+1}$ as defined in Theorem 3, and also compute

$$B_k = \sqrt{3}\left(\sum_{j=0}^{k}\hat{\gamma}_k^2(j)\right)^{1/2}\left(1+(|c_0|+2(1-c_0^2)^{1/2}\sum_{j=0}^{k}|\hat{x}_k(j)|)\right).$$

(iii) If $|\hat{c}_{k+1}| < \varepsilon B_k$, then set $\hat{c}_{k+1} = 0$.

(iv) Increase k by one (until $k = N-1$).

Indeed, if $|\hat{c}_{k+1}| < \varepsilon B_k$, then the true reflection coefficient c_{k+1} can be any number in the interval $(\hat{c}_{k+1} - \varepsilon B_k, \hat{c}_{k+1} + \varepsilon B_k)$ and zero also belongs to this interval. Having assumed the prior information about the medium, we must now conclude that the true reflection coefficient is most likely equal to zero.

In probabilistic terms, it may be assumed that the reflection coefficient sequence is composed of independent identically distributed values having a probability distribution function given by

$$p_c(c) = p_0\delta(c) + (1-p_0)/2, \qquad c\in(-1,1),$$

i.e., that we have a high probability (p_0) of having a zero reflection coefficient and a small probability of it being chosen uniformly in the interval $(-1, 1)$. If this is our a priori information on the reflection coefficients, and the measurement of depth k yields an estimate $\hat{c}_{k+1}$ that obeys the inequality

$$|c_{k+1} - \hat{c}_{k+1}| < \varepsilon B_k,$$

then it is not difficult to see that the thresholding procedure yields a maximum a posteriori (MAP) estimate of c_{k+1}. This follows if we assume that the conditional probability of obtaining $\hat{c}_{k+1}$ as an estimate (that is, $p(\hat{c}_{k+1} | c_{k+1})$) is uniform over $(\hat{c}_{k+1} - \varepsilon B_k, \hat{c}_{k+1} + \varepsilon B_k)$. Indeed the MAP estimate is defined as the c_{k+1} value that maximizes the function

$$p(c_{k+1} | \hat{c}_{k+1}) = \frac{p(\hat{c}_{k+1} | c_{k+1}) p(c_{k+1})}{\int_{-1}^{+1} p(\hat{c}_{k+1} | c_{k+1}) p(c_{k+1})}$$

and if $p(\hat{c}_{k+1} | c_{k+1})$ is not zero at $c_{k+1} = 0$ (meaning that $0 \in (\hat{c}_{k+1} - \varepsilon B_k, \hat{c}_{k+1} + \varepsilon B_k)$), then obviously $p(c_{k+1} | \hat{c}_{k+1})$ will have its maximum at $c_{k+1} = 0$. (For a discussion of MAP estimator design see, e.g., Srinath and Rajasekaran [15].)

It is also interesting to compare the thresholding strategy with the minimum entropy deconvolution (MED) method introduced by Wiggins [17] in reflection seismology (see also Walden [16]). In this approach the discrete convolutional model of the recorded seismogram is assumed:

$$\hat{v}_k = \sum_{l=0}^{m} w_l c_{k-l} + n_k,$$

or, in vector form

$$\hat{\mathbf{v}} = \mathbf{w} * \mathbf{c} + \mathbf{n},$$

where the sequence $\{n_k\}$ represents the noise in the system. (We remark that, in contrast to the scattering model developed in this paper, the deconvolution model does not admit multiple reflections.) Thus the sequence $\{w_k\}$ represents the impulse response of a discrete filter transforming the sequence of reflection coefficients into the output sequence $\{\hat{v}_k\}$.

Now consider the formation of an approximate inverse filter with impulse response $\{f_k\}$. Namely, the convolution $\mathbf{f} * \mathbf{w}$ is to be close to the first unit coordinate vector $\mathbf{e}_0$ in an appropriate sense. Once $\mathbf{f}$ is determined we naturally take $\hat{\mathbf{c}} = \mathbf{f} * \hat{\mathbf{v}}$ as the corresponding estimate of the reflection coefficient sequence.

In the MED process, the sequence $\{f_k\}$ is determined by maximization of the varimax norm of $\hat{\mathbf{e}}$:

$$\|\hat{\mathbf{c}}\|_v = \sum_{j=1}^{N} \hat{c}_j^4 \Big/ \left(\sum_{j=1}^{N} \hat{c}_j^2 \right)^2.$$

The varimax norm is proportional to the kurtosis of a zero mean process, which is a statistic that characterizes the peakedness of the corresponding probability density function (Donoho [4]). Thus maximizing the varimax norm results in suppressing most of the reflection coefficients in favor of a few large ones. This, of course, is exactly the idea behind our thresholding strategy.

It is now widely accepted that the MED processes do not perform to expectations (see Wiggins [18], for example). One of the main reasons is that the optimization criteria reduce to a highly nonlinear system of equations whose solution is approximated iteratively by local linearizations. The convergence of those iterations is problematic, in particular, because of the nonuniqueness of the local maxima.

The method of inverse scattering with thresholding does not seem to have this disadvantage. The reflectivity information recovered by this algorithm is reliable and, as the numerical experiments of Bruckstein, Koltracht, and Kailath [2], Koltracht and Lancaster [8], and the following section demonstrate, the thresholding strategy efficiently suppresses noise magnification in inverse scattering algorithms.

4. Numerical examples. The effects of the thresholding strategy are illustrated first on a synthetic reflectivity profile shown in Fig. 1. In all of the figures the vertical scale

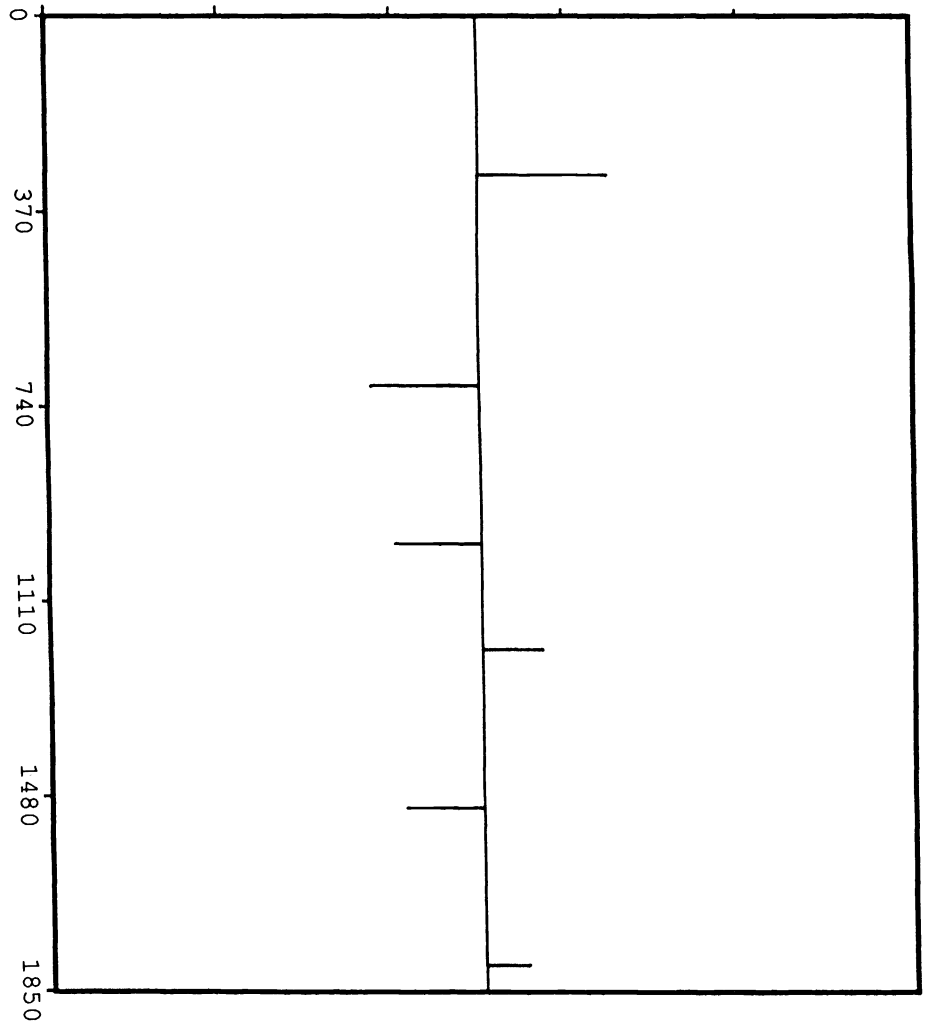


FIG. 1. *Synthetic reflectivity profile.*

denotes depth measured in the number of horizontal layers. A recursive algorithm described in Koltracht and Lancaster [7] is used to generate the "marine" response r_1, r_2, $\cdots$, r_N of a medium corresponding to this profile. As soon as a new entry r_k in the response sequence is obtained, some noise value ε_k, chosen randomly from the interval $[-\varepsilon, \varepsilon]$, is added to r_k. Since $\hat{r}_k = r_k + \varepsilon_k$ is used for the computation of $r_{k+1}, \cdots, r_N$, in the formula

$$r_{k+1} = -\left(c_{k+1} + \sum_{j=0}^{k-1} \hat{r}_{j+1}\gamma_k(j)\right) \Big/ \gamma_k(k),$$

it follows that the perturbation ε_k affects all following entries of the response sequence (a phenomenon to be expected in real-life situations). Reconstruction of the reflectivity profile with and without thresholding, as well as the corresponding "marine" responses, are shown in the following diagrams. In Fig. 2(a), the marine response corresponding to a noise level $\varepsilon = 0.02$ is presented. Figures 2(b) and 2(c) show the reconstruction without and with thresholding, respectively. We see that the thresholding algorithm gives a perfect reconstruction, whereas the algorithm without thresholding hardly reconstructs the second reflector at the depth of 700 and breaks down soon after that.

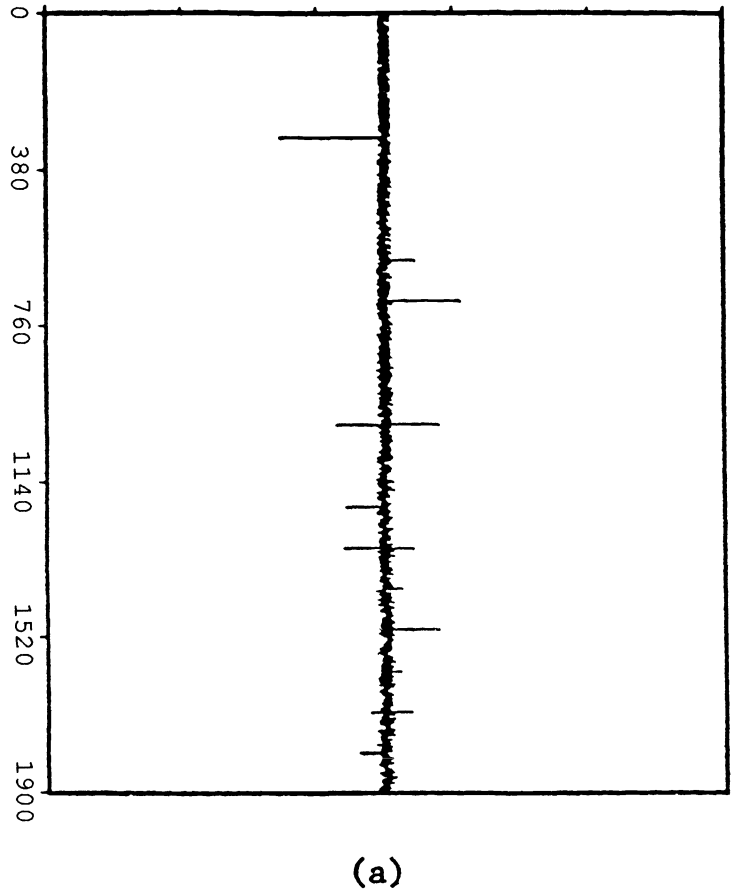
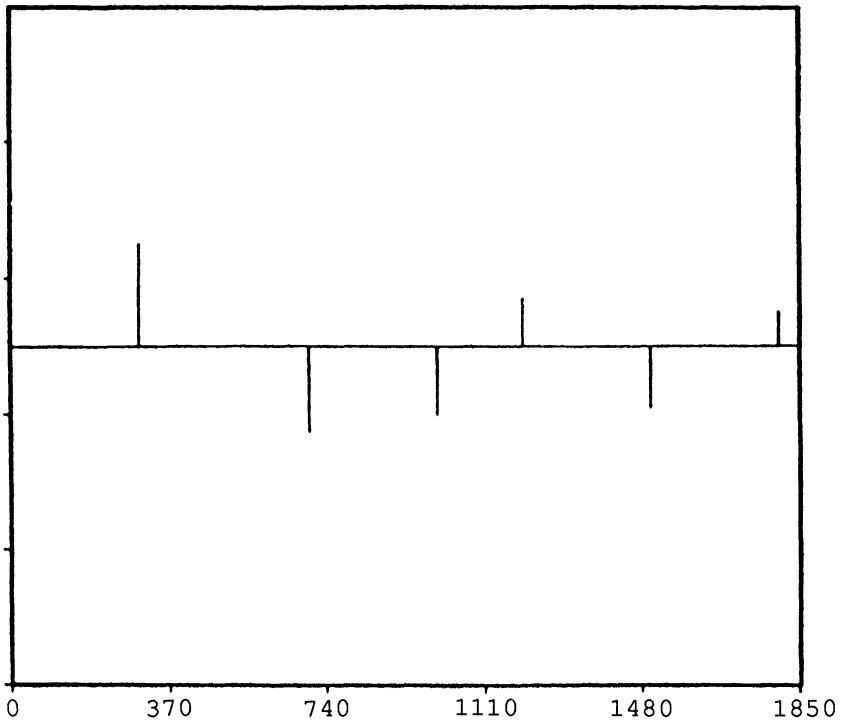
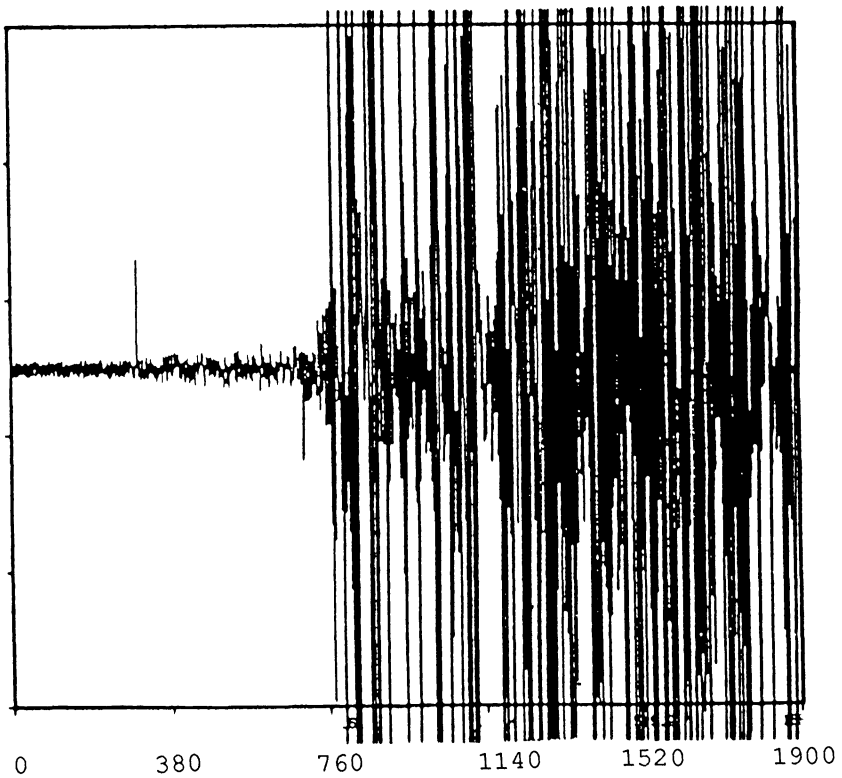


FIG. 2(a). "*Marine*" *response perturbed by noise of level* 0.02.



(c)

FIG. 2(c). *Reconstruction with threshold barrier* $\varepsilon = 0.02$.



(b)

FIG. 2(b). *Reconstruction without threshold.*

In Fig. 3(a) the marine response corresponding to the noise level $\varepsilon = 0.03$ is presented; Figures 3(b) and 3(c) show the reconstruction without and with thresholding, respectively. Again, the algorithm without thresholding breaks down before producing any reliable information, whereas the threshold algorithm recovers four out of six reflection coefficients.

In Fig. 4 we observe the effect of changing the noise barrier ε in the threshold reconstruction. This observation is important because in real-life situations, we cannot expect to have exact knowledge of ε, but rather some estimate of it. The marine response corresponding to $\varepsilon = 0.03$ (the same as in Fig. 3) is used. In Figs. 4(a) and 4(b) threshold reconstructions with $\varepsilon = 0.025$ and $\varepsilon = 0.005$, respectively, are presented. In Fig. 4(a) the fifth true reflector is recovered. The reconstruction does not change for gradually decreasing values of ε, until for $\varepsilon = 0.005$ a small ghost reflector appears just above the depth of 700. This is apparently a result of some noise going through at shallow depths. It appears to be encouraging that the reconstruction with imprecise noise levels reveals more information than the reconstruction with exactly known ε. Indeed, in practical situations (see Koltracht and Lancaster [8]) we must experiment with the noise barrier ε, which can only be roughly estimated in advance.

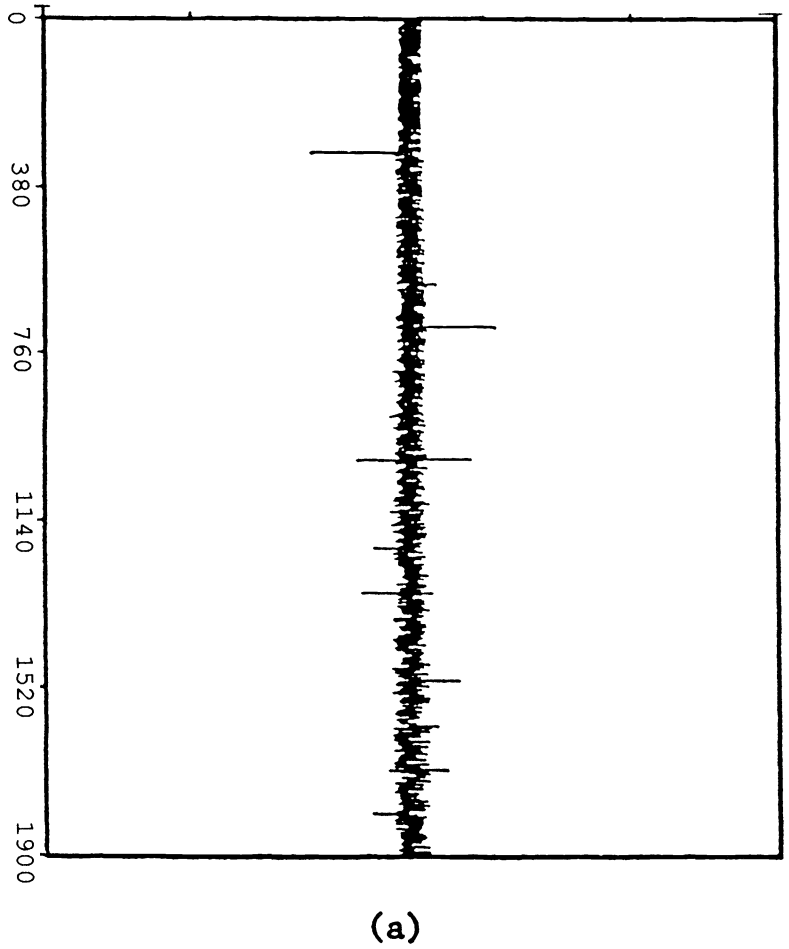
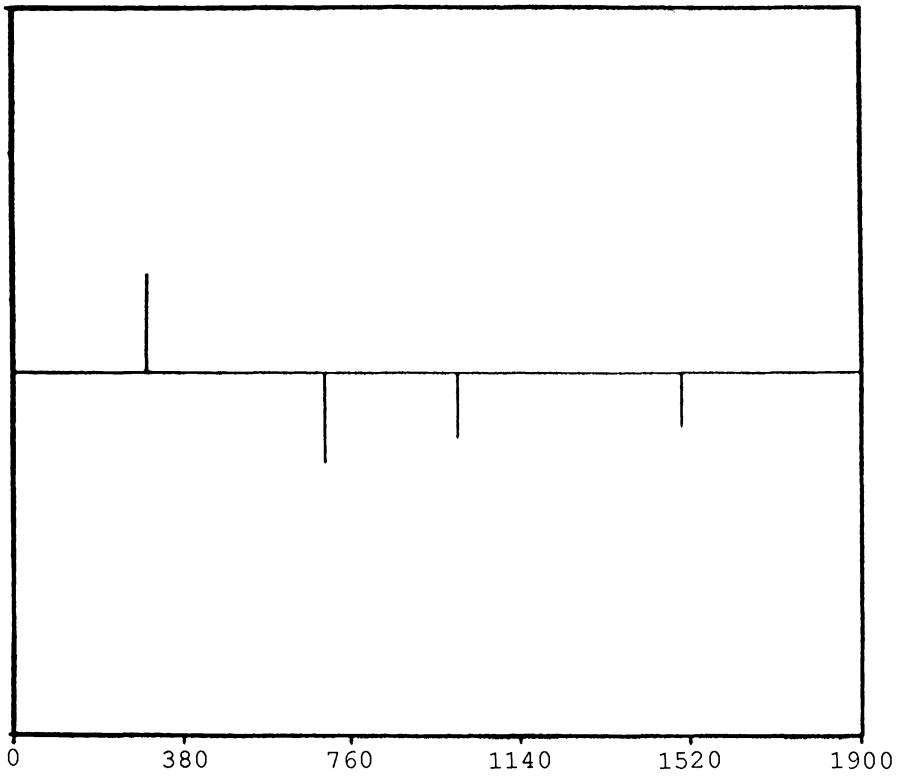
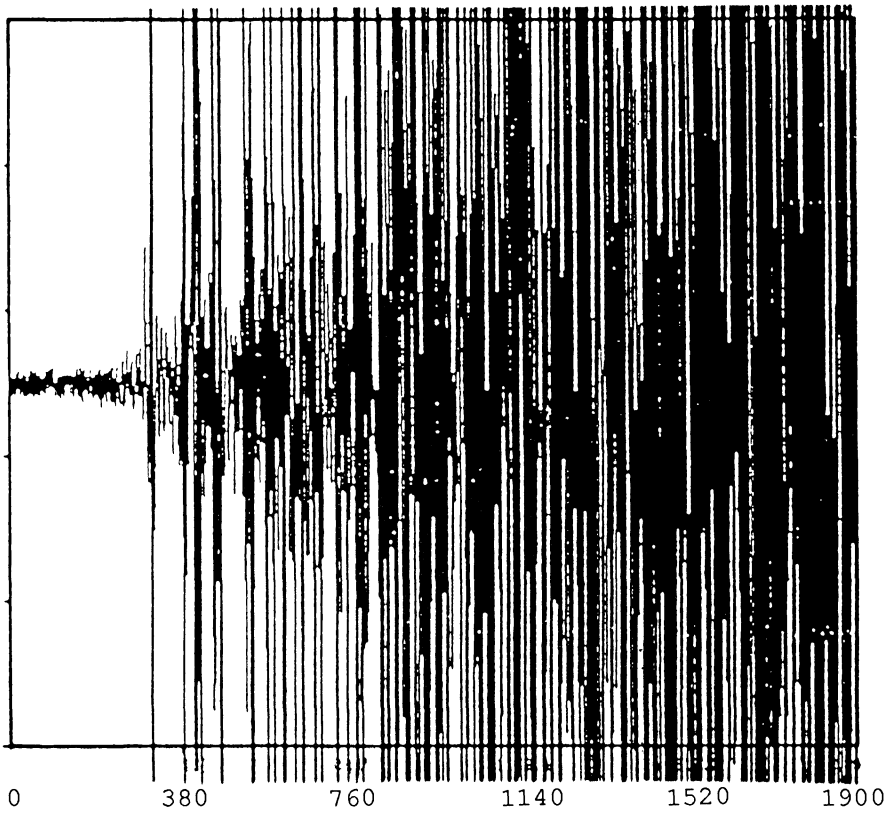


FIG. 3(a). "*Marine*" *response perturbed by noise of level* 0.03.



(c)

FIG. 3(c). *Reconstruction with threshold barrier* $\varepsilon = 0.03$.



(b)

FIG. 3(b). *Reconstruction without threshold.*

The effect of thresholding strategy is illustrated also on a set of field data. This data comes from a geophysical survey in northern Canada. (We thank K. Coffin, Department of Geology and Geophysics, University of Calgary, Calgary, British Columbia, Canada, for making this data available to us.) The data consists of about 100 traces (of 2,000 samples each) of unfiltered CDP-stack data along a horizontal survey line, as shown in Fig. 6(a). The reconstruction without thresholding breaks down at depth of about 130, as shown in Fig. 5. The reconstruction using threshold algorithms with appropriately chosen ε is shown in Fig. 6(b). It appears that the section indeed has some multiple reflections, which are eliminated with the threshold reconstruction.

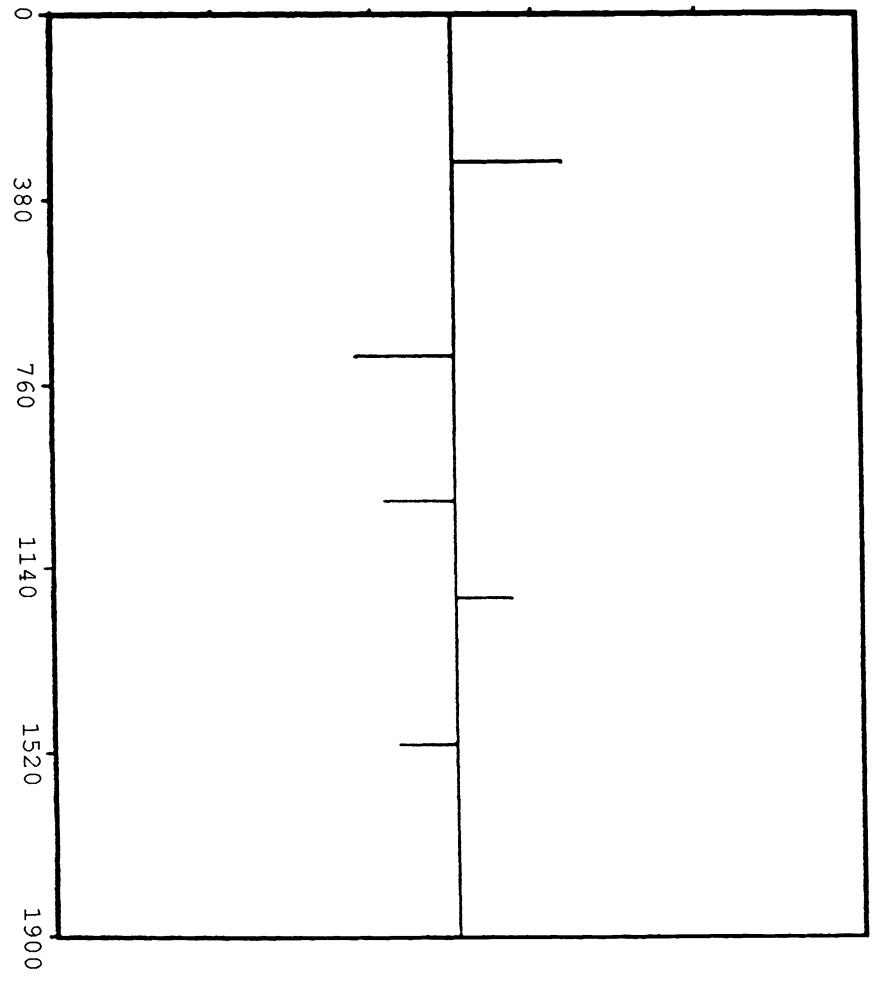


FIG. 4(a). *Reconstruction for response of Fig.* 3 *with threshold barrier* $\varepsilon = 0.025$.

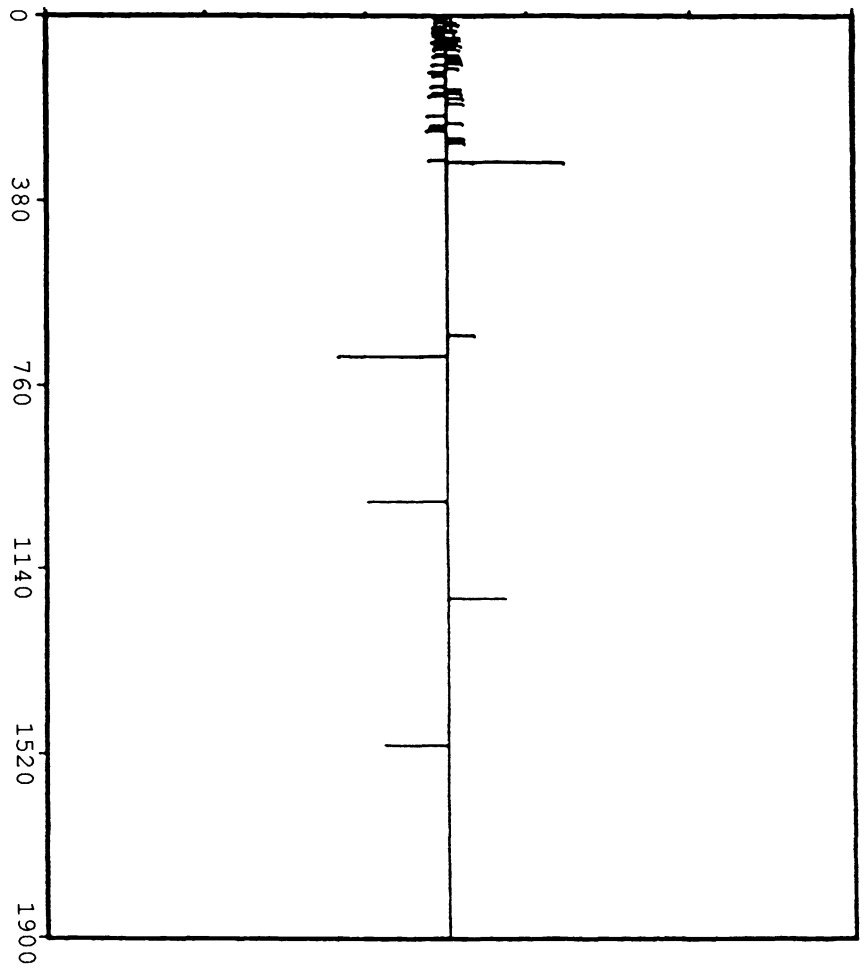
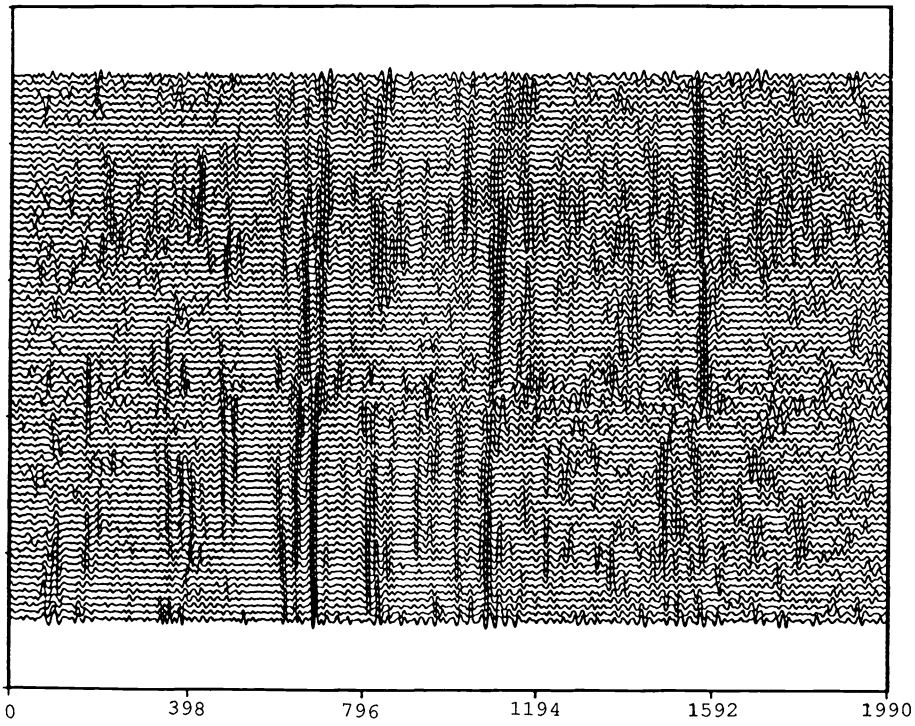


FIG. 4(b). *Reconstruction for response of Fig.* 3 *with threshold barrier* $\varepsilon = 0.005$.

5. Concluding remarks. An inverse scattering method that is stable in the presence of noise has been described. The method is based on a thresholding strategy that predicts in a statistically reliable way when small reflection coefficients are to be set to zero. Statistical interpretation of the strategy in terms of maximum a posteriori estimation has been presented. The procedure has been developed for an extended Goupillaud model of a layered medium in which the reflection coefficient characterizing the surface is a parameter. The theoretical basis for the method has been described and developed and favorable performance has been demonstrated using both synthetic and field data.



(a)

FIG. 6(a). *Seismic section consisting of new traces of unfiltered* CDP-*stack.*

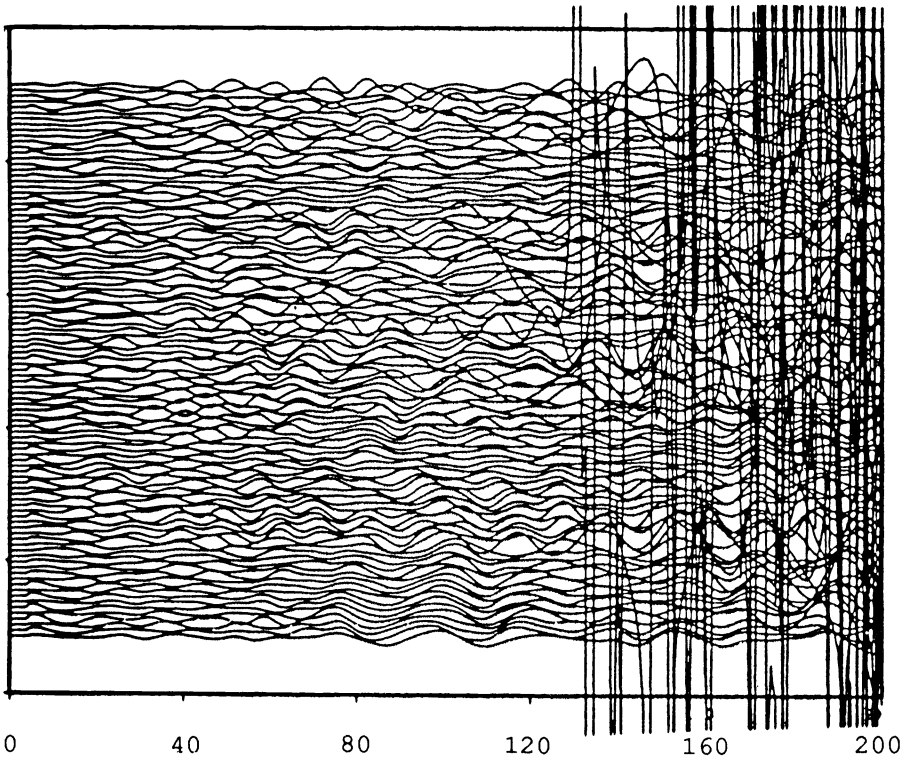
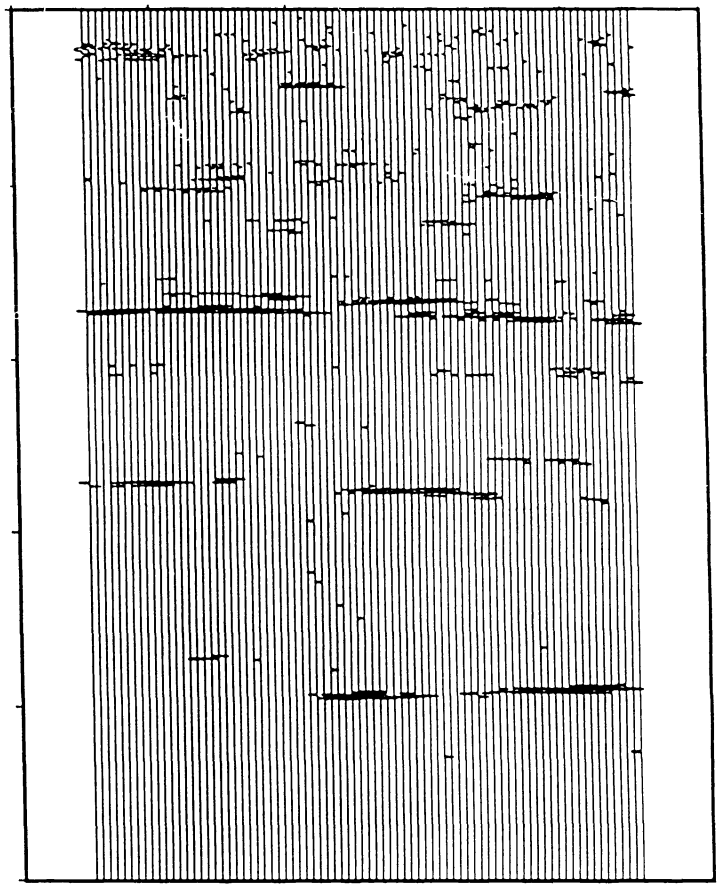


FIG. 5. *Reconstruction of seismic section of Fig.* 6(a) *without threshold.*



(b)

FIG. 6(b). *Inversion of the seismic section with the threshold algorithm.*

REFERENCES

[1] K. AKI AND P. G. RICHARDS, *Quantitative Seismology*, 2, W. H. Freeman, San Francisco, CA, 1980.

[2] A. BRUCKSTEIN, I. KOLTRACHT, AND T. KAILATH, *Inverse scattering with noisy data*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1331–1349.

[3] J. F. CLAERBOUT, *Fundamentals of Geophysical Data Processing*, McGraw-Hill, New York, 1971.

[4] D. DONOHO, *On minimum entropy deconvolution*, in Applied Time Series Analysis II, D. Findley, ed., Academic Press, New York, 1981.

[5] R. G. FERBER, *Stabilization of normal-incidence seismogram inversion removing the noise induced bias*, Geophys. Prospecting, 33 (1985), pp. 212–223.

[6] T. KAILATH, A. BRUCKSTEIN, AND D. MORGAN, *Fast matrix factorization via discrete transmission lines*, Linear Algebra Appl., 75 (1986), pp. 1–25.

[7] I. KOLTRACHT AND P. LANCASTER, *Condition Numbers of Toeplitz and Block Toeplitz Matrices*, Operator Theory Advances and Applications, 18, Birkhäuser-Verlag, Basel, Switzerland, 1986.

[8] ———, *Threshold algorithms for the prediction of reflection coefficients in a layered medium*, Geophys., 53 (1988), pp. 908–919.

[9] ———, *Generalized Schur parameters and effects of perturbations*, Linear Algebra Appl., 105 (1988), pp. 109–129.

[10] G. KUNETZ, *Generalisation des operateurs d'antiresonance a un nombre quelconque de reflecteurs*, Geophys. Prospecting, 12 (1964), pp. 283–289.
[11] H. LEV-ARI AND T. KAILATH, *Lattice filter parametrization and modelling of nonstationary processes*, IEEE Trans. Inform. Theory, 30 (1984), pp. 2–16.
[12] ———, *Triangular factorization of structured Hermitian matrices*, Operator Theory Advances and Applications, 18, Birkhäuser-Verlag, Basel, Switzerland, 1986, pp. 301–324.
[13] E. A. ROBINSON AND S. TREITEL, *Geophysical Signal Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1981.
[14] E. A. ROBINSON, *Seismic inversion and deconvolution, Part* A, *classical methods*, Seismic Exploration, 4A, Geophysical Press, London, Amsterdam, 1984.
[15] M. D. SRINATH AND P. K. RAJASEKARAN, *An Introduction to Statistical Signal Processing with Applications*, John Wiley, New York, 1979.
[16] A. T. WALDEN, *Non-Gaussian reflectivity, entropy and deconvolution*, Geophys., 50 (1985), pp. 2862–2888.
[17] R. A. WIGGINS, *Minimum entropy deconvolution*, Geoexploration, 16 (1978), pp. 21–35.
[18] ———, *Entropy guided deconvolution*, Geophys., 50 (1985), pp. 2720–2726.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 1, pp. 41–48, January 1991

005

COMPUTATION OF THE EULER ANGLES OF A SYMMETRIC 3 × 3 MATRIX*

ADAM W. BOJANCZYK† AND ADAM LUTOBORSKI‡

Abstract. Closed form formulas for computing the eigenvectors of a symmetric 3×3 matrix are presented. The matrix of the eigenvectors is computed as a product of three rotations through Euler angles. The formulas require approximately 90 arithmetic operations, six trigonometric evaluations, and two root evaluations. These formulas may be applied as a subroutine in a parallel one-sided Jacobi-type method in which three rather than two columns, as is the case in the standard Jacobi method, are operated on in each step.

Key words. Euler angles, eigenvectors

AMS(MOS) subject classifications. 15A18, 65H15

Introduction. In this paper we derive closed form formulas for computing a diagonalizing rotation matrix for a given symmetric 3×3 matrix. A standard result in the representation of the group of rotations shows that the diagonalizing matrix may be represented as a product of three plane rotations—the angles of rotations are known as Euler angles. This representation leads to a system of trigonometric equations involving rotation angles, which due to their special form can be reduced to a scalar cubic equation in cotangent of one of the angles. Thus we may use the trigonometric form of the Cardano formulas to compute all real solutions of the cubic.

The overall cost of computing eigenvectors of a 3×3 symmetric matrix from closed form formulas is approximately 90 arithmetic operations, six trigonometric evaluations, and two root evaluations. When only an approximate eigendecomposition is sought this cost may be higher than the one required by methods such as the QR or Jacobi algorithm. However, in contrast to iterative methods that, although numerically very efficient, can only produce approximate eigendecomposition, the formulas presented in this paper are of closed form.

The formulas could be utilized as a subroutine in a parallel one-sided Jacobi-type method in which three rather than two columns, as is the case in the traditional Jacobi method, are operated on in each basic step. This approach leads to open problems concerning the rate of convergence of these types of Jacobi methods and will be investigated in a forthcoming paper. It should be noted that a closed formula for the eigenvalues (but not eigenvectors) of a symmetric 3×3 matrix was given previously by Smith [Sm].

The paper is organized as follows. In § 1 preliminaries on orthogonal diagonalization are given. In § 2 the closed formulas for computing the diagonalizing rotation matrix are derived. Section 3 contains results of numerical tests.

1. Preliminaries on orthogonal diagonalization. Let $A = [a_{ij}]_{1 \le i,j \le 3}$ be a real, symmetric 3×3 matrix. Due to the spectral theorem [St, pp. 309–311], A is diagonalizable: there exists an orthogonal matrix $Q = [q_1, q_2, q_3]$ such that

$$Q^\top A Q = \Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3), \tag{1.1}$$

* Received by the editors February 27, 1989; accepted for publication (in revised form) October 17, 1989. This work was partially supported by the U.S. Army Research Office through the Mathematical Sciences Institute, Cornell University, Ithaca, New York 14853.

† Department of Electrical Engineering, Cornell University, Ithaca, New York 14853-5401 (ADAMB@TESLA.EE.CORNELL.EDU).

‡ Department of Mathematics, Syracuse University, Syracuse, New York 13244-1150 (ALUTOBOR@SUNRISE.ACS.SYR.EDU).

where $\lambda_1 \leqq \lambda_2 \leqq \lambda_3$ are the eigenvalues of the matrix A and the columns q_1, q_2, q_3 of Q are the orthonormal eigenvectors associated with these eigenvalues.

DEFINITION 1. An orthogonal matrix Q such that $\det Q = 1$ will be called a rotation matrix.

DEFINITION 2. Let $1 \leqq p < r \leqq 3$ and ϕ be a real number. An orthogonal 3×3 matrix $Q_{pr}(\phi) = [q_{ij}]_{1 \leqq i,j \leqq 3}$ given by

$$q_{pp} = q_{rr} = \cos \phi,$$

$$q_{ii} = 1 \quad \text{if } i \neq p, r,$$

$$q_{pr} = -q_{rp} = -\sin \phi,$$

$$q_{ip} = q_{pi} = q_{ir} = q_{ri} = 0 \quad \text{if } i \neq p, r$$

$$q_{ij} = 0 \quad \text{if } i \neq p, r \text{ and } j \neq p, r,$$

will be called a plane rotation through ϕ in the plane span (e_p, e_r).

Our objective is: given a symmetric 3×3 matrix A, construct a diagonalizing rotation matrix $Q_\sigma = [q_{\sigma(1)}, q_{\sigma(2)}, q_{\sigma(3)}]$ such that

$$Q_\sigma^\top A Q_\sigma = \Lambda_\sigma = \operatorname{diag}(\lambda_{\sigma(1)}, \lambda_{\sigma(2)}, \lambda_{\sigma(3)}), \tag{1.2}$$

where σ belongs to the set Σ_3 of permutations of (1, 2, 3).

As we recall in the following lemma, a rotation in R^3 may be represented as a product of three plane rotations through the Euler angles ϕ, θ, ψ.

LEMMA 1. *Let $Q = [q_{ij}]_{1 \leqq i,j \leqq 3}$ be a rotation matrix. Then there exist angles ϕ in $[0, \pi)$ and θ, ψ in $(-\pi, \pi]$ called the Euler angles of Q such that*

$$Q = Q_{12}(\phi) Q_{23}(\theta) Q_{12}(\psi). \tag{1.3}$$

Proof. The geometric definition of Euler angles (see [GMS, p. 5]) is given in Fig. 1. The algebraic proof of (1.3) is simply the QR factorization of our rotation matrix. Set $Q_1 = Q$, $Q_2 = Q_{12}(-\phi)Q_1$, $Q_3 = Q_{23}(-\theta)Q_2$, $Q_4 = Q_{12}(-\psi)Q_3$ where $Q_k = [q_{ij}^k]_{1 \leqq i,j \leqq 3}$. We then choose ϕ, θ, ψ to be the numbers that subsequently annihilate q_{13}^2, q_{23}^3, q_{12}^4, that is, such that $\cot \phi = -q_{23}^1/q_{13}^1$, $\cot \theta = -q_{33}^2/q_{23}^2$, $\cot \psi = -q_{22}^3/q_{12}^3$. Q_4 is a rotation lower triangular matrix and hence it is the identity matrix. Since $Q_{pr}(-\psi) = Q_{pr}(\psi)^{-1}$ we obtain (1.3). □

In Fig. 1, we assume that $q_3 \neq e_3$, or in other words that $\theta \neq 0$. If $\theta = 0$, then one of the remaining angles may be arbitrary. The rotation matrices $Q_{12}(\phi)$, $Q_{23}(\theta)$, and $Q_{12}(\psi)$ are called proper rotation, nutation, and precession matrices, respectively.

DEFINITION 3. Let A be a symmetric 3×3 matrix. Any angles ϕ, θ, ψ for which there exists a rotation matrix Q, $Q = Q_{p_1r_1}(\phi)Q_{p_2r_2}(\theta)Q_{p_3r_3}(\psi)$, $1 \leqq p_i, r_i \leqq 3$, $1 \leqq i \leqq 3$ that diagonalizes A are called Euler angles of the matrix A.

Remark 1. Let Q be a rotation matrix. Then there exist angles ϕ, θ, ψ in $(-\pi/4, \pi/4]$ and a rotation matrix $\hat{Q} = Q_{p_1r_1}(\phi)Q_{p_2r_2}(\theta)Q_{p_3r_3}(\psi)$ where $(p_1, r_1, p_2, r_2, p_3, r_3) \in \{(1, 2, 2, 3, 1, 2), (1, 2, 2, 3, 1, 3), (1, 2, 1, 3, 1, 2), (1, 2, 1, 3, 2, 3)\}$ such that Q may be obtained from $\hat{Q}$ by a permutation of columns and multiplication of some columns by -1.

Factorization of rotation matrices Q into Euler rotation form (1.3) has long been known and widely used in mechanics, especially in the theory of angular momentum (see [Ro]), and in algebra in the theory of representation of the rotation group SO(3) (see [GMS]).

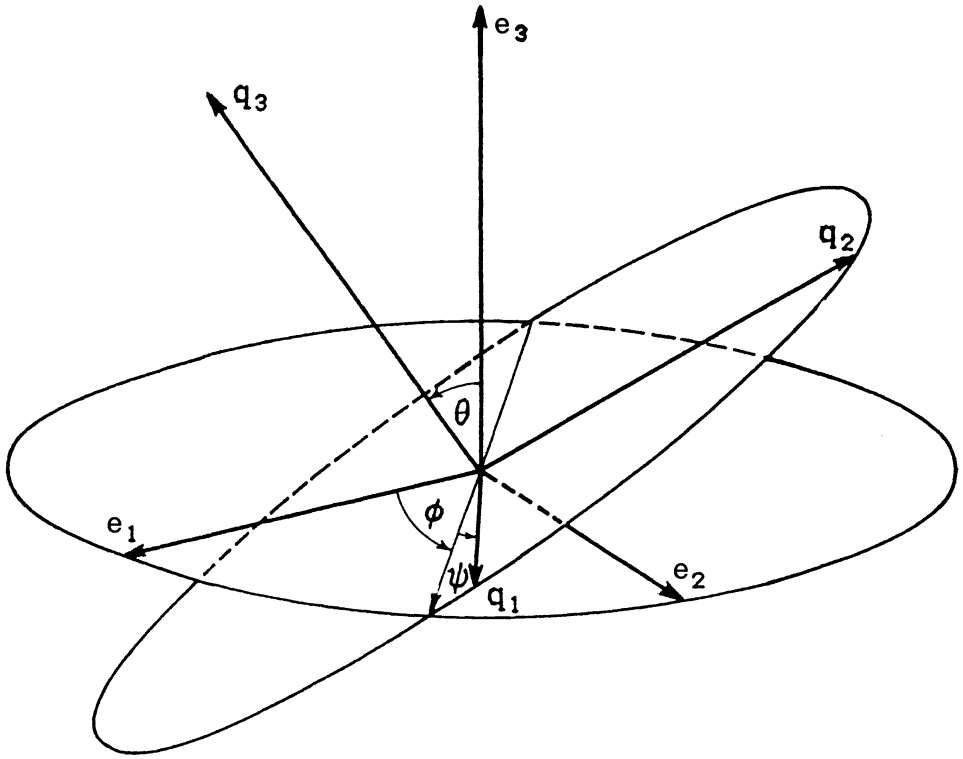


FIG. 1. *The Euler angles* ϕ, θ, ψ.

2. Computation of the Euler angles of a matrix. Our objective is to compute Euler angles ϕ, θ, ψ of a given symmetric 3×3 matrix A.

We denote

$$Q_{12}(\phi)^{\top} A Q_{12}(\phi) = B. \tag{2.1}$$

$$b_{11} = a_{11}\cos^2\phi + 2a_{12}\sin\phi\cos\phi + a_{22}\sin^2\phi = a_{11} + a_{22} - b_{22}, \tag{2.2}$$

$$b_{12} = (a_{22} - a_{11})\sin\phi\cos\phi + a_{12}(\cos^2\phi - \sin^2\phi), \tag{2.3}$$

$$b_{13} = a_{13}\cos\phi + a_{23}\sin\phi, \tag{2.4}$$

$$b_{22} = a_{11}\sin^2\phi - 2a_{12}\sin\phi\cos\phi + a_{22}\cos^2\phi, \tag{2.5}$$

$$b_{23} = -a_{13}\sin\phi + a_{23}\cos\phi, \tag{2.6}$$

$$b_{33} = a_{33}. \tag{2.7}$$

Next, we define matrix C as

$$Q_{23}(\theta)^{\top} Q_{12}(\phi)^{\top} A Q_{12}(\phi) Q_{23}(\theta) = Q_{23}(\theta)^{\top} B Q_{23}(\theta) = C. \tag{2.8}$$

$$c_{11} = a_{11} + a_{22} - b_{22}, \tag{2.9}$$

$$c_{12} = b_{12}\cos\theta + b_{13}\sin\theta, \tag{2.10}$$

$$c_{13} = -b_{12}\sin\theta + b_{13}\cos\theta, \tag{2.11}$$

$$c_{22} = b_{22}\cos^2\theta + 2b_{23}\sin\theta\cos\theta + a_{33}\sin^2\theta, \tag{2.12}$$

$$c_{23} = (a_{33} - b_{22})\sin\theta\cos\theta + b_{23}(\cos^2\theta - \sin^2\theta), \tag{2.13}$$

$$c_{33} = b_{22}\sin^2\theta - 2b_{23}\sin\theta\cos\theta + a_{33}\cos^2\theta. \tag{2.14}$$

Last, due to the spectral theorem

$$Q_{12}(\psi)^{\top} C Q_{12}(\psi) = \Lambda_{\sigma}. \tag{2.15}$$

Computation of the proper rotation and nutation angles ϕ and θ. The last column of $CQ_{12}(\psi)$ is equal to the last column of C. From (2.15) we know that

$$Q_{12}(\psi)^{\top}\begin{bmatrix} c_{13} \\ c_{23} \\ c_{32} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \lambda_{\sigma(3)} \end{bmatrix}.$$

The rotation $Q_{12}^{\top}(\psi)$ does not change the last component of a vector and therefore $c_{33} = \lambda_{\sigma(3)}$. Since $Q_{12}^{\top}(\psi)$ is an isometry

$$\begin{cases} c_{13} = 0, & (2.16) \\ c_{23} = 0. & (2.17) \end{cases}$$

Using (2.11)–(2.13), we may write (2.16), (2.17) as

$$\begin{cases} -b_{12}\sin\theta + b_{13}\cos\theta = 0, & (2.18) \\ \frac{1}{2}(a_{33} - b_{22})\sin 2\theta + b_{23}\cos 2\theta = 0. & (2.19) \end{cases}$$

Due to its very special form, the above system of trigonometric equations with respect to ϕ and θ can be reduced to a single trigonometric equation in ϕ. We exclude momentarily the degenerate matrices A that yield one or more of the following cases: $\sin\theta = 0$, $\sin 2\theta = 0$, $b_{13} = 0$, $b_{23} = 0$, since then the system uncouples and may be solved by substitution. Under this assumption we obtain that

$$\begin{cases} \cot\theta = b_{12}/b_{13}, & (2.20) \\ \cot 2\theta = (b_{22} - a_{33})/(2b_{23}), & (2.21) \end{cases}$$

for $\theta \in (-\pi/2, 0) \cup (0, \pi/2)$.

Substituting $\cot\theta$ from (2.20) into (2.21) we obtain a scalar trigonometric equation for ϕ,

$$(b_{22} - a_{33})b_{12}b_{13} + b_{23}(b_{13}^2 - b_{12}^2) = 0, \tag{2.22}$$

where $b_{ij} = b_{ij}(\phi)$ are given in (2.2)–(2.7). The explicit form of (2.22) in terms of the entries of A is

$$\begin{aligned} F(\phi) = {} & (a_{12}a_{23}a_{33} - a_{13}a_{23}^2)s^3 + (a_{11}a_{23}a_{33} \\ & - a_{22}a_{23}a_{33} + a_{12}a_{13}a_{33} + a_{23}^3 - 2a_{13}^2a_{23})s^2c \\ & + (a_{11}a_{13}a_{33} - a_{13}a_{22}a_{33} - a_{12}a_{23}a_{33} - a_{13}^3 + 2a_{13}a_{23}^2)sc^2 \\ & + (-a_{12}a_{13}a_{33} + a_{13}^2a_{23})c^3 + (-a_{11}a_{12}a_{23} + a_{12}^2a_{13})s^5 \\ & + (a_{11}a_{22}a_{23} - a_{11}^2a_{23} + a_{11}a_{12}a_{13} - 2a_{12}a_{13}a_{22} + a_{12}^2a_{23})s^4c \\ & + (a_{11}a_{12}a_{23} - a_{11}a_{13}a_{22} - a_{12}a_{22}a_{23} + a_{13}a_{22}^2)s^3c^2 \\ & + (a_{11}a_{22}a_{23} - a_{11}^2a_{23} - a_{12}a_{13}a_{22} + a_{11}a_{12}a_{13})s^2c^3 \end{aligned} \tag{2.23}$$

$$+(a_{13}a_{22}^2-a_{11}a_{13}a_{22}-a_{12}a_{22}a_{23}-a_{12}^2a_{13}+2a_{11}a_{12}a_{23})sc^4$$
$$+(a_{12}a_{13}a_{22}-a_{12}^2a_{23})c^5=0,$$

where $s = \sin\phi$ and $c = \cos\phi$. To simplify (2.23) we set

(2.24) $$d=a_{22}-a_{11},$$

(2.25) $$e=a_{12}a_{33}-a_{13}a_{23},$$

(2.26) $$f=a_{11}a_{33}+a_{23}^2-a_{22}a_{33}-a_{13}^2,$$

(2.27) $$h=a_{11}a_{23}-a_{12}a_{13},$$

(2.28) $$k=a_{13}a_{22}-a_{12}a_{23},$$

(2.29) $$y=a_{23}e-a_{12}h,$$

(2.30) $$z=a_{12}k-a_{13}e.$$

We note that F splits into two parts:

$$F(\phi)=s[a_{23}es^2+(a_{13}f-a_{23}e)c^2-a_{12}hs^4+dks^2c^2+(dk+a_{12}h)c^4]$$
$$+c[(a_{23}f+a_{13}e)s^2-a_{13}ec^2+(dh-a_{12}k)s^4+dhs^2c^2+a_{12}kc^4].$$

Eliminating c from the first part and s from the second, we obtain

(2.31) $$F(\phi)=s[a_{13}f+dk-y+(2y-a_{13}f-dk)s^2]$$
$$+c[a_{23}f+dh-z+(2z-a_{23}f-dh)c^2]=0.$$

Upon dividing (2.31) by $s \neq 0$ our resulting trigonometric equation is

(2.32) $$z\cot^3\phi+(a_{13}f+dk-y)\cot^2\phi+(a_{23}f+dh-z)\cot\phi+y=0.$$

We may now use the trigonometric form of the Cardano formulas [KK, § 1.8-4] to compute all real solutions of (2.32), which is cubic with respect to $\cot\phi$. If $z \neq 0$ we set

(2.33) $$a=\frac{a_{13}f+dk-y}{z},$$

(2.34) $$b=\frac{a_{23}f+dh-z}{z},$$

(2.35) $$c=\frac{y}{z}.$$

The standard change of variable

(2.36) $$\cot\phi=\beta-\frac{a}{3},$$

transforms (2.32) to the "reduced" form

(2.37) $$\beta^3+p\beta+q=0,$$

where

(2.38) $$p=b-\frac{a^2}{3},$$

$$q=\frac{2}{27}a^2-\frac{1}{3}ab+c. \tag{2.39}$$

We set

$$D=\frac{p^3}{27}+\frac{q^2}{4}, \tag{2.40}$$

and consider two possible cases of closed form formulas for all the solutions of (2.37).

If $D<0$, then

$$\beta_1=2\sqrt{-p/3}\cos\left(\frac{\tilde{\beta}}{3}\right), \tag{2.41}$$

$$\beta_{2,3}=2\sqrt{-p/3}\cos\left(\frac{\tilde{\beta}\pm\pi}{3}\right), \tag{2.42}$$

where

$$\cos\tilde{\beta}=-\frac{q}{2\sqrt{-(p/3)^3}}. \tag{2.43}$$

If $D\geqq 0$ and $p>0$, then

$$\beta=-2\sqrt{p/3}\cot 2\tilde{\beta}, \tag{2.44}$$

where

$$\tan\tilde{\beta}=\sqrt[3]{\tan(\tilde{\tilde{\beta}}/2)} \quad \text{and} \quad |\tilde{\beta}|\leqq\frac{\pi}{4}, \tag{2.45}$$

$$\tan\tilde{\tilde{\beta}}=\frac{2}{q}\sqrt{(p/3)^3} \quad \text{and} \quad |\tilde{\tilde{\beta}}|\leqq\frac{\pi}{2}. \tag{2.46}$$

If $D\geqq 0$ and $p<0$, then

$$\beta=-2\sqrt{-p/3}\csc 2\tilde{\beta}, \tag{2.47}$$

where

$$\tan\tilde{\beta}=\sqrt[3]{\tan(\tilde{\tilde{\beta}}/2)} \quad \text{and} \quad |\tilde{\beta}|\leqq\frac{\pi}{4}, \tag{2.48}$$

$$\sin\tilde{\tilde{\beta}}=\frac{2}{q}\sqrt{(-p/3)^3} \quad \text{and} \quad |\tilde{\tilde{\beta}}|\leqq\frac{\pi}{2}. \tag{2.49}$$

From (2.21) we compute θ by substituting ϕ given by (2.32) in (2.5), (2.6).

Computation of the precession angle ψ. From (2.15) $(\Lambda)_{12}=\frac{1}{2}(c_{22}-c_{11})\sin 2\psi+c_{12}\cos 2\psi=0$. If $c_{12}\neq 0$, then we obtain that

$$\cot 2\psi=(c_{11}-c_{22})/(2c_{12}), \tag{2.50}$$

where $-\pi/4<\psi\leqq\pi/4$.

In the following theorem we summarize our algorithm.

THEOREM 1. *Let $A=[a_{ij}]_{1\leqq i,j\leqq 3}$ be a real symmetric matrix. Then there exist Euler angles ϕ in $(-\pi,\pi]$, θ, ψ in $(-\pi/4,\pi/4]$ and a rotation matrix Q*

$$Q=Q_{12}(\phi)Q_{23}(\theta)Q_{12}(\psi),$$

such that $Q^{\top}AQ=\operatorname{diag}(\lambda_{\sigma(1)},\lambda_{\sigma(2)},\lambda_{\sigma(3)})$ where $\lambda_1\leqq\lambda_2\leqq\lambda_3$ are the eigenvalues of A

and $\sigma \in \Sigma_3$. *For nontrivial* A *with at most two off-diagonal entries equal to zero, the angles* ϕ, θ, ψ *can be computed as follows*:

Degenerate cases. If $a_{12} = 0$, $a_{11} = a_{22}$ *or* $a_{ij} \neq 0$, $1 \leqq i < j \leqq 3$ *and* $a_{13}a_{23}(a_{11} - a_{22}) = a_{12}(a_{13}^2 - a_{23}^2)$, *then* $\cot \phi = -a_{23}/a_{13}$, $\cot 2\theta = (b_{22} - a_{33})/(2b_{23})$ *and* ψ *is given by* (2.50). *If* $a_{23} = 0$, $a_{22} = a_{33}$ *or* $a_{ij} \neq 0$, $1 \leqq i < j \leqq 3$ *and* $a_{12}a_{13}(a_{22} - a_{33}) = a_{23}(a_{12}^2 - a_{13}^2)$, *then* $\phi = 0$, $\cot \theta = a_{12}/a_{13}$ *and* ψ *is given by* (2.50). *If* $a_{23} \neq 0$ *and* $a_{33} = b_{22}(\cot^{-1}(a_{13}/a_{23}))$, *then* $\cot \phi = a_{13}/a_{23}$, $\cot \theta = b_{12}/b_{13}$ *and* ψ *is given by* (2.50).

General case. ϕ, $-\pi/2 < \phi \leqq \pi/2$ *is the solution of* (2.32) *given by* (2.41), (2.36) *or* (2.44), (2.36) *or* (2.47), (2.36). θ, $-\pi/4 < \theta \leqq \pi/4$ *is given by* (2.21). ψ, $-\pi/4 < \psi \leqq \pi/4$ *is given by* (2.50). □

The computation of the proper rotation angle ϕ requires approximately 60 flops and depending on the matrix, two trigonometric evaluation and two square roots or three trigonometric evaluations, one square, and one cube root. Additional 12 flops and three trigonometric evaluations are necessary to compute the nutation angle θ and additional 24 flops and one trigonometric evaluation to compute the precession angle ψ. The total cost of computing Euler angles is approximately 90 flops, six trigonometric evaluations, and two root evaluations.

From Remark 1 we infer that for a given symmetric matrix A there exists a diagonalizing rotation matrix $\hat{Q} = Q_{p_1r_1}(\phi)Q_{p_2r_2}(\theta)Q_{p_3r_3}(\psi)$ with the Euler angles ϕ, θ, ψ in $(-\pi/4, \pi/4]$. If $(p_1, r_1, p_2, r_2) = (1, 2, 2, 3)$, then we can compute $\phi, \theta \in (-\pi/4, \pi/4]$ from (2.41), (2.42), (2.36), and (2.21). If $(p_1, r_1, p_2, r_2) = (1, 2, 1, 3)$, then we compute $\phi \in (-\pi/4, \pi/4]$ as a solution of

$$(2.51) \qquad -y \cot^3 \phi + (a_{23}f + dh - z) \cot^2 \phi + (-a_{13}f - dk + y) \cot \phi + z = 0.$$

3. Numerical experiments. We have numerically compared the Euler angles method for diagonalizing a symmetric 3×3 matrix with the QR method and the two-sided Jacobi method. The test matrices were generated in the following way. In each test a diagonal matrix Λ with eigenvalues λ_i, $i = 1, 2, 3$ was chosen. Next, the diagonal matrix was transformed into a full symmetric matrix A via a random orthogonal similarity transformation Q. The three eigensolvers were then run on symmetric 3×3 matrices generated in this way.

It was observed by Kahan [Ka] that a straightforward implementation of the Cardano formulas may lead to a loss of accuracy in finite precision arithmetic when the roots of the cubic equation differ significantly in magnitude. A stable version of the Cardano formulas may require evaluating the formulas twice. By varying λ_i's we wanted to check the sensitivity of the Cardano formulas to the magnitudes of the eigenvalues. In the tests, we never observed any significant loss of accuracy in the computed eigenvalues.

Another way of solving the cubic equation arising in the Euler angles method is to use the Newton iteration. In our tests the Newton method took on average six iterations to converge and never produced a better approximation than the direct application of the (stable) Cardano formulas.

Before running the QR method a matrix was first transformed to the tridiagonal form. The QR method (with standard shifts) required on average three iterations.

In the Jacobi method we used cyclic-by-row ordering. On average, three sweeps (for 3×3 matrices, one sweep is equivalent to computing and applying three plane rotations) were sufficient to diagonalize a matrix. We measured the accuracy of each of the methods by the magnitude of the quantity e,

$$e = \frac{\|A - \tilde{Q}\tilde{\Lambda}\tilde{Q}^{\top}\|_F}{\|A\|_F},$$

where $\tilde{Q}$ and $\tilde{\Lambda}$ denote the computed matrix of eigenvectors and eigenvalues, respectively.

For all three methods e was always of the same order of magnitude. The experiments were performed on a machine with 16 decimal digits of relative precision. A typical test would give the following results:

$\lambda_1\ \lambda_2\ \lambda_3$	Euler	Jacobi	QR
3.0, 0.0, 0.0	6.52e − 16	4.12e − 16	6.53e − 16
1.0e + 12, 1.0e + 6, 1.0	1.83e − 16	1.01e − 15	4.33e − 16
1.01e + 6, 1.0e + 6, 0.99	2.02e − 15	6.81e − 16	4.03e − 16

Acknowledgments. We thank Gene Golub, for directing our attention to the related work of Smith [Sm], and the anonymous reviewer, whose comments improved the clarity of this note.

REFERENCES

[GMS] I. M. GELFAND, R. A. MINLOS, AND Z. YA. SHAPIRO, *Representations of the Rotation and Lorenz Groups and Their Applications*, Macmillan, New York, 1963.

[Ka] W. KAHAN, *To solve a real cubic equation*, Lecture Notes for a Numerical Analysis Course, University of California, Berkeley, CA, November 10, 1986.

[KK] G. A. KORN AND T. M. KORN, *Mathematical Handbook for Scientists and Engineers*, McGraw-Hill, New York, 1961.

[Ro] M. E. ROSE, *Elementary Theory of Angular Momentum*, John Wiley, New York, 1957.

[Sm] O. K. SMITH, *Eigenvalues of a symmetric* 3×3 *matrix*, Comm. ACM, (1961), Vol. 4, No. 4, p. 168.

[St] G. STRANG, *Linear Algebra and Its Applications*, Third Edition, Harcourt, Brace, Jovanovich, San Diego, 1988.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 1, pp. 49–69, January 1991

006

HADAMARD SQUARE ROOTS*

MARVIN MARCUS† AND MARKUS SANDY‡

Abstract. If A is an n-square positive semidefinite Hermitian matrix of rank 1, then the Hadamard square root of A is the n-square matrix obtained by replacing each entry of A by the principal value of its square root. It is proved that if A has no zero or negative entries, then the Hadamard square root has odd rank and all odd ranks are possible.

Key words. matrix, rank, inertia, Hadamard product, Schur product

AMS(MOS) subject classifications. 15A03, 15A21

1. Introduction. The Hadamard product of two $m \times n$ matrices A and B is the $m \times n$ matrix C whose (i, j) entry is

$$c_{ij} = a_{ij} b_{ij}, \qquad i = 1, \cdots, m, \qquad j = 1, \cdots, n. \tag{1}$$

The matrix C in (1) is usually denoted by

$$C = A \cdot B. \tag{2}$$

The matrix (2) is also called the Schur product of A and B and, indeed, in [5, p. 458] the fact that $C \geqq 0$ (i.e., C is positive semidefinite Hermitian) whenever A and B are, is called the "Schur Product Theorem." Many years before [3, p. 173], Halmos described this result as "a remarkable theorem on positive matrices." The proofs in both [5] and [3] are the same and depend on writing A and B as sums of rank 1 positive semidefinite Hermitian matrices. Actually, the Schur Product Theorem was noted by Schur [11], and several inequalities involving $\det(A \cdot B)$ appear in [9, p. 421]. A history of the Hadamard product with an excellent accompanying bibliography can be found in [4]. This article is a valuable contribution to the matrix literature.

Thirty years ago Marcus and Khan observed that the Hadamard product of any two n-square matrices is a principal submatrix of the Kronecker product $A \otimes B$ [7]. The precise location of $A \cdot B$ in $A \otimes B$ is simple to determine. Define

$$r_t = (t-1)n + t, \qquad t = 1, \cdots, n.$$

For any p and q among $1, \cdots, n^2$ the (p, q) entry of $K = A \otimes B$ is

$$k_{pq} = a_{i_1 j_1} b_{i_2 j_2},$$

where

$$p = i_2 + n(i_1 - 1) \tag{3}$$

and

$$q = j_2 + n(j_1 - 1). \tag{4}$$

If we set $p = r_s$ and $q = r_t$ in (3) and (4), it is immediate that $i_1 = i_2 = s$ and $j_1 = j_2 = t$. Hence

$$k_{r_s r_t} = a_{st} b_{st}, \qquad s, t = 1, \cdots, n.$$

* Received by the editors May 1, 1989; accepted for publication (in revised form) October 23, 1989.

† Department of Computer Science, University of California, Santa Barbara, California 93106 (mmarcus@cs.ucsb.edu). The research of this author was supported by Air Force Office of Scientific Research grant AFOSR-88-0175.

‡ Smartstar Corporation, Santa Barbara, California 93116-1950.

The Cauchy interlacing inequalities [8, p. 203] can then be effectively used to obtain information that relates the eigenvalues of $A \cdot B$ to those of A and B. Typically, if $A \geqq 0$ then $A^T \geqq 0$ so that $H = A \cdot A^T = [|a_{ij}|^2] \geqq 0$ and

$$\sqrt{\lambda_1(H)} \leqq \lambda_1(A), \tag{5}$$

where λ_1 is the largest eigenvalue of the indicated matrix.

In [5, p. 462] there is an interesting exercise in which a 4-square $A \geqq 0$ is constructed for which

$$\operatorname{abs}(A) = [|a_{ij}|] \tag{6}$$

fails to be positive semidefinite. Since

$$[|a_{ij}|^2] = A \cdot A^T \geqq 0$$

the matrix in (6) shows that the entrywise square root of a positive semidefinite Hermitian matrix is not necessarily positive semidefinite. In order to avoid confusing the entrywise square root with the usual matrix square root we shall designate the former as

$$\sqrt{A} = [\sqrt{a_{ij}}]. \tag{7}$$

In [5, p. 462] the matrix (7) is called the "Hadamard square root." The precise definition of the square root of a complex number must be stipulated in order that (7) be well defined for any complex matrix A. The principal value of the square root function satisfies

$$\operatorname{Re} \sqrt{z} \geqq 0, \tag{8}$$

and the ambiguity in (8) for $z < 0$ is resolved with the usual

$$\sqrt{z} = |z|^{1/2} i.$$

Note that if $d > 0$ then (8) implies that

$$\sqrt{dz} = \sqrt{d}\sqrt{z}. \tag{9}$$

For angles $\omega \in [-2\pi, 2\pi]$ the definition of the principal value of the square root function implies that

$$\sqrt{e^{i\omega}} = \varepsilon(\omega) e^{i\omega/2}, \tag{10}$$

in which

$$\varepsilon(\omega) = \begin{cases} 1 & \text{if } \omega = \pi \text{ or } |\omega| < \pi, \\ -1 & \text{if } \omega = -\pi \text{ or } |\omega| > \pi. \end{cases} \tag{11}$$

This definition of the square root function is implemented in the MATLAB™ [10, pp. 3–41] function sqrt, i.e.,

$$\operatorname{sqrt}(A) = \sqrt{A}. \tag{12}$$

If $A \geqq 0$ has a negative entry, the Hermitian property will be lost in computing $\sqrt{A}$. For example, for the rank 1 matrix $A \geqq 0$,

$$A = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},$$

the Hadamard square root is

$$\sqrt{A} = \begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix},$$

a non-Hermitian matrix of rank 2.

The starting point for the results contained herein was a sequence of numerical experiments using MATLAB™ to evaluate ranks of Hadamard square roots of randomly generated $A \geqq 0$ of rank 1, i.e., $\rho(A) = 1$. In these experiments it was invariably the case that for real $A \geqq 0$ of rank 1

$$\rho(\sqrt{A})$$

was computed as either 1 or 2. As we shall see shortly, it is easy to prove that for $A \geqq 0$ and $\rho(A) = 1$,

$$\rho(\sqrt{A}) = \begin{cases} 1 & \text{if } a_{ij} \geqq 0 \text{ for all } i,j, \\ 2 & \text{if some } a_{ij} < 0. \end{cases} \tag{13}$$

Much more surprising were the results of an experiment in which

$$\rho(\sqrt{A}) \tag{14}$$

was computed for a large sample of randomly generated complex n-square matrices $A \geqq 0$ of rank 1: the rank (14) was computed as an odd integer in all cases. Despite this numerical evidence, consider the Hermitian matrix

$$A = \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 1 \end{bmatrix}. \tag{15}$$

It is simple to confirm that $A \geqq 0$ and $\rho(A) = 1$. Yet

$$\sqrt{A} = \begin{bmatrix} 1 & i & i \\ i & 1 & 1 \\ i & 1 & 1 \end{bmatrix} \tag{16}$$

is a matrix of rank 2. Of course, since the principal value of the square root function is used, the matrix (16) is no longer Hermitian. The question of making a consistent choice of arguments for the square roots (and more general powers) so that the resulting matrix is Hermitian is considered in the paper [2, pp. 640–641].

In order to state and prove the theorems that explain these phenomena we begin by writing $A \geqq 0$, $\rho(A) = 1$ as a dyad,

$$A = uu^*, \tag{17}$$

in which u is a column vector:

$$u = [r_1 e^{i\varphi_1}, r_2 e^{i\varphi_2}, \cdots, r_n e^{i\varphi_n}]^T, \tag{18}$$

$r_t = |u_t|$, $\varphi_t \in [0, 2\pi)$, $t = 1, \cdots, n$. Thus

$$A = [r_p r_q e^{i(\varphi_p - \varphi_q)}] \tag{19}$$

and since any two vectors u that serve to represent A as in (17) differ by a scalar multiple of modulus 1, it follows that the differences $\varphi_p - \varphi_q$, $p, q = 1, \cdots, n$, are the same for all such u. It will be convenient to reorder the components of u, and to do this we need only observe that if P is a permutation matrix corresponding to some $\sigma \in S_n$ then

$$\sqrt{PAP^T} = P\sqrt{A}P^T, \tag{20}$$

so that

$$\rho(\sqrt{PAP^T}) = \rho(\sqrt{A}). \tag{21}$$

Now

$$Puu^*P^T=(Pu)(Pu)^* \tag{22}$$

and thus we can assume that $r_1, \cdots, r_k$ are positive and $r_{k+1}, \cdots, r_n$ are 0. But then A consists of a k-square upper left principal submatrix bordered with 0 entries. Hence there is no loss of generality in assuming that $k = n$ in any investigation of the rank of $\sqrt{A}$. We also remark that if D is a diagonal matrix with positive entries, then in view of (9)

$$\begin{aligned}\sqrt{DAD}&=[\sqrt{d_p a_{pq} d_q}]\\&=[\sqrt{d_p}\sqrt{a_{pq}}\sqrt{d_q}]\\&=\sqrt{D}\sqrt{A}\sqrt{D}\end{aligned}$$

so that

$$\rho(\sqrt{DAD})=\rho(\sqrt{A}). \tag{23}$$

Formulas (21)–(23) enable us to normalize A in (19) as follows. We can assume that $r_1 = \cdots = r_n = 1$ by using (23) to replace A by DAD in which $D = \operatorname{diag}(r_1^{-1}, \cdots, r_n^{-1})$; formula (22) can be used to reorder the components of u so that $\varphi_1 \geqq \varphi_2 \geqq \cdots \geqq \varphi_n$. Moreover, since replacing u by λu, $|\lambda| = 1$, does not affect (17) we may assume

$$2\pi > \varphi_1 \geqq \varphi_2 \geqq \cdots \geqq \varphi_n = 0. \tag{24}$$

Henceforth we shall assume A is in normalized form as just described, so that (from (19))

$$A=[e^{i(\varphi_p-\varphi_q)}] \tag{25}$$

and (24) holds for the arguments $\varphi_1, \cdots, \varphi_n$. It is convenient to have a notation for the *angular spread* of an arbitrary matrix $A = [a_{pq}] = [|a_{pq}|e^{i\omega_{pq}}]$, $\omega_{pq} \in [-2\pi, 2\pi]$, $p, q = 1, \cdots, n$, having at least one nonzero off-diagonal entry:

$$s(A)=\max_{p\neq q}|\omega_{pq}|. \tag{26}$$

According to (26) and (24) we can assume that

$$s(A)=\varphi_1. \tag{27}$$

Referring to the definition of $\varepsilon(\omega)$ in (11) we can define an n-square matrix E_A associated with A as

$$E_A=[\varepsilon(\varphi_p-\varphi_q)]. \tag{28}$$

Note that from (25) and (11)

$$\begin{aligned}\sqrt{A}&=[\sqrt{e^{i(\varphi_p-\varphi_q)}}]\\&=[\varepsilon(\varphi_p-\varphi_q)e^{i(\varphi_p/2-\varphi_q/2)}]\\&=\Delta E_A\Delta^*\end{aligned} \tag{29}$$

where $\Delta = \operatorname{diag}(e^{i\varphi_1/2}, \cdots, e^{i\varphi_n/2})$. Hence

$$\rho(\sqrt{A})=\rho(E_A). \tag{30}$$

The principal results of this paper follow. In the statement of each result A is an n-square rank 1 positive semidefinite Hermitian matrix.

THEOREM 1. *Assume that*

$$s(A) \leqq \pi. \tag{31}$$

Then

$$\rho(\sqrt{A}) = 1 \tag{32}$$

if the inequality (31) *is strict. Otherwise,* $s(A) = \pi$ *and*

$$\rho(\sqrt{A}) = 2. \tag{33}$$

COROLLARY 1. *If* $s(A) \leqq \pi$ *and no off-diagonal entry of* A *is negative, then* $\rho(\sqrt{A}) = 1$.

COROLLARY 2. *Assume that* A *is real. Then* $\rho(\sqrt{A}) = 2$ *if any off-diagonal entry of* A *is negative. Otherwise* $\rho(\sqrt{A}) = 1$.

Theorem 1 is related to a result of Farjot [1] on infinitely divisible matrices. An n-square matrix $A \geqq 0$ is said to be infinitely divisible if $A^{(\alpha)} = [a_{ij}^{\alpha}] \geqq 0$ for all $\alpha > 0$. Farjot's theorem states that if A is infinitely divisible then $\rho(A^{(\alpha)}) = \rho(A)$ for all $\alpha > 0$.

THEOREM 2. *Assume that*

$$s(A) > \pi. \tag{34}$$

If no off-diagonal entry of A *is negative, then* $\rho(\sqrt{A})$ *is an odd integer.*

THEOREM 3. *Assume that no off-diagonal entry of* A *is negative. Then* $\rho(\sqrt{A})$ *is an odd integer.*

It is Theorem 3 that explains why MATLAB™ invariably computes $\rho(\sqrt{A})$ as an odd integer for random rank 1 matrices $A \geqq 0$. For, in generating random complex column vectors u as in (18), the probability is 0 that

$$|\varphi_p - \varphi_q| = \pi \tag{35}$$

for some $p \neq q$. But

$$\begin{aligned} A &= uu^* \\ &= [r_p r_q e^{i(\varphi_p - \varphi_q)}] \end{aligned}$$

and hence A has a negative off-diagonal entry if and only if (35) holds. (Recall that the case of some u_p being 0 was eliminated in reducing A to normalized form.)

The next two results show that any rank is possible for the Hadamard square root of an appropriate $A \geqq 0$ of rank 1. Specifically we have Theorem 4.

THEOREM 4. *Let* ν *be an odd integer,* $1 \leqq \nu \leqq n$. *Then there exists an* n*-square* $A \geqq 0$, $\rho(A) = 1$, *with no zero or negative entries such that*

$$\rho(\sqrt{A}) = \nu. \tag{36}$$

If we permit negative off-diagonal entries, then $\rho(\sqrt{A})$ *can take on any integral value between* 1 *and* n.

THEOREM 5. *Let* ν *be any integer,* $1 \leqq \nu \leqq n$. *Then there exists an* n*-square* $A \geqq 0$, $\rho(A) = 1$, *with no zero entries such that*

$$\rho(\sqrt{A}) = \nu. \tag{37}$$

Of course, in view of Theorem 3, if ν is even in Theorem 5 the corresponding A for which (37) holds must have a negative off-diagonal entry.

2. Proofs. To prove Theorem 1 we assume A is in normalized form. If $s(A) < \pi$ then every difference $\varphi_p - \varphi_q$ satisfies $|\varphi_p - \varphi_q| < \pi$ and hence from (11),

$$\varepsilon(\varphi_p - \varphi_q) = 1, \qquad p, q = 1, \cdots, n.$$

It follows that $E_A = J_n$, the n-square matrix consisting entirely of 1's, and (32) follows from (30). If $s(A) = \pi$ then there exists an integer k, $1 \leqq k < n$, and an integer m, $0 \leqq m \leqq n - (k+1)$, such that

$$\pi = \varphi_1 = \cdots = \varphi_k > \varphi_{k+1} \geqq \cdots \geqq \varphi_{k+m} > \varphi_{k+m+1} = \cdots = \varphi_n = 0.$$

From (11) again, it follows that E_A has the following block matrix form:

$$E_A = \begin{bmatrix} J_k & J_{k,m} & J_{k,n-(k+m)} \\ J_{m,k} & J_m & J_{m,n-(k+m)} \\ -J_{n-(k+m),k} & J_{n-(k+m),m} & J_{n-(k+m)} \end{bmatrix},$$

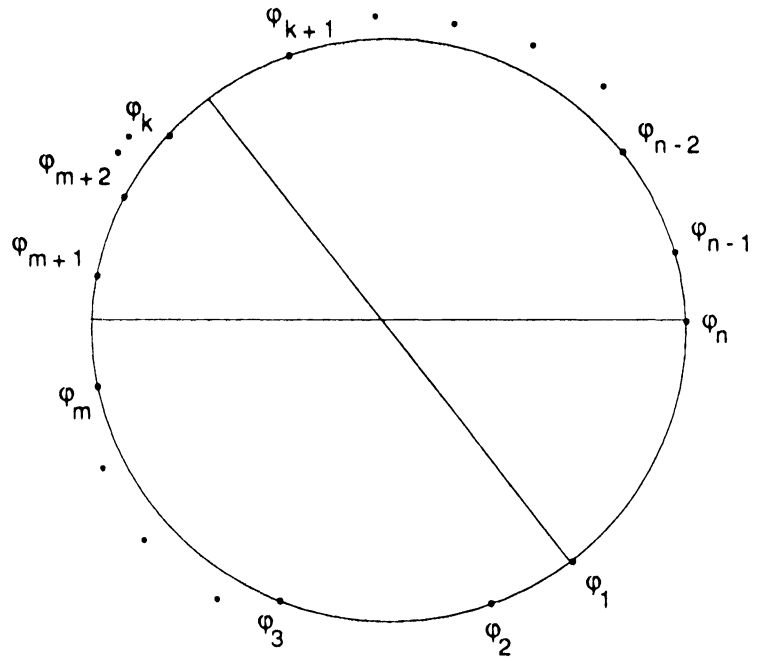
where $J_{k,m}$ is a $k \times m$ matrix of 1's, etc. Clearly, since $0 < k < n$ and $n - (k+m) \geqq 1$ it follows that $\rho(E_A) = 2$, and (33) results from (30).

Corollary 1 is an immediate consequence of Theorem 1. For, if no off-diagonal entry is negative, then $\varphi_p - \varphi_q$ can never be π for any p and q. Hence $s(A) < \pi$ and Theorem 1 implies that $\rho(\sqrt{A}) = 1$.

To prove Corollary 2 note that for a real matrix, $s(A) = 0$ or $s(A) = \pi$. If A has a negative entry (it must be off-diagonal) then $s(A) = \pi$ and $\rho(\sqrt{A}) = 2$ from Theorem 1. If all entries of A are positive, then $s(A) = 0$ and again Theorem 1 implies that $\rho(\sqrt{A}) = 1$.

We proceed to the proof of Theorem 2. Since $s(A) > \pi$ we know that $\varphi_1 > \pi$. Define k to be the largest integer such that $\varphi_1 - \varphi_k < \pi$. Note that $1 \leqq k < n$; otherwise, if k were n, the value of $s(A)$ would be at most π, contradicting $s(A) > \pi$. Also, define m to be the largest integer such that $\varphi_m > \pi$. Possibly, $m = 1$, and by definition, $m \leqq k$. It is helpful to graphically depict the points $e^{i\varphi_t}$, $t = 1, \cdots, n$, on the unit circle:

(38)

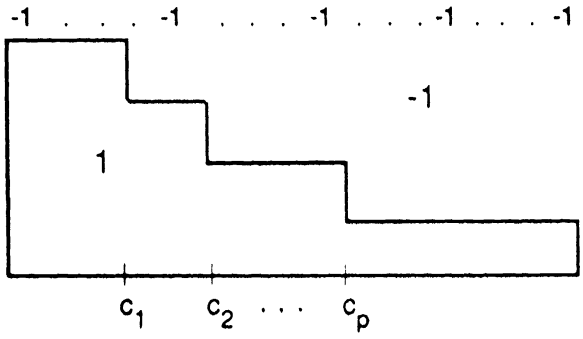


The points $e^{i\varphi_t}$ are labeled by φ_t, $t = 1, \cdots, n$. The matrix E_A defined in (28) specializes to

$$E_A = \begin{bmatrix} J_k & M \\ M^T & J_{n-k} \end{bmatrix}, \tag{39}$$

in which M is a $k \times (n-k)$ matrix of the following "staircase" form:

(40)



We observe that:

(a) The first row of M consists entirely of -1's;

(b) If $m > 1$ there exist columns $c_1 < \cdots < c_p$ such that the entries to the right of each of the steps that start at columns $c_1, c_2, \cdots, c_p$ are all -1's;

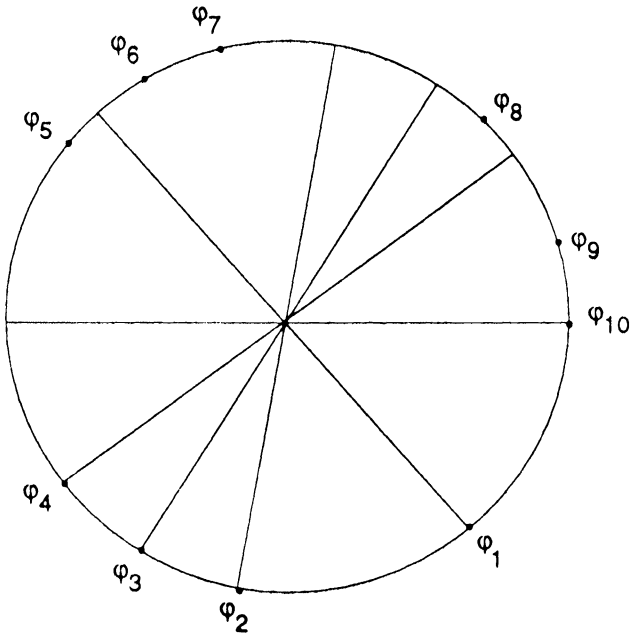
(c) If $k > m$ then rows $m+1, \cdots, k$ consist entirely of 1's.

A typical example of such an M together with a diagram of the form (38) will help fix the ideas. Consider

(41)
$$M = \begin{bmatrix} -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

For the matrix (41) the parameters are $n = 10$, $k = 5$, $m = 4$, $p = 2$, $c_1 = 3$, $c_2 = 4$. The corresponding diagram (38) specializes to

(42)



The matrix M is the submatrix of $E_A = [\varepsilon(\varphi_p - \varphi_q)]$ lying in rows $1, \cdots, 5$ and columns $6, \cdots, 10$. Since $s(A) > \pi$, φ_1 is in the lower halfplane. Since $\varphi_1 - \varphi_t > \pi$, $t = 6, \cdots,$

10, the first row of M consists of -1's (see (11)). The largest k such that $\varphi_1 - \varphi_k < \pi$ is $k = 5$ and the largest m such that $\varphi_m > \pi$ is $m = 4$. Since $\varphi_2 - \varphi_j < \pi$, $j = 6, 7$, and $\varphi_2 - \varphi_j > \pi$, $j = 8, 9, 10$, it follows from (11) again that the second row of M is

$$M_{(2)} = [1 \quad 1 \quad -1 \quad -1 \quad -1].$$

This is similar for $M_{(3)}$. The values $\varphi_4 - \varphi_j$, $j = 6, 7, 8$ are less than π while $\varphi_4 - \varphi_j > \pi$, $j = 9, 10$. Hence

$$M_{(4)} = [1 \quad 1 \quad 1 \quad -1 \quad -1].$$

Finally, $\varphi_5 - \varphi_j < \pi$, $j = 6, \cdots, 10$, so that

$$M_{(5)} = [1 \quad 1 \quad 1 \quad 1 \quad 1].$$

We remark that since A has no negative entries, no two φ_j are at opposite ends of a diameter.

Next, let V be an $(n-k) \times k$ matrix whose first column consists of 1's and whose remaining entries are 0. The following equations are simple to verify:

$$VJ_k = J_{n-k,k}, \tag{43}$$

$$J_k V^T = J_{k,n-k}, \tag{44}$$

$$VJ_k V^T = J_{n-k}, \tag{45}$$

$$VM = -J_{n-k}, \tag{46}$$

$$M^T V^T = -J_{n-k}. \tag{47}$$

Define the n-square matrix

$$L = \begin{bmatrix} I_k & 0 \\ V & I_{n-k} \end{bmatrix} \tag{48}$$

conformally partitioned with E_A in (39). Then we compute that

$$LE_A L^T = \begin{bmatrix} J_k & J_k V^T + M \\ VJ_k + M^T & VJ_k V^T + M^T V^T + VM + J_{n-k} \end{bmatrix}$$

and from (43)–(47),

$$LE_A L^T = \begin{bmatrix} J_k & J_{k,n-k} + M \\ J_{n-k,k} + M^T & 0 \end{bmatrix}. \tag{49}$$

If we refer back to the staircase matrix M in (40) we see that the matrix $J_{k,n-k} + M$ has the same form as M except that the -1 entries above the stairs are now 0 and the 1 entries below the stairs are now 2. If $k = 1$, $J_{k,n-k} + M = O_{1,n-1}$ and $\rho(E_A) = 1$. Otherwise, assume $k > 1$ and let W denote the $(k-1) \times (n-k)$ submatrix of $J_{k,n-k} + M$ obtained by deleting its zero first row:

$$J_{k,n-k} + M = \begin{bmatrix} 0 \cdots 0 \\ W \end{bmatrix}. \tag{50}$$

Define the k-square matrix K by

$$K = \begin{bmatrix} 1 & & & \\ -1 & 1 & & 0 \\ -1 & & \ddots & \\ \vdots & & 0 & \ddots \\ -1 & & & 1 \end{bmatrix} \tag{51}$$

and then let

$$S = \begin{bmatrix} K & 0 \\ 0 & I_{n-k} \end{bmatrix} \tag{52}$$

conformally partitioned with (49). It is not difficult to check that

$$KJ_kK^T = \begin{bmatrix} 1 & 0 \cdots 0 \\ 0 & \\ \vdots & 0 \\ 0 & \end{bmatrix} \tag{53}$$

and that

$$\begin{aligned} K(J_{k,n-k} + M) &= K\begin{bmatrix} 0 \cdots 0 \\ W \end{bmatrix} \\ &= J_{k,n-k} + M. \end{aligned} \tag{54}$$

Thus, from (49), (53), and (54) we have

$$S(LE_AL^T)S^T = \left[\begin{array}{c|c} E_{11} & \begin{matrix} 0 \cdots 0 \\ W \end{matrix} \\ \hline \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \quad W^T & 0 \end{array}\right] \tag{55}$$

where E_{11} is the k-square matrix with the single nonzero entry 1 in the (1, 1) position. It is obvious that the rank of the matrix (55) is

$$2\rho(W) + 1. \tag{56}$$

But then

$$\rho(E_A) = 2\rho(W) + 1 \tag{57}$$

and the proof of Theorem 2 is complete.

The proof of Theorem 3 goes as follows. Since A has no (off-diagonal) negative entries it follows that $|\varphi_p - \varphi_q|$ is never π for $p \neq q$. Thus $s(A) \neq \pi$. If $s(A) < \pi$ then Theorem 1 implies that $\rho(\sqrt{A}) = 1$. If $s(A) > \pi$ then Theorem 2 implies that $\rho(\sqrt{A})$ is odd. In any event, $\rho(\sqrt{A})$ is always odd.

As remarked above, W is obtained from the staircase matrix (40) by removing the first row, replacing the -1's by 0's and the 1's by 2's. For example, for the matrix (41)

$$W = \begin{bmatrix} 2 & 2 & 0 & 0 & 0 \\ 2 & 2 & 0 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 2 & 2 & 2 & 2 & 2 \end{bmatrix}.$$

Obviously, $\rho(W)$ is just the number of horizontal steps in W, i.e.,

(58) $$\rho(W)=p+1$$

(see (40)).

To prove Theorem 4, write ν as

(59) $$\nu=2d+1.$$

In view of (55) and (57) we need only construct an n-square A such that the resulting matrix W in (50) satisfies

(60) $$2\rho(W)+1=\nu,$$

or

(61) $$\rho(W)=d.$$

Assume first that n is even so that

(62) $$2d+1=\nu<n$$

and hence

(63) $$n-2(d+1)\geqq 0.$$

If $d=0$ so that $\nu=1$, simply take A to be J_n. Otherwise, define M to be the $(d+1)\times(n-(d+1))$ partitioned matrix:

(64) $$M = d+1\left[\begin{array}{ccccc|c} -1 & & \cdots & & -1 & \\ 1 & -1 & & & -1 & \\ 1 & 1 & -1 & \cdot & -1 & -J_{d+1,n-2(d+1)} \\ \vdots & \vdots & & \cdot & \cdot & \\ 1 & 1 & \cdots & 1 & -1 & \end{array}\right]$$

(with the first block $d+1$ columns wide, the block $-J_{d+1,n-2(d+1)}$ being $n-2(d+1)$ columns wide, and $n-(d+1)$ columns in all)

If we set $k=d+1$ so that M is $k\times(n-k)$, then it is obvious from the form of M in (64) that

(65) $$\rho(M+J_{k,n-k})=d.$$

From (30) and (39) we need only construct an A such that

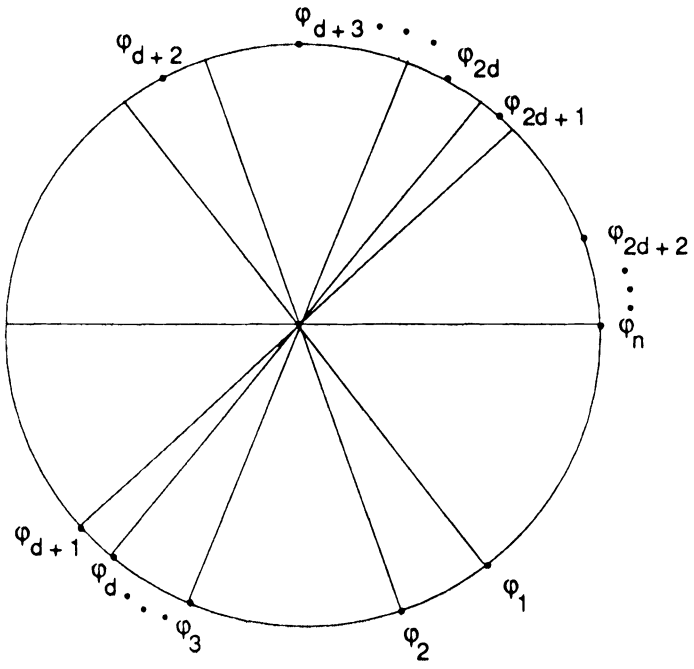
(66) $$E_A=\begin{bmatrix} J_k & M \\ M^T & J_{n-k} \end{bmatrix}$$

for the matrix M in (64). This amounts to describing a distribution of points $e^{i\varphi_t}$, $t=1,\cdots,n$, as in the diagram (38), such that

(67) $$E_A=[\varepsilon(\varphi_p-\varphi_q)]$$

is the matrix (66).

(68)



In the diagram (68)

$$2\pi > \varphi_1 > \varphi_2 > \varphi_3 > \cdots > \varphi_{d+1} > \pi > \varphi_{d+2} > \varphi_{d+3} > \cdots > \varphi_n = 0 \tag{69}$$

and

$$\begin{aligned} &\varphi_1 - \pi > \varphi_{d+2} > \varphi_2 - \pi, \\ &\varphi_2 - \pi > \varphi_{d+3} > \varphi_3 - \pi, \\ &\qquad\vdots \\ &\varphi_d - \pi > \varphi_{2d+1} > \varphi_{d+1} - \pi, \end{aligned} \tag{70}$$

$$\varphi_{d+1} - \pi > \varphi_{2d+2} > \cdots > \varphi_n = 0. \tag{71}$$

Recalling that $k = d + 1$, (69) and (11) imply that

$$\varepsilon(\varphi_p - \varphi_q) = 1, \qquad p, q = 1, \cdots, d+1 = k \tag{72}$$

so that the upper left k-square block in E_A is J_k. The first inequality in (70) implies that

$$\varepsilon(\varphi_1 - \varphi_q) = -1, \qquad q = d+2 = k+1, \cdots, n. \tag{73}$$

From the second inequality in (70),

$$\varepsilon(\varphi_2 - \varphi_{d+2}) = 1 \tag{74}$$

and

$$\varepsilon(\varphi_2 - \varphi_q) = -1, \qquad q = d+3 = k+2, \cdots, n. \tag{75}$$

We continue similarly through the last inequality in (70) which implies

$$\varepsilon(\varphi_{d+1} - \varphi_q) = 1, \qquad q = d+2 = k+1, \cdots, 2d+1 = k+d \tag{76}$$

and

$$\varepsilon(\varphi_{d+1}-\varphi_q)=-1, \qquad q=2d+2=k+d+1, \cdots, n. \tag{77}$$

Finally, the fact (in (69)) that $\pi > \varphi_{d+2}$ implies that

$$\varepsilon(\varphi_p-\varphi_q)=1, \qquad p,q=d+2=k+1, \cdots, n. \tag{78}$$

The statements (72)–(78) show that any distribution of φ_t, $t = 1, \cdots, n$, that satisfies the inequalities (69)–(71) yields a matrix E_A for which (66) holds when M is the matrix (64). Thus, for any such set of φ_t, the matrix

$$A=[e^{i(\varphi_p-\varphi_q)}]$$

has rank 1, satisfies $A \geqq 0$, and has no negative or zero entries (i.e., $(\varphi_p - \varphi_q) \neq \pi$, for any p and q). This completes the proof for the case n even. Assume next that n is odd and that $\nu = 2d + 1$. Once again we need only construct an n-square A such that the resulting W in (50) satisfies (61). The two cases $\nu < n$ and $\nu = n$ are slightly different. We dispose of $\nu = n$ first. Set $k = d + 1$ so that

$$\begin{aligned} n-k &= n-(d+1) \\ &= 2d+1-(d+1) \\ &= d. \end{aligned} \tag{79}$$

Define M to be the $(d + 1) \times d$ (i.e., $k \times (n - k)$) matrix

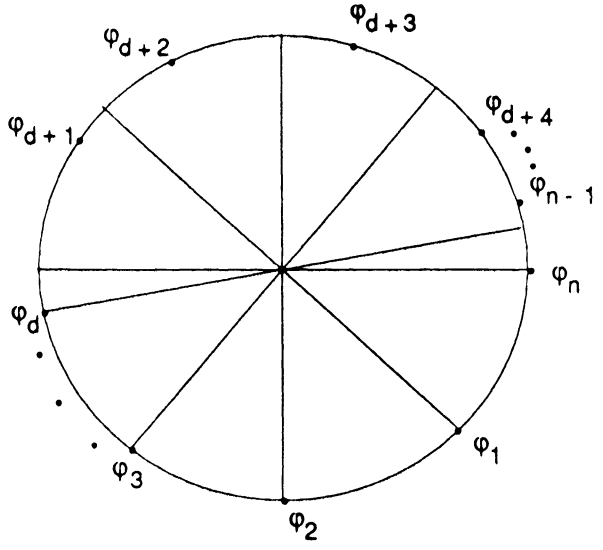
$$M=\begin{bmatrix} -1 & -1 & \cdots & -1 \\ 1 & -1 & \cdots & -1 \\ 1 & 1 & -1 & \cdots -1 \\ & & \vdots & \\ 1 & 1 & \cdots & 1 \; -1 \\ 1 & \cdot & \cdot & \cdot \; 1 \end{bmatrix} \tag{80}$$

Then

$$M+J_{k,n-k}=M+J_{d+1,d}$$

obviously has rank d. From (30) and (39) again, it is only necessary to construct A such that the matrix E_A in (67) is precisely the matrix (66). A diagram similar to (68) is useful:

(81)



In the diagram (81)

(82) $$\varphi_1 > \varphi_2 > \cdots > \varphi_d > \pi > \varphi_{d+1} > \varphi_{d+2} > \cdots > \varphi_{n-1} > \varphi_n = 0.$$

Moreover,

(83)

$$\varphi_{d+1} > \varphi_1 - \pi > \varphi_{d+2} > \varphi_2 - \pi > \varphi_{d+3} > \varphi_3 - \pi > \varphi_{d+4} > \cdots > \varphi_{n-1} > \varphi_d - \pi > \varphi_n = 0.$$

Since $k = d + 1$ and

$$|\varphi_p - \varphi_q| < \pi, \qquad p, q = 1, \cdots, k,$$

the upper left k-square block in E_A is J_k. It is routine to check that the inequalities (82) and (83) produce a matrix E_A such that the matrix M in (80) is its upper right block.

It remains to settle the case $v < n$. Set $k = d + 1$ and note that

$$v + 1 < n, \qquad n > 2(d+1),$$

and hence

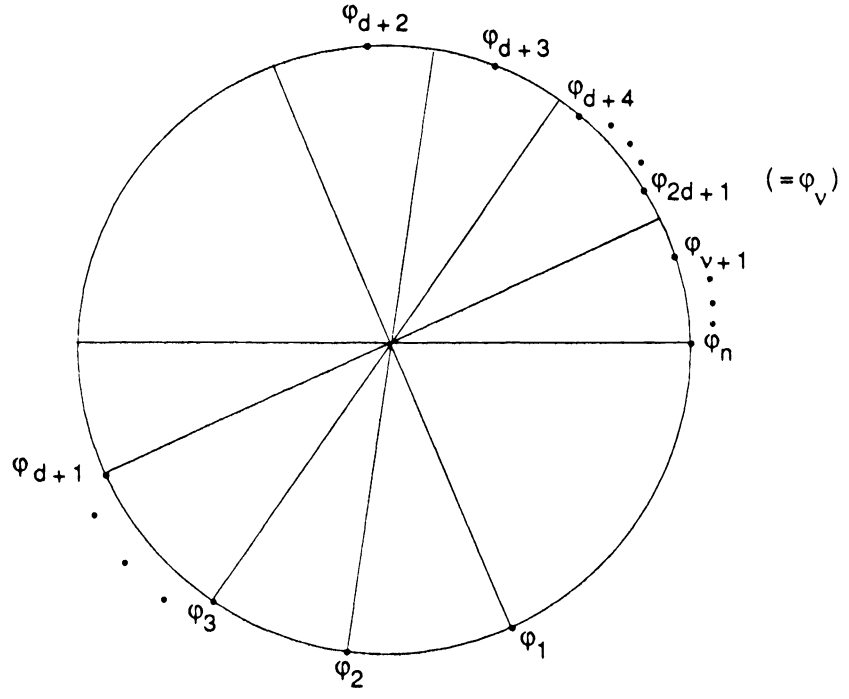
$$d + 1 < n - (d+1).$$

Define M to be the $(d + 1) \times (n - (d + 1))$ (i.e., $k \times (n - k)$) matrix

(84) $$M = \begin{bmatrix} -1 & -1 & \cdots & -1 \\ 1 & -1 & \cdots & -1 \\ 1 & 1 & -1 \cdots & -1 \\ & & \vdots & \\ 1 & \cdots & 1 & -1 \cdots -1 \end{bmatrix},$$

in which the -1's in row 2 begin in column 2, $\cdots$, the -1's in row $d + 1 = k$ begin in column $d + 1$. The appropriate set of φ's to produce an E_A with the matrix M in (84) in its upper right block is described by the following diagram:

(85)



In the diagram (85)

$$\begin{aligned}(86)\qquad &\varphi_1>\varphi_2>\varphi_3>\cdots>\varphi_{d+1}>\pi>\varphi_1-\pi>\varphi_{d+2}>\varphi_2-\pi\\ &>\varphi_{d+3}>\varphi_3-\pi>\varphi_{d+4}\cdots>\varphi_{2d+1}>\varphi_{d+1}-\pi>\varphi_{\nu+1}>\cdots>\varphi_n=0.\end{aligned}$$

It can easily be confirmed that the inequalities (86) imply that E_A has the matrix M in (84) in its upper right block. We omit the familiar details to conclude the proof of Theorem 4.

We proceed to the proof of Theorem 5. If ν is odd then Theorem 5 is precisely Theorem 4. Thus we may assume ν is even. The remainder of the proof is in two parts: $\nu = n$ and $\nu < n$. We first prove that if n is an even integer, there exists an n-square $A \geqq 0$, $\rho(A) = 1$, with no zero entries, such that

$$(87)\qquad \rho(\sqrt{A}) = n.$$

Define

$$(88)\qquad \varphi_k = (n-k)2\pi/n, \qquad k = 1, \cdots, n$$

so that

$$\varphi_1>\varphi_2>\cdots>\varphi_n=0.$$

As before, we compute the column vector $u = [e^{i\varphi_1}, \cdots, e^{i\varphi_n}]^T$ and then the matrix

$$\begin{aligned}A &= uu^*\\ &= \left[\exp\left(i\frac{2\pi}{n}(q-p)\right)\right].\end{aligned}$$

The square root of A is defined by

$$\sqrt{A} = \left[\varepsilon\left(\frac{2\pi}{n}(q-p)\right)\exp\left(i\frac{2\pi}{n}\left(\frac{q}{2}-\frac{p}{2}\right)\right)\right],$$

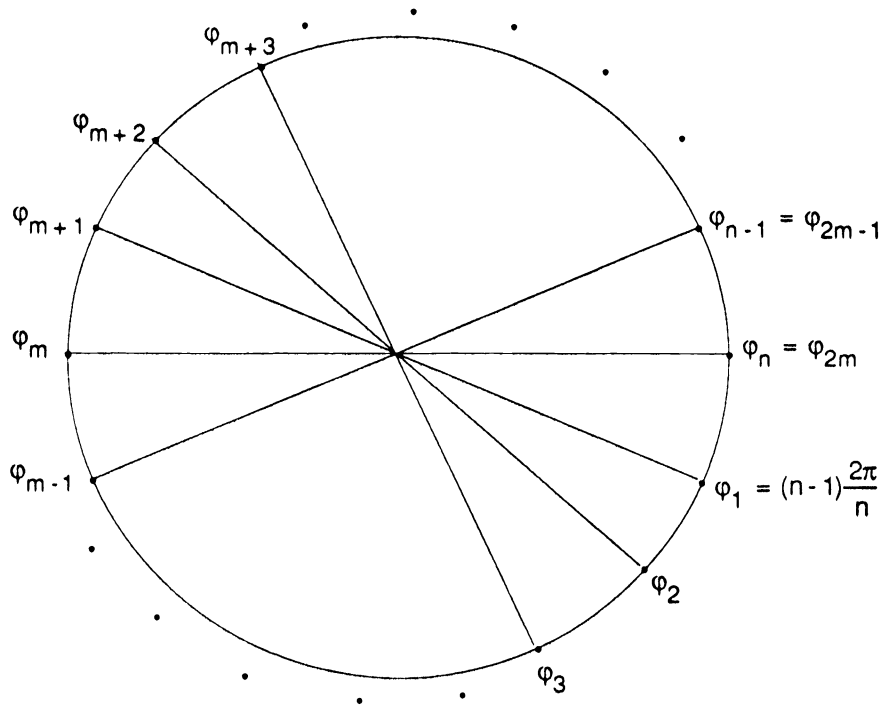
which in turn is congruent to the matrix

$$(89)\qquad E_A = \left[\varepsilon\left(\frac{2\pi}{n}(q-p)\right)\right].$$

If we set $n = 2m$, then for $p, q = 1, \cdots, n = 2m$, we have

$$(90)\qquad \varepsilon\left(\frac{2\pi}{n}(q-p)\right) = \begin{cases} 1 & \text{if } |p-q| < m \text{ or } q-p = m,\\ -1 & \text{if } |p-q| > m \text{ or } q-p = -m.\end{cases}$$

The diagram corresponding to the choice of φ_k in (88) is



The matrix E_A becomes

$$E_A = \begin{bmatrix} J_m & M \\ -M & J_m \end{bmatrix} \tag{91}$$

and

$$M = \begin{bmatrix} 1 & -1 & \cdots & -1 \\ 1 & 1 & -1 \cdots & -1 \\ & & \vdots & \\ 1 & 1 & 1 \cdots 1 & -1 \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

$$= I_m + S,$$

where S is the skew-symmetric matrix with -1 in the upper triangle. Thus, from (91)

$$E_A = I_2 \otimes J_m + \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \otimes M$$

("$\otimes$" denotes the Kronecker product), which in turn is unitarily similar to

$$R = I_2 \otimes J_m + \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} \otimes M$$

$$= \begin{bmatrix} J_m + iM & 0 \\ 0 & J_m - iM \end{bmatrix}.$$

Now

$$\det(R) = |\det(J_m + iM)|^2$$

and

$$J_m + iM = (1+i)I_m + H,$$

where H is the Hermitian matrix with $1 - i$ in the upper triangle and 0 on the main diagonal. Since the eigenvalues of H are real, it follows that R is nonsingular and hence $\sqrt{A}$ is nonsingular.

The second part of the proof begins by using the first part with ν playing the role of n and then constructing a column ν-vector u with no zero entries such that

$$\rho(\sqrt{uu^*}) = \nu. \tag{92}$$

As above, $u_\nu = 1$. Let v be the column n-vector defined by

$$v = \begin{bmatrix} u \\ e_{n-\nu} \end{bmatrix},$$

where $e_{n-\nu}$ is the column $(n - \nu)$-vector, all of whose entries are 1. We show that the required A for which (37) holds is $A = vv^*$. Note that

$$A = \begin{bmatrix} uu^* & ue_{n-\nu}^* \\ e_{n-\nu}u^* & e_{n-\nu}e_{n-\nu}^* \end{bmatrix}$$

$$= \begin{array}{c} \\ \nu \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cc} \nu & n-\nu \end{array} \\ \left[\begin{array}{c|c} uu^* & \begin{matrix} u_1 \cdots u_1 \\ \vdots \\ u_\nu \cdots u_\nu \end{matrix} \\ \hline \begin{matrix} \overline{u}_1 \cdots \overline{u}_\nu \\ \vdots \\ \overline{u}_1 \cdots \overline{u}_\nu \end{matrix} & J_{n-\nu} \end{array}\right] \end{array} \tag{93}$$

$$= \left[\begin{array}{c|c|c} \hat{u}\hat{u}^* & \begin{matrix} u_1 \\ \vdots \\ u_{\nu-1} \end{matrix} & \begin{matrix} u_1 & \cdots & u_1 \\ \vdots & & \vdots \\ u_{\nu-1} & \cdots & u_{\nu-1} \end{matrix} \\ \hline \overline{u}_1 \cdots \overline{u}_{\nu-1} & 1 & 1 \cdots 1 \\ \hline \begin{matrix} \overline{u}_1 \cdots \overline{u}_{\nu-1} \\ \vdots \\ \overline{u}_1 \cdots \overline{u}_{\nu-1} \end{matrix} & \begin{matrix} 1 \\ \vdots \\ 1 \end{matrix} & J_{n-\nu} \end{array}\right],$$

where $\hat{u}$ is the column $(\nu - 1)$-vector obtained from u by deleting the last component $u_\nu = 1$. The matrix $\sqrt{A}$ results from A by simply computing the square root of every entry in A. Since columns $\nu, \cdots, n$, in $\sqrt{A}$ are identical it follows that $\rho(\sqrt{A}) \leqq \nu$. However, the upper left ν-square block in $\sqrt{A}$ is $\sqrt{uu^*}$ and it is nonsingular. This completes the proof of Theorem 5.

3. Inertia. In this section we continue to assume that $A \geqq 0$, $\rho(A) = 1$, and that A has no zero or negative entries. As we saw in the derivation of (30), $\sqrt{A}$ is unitarily congruent to E_A and hence the inertia of $\sqrt{A}$ and $E_A = [\varepsilon(\varphi_p - \varphi_q)]$ are the same [5, p. 223]. Recall that the inertia of any n-square Hermitian matrix H is the triple of integers $In(H) = (\text{pos}, \text{neg}, \text{zer})$ where pos is the number of positive eigenvalues of H, neg is the

number of negative eigenvalues of H, and zer is the number of zero eigenvalues of H. As we saw in § 1, the arguments of the vector u for which $A = uu^*$ may be normalized so that (24) holds. If $s(A) < \pi$, then from Corollary 1, $\rho(\sqrt{A}) = 1$. Since $\sqrt{A}$ is Hermitian we conclude that

$$In(\sqrt{A}) = (1, 0, n-1). \tag{94}$$

Thus assume that $s(A) > \pi$, so that $\varphi_1 > \pi$, and, as in the proof of Theorem 2, define k to be the largest integer such that $\varphi_1 - \varphi_k < \pi$. Define m, as before, to be the largest integer such that $\varphi_m > \pi$. Possibly $m = 1$, and in any event, $m \leqq k$ (see diagram (38)). Next, define k_t to be the least integer such that

$$\varphi_{k_t} < \varphi_t - \pi, \qquad t = 1, \cdots, m. \tag{95}$$

Thus the set $\{\varphi_{k_t}, \cdots, \varphi_n\}$ are precisely all the φ_j which are less than $\varphi_t - \pi$, $t = 1, \cdots, m$. Note that $k_1 = k + 1$. Since $\varphi_1 > \pi$ and $\varphi_1 \geqq \varphi_2 \geqq \cdots \geqq \varphi_n = 0$ it follows that

$$k_1 \leqq k_2 \leqq \cdots \leqq k_m. \tag{96}$$

For example, in the diagram (42), $k_1 = 6$, $k_2 = 8$, $k_3 = 8$, $k_4 = 9$. Moreover, the number of elements in the set $\{\varphi_{k_t}, \cdots, \varphi_{10}\}$ is the number of -1's in $M_{(t)}$, $t = 1, \cdots, m$ (i.e., M is the matrix defined in (39)). If two (or more) successive k_t are the same, the corresponding rows $M_{(t)}$ are identical. If two (or more) successive k_t are distinct, then the corresponding rows $M_{(t)}$ are linearly independent. Also, rows $m + 1, \cdots, k$ (if any) in M are identical and consist entirely of 1's. There are such rows of 1's only in the case where $m < k$. Thus the rank of $M + J_{k,n-k}$ is the number, δ, of nonzero differences in the $(m - 1)$-tuple

$$[k_2 - k_1, k_3 - k_2, \cdots, k_m - k_{m-1}] \tag{97}$$

in the case where $m = k$, and the rank of $M + J_{k,n-k}$ is $\delta + 1$ if $m < k$. It is important to note that if $m = 1 < k$ then $\rho(M + J_{k,n-k}) = 1$ because the first row of $M + J_{k,n-k}$ is 0 and the remaining $k - 1$ rows consist entirely of 2's. We summarize these observations in the following theorem that expresses $\rho(W)$ (see (50)) in terms of the arguments of the vector u (see (18)).

THEOREM 6. *Let $A \geqq 0$, $\rho(A) = 1$, and assume A has no zero or negative entries. Suppose that $A = uu^*$ and that the arguments $\varphi_1, \cdots, \varphi_n$ of u are normalized so that*

$$\varphi_1 \geqq \cdots \geqq \varphi_n = 0.$$

Let k be the largest integer such that $\varphi_1 - \varphi_k < \pi$ and let m be the largest integer such that $\varphi_m > \pi$. Also, let W be the matrix in (50), and if $m > 1$ let k_t, $t = 1, \cdots, m$, be as defined in (95) and then define δ to be the number of nonzero differences among the components of (97). Then

$$\rho(W) = \begin{cases} 0 & \text{if } k = 1, \\ 1 & \text{if } k > 1 \text{ and } m = 1, \\ \delta + 1 & \text{if } 1 < m < k, \\ \delta & \text{if } 1 < m = k. \end{cases} \tag{98}$$

It is straightforward to compute the inertia of $\sqrt{A}$ in terms of $\rho(W)$.

THEOREM 7. *Let $A \geqq 0$, $\rho(A) = 1$, and assume that A has no zero or negative entries. If $s(A) < \pi$, then*

$$In(\sqrt{A}) = (1, 0, n-1). \tag{99}$$

If $s(A) > \pi$, *then*

$$In(\sqrt{A}) = (\rho(W)+1, \rho(W), n-1-2\rho(W)) \tag{100}$$

in which $\rho(W)$ *is determined entirely by the arguments* $\varphi_1, \cdots, \varphi_n$ *of* u *in the formula* (98).

Proof. If $s(A) < \pi$, then (99) was established above in (94). If $s(A) > \pi$, then from (57) and (30)

$$\begin{aligned} \rho(\sqrt{A}) &= \rho(E_A) \\ &= 2\rho(W)+1. \end{aligned} \tag{101}$$

Also, from (55), E_A is congruent to the matrix on the right side of (55), in which W is $(k-1)\times(n-k)$. If k were 1, then W does not appear in (55) and the right-hand side of that equation is the n-square matrix whose single nonzero entry is a 1 in the (1, 1) position. Then (100) holds for $\rho(W) = 0$. Thus assume $k > 1$ so that the right-hand side of (55) is

$$R = \left[\begin{array}{c|c} E_{11} & \begin{matrix} 0\cdots 0 \\ W \end{matrix} \\ \hline \begin{matrix} 0 \\ \vdots \quad W^T \\ 0 \end{matrix} & 0 \end{array}\right],$$

in which E_{11} is k-square and W is $(k-1)\times(n-k)$. Then $In(\sqrt{A}) = In(R)$ [5, Thm. 4.5.8]. Since R is the direct sum of 1 and the $(n-1)$-square matrix

$$H = \begin{bmatrix} 0 & W \\ W^T & 0 \end{bmatrix} \tag{102}$$

we write $W = UDV$, the singular value decomposition of W. Let $\rho(W) = h$ and let $\alpha_1 \geqq \cdots \geqq \alpha_h > 0$ be the positive singular values of W. Then H is unitarily similar to

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \otimes \operatorname{diag}(\alpha_1, \cdots, \alpha_h) \oplus 0_{n-1-2h}. \tag{103}$$

The eigenvalues of (103) are $\pm\alpha_t$, $t = 1, \cdots, h$, and 0, $n-1-2h$, times. Thus

$$In(R) = (h+1, h, n-1-2h).$$

Since $h = \rho(W)$, the proof is complete. □

Theorems 6 and 7 provide us with a straightforward algorithm for computing $\rho(W)$, and from this value, both $\rho(\sqrt{A})$ and $\ln(\sqrt{A})$. We start with an arbitrary $A = uu^*$, with no zero or negative entries. Following the MATLAB™ convention we use "%" for documentation.

Step 1. Take $u = A^{(1)}$, the first column of A.
Step 2. If $u_t = r_t e^{i\varphi_t}$, $t = 1, \cdots, n$, sort and translate the vector $\varphi = [\varphi_1, \cdots, \varphi_n]$ so that $\varphi_1 \geqq \cdots \geqq \varphi_n = 0$.
Step 3. If $\varphi_1 < \pi$ then set $\rho(W) = 0$ and end % $\rho(\sqrt{A}) = 1$.

Step 4. Determine the largest k such that $\varphi_1 - \varphi_k < \pi$. If $k = 1$ then set $\rho(W) = 0$ and end. % $\rho(\sqrt{A}) = 1$.

Step 5. Determine the largest m such that $\varphi_m > \pi$. If $m = 1$ then set $\rho(W) = 1$ and end. % if $k > 1$ and $m = 1$ then $\rho(W) = 1$ and $\rho(\sqrt{A}) = 3$.

Step 6. For $t = 1, \cdots, m$, determine the least k_t such that $\varphi_{k_t} < \varphi_t - \pi$. % $k > 1$ and $m > 1$.

Step 7. Determine the number, δ, of positive components of $[k_2 - k_1, \cdots, k_m - k_{m-1}]$.

Step 8. Set $\rho(W) = \delta + 1$ if $m < k$ and set $\rho(W) = \delta$ if $m = k$. % $\rho(W) = \delta + 1$ if $1 < m < k$, $\rho(W) = \delta$ if $1 < m = k$.

Step 9. end

The value of $\rho(W)$ produced by this algorithm satisfies

$$\rho(\sqrt{A}) = 2\rho(W) + 1, \tag{104}$$

$$In(\sqrt{A}) = (\rho(W) + 1, \rho(W), n - 1 - 2\rho(W)). \tag{105}$$

For, if $\varphi_1 < \pi$, then $\rho(W) = 0$ and both (104) and (105) are correct. If $\varphi_1 > \pi$ and $k = 1$, then $\rho(W) = 0$ and again (104) and (105) are correct. If $m = 1$, then the conclusion is the same. Finally, if $k \geqq m > 1$, then (98) and (100) again confirm that (104) and (105) are correct for the value of $\rho(W)$ produced by the algorithm.

As an example, for the matrix described in the diagram (42), $n = 10$, $m = 4$, $k = 5$, $k_1 = 6$, $k_2 = 8$, $k_3 = 8$, $k_4 = 9$,

$$[k_2 - k_1, k_3 - k_2, k_4 - k_3] = [2, 0, 1],$$

$\delta = 2$ and $\rho(W) = \delta + 1 = 3$. This is confirmed by (58).

4. Some computations. It is quite simple to write a MATLAB™ program to generate random complex $A \geqq 0$ of rank 1 and tabulate $\rho(\sqrt{A})$. In Table 1 the row headings indicate n, the dimension of A. The integral column headings indicate rank. For each $n = 3, \cdots, 10$, 250 random A were generated. The entry in the (p, q) position is a count of the number of p-square $A \geqq 0$, $\rho(A) = 1$, for which $\rho(\sqrt{A}) = q$. For example, the (7, 5) entry of Table 1 indicates that of the 250 random 7-square A, 87 of the matrices $\sqrt{A}$ had rank 5. The column headed t is the number of elapsed seconds required to complete the computation for each n. The computation was done on a Macintosh II™ computer with 1 MB of main memory. The program does not test for zero or negative entries in A. The listing that follows will produce an output formatted as

TABLE 1

n/ρ	1	3	5	7	9	t
3	199	51	-	-	-	17
4	122	128	-	-	-	22
5	80	158	12	-	-	28
6	51	151	48	-	-	36
7	27	133	87	3	-	46
8	12	114	105	19	-	57
9	8	90	117	32	3	70
10	5	61	116	66	2	85

indicated in the typical session following the listing (we have omitted the sum and time elapsed):

```
B=[ ];
v=[ ];
R=[ ];
i=sqrt(-1);
n=input('enter size of u: ')
j=input('enter number of iterations: ')
disp(' ')
starttime=fix(clock);
for k=1:j
    x=2*rand(1,n)-ones(1,n);
    y=2*rand(1,n)-ones(1,n);
    u=x+i*y;
    A=u'*u;
    B=sqrt(A);
    v=[v rank(B)];
end
for r=1:n
    len=length(find(v==r));
    R(r,1)=r;
    R(r,2)=len;
end
disp('Rank    Number of times ');
disp(R)
disp('The sum of the number of times is: ')
disp(sum(R(:,2)))
elapsed_time=etime(fix(clock),starttime)
```

```
enter size of u: 5
enter the number of iterations: 250
```

Rank	Number of times
1	80
2	0
3	158
4	0
5	12

5. Further work. The authors are currently investigating various extensions of the work in the present paper. These include:

1) The rank of general Hadamard powers of $A \geqq 0$, $\rho(A) = 1$;

2) The extent of the numerical range $W(\sqrt{A})$ when $A \geqq 0$, $\rho(A) = 1$, but A has negative entries so that $\sqrt{A}$ is no longer Hermitian;

3) $\rho(\sqrt{A})$ when $A \geqq 0$, $\rho(A) > 1$, and A has no negative entries;

4) $\rho(\sqrt{A})$ when A is normal, $\rho(A) \geqq 1$;

5) The distribution of the eigenvalues of $\sqrt{A}$ for $A \geqq 0$, $\rho(A) = 1$, and A has no negative entries.

6) The probability distribution of $\rho(\sqrt{A})$ for randomly generated A, $A \geqq 0$, $\rho(A) = 1$.

Acknowledgments. The authors thank Professor Roger Horn for providing us with preliminary copies of [4], the section on Hadamard Matrix Functions in his forthcoming book with C. R. Johnson [6], and a reprint of [2]. We also thank the referees for many helpful suggestions and for calling our attention to [1].

REFERENCES

[1] P. FARJOT, *Rang des matrices hermitiennes semi-définies positive indéfiniment divisibles*, Linear Algebra Appl., 15 (1976), pp. 189–196.
[2] C. H. FITZGERALD AND R. A. HORN, *On fractional Hadamard powers of positive definite matrices*, J. Math. Anal. Appl., 61 (1977), pp. 633–642.
[3] P. R. HALMOS, *Finite-Dimensional Vector Spaces*, Second Edition, Van Nostrand, Princeton, NJ, 1958.
[4] R. A. HORN, *The Hadamard product*, in Proc. Symposia in Applied Mathematics, American Mathematical Society, Providence, RI, to appear.
[5] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, 1985.
[6] ———, *Topics in Matrix Analysis*, Cambridge University Press, New York, to appear.
[7] M. MARCUS AND N. KHAN, *A note on the Hadamard product*, Canad. Math. Bull., 2 (1959), pp. 81–83.
[8] M. MARCUS AND H. MINC, *Introduction to Linear Algebra*, Dover, Mineola, New York, 1988.
[9] L. MIRSKY, *An Introduction to Linear Algebra*, Oxford, London, 1955.
[10] C. MOLER, S. HERSKOVITZ, J. LITTLE, AND S. BANGERT, MATLAB™ *for Macintosh Computers, User's Guide*, The Math Works, Inc., South Natick, MA, 1988.
[11] I. SCHUR, *Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen*, J. Reine Angew. Math., 140 (1911), pp. 1–28.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 1, pp. 70–83, January 1991

007

ON SIMULTANEOUS CONGRUENCE AND NORMS OF HERMITIAN MATRICES*

STEPHEN PIERCE† AND LEIBA RODMAN‡

Abstract. Let $A_1, \cdots, A_m, B_1, \cdots, B_m$ be $n \times n$ complex Hermitian matrices. It is said that $B_1, \cdots, B_m$ are simultaneously congruent to $A_1, \cdots, A_m$ if there exists an invertible S such that $S^*A_iS = B_i$, $i = 1, \cdots, m$. In this paper, $\inf \|I - S\|$, as S ranges over all invertible matrices which afford this simultaneous congruence, are studied. If one of the A_i is positive definite, it turns out that the growth of $\inf \|I - S\|$ is of the same magnitude as that of $\|B_1 - A_1\| + \cdots + \|B_m - A_m\|$. A counterexample with $m = 2$ is given to show that this result can be false if none of the A_i's is positive definite. An analogous result for simultaneous unitary congruence of matrices is also proved.

Key words. simultaneous congruence, norms

AMS(MOS) subject classifications. primary 15A57; secondary 15A60

1. Main results. Let $A = (A_1, \cdots, A_m)$ be an (ordered) m-tuple of (complex) Hermitian $n \times n$ matrices, and consider the set $C(A)$ of all m-tuples $B = (B_1, \cdots, B_m)$ of (necessarily Hermitian) $n \times n$ matrices that are *simultaneously congruent* to A, i.e., for some $S \in \mathrm{GL}(n, \mathbf{C})$ the equalities $B_j = S^*A_jS$, $j = 1, \cdots, m$ hold. (As usual, we denote by $\mathrm{GL}(n, \mathbf{C})$ the group of all invertible $n \times n$ matrices with complex entries.) In this paper we study the following property of the m-tuple A as above. There is a constant $K > 0$ (depending on A only) such that for every $(B_1, \cdots, B_m) \in C(A)$ there is an $S \in \mathrm{GL}(n, \mathbf{C})$ with

$$B_j = S^*A_jS, \qquad j = 1, \cdots, m,$$

$$\|I - S\| \leqq K \sum_{i=1}^{m} \|B_i - A_i\|$$

(here the norm $\|\cdot\|$ is chosen in advance). It will be convenient to use in this paper mainly the Frobenius norm

$$\|[a_{ij}]_{i,j=1}^{n}\| = \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2};$$

occasionally, the operator norm $\|A\|_H$ (the largest singular value of A) will be used; the operator norm will be distinguished by the subscript "H." If such $K > 0$ exists we say that A has the *Lipschitz property* with respect to simultaneous congruence.

Lipschitz properties of similarity of matrices have been studied in [GR]. Problems related to the Lipschitz properties of congruence and similarity have been studied in the operator-theoretic context (existence of local and global continuous cross sections); see, e.g., [DF], [P], [AFHV], and [AFHPS]. We also mention (as an inspiration to this paper) that Lipschitz behavior of solutions to various other problems recently became a subject of extensive study (see, e.g., [MS], [A]).

* Received by the editors April 19, 1989; accepted for publication October 25, 1989.

† San Diego State University, Department of Mathematical Sciences, San Diego, California 92182 (Q300085@CALSTATE.BITNET). The work of this author was partially supported by National Science Foundation grant DMS-8601959.

‡ The College of William and Mary, Department of Mathematics, Williamsburg, Virginia 23185 ($LXRODM@WMMVS). The work of this author was partially supported by National Science Foundation grant DMS-8802836.

It was shown in [PR] that in the case where $m = 1$ every Hermitian matrix has the Lipschitz property with respect to congruence. This is no longer true if $m \geqq 2$; for the reader's convenience we reproduce here the counterexample given in [PR].

Example 1.1. Let

$$A_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

It is easy to see that for every $\alpha \neq 0$ the matrices

$$B_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} (= A_1), \qquad B_2 = \begin{bmatrix} \alpha^2 & 1 \\ 1 & 0 \end{bmatrix}$$

are simultaneously congruent to A_1, A_2 (in fact,

$$\begin{bmatrix} 0 & \alpha \\ \alpha^{-1} & 0 \end{bmatrix} A_j \begin{bmatrix} 0 & \alpha^{-1} \\ \alpha & 0 \end{bmatrix} = B_j, \qquad j = 1, 2).$$

However, any invertible matrix T for which $T^* A_j T = B_j$, $j = 1, 2$ has the form

$$T = \begin{bmatrix} \pm r e^{i(\theta + \pi/2)} & |\alpha|^{-1} e^{i\theta} \\ |\alpha| e^{-i\theta} & 0 \end{bmatrix}$$

for some $r > 0$ and some real θ. For any such T we have

$$\| I - T \| \geqq |\alpha|^{-1}.$$

As $\alpha \to 0$ we see that (A_1, A_2) do not have the Lipschitz property with respect to simultaneous congruence.

Our main results are the following.

THEOREM 1.1. *Let $A = (A_1, \cdots, A_m)$ be an (ordered) m-tuple of $n \times n$ Hermitian matrices, and assume that at least one of them is positive definite. Then A has the Lipschitz property with respect to simultaneous congruence.*

In fact, A has the Lipschitz property with respect to simultaneous congruence provided that some linear combination of $A_1, \cdots, A_m$ is positive definite (we are indebted to C. R. Johnson for this observation). To prove this, apply Theorem 1.1 for the $(m + 1)$-tuple

$$\left(A_1, \cdots, A_m, \sum_{i=1}^{m} c_i A_i \right),$$

where the real numbers $c_1, \cdots, c_m$ are such that

$$\sum_{i=1}^{m} c_i A_i$$

is positive definite.

Thus, Theorem 1.1 settles in the affirmative the conjecture stated in [PR].

An analogous result (without the positive-definiteness assumption) holds for simultaneous unitary congruence of matrices. To describe this result, we introduce the following notation. For an (ordered) m-tuple $N = (N_1, \cdots, N_m)$ of $n \times n$ complex matrices, let $UC(N)$ be the set of all m-tuples $(M_1, \cdots, M_m)$ that are simultaneously unitarily similar to N, i.e., $M_j = U^* N_j U$, $j = 1, \cdots, m$ for some unitary $n \times n$ matrix U.

THEOREM 1.2. *Let $N = (N_1, \cdots, N_m)$ be an (ordered) m-tuple of $n \times n$ matrices. Then there is a constant $K \geqq 0$ (depending on N only) such that for every $(M_1, \cdots, M_m) \in UC(N)$ there is a unitary U with*

$$M_j = U^* N_j U, \qquad j = 1, \cdots, m$$

and

$$\|I - U\| \leqq K \sum_{i=1}^{m} \|N_i - M_i\|.$$

This theorem can be obtained as a corollary of Theorem 1.1. Indeed, apply Theorem 1.1 to the $(2m + 1)$-tuple of Hermitian matrices

$$(I, \operatorname{Re} N_1, \operatorname{Im} N_1, \cdots, \operatorname{Re} N_m, \operatorname{Im} N_m), \tag{1.1}$$

where we denote $\operatorname{Re} X = \frac{1}{2}(X + X^*)$; $\operatorname{Im} X = 1/2i(X - X^*)$. The $(2m + 1)$-tuple of Hermitian matrices which is simultaneously congruent to (1.1) is taken in the form

$$(I, \operatorname{Re} M_1, \operatorname{Im} M_1, \cdots, \operatorname{Re} M_m, \operatorname{Im} M_m).$$

The proof of Theorem 1.1 will be given in the next two sections.

2. A local theorem. In this section we prove a local result on which the proof of Theorem 1.1 is based.

Let $A = \{A_1, \cdots, A_m\}$ be a finite family of $n \times n$ Hermitian matrices. Let $U(n)$ be the group of all $n \times n$ complex unitary matrices. We say that $U \in U(n)$ *centralizes* A if $UA_k = A_k U$ for all $k = 1, \cdots, n$. For each $U \in U(n)$ but not centralizing A, define

$$f_A(U) = \frac{\|I - U\|^2}{\sum_{k=1}^{m} \|UA_k - A_k U\|^2}. \tag{2.1}$$

THEOREM 2.1. *There exist positive constants K, ε depending only on $A_1, \cdots, A_m$, such that if*

$$\sum_{k=1}^{m} \|UA_k - A_k U\|^2 < \varepsilon,$$

then there exists a $V \in U(n)$ such that $V^* A_k V = U^* A_k U$, $k = 1, \cdots, m$, and $f_A(V) < K$.

Remark 2.1. Note that $\|UA_k - A_k U\| = \|U^* A_k U - A_k\|$.

Remark 2.2. For V to satisfy $V^* A_k V = U^* A_k U$, $k = 1, \cdots, m$, it is necessary and sufficient that $V = WU$ where W is unitary and centralizes A. In this case, the denominators in $f_A(U)$ and $f_A(V)$ will be the same; only the numerator can change.

Remark 2.3. If the theorem is false, there must be a sequence $\{U_j\}$ in $U(n)$ such that $\lim_{j \to \infty} U_j = I_n$, $f_A(U_j)$ is increasing, and $\lim_{j \to \infty} f_A(U_j) = \infty$. Moreover, we may assume that the sequence $\{U_j\}$ is the best possible in the following sense. If V_j is any sequence of unitary matrices centralizing A, then $\|I - V_j U_j\| \geqq \|I - U_j\|$. This means that every subsequence of the $f_A(U_j)$ diverges and cannot be "repaired." We will show that there is a subsequence of the $f_A(U_j)$ that can, in fact, be "fixed" and thus prove the theorem.

The proof involves consideration of several cases.

Case I. *The irreducible case.* We assume in this section that $A = \{A_k\}_{k=1}^{m}$ forms an irreducible set, i.e., they have no common proper invariant subspace in $\mathbf{C}^n$. Then the centralizer of A consists of scalar matrices only and hence we may modify the U_j only

using multiplication by a scalar of modulus one. We will therefore assume that one occurs as an eigenvalue in every U_j. For each j, let λ_j be the eigenvalue of U_j such that $|1-\lambda_j| \geqq |1-\lambda|$ for any other eigenvalue λ of U_j.

For each j, choose $V_j \in U(n)$ such that

$$V_j^* U_j V_j = X_j \oplus Y_j$$

and if μ, λ are eigenvalues of X_j and Y_j, respectively, then

$$|\mu - \lambda| \geqq |1-\lambda_j|/n.$$

The possibility of such choice of V_j follows from elementary geometric considerations. In addition, let $B_{kj} = V_j^* A_k V_j$, $k = 1, \cdots, m$. Clearly,

$$f_A(U_j) = \frac{\|I - X_j \oplus Y_j\|^2}{\sum_{k=1}^m \|B_{kj}(X_j \oplus Y_j) - (X_j \oplus Y_j)B_{kj}\|^2}.$$

Partition B_{kj} conformally with $X_j \oplus Y_j$, obtaining

$$B_{kj} = \begin{bmatrix} B_{kj1} & B_{kj2}^* \\ B_{kj2} & B_{kj4} \end{bmatrix}.$$

Note that

$$\sum_{k=1}^m \|X_j B_{kj2} - B_{kj2} Y_j\|^2$$

is part of the denominator in the representation of $f_A(U_j)$ above.

Now $X_j \otimes I - I \otimes Y_j$ is a normal linear transformation whose smallest singular value is at least $|1-\lambda_j|/n$. Set

$$t_j = \sum_{k=1}^m \|B_{kj2}\|^2.$$

Since $\{A_1, \cdots, A_m\}$ are irreducible, we can never have $t_j = 0$.

A simple reasoning will verify that $f_A(U_j) \leqq n^3/t_j$. Indeed, the eigenvalues of $X_j \oplus Y_j$ are also those of U_j. So

$$\|I - X_j \oplus Y_j\|^2 \leqq n|1-\lambda_j|^2.$$

Also,

$$\sum_{k=1}^m \|X_j B_{kj2} - B_{kj2} Y_j\|^2 = \sum_{k=1}^m \|(X_j \otimes I - I \otimes Y_j)B_{kj2}\|^2 \geqq \alpha_j^2 t_j,$$

where α_j is the smallest singular value of $X_j \otimes I - I \otimes Y_j$. Now

$$f_A(U_j) \leqq \frac{\|I - X_j \oplus Y_j\|^2}{\sum_{k=1}^m \|X_j B_{kj2} - B_{kj2} Y_j\|^2} \leqq \frac{n|1-\lambda_j|^2}{\alpha_j^2 t_j} \leqq \frac{n^3}{t_j},$$

as claimed.

Next, we also must have $\{t_j\}$ bounded away from zero. If this were not the case, some subsequence of the V_j would converge to a unitary matrix V such that the matrices $V^* A_k V$ would have a commonly placed zero block indicating reducibility. □

Remark 2.4. The sizes of the B_{kj2} in general will not be the same. This causes no problem with the above argument. If some subsequence of the t_j approaches zero, we may assume that the corresponding subsequence of the B_{kj2} all have the same size.

Case II. *The reducible case. Preliminary results.* This case does not seem to yield to a simple induction argument. Thus we will prove a series of propositions to be used later.

PROPOSITION 2.2. *Let* $\{B_1, \cdots, B_m\}$, $\{C_1, \cdots, C_m\}$ *be two sets of Hermitian matrices. Assume that the only matrix* X *satisfying* $B_i X = XC_i$, $i = 1, \cdots, m$, *is* $X = 0$. *Then there is a constant* $K > 0$ *independent of* X *such that for all matrices* X *of suitable size the inequality*

$$\| X \|^2 \leqq K\left(\sum_{k=1}^{m} \| B_k X - X C_k \|^2 \right)$$

holds.

Proof. Define a linear transformation T on matrices the size of X by $T(X) = (B_1 X - X C_1, \cdots, B_m X - X C_m)$. Obviously the kernel of T is zero and hence its smallest singular value α_1 is positive. Thus we can take $K = 1/\alpha_1$, which depends only on the C_i and B_i. □

PROPOSITION 2.3. *Let* X *and* Y *be upper triangular matrices of sizes* m *and* n, *respectively. Suppose* $\sigma(X) = \{\lambda_1, \cdots, \lambda_m\}$ *and* $\sigma(Y) = \{\mu_1, \cdots, \mu_n\}$. *Assume further that* $0 < p = \min |\lambda_i - \mu_j|$. *Then for any* $m \times n$ *matrix* $B = [b_{ij}]_{1 \leqq i \leqq m; 1 \leqq j \leqq n}$, *we have*

$$\| XB - BY \| \geqq pq,$$

where q *is defined as follows*:

$q = |b_{m1}|$ *unless* $b_{m1} = 0$; *otherwise*,

$q = \max(|b_{m,2}|, |b_{m-1,1}|)$ *unless* $b_{m,2} = b_{m-1,1} = 0$; *otherwise*,

$q = \max(|b_{m,3}|, |b_{m-1,2}|, b_{m-2,1}|)$, *etc*.

Remark 2.5. Note the ordering of the entries in B, starting at the lower left-hand corner and working toward the upper right-hand corner in a back and forth fashion.

Proof. We can assume that the diagonal of X is $\lambda_1, \cdots, \lambda_m$ (in this order) and the diagonal of Y is $\mu_1, \cdots, \mu_n$ (in this order). Let $Z = XB - BY = [z_{ij}]_{1 \leqq i \leqq m; 1 < j \leqq n}$. Since X and Y are upper triangular, we observe that $z_{m1} = b_{m1}(\lambda_m - \mu_1)$. Thus, if $b_{m1} \neq 0$, then we are done. If $b_{m1} = 0$, then we note that

$$z_{m2} = b_{m2}(\lambda_m - \mu_2) \quad \text{and} \quad z_{m-1,1} = b_{m-1,1}(\lambda_{m-1} - \mu_1).$$

Thus we are done unless $b_{m,2} = b_{m-1,1} = 0$. Continue inductively until the result is established. □

Remark 2.6. If X and Y were diagonal, we could use an argument as in Proposition 2.2 and replace pq with the better bound $p\|B\|$, because p is then the minimum singular value of the linear map $X \otimes I - I \otimes Y$.

PROPOSITION 2.4. *Let* W *be an* $rs \times rs$ *matrix partitioned into* $s \times s$ *blocks* $W_{ij}(i, j = 1, \cdots, r)$. *Assume that* W *is a contraction*, i.e., $\|W\|_H \leqq 1$. *Then there exists an* $r \times r$ *unitary matrix* P *such that the matrix* $Z = W(P \otimes I_{s \times s})$ *has the property that the blocks* $Z_{12}, \cdots, Z_{1s}$ *are all singular*.

Moreover, given an $\varepsilon > 0$, *there exists a* $\delta > 0$, *such that if* $\|I - W\| < \delta$, *then* P *may be chosen so that* $\|I - P\| < \varepsilon$.

Proof. The validity of the second statement of the proposition will be evident in the text of the proof. Also, we will choose P so that for $j = 2, \cdots, r$ the jth column of P lies in the span of the first j standard vectors $e_1, \cdots, e_j$.

First pick p_{12} and p_{22} so that $p_{12}W_{11} + p_{22}W_{12}$ is singular, $|p_{12}|^2 + p_{22}^2 = 1$, and $p_{22} \geqq 0$. This is clearly possible, and if W_{11} is close enough to $I_{s\times s}$, then (because W is a contraction) W_{12} is close to zero, and consequently p_{22} will have to be close to one. In fact, if W_{12} is already singular, we will choose $p_{12} = 0$ and $p_{22} = 1$. Now set $p_{32} = \cdots = p_{r2} = 0$, and we have selected column 2 of P. To get column 3 of P, first choose a vector v_3 in span $\{e_1, e_2\}$ which is orthogonal to column 2 of P, say $v_3 = p_{22}e_1 - \bar{p}_{12}e_2$. Then choose numbers t and p_{33} so that $tp_{22}W_{11} - t\bar{p}_{12}W_{12} + p_{33}W_{13}$ is singular, $p_{33} \geqq 0$, and $t^2p_{22}^2 + t^2|p_{12}|^2 + p_{33}^2 = 1$. Note, as before, that if W_{11} is close to $I_{s\times s}$, then we will have to choose p_{33} close to one. Now we let column 3 of P be $tp_{22}e_1 - t\bar{p}_{12}e_2 + p_{33}e_3$. Continue in this fashion, to obtain columns 4, $\cdots$, r of P and then choose column 1 by Gram–Schmidt with the (1, 1) entry close to 1 (if W_{11} is close to I). □

PROPOSITION 2.5. *Let $\{B_1, \cdots, B_m\}$ be an irreducible set of $n \times n$ Hermitian matrices. Then there is a constant $L > 0$ depending only on the B_i such that for any singular matrix X we have*

$$\|X\|^2 \leqq L\left(\sum_{k=1}^{m} \|B_kX - XB_k\|^2\right).$$

Proof. This inequality is homogeneous in X, so we will confine our examination to those X which lie on the unit sphere S consisting of all X such that $\|X\| = 1$. Let V_0 be the matrices of determinant 0 and V_1 the scalar matrices. The $V_0 \cap S$ and $V_1 \cap S$ are disjoint closed sets in S and hence have a positive distance between them. Let T be the linear map on $n \times n$ matrices given by $T(X) = (B_1X - XB_1, \cdots, B_mX - XB_m)$. Clearly, the scalar matrices are exactly the kernel of T. Thus $\{\|T(X)\|^2, X \in V_0 \cap S\}$ must be bounded away from zero by a constant depending only on T. This completes the proof.

PROPOSITION 2.6. *Let $B_1, \cdots, B_m$ be an irreducible set of $s \times s$ Hermitian matrices. Then there is a constant $L_1 > 0$ depending only on $B_1, \cdots, B_m$ and on the positive integer r such that the following holds: Given any $rs \times rs$ contraction $W = (W_{ij})$ (the W_{ij} are all $s \times s$), there is an $r \times r$ unitary matrix P such that $Z = W(P \otimes I_s) = [Z_{ij}]_{1 \leqq i \leqq r; 1 \leqq j \leqq r}$ satisfies*

$$\sum_{i<j} \|Z_{ij}\|^2 \leqq L_1\left(\sum_{k=1}^{m} \sum_{i<j} \|B_kZ_{ij} - Z_{ij}B_k\|^2\right).$$

Proof. By Proposition 2.4, we will first assume that $W_{12}, \cdots, W_{1r}$ are all singular. If $W_{12}, \cdots, W_{1r}$ are all zero, we can finish by induction on r (observe that for $r = 2$ Proposition 2.6 is easily gotten by combining Propositions 2.4 and 2.5). Otherwise, let W_0 be the $(r-1)s \times (r-1)s$ principal submatrix of W obtained by deleting the first block row and column. By induction, choose an $(r-1) \times (r-1)$ unitary matrix P_0 such that $Z_0 = W_0(P_0 \otimes I_{s\times s}) = [Z_{ij}]_{2 \leqq i \leqq r; 2 \leqq j \leqq r}$ satisfies the result of the theorem, namely, that

$$\sum_{1<i<j} \|Z_{ij}\|^2 \leqq L_1\left(\sum_{k=1}^{m} \sum_{1<i<j} \|B_kZ_{ij} - Z_{ij}B_k\|^2\right),$$

where L_1 depends only on $B_1, \cdots, B_m$. Now let $P = 1 \oplus P_0$, and set $Z = W \cdot (P \otimes I_{s\times s})$. Now

$$[Z_{12}, \cdots, Z_{1r}] = [W_{12}, \cdots, W_{1r}][P_0 \otimes I_{s\times s}].$$

Of course $Z_{12}, \cdots, Z_{1r}$ do not have to be singular, but since $P_0 \otimes I_{s\times s}$ is unitary, $\|[Z_{12}, \cdots, Z_{1r}]\|^2 = \|[W_{12}, \cdots, W_{1r}]\|^2$. Since $W_{12}, \cdots, W_{1r}$ are all singular and not all zero, we have from Proposition 2.5 that

$$\sum_{1=i<j} \|Z_{ij}\|^2 \leqq L_2\left(\sum_{j=2}^{r}\sum_{k=1}^{m} \|B_kW_{ij} - W_{ij}B_k\|^2\right),$$

where $L_2 > 0$ depends only on $B_1, \cdots, B_m$ and r. This completes the proof. □

We will use Propositions 2.4 and 2.6 when the contraction W is a principal submatrix of a unitary matrix. In connection with this we remark that any contraction is a principal submatrix of a unitary matrix.

Case III. *Proofs for the reducible case.* We will assume that $A = \{A_1, \cdots, A_m\}$ and the sequence $\{U_j\}$ are as before. In other words, $\{U_j\}_{j=1}^{\infty}$ is a sequence of $n \times n$ unitary matrices with the following properties:

(i) $\lim_{j\to\infty} U_j = I$;

(ii) The numbers $f_A(U_j) = \|I - U_j\|^2 / \sum_{k=1}^{m} \|U_jA_k - A_kU_j\|^2$ tend to infinity;

(iii) If $\{V_j\}_{j=1}^{\infty}$ is any sequence of unitary matrices centralizing A, then

$$\|I - V_jU_j\| \geqq \|I - U_j\|, \qquad j = 1, 2, \cdots.$$

We have to prove (by contradiction) that this sequence is impossible.

Because there have to be several stages of partitioning the A_k and U_j, we will introduce the required partitioning now. First, a definition. Two ordered sets of $n \times n$ Hermitian matrices $T^{(1)} = \{T_1^{(1)}, \cdots, T_m^{(1)}\}$ and $T^{(2)} = \{T_1^{(2)}, \cdots, T_m^{(2)}\}$ are called *equivalent* if there exists $S \in \mathrm{GL}(n, \mathbf{C})$ such that $S^{-1}T_i^{(1)}S = T_i^{(2)}$ $(i = 1, \cdots, m)$. In fact, if such an S exists, it can be chosen unitary. Since $A = \{A_1, \cdots, A_m\}$ is reducible, A is equivalent to $\{B_i \oplus C_i\}_{i=1}^{m}$, where $B = \{B_1, \cdots, B_m\}$ is irreducible. If $C = \{C_1, \cdots, C_m\}$ has an irreducible component equivalent to B, then A is equivalent to $\{(I_{2\times 2} \otimes B_i) \oplus C_i'\}_{i=1}^{m}$ for some $C' = \{C_1', \cdots, C_m'\}$. Continuing this process, we may assume that every A_k has the form

$$A_k = (I_{r\times r} \otimes B_k) \oplus C_k, \qquad k = 1, \cdots, m,$$

where the B_k are $s \times s$, $\{B_1, \cdots, B_m\}$ is irreducible and no irreducible component of $\{C_k\}_{k=1}^{m}$ is equivalent to $\{B_1, \cdots, B_m\}$. By Schur's lemma, the latter condition means that there is no nonzero matrix Q such that

$$(I_{r\times r} \otimes B_k)Q = QC_k; \qquad k = 1, \cdots, m.$$

Next, partition U_j conformally with the above partition of the A_k as

$$U_j = \begin{bmatrix} W_j & Y_j \\ X_j & Z_j \end{bmatrix},$$

where W_j is the same size as $I_{r\times r} \otimes B_k$, namely, $rs \times rs$. Next, partition W_j conformally with the block diagonal matrix $I_{r\times r} \otimes B_k$ as $(W_j)_{pq}$ where each $(W_j)_{pq}$ is $s \times s$. If the ith block row of (W_j) is $(W_j)_{i1}, \cdots, (W_j)_{ir}$, then the corresponding ith block row of Y_j is denoted $(Y_j)_i$.

There is one deeper partition that we will have to make. It will be necessary to choose for each j a block diagonal $rs \times rs$ unitary matrix P_j (the blocks on the diagonal are of size $s \times s$) such that the (p, p)th block of $P_j^* W_j P_j$ is upper triangular for every

$p = 1, \cdots, r$. We will denote the (p, p)th block of $P_j^*(I \otimes B_k)P$ as $(\tilde{B}_{kj})_p$, and the (p, p)th block of $P_j^* W_j P_j$ as $(P_j^* W_j P_j)_{pp}$. These blocks will have the form

$$(\tilde{B}_{kj})_p = \begin{bmatrix} D_{kjp} & E_{kjp}^* \\ E_{kjp} & F_{kjp} \end{bmatrix}, \qquad (P_j^* W_j P_j)_{pp} = \begin{bmatrix} R_{jp} & S_{jp} \\ 0 & Q_{jp} \end{bmatrix}.$$

The sizes of R_{jp} and Q_{jp} will depend on j and p. The way R_{jp} and Q_{jp} are chosen will be indicated later in the proof.

Let α, β be increasing integer sequences (not necessarily of the same length) between one and n. If U is $n \times n$, we let $U[\alpha, \beta]$ be the submatrix of U in the intersection of rows α and columns β. If $R_j = U_j[\alpha, \beta]$, we shall say that R_j is *bounded* if there is a constant L_3 independent of j (and hence dependent only on $A_1, \cdots, A_m$), such that

$$\| R_j \|^2 \leqq L_3 \left(\sum_{k=1}^{n} \| A_k[\alpha, \alpha] R_j - R_j A_k[\beta, \beta] \| \right).$$

If $R_j = U_j[\alpha, \alpha]$ is a principal submatrix of U_j, then we shall investigate instead the boundedness of $I - R_j$. We want to demonstrate the boundedness of $I - W_j$, $I - Z_j$, X_j, and Y_j (or some subsequence thereof). As an example of this idea, we note that X_j and Y_j must be bounded by Proposition 2.2. We will use the boundedness of X_j and Y_j in our proof.

Now we consider the $rs \times rs$ matrix W_j. We will show how to investigate W_j and from this it will be clear how to treat Z_j. First we observe that any unitary matrix of the form $(P \otimes I_{s \times s}) \oplus I_{t \times t}$ commutes will all A_k. If all $(W_j)_{pq}$ are zero, we can continue to the next step and bound $\| I - (W_j)_{pp} \|$ by using the irreducible case inductively. Otherwise, we use Proposition 2.6 to allow the assumption that if $p < q$, then $(W_j)_{pq}$ is bounded. We do this by multiplying through by a matrix of the form $(P \otimes I_{s \times s}) \oplus I$.

We now observe that since U_j is unitary, we have $\| X_j \|^2 = \| Y_j \|^2$, and hence

$$\sum_{p<q} \| (W_j)_{pq} \|^2 = \sum_{p>q} \| (W_j)_{pq} \|^2.$$

Therefore, since

$$\frac{\sum_{p<q} \| (W_j)_{pq} \|^2}{\sum_{k=1}^{m} \sum_{p<q} \| A_k (W_j)_{pq} - (W_j)_{pq} A_k \|^2}$$

is bounded in terms of the A_k alone, the same is true for

$$\frac{\sum_{p>q} \| (W_j)_{pq} \|^2}{\sum_{k=1}^{m} \sum_{p>q} \| A_k (W_j)_{pq} - (W_j)_{pq} A_k \|^2}.$$

It may be that for particular $p > q$, $(W_j)_{pq}$ is not bounded; the weaker statement above, however, will suffice for our purposes.

The most difficult part is to bound $I - (W_j)_{pp}$. We assume that every eigenvalue of $(W_j)_{pp}$ has a positive real part (indeed, $\{U_j\}$ is a sequence of unitary matrices approaching I, so any principal submatrix of $\{U_j\}$ will be close to I as well (for j large enough) and, in particular, all its eigenvalues will have positive real parts).

We may also assume that each $(W_j)_{pp}$ is "adjusted" by multiplication by some scalar α_{jp} of modulus one. The choice of α_{jp} is crucial. Since $(W_j)_{pp}$ is $s \times s$, let $\lambda_1, \cdots, \lambda_s$ be the eigenvalues of $(W_j)_{pp}$ after a suitable choice of α_{jp}.

We consider two cases. In the first case, it is possible to choose α_{jp} such that

$$\sum_{v} |1-\lambda_v|^2 \leqq \sum_{v} (1-|\lambda_v|^2).$$

This would occur, for example, if all λ_v were on or in the circle $r = \cos\theta$, i.e., the circle with center $\frac{1}{2}$ and radius $\frac{1}{2}$. We now wish to examine

$$\frac{\|I-(W_j)_{pp}\|^2}{\sum_{k=1}^{m} \|B_k(W_j)_{pp}-(W_j)_{pp}B_k\|^2}.$$

Let $(U_j)_p$ be the pth block row of U_j with $(W_j)_{pp}$ removed. Since $\|(U_j)_p\|^2 = \|(Y_j)_p\|^2 + \sum_{p\neq q} \|W_{pq}\|^2$, $(U_j)_p$ is bounded. Pick a block diagonal unitary matrix $P_j(n \times n)$ such that in $P_j^* U_j P_j$, the pth $s \times s$ principal block is upper triangular with main diagonal $\lambda_1, \cdots, \lambda_s$ and $1 - |\lambda_s| \geqq 1 - |\lambda_v|$, $v = 1, \cdots, s-1$ (this holds for $p = 1, \cdots, r$). Set $\tilde{U}_j = P_j^* U_j P_j$. Let $\tilde{Y}_j, \tilde{X}_j, \tilde{W}_j$ be defined in the obvious way. Note that all $s \times s$ blocks in $\tilde{U}_j$ have the same norms as the corresponding $s \times s$ blocks in U_j.

For notational convenience, let T be the pth principal $s \times s$ block in $\tilde{U}_j$ (this is short notation for $(\tilde{W}_j)_{pp}$). The last row of T is $(0, \cdots, 0, \lambda_s)$ and hence $1 - |\lambda_s|^2$ is the norm squared of a submatrix of $(\tilde{U}_j)_p$. Thus $1 - |\lambda_s|^2$ is under control and hence so is $\sum_v (1 - |\lambda_v|^2)$. Let N be the strictly upper triangular part of T and note that $|\lambda_1|^2 + \cdots + |\lambda_s|^2 + \|N\|^2 + \|(\tilde{U}_j)_p\|^2 = s$ since this is just the sum of squares of the absolute values of the entries in s rows of a unitary matrix. But

$$\|I-T\|^2 = \sum_{v} |1-\lambda_v|^2 + \|N\|^2 = 2\sum_{v} |1-\lambda_v|^2 - \|(U_j)_p\|^2$$

$$\leqq 2\sum_{v} (1-|\lambda_v|^2) - \|(U_j)_p\|^2,$$

and hence $\|I - T\|^2$ is clearly bounded because $\sum_v (1 - |\lambda_v|^2) = \|(U_j)_p\|^2$, and $(U_j)_p$ is already bounded. This concludes the first case.

Remark 2.7. The first case above is the one in which the λ_v are clustered sufficiently close together. It includes, in fact, the case $\lambda_1 = \cdots = \lambda_s$. The next case occurs when the λ_v are close to one in modulus, but far from one in actual distance. This forces $\|I - (W_j)_{pp}\|^2$ to be relatively large. We should also note that it was not necessary to inspect the denominator in the quotient

$$\frac{\|I-(W_j)_{pp}\|^2}{\sum_{k=1}^{m} \|B_k(W_j)_{pp}-(W_j)_{pp}B_k\|^2}.$$

The numerator could be bounded by direct comparison with other bounded submatrices in U_j.

For the second case we assume that T is as before and the corresponding $s \times s$ block in $\tilde{B}_{kj}$ is $\tilde{B}_{kjp}$ as indicated previously. We further assume that λ_s is positive and greater than $\frac{1}{2}$, that $|\lambda_s - \lambda_1| \geqq |\lambda_v - \lambda_\mu|$ for all v, μ, and that $\sum_v |1 - \lambda_v|^2 \geqq \sum_v (1 - |\lambda_v|^2)$. Split the eigenvalues into two disjoint sets $\{\lambda_1, \cdots, \lambda_\mu\}, \{\lambda_{\mu+1}, \cdots, \lambda_s\}$ such that the distance π between the two sets is at least $|1 - \lambda_s|/s$. If this were not possible, all λ_v would lie in the circle centered at λ_s with radius $1 - \lambda_s$, which is interior to the circle $r = \cos\theta$ and we would be in the first case. Clearly, λ_1 and λ_s must be in different sets.

Moreover, by choosing a subsequence of the U_j if necessary, we will assume that the size μ of the first set is independent of j. We now observe that for any v,

$$\frac{|1-\lambda_v|^2}{|\lambda_1-\lambda_s|^2} \leqq 2\frac{|1-\lambda_s|^2+|\lambda_v-\lambda_s|^2}{|\lambda_v-\lambda_s|^2} \leqq 4,$$

and this inequality is independent of j as long as we are in the second case.

Next observe that we can never have $E_{kjp} = 0$ for all k for any p or j. This would contradict the irreducibility of $B_1, \cdots, B_m$. Also, for the same reason there can be no p such that there is for all $k = 1, \cdots, m$ a subsequence of the E_{kjp} approaching zero.

Order the positions in E_{kjp} starting with the lower left-hand corner and proceeding toward the upper right as in the proof of Proposition 2.3. Thus if the size of E_{kjp} is $y \times z$, then the ordering of the positions is $(y, 1), (y, 2), (y-1, 1), (y-2, 1), (y-2, 1), (y-1, 2), (y, 3), \cdots, (1, z)$. Let (y_0, z_0) be the first position in this ordering satisfying $\lim_{j\to\infty} |(E_{kjp})_{y_0z_0}| \neq 0$. By Proposition 2.3 we must have

$$\|R_{jp}E_{kjp} - E_{kjp}Q_{jp}\|^2 \geqq \pi^2 |(E_{kjp})_{y_0z_0}|/2.$$

Recalling that $\pi \geqq |\lambda_1 - \lambda_s|/s$, we now see that if D is the diagonal part of T ($D = \operatorname{diag}(\lambda_1, \cdots, \lambda_s)$), then $\|I - D\|^2$ is also under control. Let N be the nilpotent (strictly upper triangular) part of T. Then

$$\sum_v (1 - |\lambda_s|^2) + \|N\|^2 + \|(\tilde{U}_j)_p\|^2 = s.$$

Since $\sum_v |1 - \lambda_v|^2 \geqq \sum_v (1 - |\lambda_v|^2)$, and $\|I - D\|^2$ is controlled, we have concluded the second case.

At this point we have shown the existence of a subsequence of the U_j (which we will take to be $\{U_j\}$) such that (see (2.1))

$$f_A(U_j) \leqq L_j,$$

where the L_j are constants obtained in terms of the A_i only. A glance at the proofs of the propositions in this section shows that the L_j can be chosen independent of j, i.e., all L_j can be chosen the same. This concludes the proof of Theorem 2.1.

3. Proof of Theorem 1.1. We start with a lemma.

LEMMA 3.1. *Let $A = (A_1, \cdots, A_m)$ be an m-tuple of $n \times n$ Hermitian matrices, and assume that at least one of $A_1, \cdots, A_m$ is positive definite. Then there are positive constants K and ε such that for every $(B_1, \cdots, B_m) \in C(A)$ with*

$$\sum_{i=1}^{m} \|B_j - A_j\| < \varepsilon$$

there exists $S \in \mathrm{GL}(n, \mathbf{C})$ with

$$B_j = S^*A_jS, \qquad j = 1, \cdots, m,$$

and

$$\|I - S\| \leqq K \sum_{i=1}^{m} \|B_j - A_j\|.$$

Proof. It is easy to see that if Lemma 3.1 holds for an m-tuple $A = (A_1, \cdots, A_m)$, then it holds also for any m-tuple of the form $\tilde{A} = (\tilde{A}_1, \cdots, \tilde{A}_m) =$

$(T^*A_1T, \cdots, T^*A_mT)$, where T is an invertible matrix. Indeed, let $\tilde{\varepsilon} = \varepsilon/\|T^{-1}\|^2$, and let

$$(\tilde{B}_1, \cdots, \tilde{B}_m) \in C(\tilde{A})$$

with

$$\sum_{i=1}^{m} \|\tilde{B}_i - \tilde{A}_i\| < \tilde{\varepsilon}.$$

Defining $B_j = T^{*-1}\tilde{B}_jT^{-1}$, we find that

$$\sum_{i=1}^{m} \|B_i - A_i\| < \varepsilon,$$

and hence (since we have assumed Lemma 3.1 holds for $(A_1, \cdots, A_m)$) there is $S \in \mathrm{GL}(n, \mathbf{C})$ with

$$B_i = S^*A_iS, \qquad i = 1, \cdots, m$$

and with

$$\|I - S\| \leqq K \sum_{i=1}^{m} \|B_i - A_i\|.$$

Letting $\tilde{S} = T^{-1}ST$, we see that

$$\tilde{B}_i = \tilde{S}^*\tilde{A}_i\tilde{S}, \qquad i = 1, \cdots, m.$$

Moreover,

$$\|I - \tilde{S}\| = \|I - T^{-1}ST\| \leqq \|T^{-1}\|\,\|I - S\|\,\|T\|$$

$$\leqq \|T^{-1}\|\,\|T\|\,K\|T^{-1}\|^2 \sum_{i=1}^{m} \|\tilde{B}_i - \tilde{A}_i\|,$$

and the result of Lemma 3.1 is verified for the m-tuple $\tilde{A}$.

Using this fact and the assumption that one of the A_j's, say A_1, is positive definite, we shall assume without loss of generality that $A_1 = I$.

Suppose S is an invertible matrix such that $\sum_{i=1}^{m} \|S^*A_iS - A_i\| < \varepsilon$, where ε is a fixed positive number. We need to replace S with an invertible T such that

$$T^*A_iT = S^*A_iS \qquad (i = 1, \cdots, m)$$

and

$$\|I - T\| \leqq K\left(\sum_{i=1}^{m} \|T^*A_iT - A_i\|\right),$$

where the constant K depends on A only, for a suitable choice of ε. Let $S = UH$ be the polar decomposition of S, with unitary U and positive definite H. In the sequel we denote by $K_1, K_2, \cdots$ positive constants that depend on A only. As

$$S^*A_1S - A_1 = S^*S - I = H^2 - I,$$

it follows that $\|H - I\| \leqq K_1\varepsilon$.

So

$$\sum_{i=1}^{m} \|U^*A_iU-A_i\| \leqq \sum_{i=1}^{m} \|S^*A_iS-A_i\| + \sum_{i=1}^{m} \|U^*A_iU-S^*A_iS\|$$

$$< \varepsilon + \sum_{i=1}^{m} \|(I-H)U^*A_iU(I-H)\|$$

$$+ \sum_{i=1}^{m} \|(I-H)U^*A_iUH\| + \sum_{i=1}^{m} \|HU^*A_iU(I-H)\|$$

$$\leqq K_2\varepsilon.$$

Using Theorem 2.1, and choosing a suitable $\varepsilon > 0$, we find unitary W such that

$$W^*A_iW = U^*A_iU \qquad (i = 1, \cdots, m)$$

and

$$\|I-W\| \leqq K_3\left(\sum_{i=1}^{m} \|W^*A_iW-A_i\|\right).$$

Now put $T = WH$. Clearly, $S^*A_iS = T^*A_iT$ $(i = 1, \cdots, m)$.

Furthermore,

$$\|I-WH\| \leqq \|I-W\| + \|I-H\|, \tag{3.1}$$

$$\|I-H\| \leqq K_4\|H^*H-\mathrm{I}\| = K_4\|T^*T-A_1\| \leqq K_4\left(\sum_{i=1}^{m} \|T^*A_iT-A_i\|\right), \tag{3.2}$$

and

$$\|I-W\| \leqq K_3\left(\sum_{i=1}^{m} \|W^*A_iW-A_i\|\right)$$

$$\leqq K_3K_5\left(\sum_{i=1}^{m} \|HW^*A_iWH-A_i\|\right) \tag{3.3}$$

$$= K_3K_5\left(\sum_{i=1}^{m} \|T^*A_iT-A_i\|\right).$$

Combining the inequalities (3.1)–(3.3), we finish the proof of Lemma 3.1. □

Proof of Theorem 1.1. Using Lemma 3.1 choose $\varepsilon > 0$ and $K_1 > 0$ such that the inequality

$$\|I-S\| \leqq K_1 \sum_{i=1}^{m} \|B_i-A_i\| \tag{3.4}$$

is satisfied for some $S \in \mathrm{GL}(n, \mathbf{C})$ with $B_i = S^*A_iS$, $i = 1, \cdots, m$ whenever $(B_1, \cdots, B_m) \in C(A)$ and

$$\sum_{i=1}^{m} \|B_i-A_i\| < \varepsilon.$$

Since one of the A_i is assumed to be positive definite, without loss of generality we may assume that $A_1 = I$ (see the proof of Lemma 2.3 in [PR]).

Next, we note the following fact (which can be easily verified arguing by contradiction, for instance). For every $Q_1 > 0$ there exists $Q_2 > 0$ such that whenever $(B_1, \cdots, B_m) \in C(A)$, $B_i = S^*A_iS$, for $i = 1, \cdots, m$ and some $S \in \mathrm{GL}(n, \mathbf{C})$, and

$$\sum_{i=1}^{m} \|B_i - A_i\| \geqq Q_2, \tag{3.5}$$

then $\|S\| \geqq Q_1$.

Now consider the quotient

$$\frac{\|I - S\|}{\|S^*S - I\|}. \tag{3.6}$$

There is a $q > 0$ such that the quotient (3.6) is bounded when $\|S\| \geqq q$. Indeed, write the polar decomposition $S = UP$, where U is unitary and P is positive semidefinite; then

$$\frac{\|I - S\|}{\|S^*S - I\|} = \frac{\|I - UP\|}{\|P^2 - I\|} \tag{3.7}$$

and it clearly follows that (3.7) does not exceed

$$\frac{(1 + \|P\|)}{(\|P^2\| - \sqrt{n})}, \tag{3.8}$$

and using the fact that $s\|P\|^2 \leqq \|P^2\|$ (the constant s depends on n only) and $\|S\| = \|P\|$, we obtain our assertion from (3.8).

Combining the boundedness of (3.6) when $\|S\| > q$ with inequality (3.5), we note that there exist constants M, $K_2 > 0$ such that whenever

$$\sum_{i=1}^{m} \|B_i - A_i\| > M, \tag{3.9}$$

and $B_1, \cdots, B_m$ are simultaneously congruent to $A_1, \cdots, A_m$, the quotient (3.6) is bounded above by K_2.

We now consider the case where

$$\varepsilon \leqq \sum_{i=1}^{m} \|B_i - A_i\| \leqq M. \tag{3.10}$$

We show that there exists $K_3 > 0$ such that for every $(B_1, \cdots, B_m) \in C(A)$ for which (3.10) holds and every $S \in \mathrm{GL}(n, \mathbf{C})$ with

$$B_i = S^*A_iS, \qquad i = 1, \cdots, m, \tag{3.11}$$

the inequality $\|I - S\| \leqq K_3$ holds. If (3.10) and (3.11) hold, then, in particular,

$$\|S^*S - I\| \leqq M. \tag{3.12}$$

It is easy to see that the set of all $n \times n$ matrices S that satisfy (3.12) is bounded (indeed, for such S we have $\|S^*S\|_H \leqq 1 + M$ and hence $\|S\|_H \leqq \sqrt{1 + M}$). Now the

existence of K_3 with the required property is obvious. So for every $(B_1, \cdots, B_m) \in C(A)$ and for every S for which (3.10) and (3.11) hold we have

$$\|I - S\| \leqq K_3 \varepsilon^{-1} \sum_{i=1}^{m} \|B_i - A_i\|.$$

Theorem 1.1 now follows with $K = \max(K_1, K_2, K_3 \varepsilon^{-1})$. □

REFERENCES

[A] J. P. AUBIN, *Lipschitz behavior of solutions to convex optimization problems*, Math. Oper. Res., 9 (1984), pp. 87–111.

[AFHPS] E. ANDRUCHOW, L. A. FIALKOW, D. A. HERRERO, M. PECUCH HERRERO, AND D. STOJANOFF, *Joint similarity orbits with local cross sections*, Integral Equations Operator Theory, 13 (1990), pp. 1–48.

[AFHV] C. APOSTOL, L. A. FIALKOW, D. A. HERRERO, AND D. VOICULESCU, *Approximation of Hilbert Space Operators*, Vol. II, Research Notes in Mathematics, Vol. 102, Pitman, Boston, 1984.

[DF] D. DECKARD AND L. A. FIALKOW, *Characterization of Hilbert space operators with unitary cross sections*, J. Operator Theory, 2 (1979), pp. 153–158.

[GR] I. GOHBERG AND L. RODMAN, *On distance between lattices of invariant subspaces of matrices*, Linear Algebra Appl., 76 (1986), pp. 85–120.

[MS] O. L. MANGASARIAN AND T. H. SHIAU, *Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems*, SIAM J. Comput., 25 (1987), pp. 583–595.

[P] M. PECUCH HERRERO, *Global cross sections of unitary and similarity orbits of Hilbert space operators*, J. Operator Theory, 12 (1984), pp. 265–283.

[PR] S. PIERCE AND L. RODMAN, *Congruences and norms of Hermitian matrices*, Canad. J. Math., 39 (1987), pp. 1446–1458.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 1, pp. 84–114, January 1991

008

ON THE QUADRATIC CONVERGENCE OF THE FALK–LANGEMEYER METHOD*

IVAN SLAPNIČAR† AND VJERAN HARI‡

Abstract. The Falk–Langemeyer method for solving a real definite generalized eigenvalue problem, $Ax = \lambda Bx$, $x \neq 0$, is proved to be quadratically convergent under arbitrary cyclic pivot strategy if the eigenvalues of the problem are simple. The term "quadratically convergent" means roughly that the sum of squares of the off-diagonal elements of matrices from the sequence of matrix pairs generated by the method tends to zero quadratically per cycle.

Key words. generalized eigenvalue problem, Jacobi method, quadratic convergence, asymptotic convergence

AMS(MOS) subject classifications. 65F15, 65F30

1. Introduction. In this paper we study the asymptotic convergence of the method established in 1960 by Falk and Langemeyer in [2]. Their method solves generalized eigenvalue problem

$$Ax = \lambda Bx, \qquad x \neq 0, \tag{1}$$

where A and B are real symmetric matrices of order n such that the pair (A, B) is *definite*. By definition the pair (A, B) is *definite* if the matrices A and B are Hermitian or real symmetric and there exist real constants a and b such that the matrix $aA + bB$ is positive definite.

The Falk–Langemeyer method is the most commonly used Jacobi-type method for solving problem (1). Its advantage over other methods of solving problem (1) is that it applies to problem (1) for the widest class of starting pairs. Although it is not, in general, the fastest method for solving the given problem, in some cases it is the most appropriate. The QR method [11] is usually several times faster, at least on conventional computers, but it solves problem (1) only if matrix B is positive definite (or positive definitizing shift for the pair is known in advance) and if matrix B is well conditioned for Cholesky decomposition. The Falk–Langemeyer method is superior to the QR method in terms of numerical stability if matrix B is badly conditioned for Cholesky decomposition. It is also superior to the QR method if approximate eigenvectors are known, i.e., if the matrices A and B are almost diagonal. This happens in the course of modeling the parameters of a system where a sequence of matrix pairs differing only slightly from each other must be reduced. This also happens in various subspace iteration techniques (see [11]). Another reason Jacobi-type methods have attracted attention recently is that they are adaptable for parallel processing (see [12], [10]).

* Received by the editors November 23, 1988; accepted for publication (in revised form) October 27, 1989.

† Department of Mathematics, Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, R. Boškovića b.b., YU-58000 Split, Yugoslavia. Present address, Fernuniversität Gesamthochschule, Postfach 940, 5800 Hagen, Federal Republic of Germany (MA703@DHAFEU11.BITNET). Most of the results presented in this paper are part of this author's Master's thesis [University of Zagreb, Zagreb, Yugoslavia, 1988; in Croatian], done under the supervision of Professor V. Hari.

‡ Department of Mathematics, University of Zagreb, YU-41000 Zagreb, Yugoslavia.

The Jacobi-type method for solving problem (1) recently proposed by Veselić in [15] is somewhere in between previously mentioned methods in both speed and requirements. Although Veselić's method works for definite matrix pairs, a linear combination $\rho A - \lambda B$ which is reasonably well conditioned for J-symmetric Cholesky decomposition must be known in advance. This method is one of the implicit methods, i.e., it works only on the eigenvectors matrix, and is therefore approximately two times faster than the Falk–Langemeyer method.

The Jacobi-type method considered by Zimmermann in [19] is closely related to the Falk–Langemeyer method (this is briefly described in § 3) but requires positive-definite matrix B. In [19] the convergence of this method is proved under the assumption that the starting matrices are almost diagonal. The same conclusion holds for the Falk–Langemeyer method as we show in this paper.

In [4] Hari studied the asymptotic convergence of complex extension of Zimmermann's method (also for positive-definite B). He showed that his method converges quadratically under the cyclic pivot strategies if the eigenvalues of the problem are simple, while in the case of multiple eigenvalues the method can be modified so that the quadratic convergence persists. We are interested only in cyclic pivot strategies since some of them are amenable for parallel processing.

These results, the informal analysis of the convergence properties of the Falk–Langemeyer method performed by Hari in [7], and the numerical investigation suggested that the Falk–Langemeyer method behaves in the similar fashion. In this paper we prove that the Falk–Langemeyer method is quadratically convergent if the eigenvalues of the problem are simple and the pivot strategy is cyclic. The technique of the proof, originally established by the late Wilkinson in [16] (cf. [6]), is similar to that used in [4].

Two main problems that had to be solved are that neither of the matrices A and B had to be positive definite and that the transformation matrices are not orthogonal and therefore difficult to estimate. Both problems were solved using the results about almost diagonal definite matrix pairs from [7].

The paper is organized as follows. In § 2 we state the known results about almost diagonal definite matrix pairs from [7] and [14] to the extent necessary for understanding the rest of the paper. In § 3 we describe the Falk–Langemeyer method, show that it always works for definite matrix pairs (without use of definitizing shifts), and give its algorithm. We also briefly describe the Zimmermann method from [19] and [4] and relate it to the Falk–Langemeyer method. Section 4 is the central section of the paper. We first state the known result about the quadratic convergence of the Zimmermann method from [4] and show to what extent this result can be applied to the Falk–Langemeyer method. We introduce measure $\tilde{\varepsilon}_k$ which we use for defining and proving quadratic convergence. Then we prove the quadratic convergence of the Falk–Langemeyer method under the assumptions that the eigenvalues of the problem are simple, the pivot strategy is arbitrary cyclic, and the matrices A and B are almost diagonal. At the end we show that the quadratic convergence implies the convergence of the Falk–Langemeyer method. In § 5 we give the quadratic convergence results for parallel and serial strategies, briefly explain the possible modification of the Falk–Langemeyer method in the case of multiple eigenvalues, and briefly discuss numerical experiments.

2. Almost diagonal definite matrix pairs. Here we consider the structure of almost diagonal definite matrix pairs. We first state some properties of definite matrix pairs. Then we introduce chordal metric for measuring distance between eigenvalues of definite matrix pairs. We define the measures for the almost diagonality of the square matrix and of the pair of square matrices. Finally, we state an important theorem from [7]. The

theorem and its corollary reveal the structure of almost diagonal definite matrix pairs. All results are given for the general case of Hermitian matrices even though in the rest of the paper we shall consider only the case of real symmetric matrices.

Definite matrix pair (A, B) has some important properties:

(a) There exists a nonsingular matrix F such that

$$\begin{aligned} F^*AF &= \operatorname{diag}(a_1, \cdots, a_n) = D_A, \\ F^*BF &= \operatorname{diag}(b_1, \cdots, b_n) = D_B. \end{aligned} \tag{2}$$

The ratios a_i/b_i, $i = 1, \cdots, n$, of real numbers a_i, b_i are the eigenvalues of the pair (A, B) and are unique to the ordering. If $[f_1, \cdots, f_n]$ denotes the partition by columns of F, vectors f_i, $i = 1, \cdots, n$, are the corresponding eigenvectors. Matrices D_A and D_B are not uniquely determined by the pair (A, B). In the real symmetric case F^* can be changed to F^T in the relation (2).

(b) The Crawford constant $c(A, B)$,

$$c(A,B) = \inf\{|x^*(A+iB)x|; x \in \mathbf{C}^n, \|x\| = 1\} \tag{3}$$

is positive. Therefore, A and B share no common null-subspace and $|a_i| + |b_i| > 0$, $i = 1, \cdots, n$, independently of the choice of F. Note that the choice $x = e_i$ (the ith coordinate vector) in the relation (3) for $i = 1, \cdots, n$ implies

$$d_i = \sqrt[4]{(a_{ii})^2 + (b_{ii})^2} > 0, \qquad i = 1, \cdots, n, \tag{4}$$

where $A = (a_{ij})$ and $B = (b_{ij})$. Hence the matrix

$$D = \operatorname{diag}\left(\frac{1}{d_1}, \cdots, \frac{1}{d_n}\right), \tag{5}$$

is positive definite. In the real symmetric case for $n \neq 2$ only real vectors x can be taken in the relation (3).

(c) There exists a real number φ, such that the matrix B_φ from the pair (A_φ, B_φ),

$$\begin{aligned} A_\varphi &= A\cos\varphi - B\sin\varphi, \\ B_\varphi &= A\sin\varphi + B\cos\varphi, \end{aligned} \tag{6}$$

is positive definite. The matrices A and B can be simultaneously diagonalized if and only if the same holds for the matrices A_φ and B_φ.

The proofs of the above properties are simple (see [14]). If some f_i is a vector from the null-subspace of B, the eigenvalue λ_i is infinite. Such eigenvalues are not badly posed because they are zero eigenvalues of the pair (B, A) counting their multiplicities. Hence, it is better to define eigenvalues as pairs of numbers $\lambda_i = [a_i, b_i]$, $i = 1, \cdots, n$. It is also necessary to choose a finite metric for measuring the distance between eigenvalues. Such is the *chordal metric*.

Let $\mathbf{R}^2 = \mathbf{R} \times \mathbf{R}$ and $\mathbf{R}_0^2 = \mathbf{R}^2 \setminus \{[0, 0]\}$, where $\mathbf{R}$ is the set of real numbers. We say that the pairs $[a, b], [c, d] \in \mathbf{R}_0^2$ are *equivalent* if $ad - bc = 0$ and write $[a, b] \rho [c, d]$. It is easily seen that ρ is an equivalence relation on $\mathbf{R}_0^2$. Let $\mathbf{R}_0^2 / \rho$ be the set of equivalence classes. Let $\lambda, \mu \in \mathbf{R}_0^2 / \rho$ and let $[a, b]$, $[c, d]$ be their representatives, respectively. *Chordal distance* between $[a, b]$ and $[c, d]$ is defined with the formula

$$\chi([a,b],[c,d]) = \frac{|ad-bc|}{\sqrt{a^2+b^2}\sqrt{c^2+d^2}}.$$

It is easily seen that χ is constant when $[a, b]$ and $[c, d]$ vary over λ and μ, respectively. This defines metric $\tilde{\chi} : \mathbf{R}_0^2|_\rho \times \mathbf{R}_0^2|_\rho \to \mathbf{R}$ by $\tilde{\chi}(\lambda, \mu) = \chi([a, b], [c, d])$, where $[a, b]$ and $[c, d]$ are any representatives of λ and μ, respectively. However, for the sake of simplicity we shall use χ for both the functions χ and $\tilde{\chi}$. We see that $\chi(\lambda, \mu) \leqq 1$ for all $\lambda, \mu \in \mathbf{R}_0^2|_\rho$. The proof of these and some other properties of the chordal metric can be found in [11], [14], and [13].

From now on, let n denote the order of the matrices A and B and let p denote the number of distinct eigenvalues of the pair (A, B). We assume that

$$n \geqq 3, \qquad p \geqq 2,$$

and that the pair (A, B) is definite. Note that if $p = 1$ then $A = \lambda B$, so $\lambda_1 = \lambda_2 = \cdots = \lambda_n = \lambda$ and *all* vectors are eigenvectors of the pair (A, B).

The *off-norm of the square matrix A* is the quantity

$$S(A) = \sqrt{\sum_{\substack{i,j=1 \\ i \neq j}}^{n} |a_{ij}|^2} = \|A - \operatorname{diag}(A)\|,$$

where $\|\cdot\|$ denotes the Euclidean matrix norm.

The *off-norm of the pair* (A, B) is the quantity

$$\varepsilon(A, B) = \sqrt{S^2(A) + S^2(B)}. \tag{7}$$

Where no misunderstanding can arise, ε shall be used instead of $\varepsilon(A, B)$. Let

$$\lambda_1 = \cdots = \lambda_{t_1}, \qquad \lambda_{t_1+1} = \cdots = \lambda_{t_2}, \cdots, \qquad \lambda_{t_{p-1}+1} = \cdots = \lambda_{t_p}, \tag{8}$$

where

$$\lambda_{t_i} = [s_i, c_i], \qquad s_i^2 + c_i^2 = 1, \qquad i = 1, \cdots, p, \tag{9}$$

be all eigenvalues of the pair (A, B). Thus, we assume that the pair (A, B) has p distinct eigenvalues $\lambda_{t_1}, \cdots, \lambda_{t_p}$ with the appropriate multiplicities

$$n_i = t_i - t_{i-1}, \qquad i = 1, \cdots, p, \qquad t_0 = 0, \tag{10}$$

and the representatives which behave as sine and cosine are chosen. Since $p > 2$ we can define quantities

$$\delta_i = \tfrac{1}{3} \min_{\substack{1 \leqq j \leqq p \\ j \neq i}} \chi(\lambda_{t_i}, \lambda_{t_j}), \qquad \delta = \min_{1 \leqq i \leqq p} \delta_i. \tag{11}$$

Note that $\delta > 0$.

In the analysis we shall need matrices $\tilde{A}$ and $\tilde{B}$ defined as

$$\tilde{A} = DAD, \qquad \tilde{B} = DBD, \tag{12}$$

where matrix D is defined with relations (5) and (4). Since D is positive definite, the matrices $\tilde{A}$ and $\tilde{B}$ are congruent to the matrices A and B, respectively. Let us partition the matrices $\tilde{A}$ and $\tilde{B}$,

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{11} \cdots \tilde{A}_{1p} \\ \vdots \quad \vdots \\ \tilde{A}_{p1} \cdots \tilde{A}_{pp} \end{bmatrix}, \qquad \tilde{B} = \begin{bmatrix} \tilde{B}_{11} \cdots \tilde{B}_{1p} \\ \vdots \quad \vdots \\ \tilde{B}_{p1} \cdots \tilde{B}_{pp} \end{bmatrix}, \tag{13}$$

where $\tilde{A}_{ii}$ and $\tilde{B}_{ii}$ are diagonal blocks of order n_i, $i = 1, \cdots, p$, and n_i's are defined with relation (10). The relation (13) shall be written as $\tilde{A} = (\tilde{A}_{ij})$ and $\tilde{B} = (\tilde{B}_{ij})$.

Let the matrices A and B be partitioned according to the relation (13). *Departure from the block-diagonal form of the pair* (A, B) is the quantity

$$\tau(A,B)=\sqrt{\tau^2(A)+\tau^2(B)},$$

where

$$\tau^2(A)=\sum_{\substack{i=1\\ j\neq i}}^{p}\sum_{j=1}^{p}\|A_{ij}\|^2, \qquad \tau^2(B)=\sum_{\substack{i=1\\ j\neq i}}^{p}\sum_{j=1}^{p}\|B_{ij}\|^2.$$

THEOREM 1. *Let* (A, B) *be a definite pair and let the matrices* $\tilde{A}$ *and* $\tilde{B}$ *be defined by the relations* (12), (5), *and* (4). *If*

$$\varepsilon(\tilde{A},\tilde{B})<\delta, \tag{14}$$

then there exists a permutation matrix P *such that for matrices* $\tilde{A}' = P^T\tilde{A}P$ *and* $\tilde{B}' = P^T\tilde{B}P$, *partitioned according to the relation* (13), *holds*:

$$\|c_i\tilde{A}'_{ii}-s_i\tilde{B}'_{ii}\|\leqq\frac{1}{\delta_i}\sum_{\substack{j=1\\ j\neq i}}^{p}\|c_i\tilde{A}'_{ij}-s_i\tilde{B}'_{ij}\|^2, \qquad i=1,\cdots,p. \tag{15}$$

On both sides of the inequalities (15) *the Euclidean matrix norm can be substituted with the spectral norm*.

Proof. The proof of this theorem is found in [7]. □

COROLLARY 2. *Let the relation* (14) *hold for the definite pair* (A, B). *Then there exists a permutation matrix* P *such that for the matrices* $\tilde{A}' = P^T\tilde{A}P = (\tilde{a}'_{ij})$, $\tilde{B}' = P^T\tilde{B}P = (\tilde{b}'_{ij})$, $A' = P^TAP = (a'_{ij})$, *and* $B' = P^TBP = (b'_{ij})$, *partitioned according to the relation* (13), *holds*:

$$\sum_{i=1}^{p}\|c_i\tilde{A}'_{ii}-s_i\tilde{B}'_{ii}\|^2\leqq\frac{\tau^4(\tilde{A}',\tilde{B}')}{2\delta^2}, \tag{16}$$

$$\begin{aligned}\sum_{i=1}^{p}\sum_{j=t_{i-1}+1}^{t_i}\chi^2([s_i,c_i],[a'_{jj},b'_{jj}])&=\sum_{i=1}^{p}\sum_{j=t_{i-1}+1}^{t_i}|c_i\tilde{a}'_{jj}-s_i\tilde{b}'_{jj}|^2\\ &\leqq\frac{\tau^4(\tilde{A}',\tilde{B}')}{2\delta^2},\end{aligned} \tag{17}$$

$$\begin{aligned}\chi([s_i,c_i],[a'_{jj},b'_{jj}])&=|c_i\tilde{a}'_{jj}-s_i\tilde{b}'_{jj}|\leqq\frac{\tau^2(\tilde{A}',\tilde{B}')}{2\delta},\\ &j=t_{i-1}+1,\cdots,t_i, \quad i=1,\cdots,p.\end{aligned} \tag{18}$$

Proof. By Theorem 1 there exists a permutation matrix P such that the relation (15) holds for the matrices $\tilde{A}'$ and $\tilde{B}'$. The Cauchy–Schwarz inequality implies

$$\begin{aligned}\|c_i\tilde{A}'_{ij}-s_i\tilde{B}'_{ij}\|^2&\leqq(|c_i|\,\|\tilde{A}'_{ij}\|+|s_i|\,\|\tilde{B}'_{ij}\|)^2\\ &\leqq\|\tilde{A}'_{ij}\|^2+\|\tilde{B}'_{ij}\|^2, \qquad i\neq j.\end{aligned} \tag{19}$$

From the relations (15) and (19), the definition of $\tau(\tilde{A}', \tilde{B}')$, and the symmetry of matrices $\tilde{A}'$ and $\tilde{B}'$ follows:

$$\begin{aligned} \| c_i\tilde{A}'_{ii} - s_i\tilde{B}'_{ii} \| &\leqq \frac{1}{\delta_i} \sum_{\substack{j=1 \\ j\neq i}}^{p} (\|\tilde{A}'_{ij}\|^2 + \|\tilde{B}'_{ij}\|^2) \\ &\leqq \frac{1}{2\delta}\tau^2(\tilde{A}', \tilde{B}'), \qquad i = 1, \cdots, p. \end{aligned} \tag{20}$$

Finally, the relations (15), (19), and (20) and the definition of $\tau(\tilde{A}', \tilde{B}')$ imply that

$$\begin{aligned} \sum_{i=1}^{p} \| c_i\tilde{A}'_{ii} - s_i\tilde{B}'_{ii} \|^2 &\leqq \frac{1}{2\delta}\tau^2(\tilde{A}', \tilde{B}') \sum_{i=1}^{p} \sum_{\substack{j=1 \\ j\neq i}}^{p} \frac{1}{\delta_i} (\|\tilde{A}'_{ij}\|^2 + \|\tilde{B}'_{ij}\|^2) \\ &\leqq \frac{1}{2\delta^2}\tau^4(\tilde{A}', \tilde{B}'), \end{aligned}$$

which completes the proof of the relation (16).

The equalities in the relations (17) and (18) follow from the definition of the chordal metric and the fact that it does not depend upon the choice of the representatives. Inequality in the relation (17) now follows from the relation (16) and inequality in the relation (18) from the relation (20). □

Theorem 1 and Corollary 2 reveal the structure of almost diagonal definite matrix pairs in both the Hermitian and the real symmetric case. The relation (18) implies that for $i = 1, \cdots, p$, pairs $[a'_{jj}, b'_{jj}], j \in \{t_{i-1} + 1, \cdots, t_i\}$ approximate the eigenvalues λ_{t_i} with an error of order of magnitude $\tau^2(\tilde{A}', \tilde{B}')$ in the chordal metric. The relation (16) implies that the blocks $\tilde{A}'_{ii}$ and $\tilde{B}'_{ii}$, $i = 1, \cdots, p$, are proportional with the proportionality constants being λ_{t_i} also with the error of order $\tau^2(\tilde{A}', \tilde{B}')$. This proportionality becomes apparent when $\tau(\tilde{A}', \tilde{B}')$ is small enough compared to δ. Note that the relations (15) and (18) do not imply that the off-diagonal elements of blocks $\tilde{A}'_{ii}$ and $\tilde{B}'_{ii}$ tend to zero together with $\tau(\tilde{A}', \tilde{B}')$. The relation (15) shows that for fixed i the proportionality of the blocks $\tilde{A}'_{ii}$ and $\tilde{B}'_{ii}$ depends on the local separation δ_i of the eigenvalue λ_{t_i} from other eigenvalues and on quantities $\| c_i\tilde{A}'_{ij} - s_i\tilde{B}'_{ij} \|^2, j = 1, \cdots, p, j \neq i$.

3. The Falk–Langemeyer method. In this section we define the Falk–Langemeyer method, show that it always works for definite matrix pairs, and give its algorithm. At the end of the section we briefly describe the method of Zimmermann from [4] and [19], because it is closely related with the Falk–Langemeyer method. This relationship is also described.

The Falk–Langemeyer method solves problem (1) by constructing a sequence of "congruent" matrix pairs

$$(A^{(1)}, B^{(1)}), (A^{(2)}, B^{(2)}), \cdots, \tag{21}$$

where

$$\begin{aligned} A^{(1)} &= A, & B^{(1)} &= B, \\ A^{(k+1)} &= F_k^T A^{(k)} F_k, & B^{(k+1)} &= F_k^T B^{(k)} F_k, \qquad k \geqq 1. \end{aligned} \tag{22}$$

Note that the transformation (22) with nonsingular matrix F_k preserves the eigenvalues of the pair $(A^{(k)}, B^{(k)})$. This is a Jacobi-type method, hence the transformation matrices

are chosen as nonsingular *elementary plane matrices*. An *elementary plane matrix* $F = (f_{ij})$ differs from the identity matrix only at the positions (l, l), (l, m), (m, l), and (m, m), where $1 \leqq l < m \leqq n$. The matrix

$$\hat{F} = \begin{bmatrix} f_{ll} & f_{lm} \\ f_{ml} & f_{mm} \end{bmatrix}$$

is called (l, m)-*restriction* of the square matrix $F = (f_{ij})$.

For each $k \geqq 1$, the (l, m)-restriction of the matrix F_k has the form

$$\hat{F}_k = \begin{bmatrix} 1 & \alpha_k \\ -\beta_k & 1 \end{bmatrix}, \tag{23}$$

where real parameters α_k and β_k are chosen to satisfy the condition

$$a_{lm}^{(k+1)} = 0, \qquad b_{lm}^{(k+1)} = 0. \tag{24}$$

Here $A^{(k)} = (a_{ij}^{(k)})$ and $B^{(k)} = (b_{ij}^{(k)})$. Indices l and m are called *pivot indices* and the pair (l, m) is called *pivot pair*. As k varies, the pivot pair also varies, hence $l = l(k)$ and $m = m(k)$. The transition from the pair $(A^{(k)}, B^{(k)})$ to the pair $(A^{(k+1)}, B^{(k+1)})$ is called the kth *step* of the method. The manner in which we choose elements which are to be annihilated in the kth step (or just the indices (l, m) of these elements) is called *pivot strategy*. The pivot strategy is *cyclic* if every sequence of $N = n(n-1)/2$ successive pairs (l, m) contains all pairs (i, j), $1 \leqq i < j \leqq n$. A sequence of N successive steps is referred to as a *cycle*. Two most common cyclic pivot strategies are the *column-cyclic strategy* and the *row-cyclic strategy*. The former is defined by the sequence of pairs

$$(1,2),(1,3),(2,3),(1,4),(2,4),(3,4),\cdots,(1,n),(2,n),\cdots,(n-1,n),$$

and the latter by the sequence of pairs

$$(1,2),(1,3),\cdots,(1,n),(2,3),(2,4),\cdots,(2,n),(3,4),\cdots,(n-1,n).$$

These two strategies are also called *serial strategies*. *Parallel cyclic strategies* are cyclic strategies which enable simultaneous execution of approximately $n/2$ steps on parallel computers. These strategies have recently attracted considerable attention (see [12], [10]). We state the quadratic convergence results for serial and parallel strategies in § 5.

Note that if the eigenvectors are needed, we must calculate the sequence of matrices $F^{(1)}, F^{(2)}, \cdots$, where

$$F^{(1)} = I, \qquad F^{(k+1)} = F^{(k)} F_k, \qquad k \geqq 1. \tag{25}$$

From the relations (22) and (25) we obtain for $k \geqq 2$

$$F^{(k)} = F_1 \cdots F_{k-1}, \qquad A^{(k)} = (F^{(k)})^T A^{(1)} F^{(k)}, \qquad B^{(k)} = (F^{(k)})^T B^{(1)} F^{(k)}.$$

We shall now derive one step of the algorithm. Note that only (l, m)-restrictions of the involved matrices are needed. Since (22) is the congruence transformation, the pairs $(A^{(k)}, B^{(k)})$ are definite for every $k \geqq 1$ and the pairs of the corresponding (l, m)-restrictions are definite as well.

Let (index k is omitted for the sake of simplicity)

$$\hat{F}^T \hat{A} \hat{F} = \hat{A}', \qquad \hat{F}^T \hat{B} \hat{F} = \hat{B}',$$

or, respectively,

$$\begin{bmatrix} 1 & -\beta \\ \alpha & 1 \end{bmatrix}\begin{bmatrix} a_{ll} & a_{lm} \\ a_{lm} & a_{mm} \end{bmatrix}\begin{bmatrix} 1 & \alpha \\ -\beta & 1 \end{bmatrix}=\begin{bmatrix} a'_{ll} & a'_{lm} \\ a'_{lm} & a'_{mm} \end{bmatrix},$$
$$\begin{bmatrix} 1 & -\beta \\ \alpha & 1 \end{bmatrix}\begin{bmatrix} b_{ll} & b_{lm} \\ b_{lm} & b_{mm} \end{bmatrix}\begin{bmatrix} 1 & \alpha \\ -\beta & 1 \end{bmatrix}=\begin{bmatrix} b'_{ll} & b'_{lm} \\ b'_{lm} & b'_{mm} \end{bmatrix}. \tag{26}$$

Condition (24) now reads $a'_{lm} = b'_{lm} = 0$. From the relation (26) the system in unknowns α and β is obtained

$$\begin{aligned} a'_{lm} &= \alpha a_{ll} + (1-\alpha\beta)a_{lm} - \beta a_{mm} = 0, \\ b'_{lm} &= \alpha b_{ll} + (1-\alpha\beta)b_{lm} - \beta b_{mm} = 0. \end{aligned} \tag{27}$$

Eliminating nonlinear terms in both equations we obtain

$$\alpha = \frac{\mathfrak{F}_m}{\nu}, \qquad \beta = \frac{\mathfrak{F}_l}{\nu}, \tag{28}$$

where ν is solution of the equation

$$\nu^2 - \mathfrak{F}_{lm}\nu - \mathfrak{F}_l\mathfrak{F}_m = 0 \tag{29}$$

and

$$\begin{aligned} \mathfrak{F}_l &= a_{ll}b_{lm} - b_{ll}a_{lm}, \\ \mathfrak{F}_m &= a_{mm}b_{lm} - b_{mm}a_{lm}, \\ \mathfrak{F}_{lm} &= a_{ll}b_{mm} - a_{mm}b_{ll}. \end{aligned} \tag{30}$$

Defining

$$\mathfrak{F} = (\mathfrak{F}_{lm})^2 + 4\mathfrak{F}_l\mathfrak{F}_m \tag{31}$$

we obtain

$$\nu_{\pm} = \tfrac{1}{2}\operatorname{sgn}(\mathfrak{F}_{lm})(|\mathfrak{F}_{lm}| \pm \sqrt{\mathfrak{F}}).$$

The algorithm is more stable if α and β are smaller in modulus, so we take

$$\nu = \nu_+ = \tfrac{1}{2}\operatorname{sgn}(\mathfrak{F}_{lm})(|\mathfrak{F}_{lm}| + \sqrt{\mathfrak{F}}). \tag{32}$$

From the above formula we see that the necessary condition for carrying out this step is $\mathfrak{F} \geqq 0$. Let us show that this condition is fulfilled in each step due to the definiteness of the pairs $((A^{(k)}, B^{(k)}), k \geqq 1)$.

PROPOSITION 3. *Let the pair (A, B) be definite. Then the following holds*:

(i) $\mathfrak{F} \geqq 0$,

(ii) *The following statements are equivalent*:

(a) $\mathfrak{F} = 0$,

(b) $\mathfrak{F}_{lm} = \mathfrak{F}_l = \mathfrak{F}_m = 0$,

(c) *There exist real constants s and t, $|s| + |t| > 0$, such that*

$$s\hat{A} + t\hat{B} = 0.$$

Proof. The proof can be found in [4] and [13], but for the completeness of exposition we present it below.

Using the relation (6) we can define the pair (A_φ, B_φ) such that the matrix B_φ is positive definite. Let us calculate the quantities $(\mathfrak{F}_{lm})_\varphi$, $(\mathfrak{F}_l)_\varphi$, $(\mathfrak{F}_m)_\varphi$, and $(\mathfrak{F})_\varphi$ from the pair (A_φ, B_φ) using the relations (30) and (31). It is easy to verify that

$$\mathfrak{F}_l = (\mathfrak{F}_l)_\varphi, \qquad \mathfrak{F}_m = (\mathfrak{F}_m)_\varphi,$$

$$\mathfrak{F}_{lm} = (\mathfrak{F}_{lm})_\varphi, \qquad \mathfrak{F} = (\mathfrak{F})_\varphi.$$

Therefore, without loss of generality, we can assume that the matrix B from the pair (A, B) is positive definite. The statement (c) is now equivalent to the statement $\hat{A} = c\hat{B}$, $c \in \mathbf{R}$.

(i) With the notation

$$x = a_{ll}\sqrt{\frac{b_{mm}}{b_{ll}}}, \qquad y = a_{mm}\sqrt{\frac{b_{ll}}{b_{mm}}}, \qquad z = \frac{b_{lm}}{\sqrt{b_{ll}b_{mm}}},$$

the following identity holds:

$$\mathfrak{F} = b_{ll}b_{mm}[(x-y)^2 + 4(xz - a_{lm})(yz - a_{lm})].$$

Since the right side of the above relation is the square polynomial in a_{lm}, we have

$$\begin{aligned} \mathfrak{F} &= b_{ll}b_{mm}P_2(a_{lm}) \geqq b_{ll}b_{mm}P_2\left(\frac{x+y}{2}z\right) \\ &= b_{ll}b_{mm}(x-y)^2(1-z^2) = \mathfrak{F}_{lm}^2\left(1 - \frac{b_{lm}^2}{b_{ll}b_{mm}}\right) \geqq 0. \end{aligned} \tag{33}$$

In the last inequality we have used the assumption that the matrix B, and therefore the matrix $\hat{B}$, is positive definite.

(ii) Let (a) hold. The relation (33) implies that $\mathfrak{F}_{lm} = 0$. Matrix B is by the assumption positive definite. Therefore $b_{ll} > 0$, $b_{mm} > 0$, and the equality $\mathfrak{F}_{lm} = 0$ can be written as

$$\frac{b_{mm}}{b_{ll}}a_{ll} = a_{mm}.$$

Using this relation we can write

$$\mathfrak{F}_m = a_{mm}b_{lm} - b_{mm}a_{lm} = \frac{b_{mm}}{b_{ll}}(a_{ll}b_{lm} - b_{ll}a_{lm}) = \frac{b_{mm}}{b_{ll}}\mathfrak{F}_l,$$

or $b_{ll}\mathfrak{F}_m = b_{mm}\mathfrak{F}_l$. From the definition of $\mathfrak{F}$, since $\mathfrak{F}_{lm} = 0$, we conclude that $\mathfrak{F}_l = \mathfrak{F}_m = 0$. This gives (b).

Let (b) hold. Then

$$a_{mm} = b_{mm}\frac{a_{ll}}{b_{ll}}, \qquad a_{lm} = b_{lm}\frac{a_{ll}}{b_{ll}}.$$

Therefore, $\hat{A} = c\hat{B}$, where

$$c = \frac{a_{ll}}{b_{ll}} = \frac{a_{mm}}{b_{mm}},$$

and (c) holds.

Let (c) hold. Then obviously (b) holds, and if (b) holds, then (a) holds. □

Now we see that the Falk–Langemeyer method can be applied to all definite matrix pairs. Note that definitizing shifts are not used and need not be known.

We have two special cases in the algorithm. If $\mathfrak{F} = 0$, then the matrices $\hat{A}$ and $\hat{B}$ are proportional as shown in Proposition 3. Therefore, the two equations in the system (27) are linearly dependent and the system has a parametric solution in one of the following forms:

$$(\alpha,\beta)=\left(\frac{cb_{mm}-b_{lm}}{b_{ll}-cb_{lm}},c\right), \qquad (\alpha,\beta)=\left(\frac{ca_{mm}-a_{lm}}{a_{ll}-ca_{lm}},c\right),$$

$$(\alpha,\beta)=\left(c,\frac{cb_{ll}+b_{lm}}{b_{mm}+cb_{lm}}\right), \qquad (\alpha,\beta)=\left(c,\frac{ca_{ll}+a_{lm}}{a_{mm}+ca_{lm}}\right),$$

where c is real. For every c at least one of the quotients is well defined due to the definiteness of the pair $(\hat{A}, \hat{B})$. It is best to set $c = 0$ to ensure that α_k and β_k tend to zero together with $\varepsilon(A^{(k)}, B^{(k)})$ as $k \to \infty$ (see step (5)(a) in Algorithm 4). Setting $c = 0$ also reduces the operation count. This choice yields four possibilities for (α, β):

$$\left(-\frac{b_{lm}}{b_{ll}},0\right), \qquad \left(-\frac{a_{lm}}{a_{ll}},0\right), \tag{34}$$

$$\left(0,\frac{b_{lm}}{b_{mm}}\right), \qquad \left(0,\frac{a_{lm}}{a_{mm}}\right). \tag{35}$$

Due to the definiteness of the pair (A, B), we have

$$|a_{ii}| + |b_{ii}| > 0, \qquad i = 1, \cdots, n, \tag{36}$$

so at least one quotient is defined in each of the relations (34) and (35). In order to obtain better condition of the transformation matrix, we choose the relation in which the defined quotient has smaller absolute value. If both quotients in the chosen relation are defined, then they are equal, and for numerical reasons it is better to choose one in which the sum of squares of the numerator and the denominator is greater.

The second special case is when $\mathfrak{F} > 0$ and $\mathfrak{F}_{lm} = 0$. This means that diagonals of the matrices $\hat{A}$ and $\hat{B}$ are proportional, while the matrices themselves are not. Then $\operatorname{sgn}(\mathfrak{F}_{lm})$ is not defined. Since $\mathfrak{F}_l\mathfrak{F}_m > 0$, we have $\operatorname{sgn}(\mathfrak{F}_l) = \operatorname{sgn}(\mathfrak{F}_m)$. Substituting $\operatorname{sgn}(\mathfrak{F}_{lm})$ with $\operatorname{sgn}(\mathfrak{F}_l)$ in (32) gives

$$\nu = \operatorname{sgn}(\mathfrak{F}_l)\sqrt{\mathfrak{F}_l\mathfrak{F}_m}.$$

Inserting this in (28) gives, after simple calculation,

$$\alpha = \sqrt{\frac{b_{mm}}{b_{ll}}} = \sqrt{\frac{a_{mm}}{a_{ll}}}, \qquad \beta = \frac{1}{\alpha}. \tag{37}$$

The relation (36) implies that at least one of the quotients b_{mm}/b_{ll} and a_{mm}/a_{ll} is defined and different from zero. If both quotients are defined then they are equal and it is better to choose one in which the sum of squares of the numerator and the denominator is greater.

We can now define an algorithm of the method.

ALGORITHM 4. Definite matrix pair (A, B) is given.

(1) Set $k = 1$, $A^{(1)} = A$, $B^{(1)} = B$, $F^{(1)} = I$ and choose the pivot strategy.

(2) Choose the pivot pair $(l, m) = (l(k), m(k))$ according to the strategy.

(3) If $a_{lm}^{(k)} = b_{lm}^{(k)} = 0$, then set $k = k + 1$, $A^{(k+1)} = A^{(k)}$, $B^{(k+1)} = B^{(k)}$, $F^{(k+1)} = F^{(k)}$ and go to step (2). Otherwise go to step (4).

(4) Calculate the quantities $\mathfrak{F}_l^{(k)}$, $\mathfrak{F}_m^{(k)}$, $\mathfrak{F}_{lm}^{(k)}$, and $\mathfrak{F}^{(k)}$ from formulas

$$\mathfrak{F}_l^{(k)} = a_{ll}^{(k)} b_{lm}^{(k)} - b_{ll}^{(k)} a_{lm}^{(k)}, \qquad \mathfrak{F}_m^{(k)} = a_{mm}^{(k)} b_{lm}^{(k)} - b_{mm}^{(k)} a_{lm}^{(k)},$$

$$\mathfrak{F}_{lm}^{(k)} = a_{ll}^{(k)} b_{mm}^{(k)} - a_{mm}^{(k)} b_{ll}^{(k)}, \qquad \mathfrak{F}^{(k)} = (\mathfrak{F}_{lm}^{(k)})^2 + 4\mathfrak{F}_l^{(k)} \mathfrak{F}_m^{(k)}.$$

(5) (a) If $\mathfrak{F}^{(k)} = 0$ perform the following steps: If $|b_{ll}^{(k)}| \geqq |a_{ll}^{(k)}|$, then set $\alpha_k = -b_{lm}^{(k)} / b_{ll}^{(k)}$;

otherwise set $\alpha_k = -a_{lm}^{(k)} / a_{ll}^{(k)}$.

If $|b_{mm}^{(k)}| \geqq |a_{mm}^{(k)}|$, then set $\beta_k = b_{lm}^{(k)} / b_{mm}^{(k)}$;

otherwise set $\beta_k = a_{lm}^{(k)} / a_{mm}^{(k)}$.

Finally, if $|\alpha_k| \geqq |\beta_k|$, then set $\alpha_k = 0$; otherwise set $\beta_k = 0$.

(b) If $\mathfrak{F}^{(k)} > 0$ perform the following steps:

(i) If $\mathfrak{F}_{lm}^{(k)} \neq 0$, then calculate

$$\nu_k = \tfrac{1}{2} \operatorname{sgn}(\mathfrak{F}_{lm}^{(k)})(|\mathfrak{F}_{lm}^{(k)}| + \sqrt{\mathfrak{F}^{(k)}}),$$

$$\alpha_k = \frac{\mathfrak{F}_m^{(k)}}{\nu_k}, \qquad \beta_k = \frac{\mathfrak{F}_l^{(k)}}{\nu_k}.$$

(ii) If $\mathfrak{F}_{lm}^{(k)} = 0$, then, according to the relation (37), calculate

$$\alpha_k = \sqrt{\frac{b_{mm}^{(k)}}{b_{ll}^{(k)}}} = \sqrt{\frac{a_{mm}^{(k)}}{a_{ll}^{(k)}}}, \qquad \beta_k = \frac{1}{\alpha_k}.$$

If both quotients for α_k are defined, then choose one in which the sum of squares of the numerator and the denominator is greater.

(6) Perform the transformation

$$A^{(k+1)} = F_k^T A^{(k)} F_k, \qquad B^{(k+1)} = F_k^T B^{(k)} F_k, \tag{38}$$

$$F^{(k+1)} = F^{(k)} F_k. \tag{39}$$

(7) Set $k = k + 1$ and move to step (2). □

Since matrices $A^{(k)}$, $B^{(k)}$, $A^{(k+1)}$, and $B^{(k+1)}$ are symmetric, it is enough to store and to transform only upper triangles. In the transformation (38) only lth and mth row and column of the matrices $A^{(k)}$ and $B^{(k)}$ are changed and in the transformation (39) only lth and mth columns of the matrix $F^{(k)}$ are changed. Note that the eigenvalues can be found without calculating the matrices $F^{(k)}$, $k \geqq 1$, and therefore the transformation (39) can be omitted. This reduces the operational count about 50 percent.

Stopping criteria of the infinite iterative procedure defined with this algorithm are described in § 5.

From now on, the term "Falk–Langemeyer method" denotes the method described by Algorithm 4.

The Zimmermann method. We shall now relate the Falk–Langemeyer method with another method for solving the generalized eigenvalue problem. This method is due to

Zimmermann, who roughly described it in her thesis [19]. Later on, in his thesis [4], Hari derived its algorithm and proved its quadratic convergence.

The Zimmermann method is defined for symmetric matrix pairs (A, B) where matrix B is positive definite. We shall denote this fact as $B > 0$. At the beginning of the iterative procedure the initial pair (A, B) is normalized such that

$$A^{(1)} = DAD, \qquad B^{(1)} = DBD,$$

where

$$D = \operatorname{diag}\left(\frac{1}{\sqrt{b_{11}}}, \cdots, \frac{1}{\sqrt{b_{nn}}}\right).$$

Therefore, $b_{ii}^{(k)} = 1$, $i = 1, \cdots, n$. The Zimmermann method constructs a sequence of pairs $((A^{(k)}, B^{(k)}), k \geqq 1)$ by the rule

$$A^{(k+1)} = Z_k^T A^{(k)} Z_k, \qquad B^{(k+1)} = Z_k^T B^{(k)} Z_k, \qquad k \geqq 1.$$

The nonsingular matrices Z_k are chosen to preserve the units on the diagonal of $B^{(k+1)}$ (automatic normalization at each step) and to annihilate the pivot elements. In [4] it is shown that for $k \geqq 1$ holds

$$\hat{Z}_k = \frac{1}{\sqrt{1-(b_{lm}^{(k)})^2}} \begin{bmatrix} \cos\varphi_k & \sin\varphi_k \\ -\sin\psi_k & \cos\psi_k \end{bmatrix},$$

where

$$\cos\varphi_k = \cos\theta_k + \xi_k(\sin\theta_k - \eta_k\cos\theta_k),$$

$$\sin\varphi_k = \sin\theta_k - \xi_k(\cos\theta_k + \eta_k\sin\theta_k),$$

$$\cos\psi_k = \cos\theta_k - \xi_k(\sin\theta_k + \eta_k\cos\theta_k),$$

$$\sin\psi_k = \sin\theta_k + \xi_k(\cos\theta_k - \eta_k\sin\theta_k),$$

$$\xi_k = \frac{b_{lm}^{(k)}}{\sqrt{1+b_{lm}^{(k)}} + \sqrt{1-b_{lm}^{(k)}}},$$

$$\eta_k = \frac{b_{lm}^{(k)}}{(1+\sqrt{1+b_{lm}^{(k)}})(1+\sqrt{1-b_{lm}^{(k)}})},$$

$$\tan 2\theta_k = \frac{2a_{lm}^{(k)} - (a_{ll}^{(k)} + a_{mm}^{(k)})b_{lm}^{(k)}}{(a_{mm}^{(k)} - a_{ll}^{(k)})\sqrt{1-(b_{lm}^{(k)})^2}},$$

$$-\frac{\pi}{4} \leqq \theta_k \leqq \frac{\pi}{4}.$$

If $a_{lm}^{(k)} = b_{lm}^{(k)} = 0$, we set $\theta_k = 0$. If the (l, m)-restrictions of $A^{(k)}$ and $B(k)$ are proportional and $a_{lm}^{(k)}$ and $b_{lm}^{(k)}$ are not both equal to zero, we set $\theta_k = \pi/4$.

If the matrix B is not positive definite but the pair (A, B) is, then there exists a definitizing shift μ such that the matrix $A - \mu B$ is positive definite. If this shift is known in advance, then the Zimmermann method can be applied to the pair (A, B) in the sense that each Z_k is computed from the pair $(B^{(k)}, A^{(k)} - \mu B^{(k)})$.

Although the Zimmermann method seems quite different from the Falk–Langemeyer method, the two methods are closely related. The following theorem gives precise for-

mulation of this relationship. For this occasion only we assume that in step (5)(a) of Algorithm 4 (that is, when $\mathfrak{F}^{(k)} = 0$), parameters α_k and β_k are computed according to (37). For this version of the Falk–Langemeyer method the following theorem holds.

THEOREM 5. *Let A and B be symmetric matrices of order n and let B be positive definite. Let the sequences $((A^{(k)}, B^{(k)}), k \geqq 1)$ and $((A^{(k)'}, B^{(k)'}), k \geqq 1)$ be generated from the starting pair (A, B) with the Falk–Langemeyer and the Zimmermann method, respectively. If the corresponding pivot strategies are the same, then*

$$A^{(k)'} = D^{(k)} A^{(k)} D^{(k)}, \qquad B^{(k)'} = D^{(k)} B^{(k)} D^{(k)}, \qquad k \geqq 1,$$

where

$$D^{(k)} = \operatorname{diag}\left(\frac{1}{\sqrt{b_{11}^{(k)}}}, \cdots, \frac{1}{\sqrt{b_{nn}^{(k)}}} \right), \qquad k \geqq 1.$$

Proof. The proof of this theorem is found in [4, § 2.3]. □

Let us again suppose that the matrix B is not positive definite while the pair (A, B) is, and that a positive definitizing shift μ is known in advance. Let us apply to the pair (A, B) the Zimmermann method in the sense mentioned above and the version of the Falk–Langemeyer method which we used in Theorem 5. It is easy to see that the parameters α_k and β_k from the Falk–Langemeyer method are invariant under the transformations $(A, B) \to (B, A - \mu B)$. Therefore, Theorem 5 holds in this case, as well, with

$$D^{(k)} = \operatorname{diag}\left(\frac{1}{\sqrt{a_{11}^{(k)} - \mu b_{11}^{(k)}}}, \cdots, \frac{1}{\sqrt{a_{nn}^{(k)} - \mu b_{nn}^{(k)}}} \right), \qquad k \geqq 1.$$

We can conclude that *if the starting pair is positive definite or the definitizing shift is known in advance, then the Falk–Langemeyer (Zimmermann) method is the fast scaled (normalized) version of the Zimmermann (Falk–Langemeyer) method.*

4. Quadratic convergence. In this section we prove that the Falk–Langemeyer method is quadratically convergent if the starting definite pair has simple eigenvalues and the pivot strategy is cyclic. Definitizing shifts are not used and need not be known. We first state the result about the quadratic convergence of the Zimmermann method, and show to what extent this result can be applied to the Falk–Langemeyer method if the matrix B is positive definite. Then we define the quadratic convergence for the Falk–Langemeyer method. In § 4.1 we prove preliminary results which we use in the proof of the quadratic convergence of the Falk–Langemeyer method in § 4.2.

The result about the quadratic convergence of the Zimmermann method can be summarized as follows. Let the sequence $((A^{(k)}, B^{(k)}), k \geqq 1)$ be generated by the Zimmermann method from the pair (A, B), $B > 0$, and let $\varepsilon_k = \varepsilon(A^{(k)}, B^{(k)})$, where ε is defined with the relation (7). Note that ε_k is the natural measure for convergence of the Zimmermann method since each matrix $B^{(k)}$ has units along the diagonal.

We say that the Zimmermann method is *quadratically convergent* on the pair (A, B) if $\varepsilon_k \to 0$ as $k \to \infty$ and there exist a constant $c_0 > 0$ and an integer r_0 such that for $r \geqq r_0$ holds

$$\varepsilon_{(r+1)N+1} \leqq c_0 \varepsilon_{rN+1}^2.$$

Hence of special importance are conditions under which the above relation holds for $r = 1$. We call them *asymptotic assumptions*. Let

$$\sigma = \operatorname{spr}(A, B) = \max_{1 \leqq i \leqq n} |\lambda_i|, \qquad \gamma = \tfrac{1}{3} \min_{i \neq j} |\lambda_i - \lambda_j|.$$

THEOREM 6. *Let the sequence* $((A^{(k)}, B^{(k)}), k \geqq 1)$ *be generated by the Zimmermann method from the starting pair* (A, B), $B > 0$, *and let the asymptotic assumptions*

$$S(B^{(1)}) \leqq \frac{1}{2N}, \qquad 2\sqrt{1+\sigma^2}\varepsilon_1 < \gamma, \tag{40}$$

hold. If the eigenvalues of the pair (A, B) *are simple and the pivot strategy is cyclic, then*

$$\varepsilon_{N+1} \leqq \sqrt{N(1+\sigma^2)}\frac{\varepsilon_1^2}{\gamma}. \tag{41}$$

Proof. The proof of this theorem is found in [4, § 3.3]. □

In Theorem 6 the term σ appears in the assumption (40) and in the assertion (41) because matrix B is not diagonal and matrix A is not normalized. From Theorem 5 we see that Theorem 6 holds for the Falk–Langemeyer method provided that step (5)(a) of Algorithm 4 is appropriately changed, the matrix B is positive definite, and the pairs $(A^{(k)}, B^{(k)})$ generated by the Falk–Langemeyer method are normalized so that $b_{ii}^{(k)} = 1$, $i = 1 \cdots n$, $k \geqq 1$.

In the rest of this section we prove that the Falk–Langemeyer method defined with Algorithm 4 is quadratically convergent on definite matrix pairs with simple eigenvalues if the pivot strategy is cyclic. We must first define the measure for the quadratic convergence.

Let (A, B) be a definite pair. We shall use the measure $\tilde{\varepsilon} = \tilde{\varepsilon}(A, B)$ defined by

$$\tilde{\varepsilon}(A, B) = \varepsilon(\tilde{A}, \tilde{B}),$$

where $\tilde{A}$ and $\tilde{B}$ are given by the relations (12), (5), and (4). The measure $\tilde{\varepsilon}$ enables us to use the results of Corollary 2 and it takes into account the diagonal elements of matrices A and B. Note that the measure $\varepsilon(A, B)$ is generally not the proper measure for almost diagonality of the pair (A, B) since it takes no account of the diagonals of matrices A and B.

Let the sequence of pairs

$$(A^{(1)}, B^{(1)}), (A^{(2)}, B^{(2)}), \cdots \tag{42}$$

be generated by the Falk–Langemeyer method from the starting definite pair (A, B). For $k \geqq 1$ we set

$$\tilde{\varepsilon}_k = \tilde{\varepsilon}(A^{(k)}, B^{(k)}) = \varepsilon(\tilde{A}^{(k)}, \tilde{B}^{(k)}), \tag{43}$$

$$\tilde{A}^{(k)} = D_k A^{(k)} D_k, \qquad \tilde{B}^{(k)} = D_k B^{(k)} D_k, \tag{44}$$

$$D_k = \operatorname{diag}\left(\frac{1}{d_1^{(k)}}, \cdots, \frac{1}{d_n^{(k)}}\right), \tag{45}$$

$$d_i^{(k)} = \sqrt[4]{(a_{ii}^{(k)})^2 + (b_{ii}^{(k)})^2}, \qquad i = 1, \cdots, n. \tag{46}$$

From the relations (44), (45), and (46) we see that the pairs $(\tilde{A}^{(k)}, \tilde{B}^{(k)})$ are normalized in the sense that

$$(\tilde{a}_{ii}^{(k)})^2 + (\tilde{b}_{ii}^{(k)})^2 = 1, \qquad i = 1, \cdots, n. \tag{47}$$

DEFINITION 7. The Falk–Langemeyer method is *quadratically convergent* on the pair (A, B) if $\tilde{\varepsilon}_k \to 0$ as $k \to 0$ and there exist a constant $c_0 > 0$ and an integer r_0 such that for $r \geqq r_0$ holds

$$\tilde{\varepsilon}_{(r+1)N+1} \leqq c_0 \tilde{\varepsilon}_{rN+1}^2. \tag{48}$$

From Definition 7 we see that ultimately $\tilde{\varepsilon}_k$ decreases quadratically per cycle. At the end of § 4.2 we shall show that the quadratic convergence implies the convergence of the sequence (42) towards the pair of diagonal matrices (D_A, D_B), where

$$D_A = \operatorname{diag}(a_1, \cdots, a_n), \qquad D_B = \operatorname{diag}(b_1, \cdots, b_n). \tag{49}$$

Here $\lambda_i = [a_i, b_i]$, $i = 1, \cdots, n$, are the eigenvalues of the pair (A, B). Finally, we shall show that ultimately the quadratic reduction of $\tilde{\varepsilon}_{rN+1}$ implies the quadratic reduction of ε_{rN+1} and vice versa.[1]

In order to be able to observe the measure $\tilde{\varepsilon}$ we must solve one more problem. The transformation matrices F_k are calculated from unnormalized pairs $(A^{(k)}, B^{(k)})$ and are therefore difficult to estimate. To solve this problem we shall observe the sequence obtained from the pair (A, B) with the following process:

normalization, step of the method, normalization, step of the method, $\cdots$

This sequence reads

$$(\tilde{\bar{A}}^{(1)}, \tilde{\bar{B}}^{(1)}), (\bar{A}^{(2)}, \bar{B}^{(2)}), (\tilde{\bar{A}}^{(2)}, \tilde{\bar{B}}^{(2)}), (\bar{A}^{(3)}, \bar{B}^{(3)}), (\tilde{\bar{A}}^{(3)}, \tilde{\bar{B}}^{(3)}), \cdots, \tag{50}$$

where

$$(\tilde{\bar{A}}^{(1)}, \tilde{\bar{B}}^{(1)}) = (\tilde{A}^{(1)}, \tilde{B}^{(1)}), \tag{51}$$

and for $k \geqq 1$ holds

$$\bar{A}^{(k+1)} = \tilde{\bar{F}}_k^T \tilde{\bar{A}}^{(k)} \tilde{\bar{F}}_k, \qquad \bar{B}^{(k+1)} = \tilde{\bar{F}}_k^T \tilde{\bar{B}}^{(k)} \tilde{\bar{F}}_k, \tag{52}$$

$$\tilde{\bar{A}}^{(k+1)} = \bar{D}_{k+1} \bar{A}^{(k+1)} \bar{D}_{k+1}, \qquad \tilde{\bar{B}}^{(k+1)} = \bar{D}_{k+1} \bar{B}^{(k+1)} \bar{D}_{k+1}, \tag{53}$$

$$\bar{D}_{k+1} = \operatorname{diag}\left(\frac{1}{\bar{d}_1^{(k+1)}}, \cdots, \frac{1}{\bar{d}_n^{(k+1)}} \right), \tag{54}$$

$$\bar{d}_i^{(k+1)} = \sqrt[4]{(\bar{a}_{ii}^{(k+1)})^2 + (\bar{b}_{ii}^{(k+1)})^2}, \qquad i = 1, \cdots, n. \tag{55}$$

Of course, the sequences (42) and (50) are generated using the same pivot strategy. The matrices $\tilde{\bar{F}}_k$ are calculated according to Algorithm 4, but from the pairs $(\tilde{\bar{A}}^{(k)}, \tilde{\bar{B}}^{(k)})$. Since, in the transition from $(\tilde{\bar{A}}^{(k)}, \tilde{\bar{B}}^{(k)})$ to $(\bar{A}^{(k+1)}, \bar{B}^{(k+1)})$ of all diagonal elements only those at positions (l, l) and (m, m) are being changed, we conclude that

$$\bar{d}_i^{(k+1)} = \sqrt[4]{(\bar{a}_{ii}^{(k+1)})^2 + (\bar{b}_{ii}^{(k+1)})^2} = \sqrt[4]{(\tilde{a}_{ii}^{(k)})^2 + (\tilde{b}_{ii}^{(k)})^2} = 1, \tag{56}$$
$$i = 1, \cdots, n, \qquad i \neq l, m.$$

We will now show that the operations of normalization and of carrying out one step of the algorithm commute. This is equivalent to showing that $\tilde{\bar{A}}^{(k)} = \tilde{A}^{(k)}$ and $\tilde{\bar{B}}^{(k)} = \tilde{B}^{(k)}$ for $k \geqq 1$.

Let $\tilde{F}_k$ be the transformation matrices obtained according to Algorithm 4 from the pairs $(\tilde{A}^{(k)}, \tilde{B}^{(k)})$, $k \geqq 1$. The following proposition shows that the matrices F_k and $\tilde{F}_k$ are simply related.

---

[1] Here ε_k measures off-diagonal elements of the pairs from the sequence (42) and should not be confused with the quantity used in connection with the Zimmermann method.

PROPOSITION 8. *For* $k \geqq 1$, $\tilde{F}_k = D_k^{-1} F_k D_k$ *holds.*

Proof. Because of the relations (44), (45), and (46) we have

$$\hat{\tilde{A}}^{(k)} = \begin{bmatrix} \dfrac{a_{ll}^{(k)}}{(d_l^{(k)})^2} & \dfrac{a_{lm}^{(k)}}{d_l^{(k)} d_m^{(k)}} \\ \dfrac{a_{lm}^{(k)}}{d_l^{(k)} d_m^{(k)}} & \dfrac{a_{mm}^{(k)}}{(d_m^{(k)})^2} \end{bmatrix}, \qquad \hat{\tilde{B}}^{(k)} = \begin{bmatrix} \dfrac{b_{ll}^{(k)}}{(d_l^{(k)})^2} & \dfrac{b_{lm}^{(k)}}{d_l^{(k)} d_m^{(k)}} \\ \dfrac{b_{lm}^{(k)}}{d_l^{(k)} d_m^{(k)}} & \dfrac{b_{mm}^{(k)}}{(d_m^{(k)})^2} \end{bmatrix}. \tag{57}$$

The assertion is now obtained by simply using the relation (57) in Algorithm 4 and calculating the matrix $\tilde{F}_k$. □

PROPOSITION 9. *For* $k \geqq 1$ *the following holds*:

(i) $(\tilde{\bar{A}}^{(k)}, \tilde{\bar{B}}^{(k)}) = (\tilde{A}^{(k)}, \tilde{B}^{(k)})$,

(ii) $D_k = D_1 \bar{D}_2 \bar{D}_3 \cdots \bar{D}_k$.

Proof. The proof is by induction in respect to k.

(i) For $k = 1$ the assertion holds due to the relation (51). Suppose that the assertion holds for some $k \geqq 1$. This means that

$$\tilde{\bar{A}}^{(k)} = \tilde{A}^{(k)}, \qquad \tilde{\bar{B}}^{(k)} = \tilde{B}^{(k)}, \qquad \tilde{\bar{F}}_k = \tilde{F}_k. \tag{58}$$

From the relation (52) it follows that $\bar{A}^{(k+1)} = \tilde{\bar{F}}_k^T \tilde{\bar{A}}^{(k)} \tilde{\bar{F}}_k$, which, because of the relation (58), implies that $\bar{A}^{(k+1)} = \tilde{F}_k^T \tilde{A}^{(k)} \tilde{F}_k$. Since the relation (44) and Proposition 8 imply

$$\begin{aligned} \bar{A}^{(k+1)} &= D_k F_k^T D_k^{-1} D_k A^{(k)} D_k D_k^{-1} F_k D_k = D_k F_k^T A^{(k)} F_k D_k \\ &= D_k A^{(k+1)} D_k, \end{aligned} \tag{59}$$

we conclude that normalizations of the matrices $A^{(k+1)}$ and $\bar{A}^{(k+1)}$ give the same matrix. Now we use the same argument to show that $\tilde{\bar{B}}^{(k+1)} = \tilde{B}^{(k+1)}$ for $k \geqq 1$ and to prove (i).

(ii) For $k = 1$ the assertion is trivially fulfilled. Let the assertion hold for some $k \geqq 1$. From the relations (59), (53) and the assertion (i) we obtain

$$\begin{aligned} \bar{D}_{k+1} D_k A^{(k+1)} D_k \bar{D}_{k+1} &= \bar{D}_{k+1} \bar{A}^{(k+1)} \bar{D}_{k+1} \\ &= \tilde{\bar{A}}^{(k+1)} = \tilde{A}^{(k+1)} = D_{k+1} A^{(k+1)} D_{k+1}. \end{aligned}$$

It is obvious that $D_{k+1} = D_k \bar{D}_{k+1}$ and inserting the induction assumption we conclude that (ii) holds. □

From Proposition 9 we see that the relations (50), (52), and (53) can be written as

$$(\tilde{A}^{(1)}, \tilde{B}^{(1)}), (\bar{A}^{(2)}, \bar{B}^{(2)}), (\tilde{A}^{(2)}, \tilde{B}^{(2)}), (\bar{A}^{(3)}, \bar{B}^{(3)}), (\tilde{A}^{(3)}, \tilde{B}^{(3)}), \cdots, \tag{60}$$

$$\bar{A}^{(k+1)} = \tilde{F}_k^T \tilde{A}^{(k)} \tilde{F}_k, \qquad \bar{B}^{(k+1)} = \tilde{F}_k^T \tilde{B}^{(k)} \tilde{F}_k, \tag{61}$$

$$\tilde{A}^{(k+1)} = \bar{D}_{k+1} \bar{A}^{(k+1)} \bar{D}_{k+1}, \qquad \tilde{B}^{(k+1)} = \bar{D}_{k+1} \bar{B}^{(k+1)} \bar{D}_{k+1}. \tag{62}$$

The relations (60), (61), (62), (54), and (55) define the normalized Falk–Langemeyer method. We use the normalized method only as an aid to obtain information about the quantity $\tilde{\varepsilon}_k$.

4.1. Preliminaries. Here we define asymptotic assumptions and prove several lemmas which are used later in the proof of the quadratic convergence of the Falk–Langemeyer method. The quadratic convergence proof is based on the idea of Wilkinson (see [18]) which consists in estimating the growth of already annihilated elements in the current

cycle. To this end we must estimate the transformation parameters $\tilde{\alpha}_k$ and $\tilde{\beta}_k$ and also the growth of all off-diagonal elements in the current cycle. These two tasks are solved in Lemmas 11, 13, 14, and 15. Lemma 10 gives us two numeric relations which are used in the proof. Lemmas 11 and 12 estimate the transformation parameters $\tilde{\alpha}_k$ and $\tilde{\beta}_k$, and the measure $\tilde{\varepsilon}_k$ in one step. Lemmas 13–15 estimate the growth of $\tilde{\alpha}_k$, $\tilde{\beta}_k$, and $\tilde{\varepsilon}_k$ during N consecutive steps. Lemma 15 is the most important for the proof of the quadratic convergence. In this section we do not assume that the pivot strategy is cyclic. Therefore the results of this section hold for any pivot strategy. However, if the pivot strategy is cyclic, then Lemmas 13–15 explain the behaviour of $\tilde{\alpha}_k$, $\tilde{\beta}_k$, and $\tilde{\varepsilon}_k$ during one cycle.

As we said in § 1, the quadratic convergence can always be expected if the eigenvalues of problem (1) are simple. We will therefore use two quadratic convergence assumptions:

(A1) The eigenvalues of the pair (A, B) are simple, i.e.,

$$p = n \geqq 3.$$

(A2) The pair (A, B) is almost diagonal, i.e.,

$$\frac{\tilde{\varepsilon}_1}{\delta} < \frac{1}{2N}.$$

Asymptotic assumption (A2) is similar to the assumptions used in Theorem 6 and in convergence results of some other Jacobi-type methods (see [4], [1]). Assumption (A1) implies

$$N \geqq 3 \tag{63}$$

and

$$\varepsilon_k = \tau_k, \qquad \tilde{\varepsilon}_k = \tilde{\tau}_k, \qquad k \geqq 1, \tag{64}$$

where $\tau_k = \tau(A^{(k)}, B^{(k)})$ and $\tau_k = \tau(\tilde{A}^{(k)}, \tilde{B}^{(k)})$. We shall use the notation

$$\tilde{a}_k = |\tilde{a}_{lm}^{(k)}|, \qquad \tilde{b}_k = |\tilde{b}_{lm}^{(k)}|, \qquad k \geqq 1. \tag{65}$$

LEMMA 10. *Let r be an integer such that $r \geqq 3$ and let x be a nonnegative real number satisfying $2xr < 1$. Then the following inequalities hold:*

$$(1-x)^{-r} \leqq 1 + \tfrac{12}{7} \cdot r \cdot x, \qquad (1+x)^r \leqq 1 + \tfrac{4}{3} \cdot r \cdot x.$$

Proof. The proof of this lemma is elementary and can be found in [4]. □

The following lemma shows how the transformation parameters $\tilde{\alpha}_k$ and $\tilde{\beta}_k$ from matrices $\tilde{F}_k$ are bounded with $\tilde{\varepsilon}_k$.

LEMMA 11. *Let the assumption* (A1) *hold. If for some $k \geqq 1$ holds*

$$\tilde{\varepsilon}_k < \frac{2}{3N}\delta, \tag{66}$$

then

$$\max\{|\tilde{\alpha}_k|, |\tilde{\beta}_k|\} \leqq 0.34 \cdot \frac{\sqrt{(\tilde{a}_k)^2 + (\tilde{b}_k)^2}}{\delta}. \tag{67}$$

Proof. Suppose that for some $k \geqq 1$ the relation (66) holds. Then Theorem 1 and Corollary 2 hold for the pair $(\tilde{A}^{(k)}, \tilde{B}^{(k)})$ as well. Assumption (A1) and the relations

(63), (64), and (18) imply that there exists an ordering of the eigenvalues of the pair (A, B),[2] such that

$$\chi(\lambda_i, [\tilde{a}_{ii}^{(k)}, \tilde{b}_{ii}^{(k)}]) \leqq \frac{\tilde{\varepsilon}_k^2}{2\delta} < \frac{4\delta^2}{9N^2} \cdot \frac{1}{2\delta} < \frac{2}{81} \cdot \delta < 0.025 \cdot \delta, \qquad i = 1, \cdots, n. \tag{68}$$

Applying twice the triangle inequality and using the definition (11) and the relation (68), we obtain

$$\begin{aligned} |\tilde{\mathfrak{F}}_{lm}^{(k)}| &= |\tilde{a}_{ll}^{(k)}\tilde{b}_{mm}^{(k)} - \tilde{a}_{mm}^{(k)}\tilde{b}_{ll}^{(k)}| = \chi([\tilde{a}_{ll}^{(k)}, \tilde{b}_{ll}^{(k)}], [\tilde{a}_{mm}^{(k)}, \tilde{b}_{mm}^{(k)}]) \\ &\geqq \chi(\lambda_l, \lambda_m) - \chi(\lambda_l, [\tilde{a}_{ll}^{(k)}, \tilde{b}_{ll}^{(k)}]) - \chi(\lambda_m, [\tilde{a}_{mm}^{(k)}, \tilde{b}_{mm}^{(k)}]) \\ &> 3 \cdot \delta - 2 \cdot 0.025 \cdot \delta = 2.95 \cdot \delta. \end{aligned} \tag{69}$$

It is obvious that $|\tilde{\mathfrak{F}}_{lm}^{(k)}| \neq 0$. This excludes cases (5)(a) and (5)(b)(ii) of Algorithm 4. Therefore, we have

$$\max\{|\tilde{\alpha}_k|, |\tilde{\beta}_k|\} \leqq \frac{2}{|\tilde{\mathfrak{F}}_{lm}^{(k)}| + \sqrt{\tilde{\mathfrak{F}}^{(k)}}} \cdot \max\{|\tilde{\mathfrak{F}}_l^{(k)}|, |\tilde{\mathfrak{F}}_m^{(k)}|\}. \tag{70}$$

From the Cauchy–Schwarz inequality and the relations (47) and (65) we have

$$\begin{aligned} |\tilde{\mathfrak{F}}_l^{(k)}| &= |\tilde{a}_{ll}^{(k)}\tilde{b}_{lm}^{(k)} - \tilde{b}_{ll}^{(k)}\tilde{a}_{lm}^{(k)}| \leqq \sqrt{(\tilde{a}_{ll}^{(k)})^2 + (\tilde{b}_{ll}^{(k)})^2}\sqrt{(\tilde{a}_{lm}^{(k)})^2 + (\tilde{b}_{lm}^{(k)})^2} \\ &= \sqrt{(\tilde{a}_k)^2 + (\tilde{b}_k)^2}. \end{aligned}$$

The same estimate holds for $\tilde{\mathfrak{F}}_m^{(k)}$ and therefore

$$\max\{|\tilde{\mathfrak{F}}_l^{(k)}|, |\tilde{\mathfrak{F}}_m^{(k)}|\} \leqq \sqrt{(\tilde{a}_k)^2 + (\tilde{b}_k)^2}. \tag{71}$$

Since

$$(\tilde{a}_k)^2 + (\tilde{b}_k)^2 \leqq \tfrac{1}{2} \cdot \tilde{\varepsilon}_k^2, \tag{72}$$

the relations (69), (71), and (72) imply

$$\begin{aligned} \sqrt{\tilde{\mathfrak{F}}^{(k)}} &= \sqrt{(\tilde{\mathfrak{F}}_{lm}^{(k)})^2 + 4\tilde{\mathfrak{F}}_l^{(k)}\tilde{\mathfrak{F}}_m^{(k)}} \geqq \sqrt{(2.95\delta)^2 - 4 \cdot \tilde{\varepsilon}_k^2/2} \\ &\geqq \sqrt{(2.95\delta)^2 - 2 \cdot (4/9N^2)\delta^2} \geqq 2.933\delta. \end{aligned}$$

The assertion (67) now follows from the relations (70), (69), (71), and the above relation. □

The following lemma gives the relation between $\tilde{\varepsilon}_k$ and $\tilde{\varepsilon}_{k+1}$. It is used later in the proof of Lemma 15.

LEMMA 12. *Let the assumption* (A1) *hold. If for some* $k \geqq 1$ *the relation* (66) *holds, then*

$$\tilde{\varepsilon}_{k+1}^2 \leqq \frac{1 + 0.494 \cdot \tilde{\varepsilon}_k/\delta}{1 - 0.077 \cdot \tilde{\varepsilon}_k/\delta} [\tilde{\varepsilon}_k^2 - 2(\tilde{a}_k^2 + \tilde{b}_k^2)]. \tag{73}$$

Proof. Suppose that the relation (66) holds for some $k \geqq 1$. The relation (62), together with the definition of $\tilde{\varepsilon}_k$, implies

$$\tilde{\varepsilon}_{k+1}^2 = S^2(\bar{D}_{k+1}\bar{A}^{(k+1)}\bar{D}_{k+1}) + S^2(\bar{D}_{k+1}\bar{B}^{(k+1)}\bar{D}_{k+1}). \tag{74}$$

---

[2] Since $p = n$, the eigenvalues can be ordered so that the matrix P from Theorem 1 and Corollary 2 is the identity matrix.

If

$$m_{k+1} = \min \{ \bar{d}_1^{(k+1)}, \cdots, \bar{d}_n^{(k+1)} \},$$

then the relation (74) implies

$$\tilde{\varepsilon}_{k+1}^2 \leq S^2\left(\frac{1}{(m_{k+1})^2} \bar{A}^{(k+1)}\right) + S^2\left(\frac{1}{(m_{k+1})^2} \bar{B}^{(k+1)}\right) = \frac{1}{(m_{k+1})^4} \cdot \bar{\varepsilon}_{k+1}^2. \tag{75}$$

Let us define vectors

$$\tilde{a}^l = (\tilde{a}_{l,1}^{(k)}, \tilde{a}_{l,2}^{(k)}, \cdots, \tilde{a}_{l,l-1}^{(k)}, \tilde{a}_{l,l+1}^{(k)}, \cdots, \tilde{a}_{l,m-1}^{(k)}, \tilde{a}_{l,m+1}^{(k)}, \cdots, \tilde{a}_{l,n}^{(k)}),$$

$$\tilde{a}^m = (\tilde{a}_{m,1}^{(k)}, \tilde{a}_{m,2}^{(k)}, \cdots, \tilde{a}_{m,l-1}^{(k)}, \tilde{a}_{m,l+1}^{(k)}, \cdots, \tilde{a}_{m,m-1}^{(k)}, \tilde{a}_{m,m+1}^{(k)}, \cdots, \tilde{a}_{m,n}^{(k)}),$$

$$\tilde{a}_l^T = (\tilde{a}_{1,l}^{(k)}, \tilde{a}_{2,l}^{(k)}, \cdots, \tilde{a}_{l-1,l}^{(k)}, \tilde{a}_{l+1,l}^{(k)}, \cdots, \tilde{a}_{m-1,l}^{(k)}, \tilde{a}_{m+1,l}^{(k)}, \cdots, \tilde{a}_{n,l}^{(k)}),$$

$$\tilde{a}_m^T = (\tilde{a}_{1,m}^{(k)}, \tilde{a}_{2,m}^{(k)}, \cdots, \tilde{a}_{l-1,m}^{(k)}, \tilde{a}_{l+1,m}^{(k)}, \cdots, \tilde{a}_{m-1,m}^{(k)}, \tilde{a}_{m+1,m}^{(k)}, \cdots, \tilde{a}_{n,m}^{(k)}),$$

where generally a^T denotes the transposed vector a. Let $\bar{a}^l$, $\bar{a}^m$, $\bar{a}_l$, and $\bar{a}_m$ be row and column vectors defined in the same way, but from elements of the matrix $\bar{A}^{(k+1)}$. Relation (61) implies that

$$\begin{bmatrix} \bar{a}^l \\ \bar{a}^m \end{bmatrix} = \hat{\hat{F}}_k^T \begin{bmatrix} \tilde{a}^l \\ \tilde{a}^m \end{bmatrix}, \qquad [\bar{a}_l, \bar{a}_m] = [\tilde{a}_l, \tilde{a}_m] \hat{\hat{F}}_k.$$

Therefore,

$$\left\| \begin{bmatrix} \bar{a}^l \\ \bar{a}^m \end{bmatrix} \right\|^2 \leq \| \hat{\hat{F}}_k^T \|_2^2 \left\| \begin{bmatrix} \tilde{a}^l \\ \tilde{a}^m \end{bmatrix} \right\|^2, \qquad \| [\bar{a}_l, \bar{a}_m] \|^2 \leq \| \hat{\hat{F}}_k \|_2^2 \| [\tilde{a}_l, \tilde{a}_m] \|^2.$$

The off-diagonal elements of the matrix $\tilde{A}^{(k)}$ which are changed in the transformation (61), are exactly the elements of vectors $\tilde{a}^l$, $\tilde{a}^m$, $\tilde{a}_l$, and $\tilde{a}_m$ with the exception of $\tilde{a}_{lm}^{(k)}$ and $\tilde{a}_{ml}^{(k)}$ which are annihilated. Since $\| \hat{\hat{F}}_k^T \|_2 = \| \hat{\hat{F}}_k \|_2$, we conclude that

$$S^2(\bar{A}^{(k+1)}) \leq S^2(\tilde{A}^{(k)}) + (\| \hat{\hat{F}}_k \|_2^2 - 1)(\|\tilde{a}^l\|^2 + \|\tilde{a}^m\|^2 + \|\tilde{a}_l\|^2 + \|\tilde{a}_m\|^2) - 2\tilde{a}_k^2.$$

Since $\| \hat{\hat{F}}_k \|_2 \geq 1$ (see further in the proof), we conclude that

$$S^2(\bar{A}^{(k+1)}) \leq \| \hat{\hat{F}}_k \|_2^2 (S^2(\tilde{A}^{(k)}) - 2\tilde{a}_k^2).$$

By applying the similar analysis to matrix $\tilde{B}^{(k)}$, we obtain

$$S^2(\bar{B}^{(k+1)}) \leq \| \hat{\hat{F}}_k \|_2^2 (S^2(\tilde{B}^{(k)}) - 2\tilde{b}_k^2).$$

Adding two previous inequalities and using the definitions of $\bar{\varepsilon}_{k+1}$ and $\tilde{\varepsilon}_k$, gives

$$\bar{\varepsilon}_{k+1}^2 \leq \| \hat{\hat{F}}_k \|_2^2 [\tilde{\varepsilon}_k^2 - 2(\tilde{a}_k^2 + \tilde{b}_k^2)].$$

Inserting this inequality into relation (75), we obtain

$$\tilde{\varepsilon}_{k+1}^2 \leq \frac{\| \hat{\hat{F}}_k \|_2^2}{(m_{k+1})^4} \cdot [\tilde{\varepsilon}_k^2 - 2(\tilde{a}_k^2 + \tilde{b}_k^2)]. \tag{76}$$

To complete the proof we must find the upper bound for $\| \hat{\hat{F}}_k \|_2^2$ and the lower bound for m_{k+1}.

The relation (56) implies that

$$m_{k+1} = \min \{1, \bar{d}_l^{(k+1)}, \bar{d}_m^{(k+1)}\}. \tag{77}$$

Relation (61) implies that

$$\bar{a}_{ll}^{(k+1)} = \tilde{a}_{ll}^{(k)} - 2\tilde{\beta}_k \tilde{a}_{lm}^{(k)} + \tilde{\beta}_k^2 \tilde{a}_{mm}^{(k)}, \qquad \bar{b}_{ll}^{(k+1)} = \tilde{b}_{ll}^{(k)} - 2\tilde{\beta}_k \tilde{b}_{lm}^{(k)} + \tilde{\beta}_k^2 \tilde{b}_{mm}^{(k)},$$

$$\bar{a}_{mm}^{(k+1)} = \tilde{\alpha}_k^2 \tilde{a}_{ll}^{(k)} + 2\tilde{\alpha}_k \tilde{a}_{lm}^{(k)} + \tilde{a}_{mm}^{(k)}, \qquad \bar{b}_{mm}^{(k+1)} = \tilde{\alpha}_k^2 \tilde{b}_{ll}^{(k)} + 2\tilde{\alpha}_k \tilde{b}_{lm}^{(k)} + \tilde{b}_{mm}^{(k)}.$$

Therefore

$$\begin{aligned}(\bar{d}_l^{(k+1)})^4 &= (\bar{a}_{ll}^{(k+1)})^2 + (\bar{b}_{ll}^{(k+1)})^2 = 1 + 4\beta_k^2[(\tilde{a}_{lm}^{(k)})^2 + (\tilde{b}_{lm}^{(k)})^2] \\ &\quad + \tilde{\beta}_k^4 - 4\tilde{\beta}_k(\tilde{a}_{ll}^{(k)}\tilde{a}_{lm}^{(k)} + \tilde{b}_{ll}^{(k)}\tilde{b}_{lm}^{(k)}) - 4\tilde{\beta}_k^3(\tilde{a}_{mm}^{(k)}\tilde{a}_{lm}^{(k)} + \tilde{b}_{mm}^{(k)}\tilde{b}_{lm}^{(k)}) \\ &\quad + 2\tilde{\beta}_k^2(\tilde{a}_{ll}^{(k)}\tilde{a}_{mm}^{(k)} + \tilde{b}_{ll}^{(k)}\tilde{b}_{mm}^{(k)}) \\ &\geqq 1 - 4|\tilde{\beta}_k|\,|\tilde{a}_{ll}^{(k)}\tilde{a}_{lm}^{(k)} + \tilde{b}_{ll}^{(k)}\tilde{b}_{lm}^{(k)}| - 4|\tilde{\beta}_k|^3\,|\tilde{a}_{mm}^{(k)}\tilde{a}_{lm}^{(k)} + \tilde{b}_{mm}^{(k)}\tilde{b}_{lm}^{(k)}| \\ &\quad - 2\tilde{\beta}_k^2\,|\tilde{a}_{ll}^{(k)}\tilde{a}_{mm}^{(k)} + \tilde{b}_{ll}^{(k)}\tilde{b}_{mm}^{(k)}|.\end{aligned} \tag{78}$$

Using the relation (47) and the Cauchy–Schwarz inequality in the relation (78), we obtain

$$(\bar{d}_l^{(k+1)})^4 \geqq 1 - 4|\tilde{\beta}_k|(1 + |\tilde{\beta}_k|^2)\sqrt{(\tilde{a}_{lm}^{(k)})^2 + (\tilde{b}_{lm}^{(k)})^2} - 2\tilde{\beta}_k^2. \tag{79}$$

Using a similar argument we obtain

$$(\bar{d}_m^{(k+1)})^4 \geqq 1 - 4|\tilde{\alpha}_k|(1 + |\tilde{\alpha}_k|^2)\sqrt{(\tilde{a}_{lm}^{(k)})^2 + (\tilde{b}_{lm}^{(k)})^2} - 2\tilde{\alpha}_k^2. \tag{80}$$

Relations (77), (79), (80), (72), and Lemma 11 now imply

$$m_{k+1}^4 \geqq 1 - 4 \cdot \frac{0.34}{\sqrt{2}} \cdot \frac{\tilde{\varepsilon}_k}{\delta} \cdot \left[1 + \left(\frac{0.34}{\sqrt{2}} \cdot \frac{\tilde{\varepsilon}_k}{\delta}\right)^2\right] \cdot \frac{\tilde{\varepsilon}_k}{\sqrt{2}} - 2 \cdot \left(\frac{0.34}{\sqrt{2}} \cdot \frac{\tilde{\varepsilon}_k}{\delta}\right)^2. \tag{81}$$

Since χ is chordal metric, from the relation (11) we see that $\delta \leqq \frac{1}{3}$. The relations (66) and (63) therefore imply

$$\tilde{\varepsilon}_k < \frac{2}{9N} < \frac{2}{27}. \tag{82}$$

Inserting the relation (82) and the assumption (66) into the relation (81) we obtain

$$m_{k+1}^4 > 1 - \frac{0.34}{\sqrt{2}} \cdot \frac{\tilde{\varepsilon}_k}{\delta}\left[4 \cdot \left(1 + \left(\frac{0.34}{\sqrt{2}} \cdot \frac{2}{3N}\right)^2\right) \cdot \frac{2}{27\sqrt{2}} + 0.34 \cdot \sqrt{2} \cdot \frac{2}{3N}\right].$$

Finally, taking into account that $N \geqq 3$ we obtain

$$m_{k+1}^4 \geqq 1 - 0.077 \cdot \frac{\tilde{\varepsilon}_k}{\delta}. \tag{83}$$

We shall now estimate $\|\hat{\tilde{F}}_k\|_2^2$. Since

$$\|\hat{\tilde{F}}_k\|_2^2 \leqq \|\hat{\tilde{F}}_k\|_1 \cdot \|\hat{\tilde{F}}_k\|_\infty,$$

where $\|A\|_1 = \max_j \sum_i |a_{ij}|$, $\|A\|_\infty = \max_i \sum_j |a_{ij}|$ for $A = (a_{ij})$, we obtain

$$\|\hat{\tilde{F}}_k\|_2^2 \leqq (1 + \max\{|\tilde{\alpha}_k|, |\tilde{\beta}_k|\})^2.$$

Lemma 11 and the relation (66) now imply

$$
\begin{aligned}
\|\hat{\tilde{F}}_k\|_2^2 &\leq \left(1+\frac{0.34}{\sqrt{2}}\frac{\tilde{\varepsilon}_k}{\delta}\right)^2 \\
&\leq 1+\frac{0.34}{\sqrt{2}}\frac{\tilde{\varepsilon}_k}{\delta}\left(2+\frac{0.34}{\sqrt{2}}\frac{2}{3N}\right) \qquad (84)\\
&\leq 1+0.494\frac{\tilde{\varepsilon}_k}{\delta}.
\end{aligned}
$$

The relation (73) now follows from the relations (76), (83), and (84). □

We shall now prove that if the assumptions (A1) and (A2) hold, then Lemmas 11 and 12 hold during N consecutive steps.

LEMMA 13. *Let the asymptotic assumptions* (A1) *and* (A2) *hold. Then for each* $k \in \{1, \cdots, N\}$ *holds*

$$
\tilde{\varepsilon}_k \leq \frac{1}{1-0.3\cdot(k-1)\tilde{\varepsilon}_1/\delta}\cdot\tilde{\varepsilon}_1, \qquad \frac{\tilde{\varepsilon}_k}{\delta} < \frac{2}{3N}.
$$

Proof. The proof is by induction. For $k = 1$, the lemma is trivially fulfilled. Suppose the lemma holds for some $k \in \{1, \cdots, N-1\}$. From the second inequality in the induction assumption we conclude that, for the chosen k, Lemmas 11 and 12 hold. From Lemma 12 it follows that

$$
\begin{aligned}
\tilde{\varepsilon}_{k+1}^2 &\leq \frac{1+0.494\cdot\tilde{\varepsilon}_k/\delta}{1-0.077\cdot\tilde{\varepsilon}_k/\delta}\cdot\tilde{\varepsilon}_k^2 \\
&\leq \frac{1}{(1-0.494(\tilde{\varepsilon}_k/\delta))(1-0.077(\tilde{\varepsilon}_k/\delta))}\tilde{\varepsilon}_k^2 \qquad (85)\\
&\leq \frac{1}{(1-0.3(\tilde{\varepsilon}_k/\delta))^2}\tilde{\varepsilon}_k^2.
\end{aligned}
$$

Hence

$$
\tilde{\varepsilon}_{k+1} \leq \frac{1}{1-0.3\cdot\tilde{\varepsilon}_k/\delta}\cdot\tilde{\varepsilon}_k.
$$

Inserting the induction assumption in this inequality, we obtain

$$
\begin{aligned}
\tilde{\varepsilon}_{k+1} &\leq \frac{1}{1-0.3\dfrac{1}{1-0.3(k-1)\tilde{\varepsilon}_1/\delta}\tilde{\varepsilon}_1/\delta}\cdot\frac{1}{1-0.3(k-1)\tilde{\varepsilon}_1/\delta}\cdot\tilde{\varepsilon}_1 \\
&\leq \frac{1}{1-0.3(k-1)\tilde{\varepsilon}_1/\delta-0.3(\tilde{\varepsilon}_1/\delta)}\cdot\tilde{\varepsilon}_1 = \frac{1}{1-0.3\cdot k\cdot\tilde{\varepsilon}_1/\delta}\cdot\tilde{\varepsilon}_1,
\end{aligned}
$$

and the first assertion of the lemma is proved. From this assertion for $k+1$, because of the asymptotic assumption (A2), we now have

$$
\frac{\tilde{\varepsilon}_{k+1}}{\delta} \leq \frac{1}{1-0.3\cdot k\cdot\tilde{\varepsilon}_1/\delta}\cdot\frac{\tilde{\varepsilon}_1}{\delta} < \frac{1}{1-0.3(N-1)/2N}\cdot\frac{1}{2N} = \frac{1}{2-0.3}\cdot\frac{1}{N} < \frac{2}{3N},
$$

which completes the proof. □

LEMMA 14. *If the asymptotic assumptions* (A1) *and* (A2) *hold, then the assertions* (67) *and* (73) *of Lemmas* 11 *and* 12 *hold for every* $k \in \{1, \cdots, N\}$.

Proof. The assertion follows immediately from second assertion of Lemma 13. □

The next lemma explains the behaviour of $S(\tilde{A}^{(k)})$, $S(\tilde{B}^{(k)})$ and $\tilde{\varepsilon}_k$, and of the transformation parameters $\tilde{\alpha}_k$ and $\tilde{\beta}_k$ during N consecutive steps. Let us define the quantity

$$c_N = \frac{1 + 0.494 \cdot 2/3N}{1 - 0.077 \cdot 2/3N}. \tag{86}$$

LEMMA 15. *Let the asymptotic assumptions* (A1) *and* (A2) *hold. Then*:

(i) *For* $k = 1, \cdots, N$ *holds*

$$\begin{bmatrix} S^2(\tilde{A}^{(k+1)}) \\ S^2(\tilde{B}^{(k+1)}) \\ \tilde{\varepsilon}_{k+1}^2 \end{bmatrix} \leqq (c_N)^k \begin{bmatrix} S^2(\tilde{A}^{(1)}) \\ S^2(\tilde{B}^{(1)}) \\ \tilde{\varepsilon}_1^2 \end{bmatrix} \leqq 1.566 \begin{bmatrix} S^2(\tilde{A}^{(1)}) \\ S^2(\tilde{B}^{(1)}) \\ \tilde{\varepsilon}_1^2 \end{bmatrix}.$$

(ii) *For any choice* $\tilde{\omega}_k \in \{\tilde{\alpha}, \tilde{\beta}_k\}$, $1 \leqq k \leqq N$, *holds*

$$\sum_{k=1}^{N} \tilde{\omega}_k^2 \leqq 0.091 \cdot \frac{\tilde{\varepsilon}_1^2}{\delta^2}.$$

Proof. (i) Because of Lemmas 12 and 14, for $k = 1, \cdots, N$ holds

$$\begin{aligned} \tilde{\varepsilon}_{k+1}^2 &\leqq c_N(\tilde{\varepsilon}_k^2 - 2(\tilde{a}_k^2 + \tilde{b}_k^2)) \\ &\leqq c_N\{c_N[\tilde{\varepsilon}_{k-1}^2 - 2(\tilde{a}_{k-1}^2 + \tilde{b}_{k-1}^2)] - 2(\tilde{a}_k^2 + \tilde{b}_k^2)\} \\ &\leqq \cdots \leqq (c_N)^k \tilde{\varepsilon}_1^2 - 2 \sum_{j=1}^{k} (c_N)^{k-j+1}(\tilde{a}_j^2 + \tilde{b}_j^2). \end{aligned} \tag{87}$$

From the relation (87) immediately follows

$$\tilde{\varepsilon}_{k+1}^2 \leqq (c_N)^k \tilde{\varepsilon}_1^2 \leqq (c_N)^N \tilde{\varepsilon}_1^2, \qquad k = 1, \cdots, N. \tag{88}$$

Using Lemma 10 we obtain

$$(c_N)^N < \left(1 + \frac{4}{3} \cdot 0.494 \cdot \frac{2}{3N} \cdot N\right)\left(1 + \frac{12}{7} \cdot 0.077 \cdot \frac{2}{3N} \cdot N\right) < 1.566. \tag{89}$$

Inserting this inequality into relation (88), we obtain

$$\tilde{\varepsilon}_{k+1}^2 \leqq 1.566 \cdot \tilde{\varepsilon}_1^2, \qquad k = 1, \cdots, N.$$

From the proof of Lemma 12 we see that the above estimates hold for the quantities $S^2(\tilde{A}^{(k+1)})$ and $S^2(\tilde{B}^{(k+1)})$, as well. Therefore (i) is proved.

(ii) Since $c_N > 1$, from the relation (87) for $k = N$ we have

$$\tilde{\varepsilon}_{N+1}^2 \leqq (c_N)^N \tilde{\varepsilon}_1^2 - 2 \sum_{k=1}^{N} (\tilde{a}_k^2 + \tilde{b}_k^2).$$

Since $\tilde{\varepsilon}_{N+1}^2 \geqq 0$, this inequality implies

$$\sum_{k=1}^{N} (\tilde{a}_k^2 + \tilde{b}_k^2) \leqq \frac{1}{2} \cdot (c_N)^N \tilde{\varepsilon}_1^2 \leqq 0.783 \cdot \tilde{\varepsilon}_1^2.$$

The above inequality, together with Lemmas 11 and 14, implies

$$\sum_{k=1}^{N} \tilde{\omega}_k^2 \leqq \sum_{k=1}^{N} \max\{\tilde{\alpha}_k^2, \tilde{\beta}_k^2\} \leqq \sum_{k=1}^{N} 0.34^2 \cdot (\tilde{a}_k^2 + \tilde{b}_k^2) \cdot \frac{1}{\delta^2}$$

$$\leqq 0.1156 \cdot 0.783 \cdot \frac{\tilde{\varepsilon}_1^2}{\delta^2} \leqq 0.091 \cdot \frac{\tilde{\varepsilon}_1^2}{\delta^2}$$

and the lemma is proved. □

4.2. The proof. Here we prove that the Falk–Langemeyer method is quadratically convergent if the assumptions (A1) and (A2) are fulfilled and the pivot strategy is cyclic. Then we prove that the quadratic convergence implies the convergence of the sequence of pairs (42) towards the pair of diagonal matrices. At the end we prove that the measures $\tilde{\varepsilon}_k$ and ε_k are equivalent in the sense that ultimately the quadratic reduction of $\tilde{\varepsilon}_{kN+1}$ implies the quadratic reduction of of ε_{kN+1} and vice versa.

We can now prove our paper's central theorem.

THEOREM 16. *Let the asymptotic assumptions* (A1) *and* (A2) *hold and let the sequence* $((A^{(k)}, B^{(k)}), k \geqq 1)$ *be generated with the Falk–Langemeyer method from the pair* (A, B). *Then for any cyclic strategy holds*

$$\tilde{\varepsilon}_{N+1} \leqq \sqrt{N} \cdot \frac{\tilde{\varepsilon}_1^2}{\delta}.$$

Proof. Let us fix some $k \in \{1, \cdots, N\}$. Then the pivot pair (l, m) is also fixed. We want to know what happens with the element on this position until the end of cycle. Therefore, we will observe the elements $\tilde{a}_{lm}^{(r)}$, $r = k+1, \cdots, N$. We know that $\tilde{a}_{lm}^{(k+1)} = 0$ and that the elements $\tilde{a}_{ml}^{(r)}$ actually change at most $2(n-2)$ times. Let $r_1, \cdots, r_s$, $s \leqq 2n-4$, denote those values of r for which $\tilde{a}_{lm}^{(r)}$ changes in the rth step. Let us introduce the notation:

$$\begin{aligned} &\bar{z}_i = \bar{a}_{lm}^{(r_i+1)}, \qquad \tilde{z}_i = \tilde{a}_{lm}^{(r_i+1)}, \\ &\bar{d}_j^{(i)} = \sqrt[4]{(\bar{a}_{jj}^{(r_i+1)})^2 + (\bar{b}_{jj}^{(r_i+1)})^2}, \qquad j \in \{l, m\}, \\ &\bar{m}_{lm}^{(i)} = \min\{\bar{d}_l^{(i)}, \bar{d}_m^{(i)}\}, \qquad d_N = \sqrt{1 - 0.077 \cdot 2/3N}. \end{aligned} \tag{90}$$

Performing the r_1th step according to Algorithm 4 gives

$$\bar{z}_1 = (0 \cdot 1 \pm \tilde{a}^{(r_1)} \tilde{\omega}_{r_1}),$$

where $\tilde{\omega}_{r_1} \in \{\tilde{\alpha}_{r_1}, \tilde{\beta}_{r_1}\}$ and $\tilde{a}^{(r_1)}$ is some off-diagonal element of the matrix $\tilde{A}^{(r_1)}$. Since

$$\tilde{z}_i = \frac{\bar{z}_i}{\bar{d}_l^{(i)} \bar{d}_m^{(i)}}, \qquad i = 1, \cdots, s, \tag{91}$$

from the relations (90), (83), and Lemma 12, it follows that

$$|\tilde{z}_1| \leqq \frac{1}{(\bar{m}_{lm}^{(i)})^2} |\tilde{a}^{(r_1)}| |\tilde{\omega}_{r_1}| \leqq \frac{1}{d_N} |\tilde{a}^{(r_1)}| |\bar{\omega}_{r_1}|. \tag{92}$$

Furthermore, in the r_2th step, we have

$$\bar{z}_2 = (1 \cdot \tilde{z}_1 \pm \tilde{a}^{(r_2)} \tilde{\omega}_{r_2}), \tag{93}$$

where $\tilde{\omega}_{r_2} \in \{\tilde{\alpha}_{r_2}, \tilde{\beta}_{r_2}\}$, and $\tilde{a}^{(r_2)}$ is some off-diagonal element of the matrix $\tilde{A}^{(r_2)}$. The relations (93), (92), and (91) imply

$$|\bar{z}_2| \leqq \frac{1}{d_N} \cdot \left(\frac{1}{d_N} |\tilde{a}^{(r_1)}| \, |\tilde{\omega}_{r_1}| + |\tilde{a}^{(r_2)}| \, |\tilde{\omega}_{r_2}| \right).$$

By induction we obtain

$$|\bar{z}_j| \leqq \sum_{i=1}^{j} \frac{1}{(d_N)^{j-i+1}} |\tilde{a}^{(r_i)}| \, |\tilde{\omega}_{r_i}|, \qquad j = 1, \cdots, s. \tag{94}$$

For $k = 1, \cdots, N+1$, the following notation is introduced:

$$\tilde{A}^{(k)} = \tilde{D}_A^{(k)} + \tilde{E}^{(k)}, \qquad \tilde{D}_A^{(k)} = \operatorname{diag}(\tilde{a}_{ii}^{(k)}). \tag{95}$$

Matrix $\tilde{E}^{(N+1)}$ obviously consists of elements which have undergone the maximal number of changes. If $s(i, j)$ denotes the number of changes of the element on position (i, j), then

$$s(i,j) \leqq 2n-4, \qquad i, j \in \{1, \cdots, n\}, \qquad i \neq j.$$

The quantity $s(i, j)$ depends on (i, j) and the pivot strategy. Elements of the matrix $\tilde{E}^{(N+1)}$ can therefore be denoted as $\tilde{z}_{s(i,j)}$.

Having in mind relation (94), we can now write

$$|\tilde{E}^{(N+1)}| \leqq \frac{1}{(d_N)^{2n-4}} (|\tilde{P}^{(2)}| \, |\tilde{\omega}_2| + |\tilde{P}^{(3)}| \, |\tilde{\omega}_3| + \cdots + |\tilde{P}^{(N)}| \, |\tilde{\omega}_N|). \tag{96}$$

Here the notation $|C| = (|c_{ij}|)$ for $C = (c_{ij})$ is used. Matrix $\tilde{P}^{(k)}$ consists precisely of those elements of $l(k)$th and $m(k)$th row and column of the matrix $\tilde{E}^{(k)}$ which already were pivot elements,[3] i.e., of elements which contribute to the final estimate. All other elements of the matrix $\tilde{P}^{(k)}$ are zeros.

Assertion (i) of Lemma 15 gives us

$$\| |\tilde{P}^{(k)}| \| = \|\tilde{P}^{(k)}\| \leqq S(\tilde{A}^{(k)}) \leqq \sqrt{1.566} \cdot S(\tilde{A}^{(1)}), \qquad k = 2, \cdots, N. \tag{97}$$

From the relations (96) and (97), Lemma 15, and the Cauchy–Schwarz inequality we obtain

$$\begin{aligned} S(\tilde{A}^{(N+1)}) &= \|\tilde{E}^{(N+1)}\| = \| |\tilde{E}^{(N+1)}| \| \\ &\leqq \frac{1}{(d_N)^{2n-4}} \sqrt{1.566} \cdot S(\tilde{A}^{(1)}) \sum_{k=2}^{N} |\tilde{\omega}_k| \\ &\leqq \frac{1.252}{(d_N)^{2n-4}} \cdot S(\tilde{A}^{(1)}) \cdot \left[(N-1) \sum_{k=2}^{N} \tilde{\omega}_k^2 \right]^{1/2} \\ &\leqq \frac{1.242}{(d_N)^{2n-4}} \cdot S(\tilde{A}^{(1)}) \cdot \left[N \sum_{k=1}^{N} \tilde{\omega}_k^2 \right]^{1/2}. \end{aligned} \tag{98}$$

---

[3] Here $(l(k), m(k))$ denotes pivot pair in the kth step, so this k should not be confused with the k that was fixed at the beginning of the proof.

Since $N \geqq 3$, from Lemma 10 it follows that

$$\frac{1}{(d_N)^{2n-4}} = \frac{1}{(1-0.077(2/3N))^{n-2}}$$

$$< 1 + \frac{12}{7} \cdot 0.077 \cdot \frac{2}{3N}(n-2)$$

$$\leqq 1 + \frac{12}{7} \cdot 0.077 \cdot \frac{4}{3} \cdot \frac{1}{n} \leqq 1.059.$$

Finally, inserting this inequality and assertion (ii) of Lemma 15 into relation (98), we obtain

$$S(\tilde{A}^{(N+1)}) \leqq 0.4 \cdot S(\tilde{A}^{(1)}) \sqrt{N} \cdot \frac{\tilde{\varepsilon}_1}{\delta}.$$

Applying a similar analysis to matrices $\tilde{B}^{(k)}$ yields

$$S(\tilde{B}^{(N+1)}) \leqq 0.4 \cdot S(\tilde{B}^{(1)}) \sqrt{N} \cdot \frac{\tilde{\varepsilon}_1}{\delta}.$$

From the last two inequalities and the definitions of $\tilde{\varepsilon}_{N+1}$ and $\tilde{\varepsilon}_1$ follows

$$\tilde{\varepsilon}_{N+1} \leqq 0.4 \cdot \sqrt{N} \cdot \frac{\tilde{\varepsilon}_1^2}{\delta},$$

and the theorem is proved. □

Note that in the proof of Theorem 16 it is not necessary to assume that the affiliation is preserved, i.e., that the pairs $[a_{ii}^{(k)}, b_{ii}^{(k)}]$ approximate the eigenvalues λ_i for $i = 1, \cdots, n$, $k = 1, \cdots, N$. However, for large enough k this fact follows from Theorem 17.

From Theorem 16 and the assumptions (A1) and (A2) it follows that

$$\tilde{\varepsilon}_{N+1} < \sqrt{N} \cdot \frac{1}{2N} \cdot \tilde{\varepsilon}_1 = \frac{1}{2\sqrt{N}} \cdot \tilde{\varepsilon}_1 < 0.3 \cdot \tilde{\varepsilon}_1. \tag{99}$$

Applying inductively the relation (99) we obtain

$$\tilde{\varepsilon}_{rN+1} \leqq (0.3)^r \cdot \tilde{\varepsilon}_1, \qquad r \geqq 1. \tag{100}$$

Therefore,

$$\lim_{r \to \infty} \tilde{\varepsilon}_{rN+1} = 0. \tag{101}$$

From the relation (101) and the assertion (i) of Lemma 15 we conclude that

$$\lim_{k \to \infty} \tilde{\varepsilon}_k = 0. \tag{102}$$

The relation (102) *and Theorem* 16 *imply the quadratic convergence of the Falk–Langemeyer method according to Definition* 7 *if the eigenvalues are simple and the pivot strategy is cyclic.*

Next we prove that under assumptions of Theorem 16 the sequences of matrices $(A^{(k)}, k \geqq 1)$ and $(B^{(k)}, k \geqq 1)$, generated by the Falk–Langemeyer method, converge towards diagonal matrices.

THEOREM 17. *Let the assumptions of Theorem* 16 *hold. Then*

$$\lim_{k\to\infty} A^{(k)} = D_A, \qquad \lim_{k\to\infty} B^{(k)} = D_B,$$

where D_A and D_B are diagonal matrices.

Proof. The relation (44) implies that

$$A^{(k)} = (D_k)^{-1}\tilde{A}^{(k)}(D_k)^{-1}, \qquad B^{(k)} = (D_k)^{-1}\tilde{B}^{(k)}(D_k)^{-1}, \tag{103}$$

where diagonal matrices D_k are defined with the relations (45) and (46). It is therefore sufficient to prove that the sequences $(\tilde{A}^{(k)}, k \geqq 1)$, $(\tilde{B}^{(k)}, k \geqq 1)$, and $((D_k)^{-1}, k \geqq 1)$ converge towards diagonal matrices. The relation (102) implies that the off-diagonal elements of matrices $\tilde{A}^{(k)}$ and $\tilde{B}^{(k)}$ tend to zero as $k \to \infty$. Therefore, it remains to prove that for $i = 1, \cdots, n$ the sequences $(\tilde{a}_{ii}^{(k)}, k \geqq 1)$ and $(\tilde{b}_{ii}^{(k)}, k \geqq 1)$ converge. The relation (18) and the assumption (A1) imply that for each $k \geqq 1$ there exists an ordering of the eigenvalues $\lambda_i = [s_i, c_i]$, $i = 1, \cdots, n$, such that

$$| c_i\tilde{a}_{ii}^{(k)} - s_i\tilde{b}_{ii}^{(k)} | \leqq \frac{\tilde{\varepsilon}_k^2}{2\delta}, \qquad i = 1, \cdots, n. \tag{104}$$

Let us consider unit vectors $[s_i, c_i]^T$ and $[\tilde{a}_{ii}^{(k)}, \tilde{b}_{ii}^{(k)}]^T$ in $\mathbf{R}^2$. The left-hand side of the inequality (104) is $|\sin \varphi_i^{(k)}|$ where $\varphi_i^{(k)}$ is the angle between these two vectors. The relations (102) and (104) imply

$$\lim_{k\to\infty} \sin \varphi_i^{(k)} = 0, \qquad i = 1, \cdots, n.$$

Hence, for each i the sequence of vectors $([\tilde{a}_{ii}^{(k)}, \tilde{b}_{ii}^{(k)}]^T, k \geqq 1)$ has only finite number of accumulation points in $\mathbf{R}^2$. Therefore, it suffices to show that for large enough k the changes in $\tilde{a}_{ii}^{(k)}$ and $\tilde{b}_{ii}^{(k)}$ are arbitrarily small. From the relation (102) and Lemma 11 we see that $\tilde{\alpha}_k \to 0$ and $\tilde{\beta}_k \to 0$ as $k \to \infty$. Therefore, the changes in $\tilde{a}_{ii}^{(k)}$ and $\tilde{b}_{ii}^{(k)}$ tend to zero as $k \to \infty$. This proves that for each $i \in \{1, \cdots, n\}$ limits $\lim_{k\to\infty} \tilde{a}_{ii}^{(k)}$ and $\lim_{k\to\infty} \tilde{b}_{ii}^{(k)}$ exist.

We shall now prove that $((D_k)^{-1}, k \geqq 1)$ is a convergent sequence. Looking at the definition of D_k (relation (45)) we see that it suffices to prove that for each $i \in \{1, \cdots, n\}$ the sequence $(d_i^{(k)}, k \geqq 1)$ converges to a nonzero number. From Proposition 9 we have

$$d_i^{(k)} = d_i^{(1)}\bar{d}_i^{(2)} \cdots \bar{d}_i^{(k)}, \qquad i = 1, \cdots, n, \qquad k \geqq 2.$$

From the definiteness of pairs $(A^{(1)}, B^{(1)})$ and $(\bar{A}^{(k)}, \bar{B}^{(k)})$ we conclude that $d_i^{(1)}$ and $\bar{d}_i^{(k)}$ are different from zero for all i and k. Hence it suffices to prove that the infinite product $\prod_{k=2}^{\infty} \bar{d}_i^{(k)}$ converges.[4] This product converges if and only if the product $\prod_{k=2}^{\infty} (\bar{d}_i^{(k)})^4$ converges. Therefore, it suffices to show that the latter product is absolutely convergent. From the relation (78) we see that for $i \in \{1, \cdots, n\}$ and $k \geqq 2$ we can write $(\bar{d}_i^{(k)})^4 = 1 + u_i^{(k)}$, so it suffices to show that the series $\sum_{k=2}^{\infty} u_i^{(k)}$ are absolutely convergent for all $i \in \{1, \cdots, n\}$.

The relation (83) of Lemma 12 implies that

$$(\bar{d}_i^{(k+1)})^4 \geqq 1 - 0.077 \cdot \frac{\tilde{\varepsilon}_k}{\delta}, \qquad 1 = 1, \cdots, n, \qquad k \geqq 1.$$

---

[4] Since all factors in the product are nonzero, the limit, if it exists, is also nonzero.

Looking for upper bound instead of lower bound in the relation (78) and making similar estimates as in the relation (83), we obtain

$$(\bar{d}_i^{(k+1)})^4 \leqq 1 + 0.077 \cdot \frac{\tilde{\varepsilon}_k}{\delta}, \qquad i = 1, \cdots, n, \qquad k \geqq 1.$$

Therefore,

$$|u_i^{(k+1)}| = |(\bar{d}_i^{(k+1)})^4 - 1| \leqq 0.077 \cdot \frac{\tilde{\varepsilon}_k}{\delta}, \qquad i = 1, \cdots, n, \qquad k \geqq 1.$$

Hence it suffices to show that the series $\sum_{k=1}^{\infty} \tilde{\varepsilon}_k$ converges. From the assertion (i) of Lemma 15 we have

$$\tilde{\varepsilon}_{rN+i} \leqq 1.3 \cdot \tilde{\varepsilon}_{rN+1}, \quad 1 \leqq i \leqq N, \quad r \geqq 1,$$

hence it suffices to prove the convergence of the sequence $\sum_{r=1}^{\infty} \tilde{\varepsilon}_{rN+1}$. From the relation (100) we see that the later series is majorized by the convergent series $\sum_{r=1}^{\infty} (0.3)^r \cdot \tilde{\varepsilon}_1$. This proves the absolute convergence of the series $\sum_{k=2}^{\infty} u_i^{(k)}$ for $i \in \{1, \cdots, n\}$ and therefore the convergence of the sequence $((D_k)^{-1}, k \geqq 1)$. □

Note that the global convergence (i.e., the convergence for all definite pairs (A, B)) of the Falk–Langemeyer method in the case of cyclic pivot strategies is not yet proved.

We end this section by showing that our asymptotic assumptions also imply ultimate quadratic reduction of ε_{rN+1}. Indeed, for $r \geqq 1$ the relation (103) implies

$$\varepsilon_{rN+1} \leqq (d_{\max}^{(rN+1)})^2 \cdot \tilde{\varepsilon}_{rN+1}, \qquad \tilde{\varepsilon}_{rN+1} \leqq \frac{1}{(d_{\min}^{(rN+1)})^2} \varepsilon_{rN+1},$$

where

$$d_{\max}^{(rN+1)} = \max \{d_1^{(rN+1)}, \cdots, d_n^{(rN+1)}\},$$
$$d_{\min}^{(rN+1)} = \min \{d_1^{(rN+1)}, \cdots, d_n^{(rN+1)}\}.$$

Theorem 16 implies

$$\varepsilon_{(r+1)N+1} \leqq (d_{\max}^{((r+1)N+1)})^2 \tilde{\varepsilon}_{(r+1)N+1} \leqq (d_{\max}^{((r+1)N+1)})^2 \frac{\sqrt{N}}{\delta} \tilde{\varepsilon}_{rN+1}^2$$
$$\leqq \left[\frac{d_{\max}^{((r+1)N+1)}}{(d_{\min}^{(rN+1)})^2} \right]^2 \frac{\sqrt{N}}{\delta} \varepsilon_{rN+1}^2 \leqq c \cdot \frac{\sqrt{N}}{\delta} \varepsilon_{rN+1}^2, \qquad r \geqq 1,$$

where c is an upper bound of the convergent sequence $([d_{\max}^{((r+1)N+1)}/(d_{\min}^{(rN+1)})^2]^2, r \geqq 1)$. In a similar way we can prove that the quadratic reduction of ε_{rN+1} ultimately implies the quadratic reduction of $\tilde{\varepsilon}_{rN+1}$.

The techniques described in this section can be used for studying asymptotic convergence properties of various different Jacobi-type algorithms.

5. Concluding remarks. In Algorithm 4 only (l, m)-restrictions of the pair $(A^{(k)}, B^{(k)})$ are used in each step. Therefore, parallel strategies are in fact cyclic (see [10]) and Theorems 16 and 17 hold for them as well.

In [13] it is proved that if the assumptions of Theorem 16 hold and the pivot strategy is serial, then

$$\tilde{\varepsilon}_{N+1} \leqq \frac{\tilde{\varepsilon}_1^2}{\delta}.$$

Modified method. If the problem (1) has multiple eigenvalues, the method can fail to be quadratically convergent. This failure occurs because when pairs $[a_{ll}^{(k)}, b_{ll}^{(k)}]$ and $[a_{mm}^{(k)}, b_{mm}^{(k)}]$ (here (l, m) is the pivot pair in the kth step) approximate the same eigenvalues, then parameters $\tilde{\alpha}_k$ and $\tilde{\beta}_k$ can be of order $O(1)$ and, therefore, some previously annihilated elements can become of order $O(\tilde{\varepsilon}_k)$ again. This situation is described in detail in [7] and [13]. Simple omitting of these critical steps does not yield to the quadratic convergence, even though the measure $\tilde{\tau}_k = \tau(\tilde{A}^{(k)}, \tilde{B}^{(k)})$, $k \geqq 1$, from Corollary 2 tends to zero. The relation (16) does not imply that the off-diagonal elements of diagonal blocks tend to zero together with $\tilde{\tau}_k$, but merely that the diagonal blocks become more and more proportional. Therefore, $\tilde{\varepsilon}_k$ does not have to tend to zero at all and the convergence of $\tilde{\tau}_k$ can considerably slow down. If we modify the method so that in such cases we use triangular transformation matrices similar to the matrix from step (5)(a) of Algorithm 4, the quadratic convergence persists.

Modification of the Falk–Langemeyer method and the proof of quadratic convergence of the modified method will be topics of our subsequent paper.

Numerical results. Our test program is written in FORTRAN in double precision. Test pairs were generated in the manner that $A = G^T D_A G$ and $B = G^T D_B G$, where diagonal matrices D_A and D_B are being read and G is random. For elements of matrix G only numbers which are sums of the powers of 2 were used, so the test pairs were stored as accurately as possible.

The iterative process is terminated when, after some cycle r, inequality

$$\varepsilon_{rN+1} < eps \cdot \sqrt{\|A\|^2 + \|B\|^2} \cdot 2N$$

is fulfilled, where *eps* is machine precision. After the end of the process, the maximal error of the residual

$$\max_{1 \leqq i \leqq n} \left\{ \frac{\|b_i' A f_i' - a_i' B f_i'\|_{\max}}{\sqrt{(a_i')^2 + (b_i')^2}\sqrt{\|A f_i'\|^2 + \|B f_i'\|^2}} \right\},$$

is calculated. Here $[a_i', b_i']$ are the calculated eigenvalues of the pair (A, B) and f_i' are the corresponding eigenvectors. Also the maximal absolute values of the off-diagonal elements of matrices $(F')^T A F'$ and $(F')^T B F'$ are calculated. Those three quantities were usually of order of stopping criterion. Infinite eigenvalues were represented with numbers of order of magnitude $O(1/\text{machine precision})$.

We observed the convergence of both measures ε_k and $\tilde{\varepsilon}_k$. Observations confirmed all theoretical results. For starting pairs that were not almost diagonal, convergence was in the beginning linear and several cycles were needed before quadratic convergence started. The asymptotic assumption (A2) appears to be very adequate because in almost all cases quadratic convergence started after it was fulfilled. Algorithm behaved very regularly in the sense that the condition $\mathfrak{I}^{(k)} \geqq 0$, $k \geqq 1$ (see assertion (i) of Proposition 3) was always fulfilled for definite starting pairs. This condition was fulfilled even in some cases when the starting pair was semidefinite, or slightly indefinite.

Average number of cycles for smaller matrices ($n \leqq 15$) was around 10 and for larger matrices ($n \leqq 100$) around 15. Last cycles were usually empty, i.e., not all N steps were executed. For orientation, the approximate duration of the process is five minutes for $n = 40$ and one and a half hours for $n = 100$ on IBM PC/AT with a coprocessor, and about 30 times shorter on IBM 4371.

In the presence of very close eigenvalues several additional cycles were usually needed because the quadratic convergence was delayed. The existence of additional cycles does not disagree with theoretical results since the quantity δ from the asymptotic assumption (A2) is in this case very small.

We observed that the results are generally better if increasing or decreasing order of numbers defined with diagonal pairs $[a_{ii}^{(k)}, b_{ii}^{(k)}]$ is preserved by interchanging pivot rows and columns if necessary. However, interchanging must be stopped after the asymptotic assumption (A2) is fulfilled. Otherwise some off-diagonal element which was not yet annihilated can "run away" from annihilation and therefore terminate quadratic convergence.

Example. We give an example of the pair of order 10 generated in the previously described manner. Elements of the matrices D_A and D_B are

$$-2, 1, 10, 0, -0.001, 10, 1, 5, 5, 4$$

and

$$-1, 0.1, -1, -100, -100, 0, -1, 0.1, 1, 1,$$

respectively, so the exact eigenvalues of the problem are

$$2, 10, -10, 0, 0.00001, \infty, -1, 50, 5, 4.$$

Elements of the matrix G are uniformly distributed integers from the interval $[-10, 10]$. Note that both matrices A and B are indefinite, while the pair (A, B) itself is definite (for example, $A - 3B > 0$). In order to increase the stability of the computation, the process started from the normalized pair $(\tilde{A}, \tilde{B})$.

Only upper triangles of the matrices A and B are displayed. Each row begins with the diagonal element. Asymptotic convergence is described as follows: in column CYC is the number of cycle; in column ROT is the number of rotations performed in the cycle; columns SUMA, SUMB, SUM and SUMT display values of $S(A^{(k)})$, $S(B^{(k)})$, ε_k and $\tilde{\varepsilon}_k$ after the cycle, respectively.

```
ORDER OF MATRICES N = 10
COLUMN CYCLIC PIVOT STRATEGY
STOPPING CRITERION: SUM(K) < .49D-13

MATRIX A
ROW
  1     .21350D+04     .41900D+03     .11600D+03    -.11430D+04    -.10490D+04
        .44002D+03    -.13750D+04     .20027D+02     .51903D+03    -.60802D+03
  2     .14310D+04    -.32700D+03    -.34200D+03    -.26100D+03    -.10390D+04
       -.29600D+03    -.43200D+03     .50000D+03    -.13100D+03
  3     .18320D+04     .28000D+03     .93300D+03     .64100D+03    -.91000D+03
       -.38500D+03    -.74500D+03    -.58200D+03
  4     .11860D+04     .85799D+03     .34099D+03     .56100D+03    -.40001D+03
       -.49101D+03     .57401D+03
  5     .99295D+03     .68695D+03     .22402D+03    -.53006D+03    -.69606D+03
       -.90965D+02
  6     .13360D+04    -.57298D+03    -.43106D+03    -.97606D+03     .43035D+02
  7     .13470D+04     .29703D+03     .10027D+02     .67399D+03
  8     .87292D+03     .32992D+03    -.49550D+01
  9     .88792D+03     .40105D+03
 10     .82798D+03
```

```
MATRIX B
ROW
  1   -.72420D+04   -.81550D+04    .40130D+04   -.20630D+04   -.39800D+03
       .31460D+04   -.73170D+04    .83650D+04    .27080D+04   -.78930D+04
  2   -.99425D+04    .50853D+04   -.28814D+04   -.28007D+04    .93320D+03
      -.79655D+04    .68774D+04   -.11530D+03   -.79841D+04
  3   -.26020D+04    .14878D+04    .14297D+04   -.63790D+03    .39712D+04
      -.34409D+04   -.16500D+02    .39654D+04
  4   -.10848D+04   -.15558D+04   -.55260D+03   -.20896D+04    .11306D+04
      -.83980D+03   -.18888D+04
  5   -.59091D+04   -.46836D+04   -.23130D+03   -.41948D+04   -.61837D+04
       .10755D+04
  6   -.50607D+04    .28013D+04   -.69463D+04   -.62826D+04    .43479D+04
  7   -.72920D+04    .82531D+04    .26803D+04   -.78484D+04
  8   -.12938D+05   -.81476D+04    .10050D+05
  9   -.82005D+04    .46091D+04
 10   -.88056D+04

ASYMPTOTIC CONVERGENCE
CYC        ROT          SUMA          SUMB          SUM          SUMT

  1         45          .60D+00       .35D+01       .35D+01       .18D+01
  2         45          .72D+00       .32D+00       .79D+00       .12D+01
  3         45          .57D+00       .28D+00       .64D+00       .72D+00
  4         45          .31D+00       .21D+00       .38D+00       .24D+00
  5         45          .18D-01       .23D-01       .30D-01       .25D-01
  6         45          .39D-02       .13D-01       .13D-01       .93D-02
  7         45          .56D-04       .15D-03       .16D-03       .92D-04
  8         44          .18D-09       .61D-09       .63D-09       .19D-09
  9         29          .26D-20       .19D-20       .32D-20       .93D-20
TOTAL NO. OF ROTATIONS  388    TIME(sec)  5.68

CALCULATED EIGENVALUES
 I          A(I,I)                B(I)                    D(I)
 1       .36774301D+01         .91935754D+00          .40000000000001D+01
 2       .29605919D+01         .59211837D-01          .50000000000001D+02
 3       .24951699D+00         .24951699D-01          .10000000000015D+02
 4       .62005903D+00         .12401181D+00          .50000000000011D+01
 5      -.16044487D+00        -.80222433D-01          .19999999999999D+01
 6      -.94464581D-04        -.94464581D+01          .10000000000004D-04
 7      -.43510239D-16        -.23229057D+01          .18730953468577D-16
 8       .62365226D-01        -.62365226D-01         -.99999999999990D+00
 9       .42815492D+01        -.25722206D-13         -.16645341957511D+15
10       .32339485D+01        -.32339485D+00         -.99999999999999D+01

MAXIMAL(relative) ERROR = .15D-13 FOR I = 3
MAXIMAL OFF-DIAGONAL ELEMENTS:
Ft A F = .35D-14    Ft B F = .30D-13
```

Acknowledgments. We would like to thank Professor K. Veselić from Fernuniversität Hagen for his helpful suggestions. We also thank both reviewers for their comments, which helped us clarify some important parts of the paper.

REFERENCES

[1] J.-P. CHARLIER AND P. VAN DOOREN, *A Jacobi-like algorithm for computing the generalized Schur form of a regular pencil*, Tech. Report, Philips Research Laboratory, Brussels, Belgium, 1988.

[2] S. FALK AND P. LANGEMEYER, *Das Jacobische Rotations-Verfahren für reelsymmetrische Matrizen-Paare* I, II, Elektron. Datenverarbeitung, (1960), pp. 30–43.
[3] G. GOSE, *Das Jacobi Verfahren für* $Ax = \lambda Bx$, Z. Angew. Math. Mech., 59 (1979), pp. 93–101.
[4] V. HARI, *On cyclic Jacobi methods for the positive definite generalized eigenvalue problem*, Ph.D. thesis, Department of Mathematics and Computer Science, Fernuniversität Hagen, Hagen, Federal Republic of Germany, 1984.
[5] ———, *On the convergence of cyclic Jacobi-like processes*, Linear Algebra Appl., 81 (1986), pp. 105–127.
[6] ———, *On the quadratic convergence of Jacobi algorithms*, Radovi Mathematički, 2 (1986), pp. 127–146.
[7] ———, *On pairs of almost diagonal matrices*, Linear Algebra Appl., to appear.
[8] P. HENRICI AND K. ZIMMERMANN, *An estimate for the norms of certain cyclic Jacobi operators*, Linear Algebra Appl., 1 (1968), pp. 489–501.
[9] H. P. M. VAN KEMPEN, *On the quadratic convergence of the serial cyclic Jacobi method*, Numer. Math., 9 (1966), pp. 19–22.
[10] F. LUK AND H. PARK, *On the equivalence and convergence of parallel Jacobi algorithms*, in Proc. SPIE Conference on Advanced Algorithms and Architectures for Signal Processing II, San Diego, 1987.
[11] B. N. PARLETT, *Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs, NJ, 1980.
[12] A. H. SAMEH, *On Jacobi and Jacobi-like algorithms for a parallel computer*, Math. Comp., 25 (1971), pp. 579–590.
[13] I. SLAPNIČAR, *Quadratic convergence of the Falk–Langemeyer method*, Master's thesis, University of Zagreb, Zagreb, Yugoslavia, 1988. (In Croatian.)
[14] G. W. STEWART, *Perturbation bounds for the definite generalized eigenvalue problem*, Linear Algebra Appl., 23 (1960), pp. 69–85.
[15] K. VESELIĆ, *An eigenreduction algorithm for definite matrix pairs and its applications to overdamped linear systems*, Fernuniversität Hagen, Hagen, Federal Republic of Germany, preprint, 1989.
[16] J. H. WILKINSON, *Note on the quadratic convergence of the cyclic Jacobi processes*, Numer. Math., 4 (1962), pp. 296–300.
[17] ———, *Almost diagonal matrices with multiple or close eigenvalues*, Linear Algebra Appl., 1 (1968), pp. 1–12.
[18] ———, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.
[19] K. ZIMMERMANN, *On the convergence of a Jacobi process for ordinary and generalized eigenvalue problems*, Ph.D. No. 4305, Eidgenossische Technische Hochschule, Zürich, Switzerland, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 1, pp. 115–127, January 1991

009

LEAST SQUARES APPROXIMATION BY REAL NORMAL MATRICES WITH SPECIFIED SPECTRUM*

MOODY T. CHU†

Abstract. The problem of best approximating a given real matrix in the Frobenius norm by real, normal matrices subject to a prescribed spectrum is considered. The approach is based on using the projected gradient method. The projected gradient of the objective function on the manifold of constraints can be formulated explicitly. This gives rise to a descent flow that can be followed numerically. The explicit form also facilitates the computation of the second-order optimality condition from which some interesting properties of the stationary points are related to the well-known Wielandt–Hoffman theorem.

Key words. least squares, projected gradient, normal matrix, spectral constraint

AMS(MOS) subject classifications. 65F15, 49D10

1. Introduction. A matrix $A \in C^{n\times n}$ is normal if and only if $A^*A = AA^*$. Normality, as it includes the Hermitian, unitary, and skew-Hermitian matrices, defines a rather general and important class of matrices. In [7], 70 equivalent conditions are listed to characterize a normal matrix. This again reflects that normality may arise in many different ways.

One interesting question that has received considerable attention is the determination of a closest normal matrix to a given square complex matrix. This problem has only recently been completely solved (in the Frobenius norm) in [4], and independently in [12]. It turns out that finding a nearest normal matrix is equivalent to finding a unitary similarity transformation which makes the sum of squares of moduli of the diagonal elements as large as possible [8]. The Jacobi algorithm, therefore, may be derived from this perspective to solve the nearest to normality problem.

In this paper we assume the following situation happens. Experimental data has been collected in the matrix A which, by some prior knowledge, should be a normal matrix with known spectrum. Generally, due to measurement errors, A will not satisfy these requirements. Since A still contains some useful information, we would like to retrieve its least squares approximation that satisfies these requirements.

In practice, we may well be interested in real matrices. It is well known [5, p. 284] that a real normal matrix is always orthogonally similar to a real quasi-diagonal matrix

$$\text{(1)} \qquad \operatorname{diag}\left\{\begin{bmatrix} \lambda_1 & \nu_1 \\ -\nu_1 & \lambda_1 \end{bmatrix}, \cdots, \begin{bmatrix} \lambda_q & \nu_q \\ -\nu_q & \lambda_q \end{bmatrix}, \lambda_{2q+1}, \cdots, \lambda_n\right\}$$

where λ_k, ν_k are real numbers and $\nu_k \neq 0$ $(k = 1, 2, \cdots, q)$. Therefore, we consider the following problem in this paper.

Problem A. Given a matrix $A \in R^{n\times n}$ and a set of eigenvalues $\{\lambda_1 \pm i\nu_1, \cdots, \lambda_q \pm i\nu_q, \lambda_{2q+1}, \cdots, \lambda_n\}$ where λ_k, ν_k are real numbers and $\nu_k \neq 0$ $(k = 1, 2, \cdots, q)$, find an orthogonal matrix Q that minimizes the function

$$\text{(2)} \qquad F(Q) := \tfrac{1}{2}\|Q^T\Lambda Q - A\|^2$$

* Received by the editors March 27, 1989; accepted for publication (in revised form) January 15, 1990. This work was supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U.S. Department of Energy, under contract W-31-109-Eng-38, while the author was spending his sabbatical leave at Argonne National Laboratory, Argonne, Illinois.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695–8205 (chu@matmtc.ncsu.edu).

where Λ is the quasi-diagonal matrix given by (1) and $\|\cdot\|$ means the Frobenius matrix norm.

A special case of Problem A has been considered in [3]. There it is shown that when A is symmetric and when Λ is diagonal with distinct elements arranged in descending order, the columns of the optimal Q^T should be the normalized eigenvectors of A corresponding to eigenvalues arranged in the descending order. In this paper we study the extension to more general classes of matrices.

Our idea is closely related to the setting in [1]. Our approach is parallel to that in [3]. Without using the Lagrangian function, we first formulate explicitly the projection of the gradient of the objective function F onto the feasible set $O(n) := \{Q \in R^{n\times n} | Q^TQ = I\}$. This formula gives rise to the construction of a descent flow that can be followed numerically. We then derive the so-called projected Hessian on the tangent space of $O(n)$. Wherever possible, we classify the stationary points from the second-order condition. Finally, we discuss the connection between our results and the well-known Wielandt–Hoffman theorem [9].

2. Preliminaries. Let $\langle A, B\rangle$ denote the Frobenius inner product of two matrices $A, B \in R^{n\times n}$:

$$\langle A, B\rangle := \text{trace}\,(AB^T) = \sum_{i,j} a_{ij}b_{ij}. \tag{3}$$

We first consider the function F in (2) to be defined everywhere in $R^{n\times n}$. For $Z, H \in R^{n\times n}$, the Fréchet derivative of F at Z acting on H is calculated to be

$$\begin{aligned} F'(Z)H &= \langle Z^T\Lambda Z - A, H^T\Lambda Z + Z^T\Lambda H\rangle \\ &= \langle (\Lambda Z)(Z^T\Lambda Z - A)^T, H\rangle + \langle (\Lambda^T Z)(Z^T\Lambda Z - A), H\rangle. \end{aligned} \tag{4}$$

In the second equation above we have used the adjoint property

$$\langle A, BC\rangle = \langle B^TA, C\rangle = \langle AC^T, B\rangle$$

to rearrange terms. With respect to the Frobenius inner product, the equation (4) suggests that the gradient of F at a general matrix $Z \in R^{n\times n}$ may be interpreted as the matrix

$$\nabla F(Z) := (\Lambda Z)(Z^T\Lambda Z - A)^T + (\Lambda^T Z)(Z^T\Lambda Z - A). \tag{5}$$

Let $S(n)$ denote the subspace of all symmetric matrices in $R^{n\times n}$. It is easy to see that the tangent space $T_QO(n)$ of the feasible set $O(n)$ is given by [3]:

$$T_QO(n) := QS(n)^\perp \tag{6}$$

where $S(n)^\perp$, the orthogonal complement of $S(n)$ in $R^{n\times n}$, is precisely the subspace of all skew-symmetric matrices. It is also easy to see that the orthogonal complement of $T_QO(n)$ is the subspace

$$N_QO(n) := QS(n). \tag{7}$$

Therefore, an orthogonal matrix Q is a stationary point of Problem A only if

$$(\Lambda Q)(Q^T\Lambda Q - A)^T + (\Lambda^T Q)(Q^T\Lambda Q - A) \in QS(n). \tag{8}$$

For convenience, we define in the sequel

$$X := Q^T\Lambda Q. \tag{9}$$

Then (8) is equivalent to

$$X(X^T - A^T) + X^T(X - A) \in S(n), \tag{10}$$

or

$$XA^T+X^TA=AX^T+A^TX. \tag{11}$$

Let $[A, B] := AB - BA$ denote the Lie bracket. Lemma 2.1 follows.

LEMMA 2.1. *A necessary condition for $Q \in O(n)$ to be a stationary point for Problem* A *is that the matrix $[X, A^T]$ with X defined by* (9) *is symmetric.*

We remark that if A is symmetric and Λ is diagonal, then X is symmetric and $[X, A^T]$ is skew symmetric. In this case, we conclude, from Lemma 2.1, that at a stationary point the matrix X must commute with A. This is one of the results discussed in [3].

The projected gradient of F on the manifold $O(n)$ can be calculated without any difficulty. Mainly this is due to the understanding that for any fixed $Q \in O(n)$,

$$R^{n\times n} = T_QO(n)\oplus N_QO(n) = QS(n)^{\perp}\oplus QS(n). \tag{12}$$

Any matrix $Z \in R^{n\times n}$ has a unique orthogonal splitting

$$Z = Q\{\tfrac{1}{2}(Q^TZ - Z^TQ)\} + Q\{\tfrac{1}{2}(Q^TZ + Z^TQ)\} \tag{13}$$

as the sum of elements from $T_QO(n)$ and $N_QO(n)$. Accordingly, the projection $g(Q)$ of $\nabla F(Q)$ onto the tangent space $T_QO(n)$ can be calculated explictly as follows:

$$\begin{aligned} g(Q) &= \frac{Q}{2}\{Q^T\nabla F(Q) - \nabla F(Q)^TQ\} \\ &= \frac{Q}{2}\{Q^T\{(\Lambda Q)(Q^T\Lambda Q - A)^T + (\Lambda^TQ)(Q^T\Lambda Q - A)\} \\ &\qquad -\{(\Lambda Q)(Q^T\Lambda Q - A)^T + (\Lambda^TQ)(Q^T\Lambda Q - A)\}^TQ\} \\ &= -\frac{Q}{2}\{[Q^T\Lambda Q, A^T] - [Q^T\Lambda Q, A^T]^T\}. \end{aligned} \tag{14}$$

It is clear that the vector field

$$\frac{dQ}{dt} := -g(Q) = \frac{Q}{2}\{[Q^T\Lambda Q, A^T] - [Q^T\Lambda Q, A^T]^T\} \tag{15}$$

defines a steepest descent flow $Q(t)$ on the manifold $O(n)$ for the objective function F in (2). Upon substitution, the corresponding $X(t)$ is governed by the ordinary differential equation

$$\begin{aligned} \frac{dX}{dt} &:= \frac{dQ^T}{dt}\Lambda Q + Q^T\Lambda\frac{dQ}{dt} \\ &= \left[X, \frac{[X, A^T] - [X, A^T]^T}{2}\right]. \end{aligned} \tag{16}$$

Starting with an appropriate initial value, say $X(0) = \Lambda$, the positive orbit of (16) marches to a limit point which is a (local) least squares normal matrix approximation to A.

We remark again that if A is symmetric and Λ is diagonal, then the flow (16) is reduced to

$$\frac{dX}{dt} = [X, [X, A]], \tag{17}$$

which is analyzed in [3].

It is worth mentioning that the second term in the bracket of (16) is skew symmetric. Therefore, the solution flow $X(t)$ of (16) naturally is isospectral [2] to the initial value $X(0)$. In particular, we have $\|X(t)X(t)^T - X(t)^TX(t)\| = \|X(0)X(0)^T - X(0)^TX(0)\|$ for all t. Thus, apart from numerical errors induced when solving the differential equation (16) on computers, the deviation of normality of $X(t)$ will remain the same as that of $X(0)$.

The function g in (14) is defined for orthogonal matrices only. We now derive an explicit formula for the projected Hessian of the objective function F without utilizing the Lagrange multiplier. Readers are referred to [3] for an explanation of why this technique works. Obviously we may extend g smoothly to cover the entire space $R^{n\times n}$ simply by defining

$$G(Z) := \frac{Z}{2}\{[Z^T\Lambda Z, A^T]^T - [Z^T\Lambda Z, A^T]\}. \tag{18}$$

The Fréchet derivative of G can easily be calculated. In particular, at any stationary point Q of Problem A and for every tangent vector QK where $K \in S(n)^\perp$, it holds that

$$\begin{aligned}\langle G'(Q)QK, QK\rangle &= \left\langle \frac{[[X,K],A^T]^T - [[X,K],A^T]}{2}, K\right\rangle \\ &= -\langle [[X,K],A^T], K\rangle \\ &= \langle [X,K],[A,K]\rangle.\end{aligned} \tag{19}$$

It can be proved that formula (19) is precisely the evaluation of the projected Hessian of the Lagrangian function of Problem A [6, p. 80]. Thus a necessary condition (and a sufficient condition if the strict inequality holds) for a stationary point Q to be a local minimum is that

$$\langle [X,K],[A,K]\rangle \geqq 0 \quad \text{for every } K\in S(n)^\perp. \tag{20}$$

3. Application I—real eigenvalues. We now apply the first-order condition (11) and the second-order condition (20) to classify the stationary points for Problem A. It will prove useful if we define

$$E := QAQ^T. \tag{21}$$

We observe that the first-order condition (11) and the second-order condition (20) are equivalent to

$$\Lambda E^T + \Lambda^T E = E\Lambda^T + E^T\Lambda \tag{22}$$

and

$$\langle [\Lambda,K],[E,K]\rangle \geqq 0 \quad \text{for every } K\in S(n)^\perp, \tag{23}$$

respectively.

In this section we first consider the case when Λ has only real eigenvalues. It follows that the matrix $X = Q^T\Lambda Q$ must be symmetric for any $Q \in O(n)$. For any general matrix $A \in R^{n\times n}$, let

$$A_S := \tfrac{1}{2}(A + A^T) \tag{24}$$

and

$$A_K := \tfrac{1}{2}(A - A^T) \tag{25}$$

denote the symmetric and skew-symmetric parts of A, respectively. We observe that

$$\|X-A\|^2 = \|X-A_S\|^2 + \|A_K\|^2. \tag{26}$$

Since the second term in (26) is fixed once A is given, a least squares approximation to A amounts to a least square approximation to A_S. Therefore, it suffices to consider the case when A is symmetric.

Suppose A is symmetric. We shall arrange eigenvalues of A in the natural ordering

$$\mu_1 \geqq \mu_2 \geqq \cdots \geqq \mu_n. \tag{27}$$

We further divide our discussions according to whether or not Λ has simple eigenvalues.

Case 1 (Λ has only distinct eigenvalues). For clarity, we shall assume the diagonal elements of Λ are arranged in the descending order

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n. \tag{28}$$

The following theorem completely classifies all the stationary points.

THEOREM 3.1. *Suppose A is symmetric and has eigenvalues arranged as in* (27). *Suppose Λ is diagonal and has elements arranged as in* (28). *Then the stationary points of Problem* A *are classified as follows*:

1. *An orthogonal matrix Q is a stationary point of F only if columns $q_1, \cdots, q_n$ of Q^T are orthonormal eigenvectors of A.*
2. *A stationary point Q is a local minimizer (or, a local maximizer) of F only if columns $q_1, \cdots, q_n$ of Q^T correspond with eigenvalues $\mu_1, \cdots, \mu_n$ (or, the reverse order), respectively. All other stationary points are saddle points.*
3. *Any least squares approximation X to A is of the form*

$$X = \lambda_1 q_1 q_1^T + \cdots + \lambda_n q_n q_n^T. \tag{29}$$

   The least squares approximation X is unique if A itself has distinct eigenvalues.
4. *The minimal value of F is equal to $\frac{1}{2}\sum_{i=1}^{n}(\lambda_i - \mu_i)^2$.*
5. *Local extreme points are also global extreme points.*

Proof. The proof of this theorem can be found in [3]. The main point is that the simplicity of eigenvalues of Λ and the condition (22) require that E be a diagonal matrix [11, p. 416]. Also, part 5 follows from the fact that all extreme points yield the same function value as specified in part 4.

Case 2 (Λ has multiple eigenvalues). When multiple eigenvalues occur, the analysis becomes more complicated because the matrix E is not necessarily a diagonal matrix. For demonstration purpose, we shall only consider the special case when all eigenvalues, except the one which has multiplicity two, of Λ are simple.

We shall assume the diagonal elements of Λ are arranged in the ordering

$$\lambda_1 > \cdots > \lambda_k = \lambda_{k+1} > \cdots > \lambda_n \tag{30}$$

with $1 \leqq k \leqq n-1$. Then the first-order condition (22) implies that at a stationary point E must be a quasi-diagonal matrix of the form [11]

$$E = \operatorname{diag}\left\{e_1, \cdots, e_{k-1}, \begin{bmatrix} e_k & e_* \\ e_* & e_{k+1} \end{bmatrix}, e_{k+2}, \cdots, e_n\right\}. \tag{31}$$

It follows from (21) that $e_1, \cdots, e_{k-1}, e_{k+2}, \cdots, e_n$ must be $n-2$ eigenvalues of A (note that we are assuming that A is symmetric), and that columns $q_1, \cdots, q_{k-1}$,

$q_{k+2}, \cdots, q_n$ of the matrix Q^T must be the corresponding orthonormal eigenvectors. Obviously, the 2×2 matrix

$$R := \begin{bmatrix} e_k & e_* \\ e_* & e_{k+1} \end{bmatrix} \tag{32}$$

determines the remaining two eigenvalues, denoted by μ_s and μ_t, of A. The columns q_k and q_{k+1} are two orthonormal vectors in the spaced spanned by eigenvectors of μ_s and μ_t.

It is not difficult to see that

(33)

$$\begin{aligned} \langle [\Lambda, K], [E, K] \rangle = 2 \sum_{\substack{i<j \\ i \neq k, k+1 \\ j \neq k, k+1}} (\lambda_i - \lambda_j)(e_i - e_j) k_{ij}^2 \\ + 2 \sum_{k+1<j} (\lambda_k - \lambda_j)\{(e_k - e_j)k_{kj}^2 + 2e_* k_{kj} k_{k+1,j} + (e_{k+1} - e_j)k_{k+1,j}^2\} \\ + 2 \sum_{i<k} (\lambda_i - \lambda_k)\{(e_i - e_k)k_{ik}^2 - 2e_* k_{ik} k_{i,k+1} + (e_i - e_{k+1})k_{i,k+1}^2\}. \end{aligned}$$

We note that the three summations in (33) are mutually exclusive. Therefore, $\langle [\Lambda, K], [E, K] \rangle \geqq 0$ for every $K \in S(n)^{\perp}$ if and only if every single term in (33) is nonnegative. Because of the specified ordering of the eigenvalues λ_i, we conclude that for a stationary point Q to be a local minimizer, it is necessary that

$$e_1 \geqq e_2 \geqq \cdots \geqq e_{k-1} \geqq e_{k+2} \geqq \cdots \geqq e_n, \tag{34}$$

and that the matrices

$$\begin{aligned} \begin{bmatrix} e_i - e_k & -e_* \\ -e_* & e_i - e_{k+1} \end{bmatrix} &= e_i I - R \quad \text{for every } i < k, \\ \begin{bmatrix} e_k - e_j & e_* \\ e_* & e_{k+1} - e_j \end{bmatrix} &= R - e_j I \quad \text{for every } k+1 < j \end{aligned} \tag{35}$$

be positive semidefinite. From the above, we have proved the following theorem.

THEOREM 3.2. *Suppose A is symmetric and has eigenvalues arranged as in* (27). *Suppose Λ is diagonal and has elements arranged as in* (30). *Then the stationary points of Problem* A *are classified as follows*:

1. *An orthogonal matrix Q is a stationary point of F only if columns $q_1, \cdots, q_{k-1}, q_{k+2}, \cdots, q_n$ of the matrix Q^T are $n-2$ orthonormal eigenvectors of A, and q_k, q_{k+1} are linear combinations of the remaining two orthonormal eigenvectors.*
2. *A stationary point Q is a local minimizer of F only if columns $q_1, \cdots, q_{k-1}$ of Q^T correspond with eigenvalues $\mu_1, \cdots, \mu_{k-1}$, and $q_{k+2}, \cdots, q_n$ correspond with eigenvalues $\mu_{k+2}, \cdots, \mu_n$, and q_k, q_{k+1} are linear combinations of eigenvectors corresponding with eigenvalues μ_k, μ_{k+1}. Similarly, a stationary point Q is a local maximizer of F only if the above correspondence is in the reverse order. All other stationary points are saddle points.*
3. *Any least squares approximation X to A is of the form*

$$X = \lambda_1 q_1 q_1^T + \cdots + \lambda_k (q_k q_k^T + q_{k+1} q_{k+1}^T) + \cdots + \lambda_n q_n q_n^T. \tag{36}$$

The choice of q_k and q_{k+1} is immaterial. The least squares approximation is unique if the first $k-1$ and the last $n-k-1$ eigenvalues of A are distinct.

4. *The minimal value of F is equal to $\frac{1}{2}\sum_{i=1}^{n}(\lambda_i-\mu_i)^2$.*
5. *Local extreme points are also global extreme points.*

We remark that the proof for the above theorem can be generalized to cover other cases of multiple eigenvalues. The details are left to the readers.

4. Application II—complex eigenvalues. One of the difficulties associated with this case is that there is no clear way to order the eigenvalues. Even so, we have made some interesting observations.

Case 3 (A is a 2×2 matrix). The simple 2×2 case offers considerable insights into the understanding of higher-dimensional problems. Let

$$\Lambda=\begin{bmatrix}\lambda & \nu\\ -\nu & \lambda\end{bmatrix}. \tag{37}$$

For any $E\in R^{2\times 2}$, it is easy to see that the matrix $\Lambda E^T+\Lambda^T E$ is always symmetric. This is to say that any $Q\in O(2)$ is a stationary point. Indeed, we find that

$$X:=Q^T\Lambda Q\equiv\begin{cases}\Lambda & \text{if } \det Q=1,\\ \Lambda^T & \text{if } \det Q=-1.\end{cases} \tag{38}$$

So the least squares approximation problem is trivial. The objective function value is given by

$$F(Q)\equiv\tfrac{1}{2}((a_{11}-\lambda)^2+(a_{22}-\lambda)^2+(a_{12}\mp\nu)^2+(a_{21}\pm\nu)^2) \tag{39}$$

depending upon $\det Q=\pm 1$, respectively. It is readily seen from (39) that the signs of ν and $a_{12}-a_{21}$ determine which one of Λ or Λ^T better approximates A.

Case 4 (A is a symmetric matrix). Again, for demonstration purpose, we shall consider only the case when Λ is of the form

$$\Lambda=\operatorname{diag}\left\{\lambda_1,\cdots,\begin{bmatrix}\lambda_k & \nu_*\\ -\nu_* & \lambda_k\end{bmatrix},\cdots,\lambda_n\right\} \tag{40}$$

where

$$\lambda_1>\lambda_2>\cdots>\lambda_n \tag{41}$$

and $\nu_*>0$. Since A is symmetric, so is E. We write $\Lambda=\Lambda_S+\Lambda_K$ as the sum of its own symmetric and skew-symmetric parts. The first-order condition (22) requires

$$(\Lambda^T+\Lambda)E=E(\Lambda^T+\Lambda). \tag{42}$$

Because $\Lambda^T+\Lambda=2\Lambda_S$ is diagonal, it follows that E must be a quasi-diagonal matrix of the form (31). Furthermore, we know that

$$\langle[\Lambda,K],[E,K]\rangle=\langle[\Lambda_S,K],[E,K]\rangle \tag{43}$$

since $[\Lambda_K, K]$ is skew symmetric and $[E, K]$ is symmetric. We state Theorem 4.1.

THEOREM 4.1. *Suppose A is symmetric and has eigenvalues arranged as in* (27). *Suppose Λ is quasi-diagonal and has elements arranged as in* (40) *and* (41). *Then the stationary points of Problem* A *are classified as follows*:

1. *An orthogonal matrix Q is a stationary point of F only if columns $q_1,\cdots,q_{k-1},q_{k+2},\cdots,q_n$ of the matrix Q^T are $n-2$ orthonormal eigenvectors of A,*

and q_k, q_{k+1} are linear combinations of the remaining two orthonormal eigenvectors.

2. *A stationary point Q is a local minimizer of F only if columns $q_1, \cdots, q_{k-1}$ of Q^T correspond with eigenvalues $\mu_1, \cdots, \mu_{k-1}$, and $q_{k+2}, \cdots, q_n$ correspond with eigenvalues $\mu_{k+2}, \cdots, \mu_n$, and q_k, q_{k+1} are linear combinations of eigenvectors corresponding with eigenvalues μ_k, μ_{k+1}. Similarly, a stationary point Q is a local maximizer of F only if the correspondence above is in the reverse order. All other stationary points are saddle points.*
3. *Any least squares approximation X to A is of the form*

$$X = \lambda_1 q_1 q_1^T + \cdots + \lambda_k (q_k q_k^T + q_{k+1} q_{k+1}^T) + \nu_* (q_k q_{k+1}^T - q_{k+1} q_k^T) + \cdots + \lambda_n q_n q_n^T. \tag{44}$$

The choice of q_k and q_{k+1} is immaterial. The least squares approximation is unique if the first $k-1$ and the last $n-k-1$ eigenvalues of A are distinct.

4. *The minimal value of F is equal to $\nu_*^2 + \frac{1}{2}\sum_{i=1}^n (\lambda_i - \mu_i)^2$.*
5. *Local extreme points are also global extreme points.*

Proof. The analysis of stationary points for this case is essentially identical to that of Case 2 in the preceding section.

Case 5 (A is a normal matrix). Obviously we should suppose A has complex eigenvalues, otherwise A would be symmetric. Now we have real difficulty in the analysis of the stationary points. In fact, we do not even have a clear way of identifying all stationary points. We can only report some partial results.

For simplicity, we shall assume that Λ is given by (40) and that (41) holds. We partition Λ into three blocks $\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3$ where

$$\Lambda_1 = \operatorname{diag}\{\lambda_1, \cdots, \lambda_{k-1}\},$$

$$\Lambda_2 = \begin{bmatrix} \lambda_k & \nu_* \\ -\nu_* & \lambda_k \end{bmatrix}, \tag{45}$$

$$\Lambda_3 = \operatorname{diag}\{\lambda_{k+2}, \cdots, \lambda_n\}. \tag{46}$$

It can be verified easily that any E of the form

$$E = E_1 \oplus E_2 \oplus E_3 \tag{47}$$

satisfies the first-order condition (22) if $E_i + E_i^T$ is a diagonal matrix for $i = 1, 3$ and $E_2 \in R^{2\times 2}$. This, of course, is only a sufficient condition of being a stationary point.

We consider a simple 3×3 example. Let

$$A = \begin{bmatrix} -0.44910244205626 & -2.69770357656912 & -0.84185971635958 \\ 0.02746606843380 & -0.23010080980457 & -2.76631903691207 \\ -2.82587649838907 & -0.61291990656488 & -1.32079674813917 \end{bmatrix}$$

and

$$\Lambda = \begin{bmatrix} 15.0 & 0.0 & 0.0 \\ 0.0 & -3.0 & 12.0 \\ 0.0 & -12.0 & -3.0 \end{bmatrix}.$$

We calculate that $\|AA^T - A^TA\| \approx 4.5540 \times 10^{-14}$. So up to the fourteenth digit A is a normal matrix whose eigenvalues are $\{1 \pm 2i, -4\}$. Starting with $X(0) = \Lambda$, we follow the descent flow (16) by using the subroutine ODE in [13]. The local error tolerance is

set at 10^{-13}. We regard that the flow has converged to its limit point and the integration is terminated automatically whenever the difference between two consecutive output values is less than 10^{-12}. At $t \approx 0.5$, we obtain an approximate limit point

$$X = \begin{bmatrix} 5.047565112549 & -12.481140871140 & -1.983297617463 \\ 1.946294703163 & 0.447719348364 & 12.759402874230 \\ -12.486964555620 & -3.288091746472 & 3.504715539087 \end{bmatrix}$$

for the flow (16). The corresponding stationary point is approximated by

$$Q = \begin{bmatrix} 0.668645609196 & -0.437652789090 & -0.601143148929 \\ 0.437652789090 & 0.885212316658 & -0.157667975945 \\ 0.601143148929 & -0.157667975945 & 0.783433292538 \end{bmatrix}.$$

We calculate that $\|XX^T - X^TX\| \approx 2.7084 \times 10^{-10}$, $\|Q^TQ - I\| \approx 1.3866 \times 10^{-13}$. So X and Q are reasonably normal and orthogonal, respectively. The corresponding matrix $E := QAQ^T$ is given by

$$E = \begin{bmatrix} 0.644444444445 & -0.801988510684 & 2.173413906502 \\ 2.314685340881 & -0.608926976624 & -1.676627286676 \\ -0.095631338793 & -2.743293953342 & -2.035517467820 \end{bmatrix}.$$

We calculate that $\|\Lambda E^T + \Lambda^T E - E\Lambda^T + E^T\Lambda\| \approx 1.2299 \times 10^{-11}$. So we may say that up to the numerical error the matrix E satisfies the equation (22). But obviously E is not of the form (47). We think this complication is due to the fact that the spectra of A and Λ are "incompatible," i.e., the two triangles in the complex plane connecting eigenvalues of A and Λ, respectively, point to opposite directions.

In perturbation theory, we should not expect the spectrum of Λ to be distributed in a significantly different pattern from that of A. In part, this is because eigenvalues depend continuously upon components of the matrix. In part, this is because A, representing a sensible empirical data, should more or less reflect the physical reality. Now that Λ is assumed to be of the form (40), let us suppose that A also has only one pair of complex conjugate eigenvalues. Thus A can be reduced to the matrix

$$E := \operatorname{diag}\left\{e_1, \cdots, \begin{bmatrix} e_k & e_* \\ -e_* & e_k \end{bmatrix}, \cdots, e_n\right\}. \tag{48}$$

Now we shall see how the ordering of $\{e_1, \cdots, e_n\}$ affects the definiteness of the projected Hessian of F at such a point. By direct computation, we obtain

$$\begin{aligned} \langle [\Lambda, K], [E, K] \rangle = 2 &\sum_{\substack{i<j \\ i \neq k, k+1 \\ j \neq k, k+1}} (\lambda_i - \lambda_j)(e_i - e_j)k_{ij}^2 \\ &+ 2 \sum_{i<k} (k_{ik}^2 + k_{i,k+1}^2)((e_i - e_k)(\lambda_i - \lambda_k) + v_* e_*) \\ &+ 2 \sum_{k+1<j} (k_{kj}^2 + k_{k+1,j}^2)((e_k - e_j)(\lambda_k - \lambda_j) + v_* e_*). \end{aligned} \tag{49}$$

Every single term in (49) needs to be nonnegative in order that the projected Hessian of F is positive semidefinite. This, of course, will be the case if the ordering of $\{e_1, \cdots, e_n\}$ is "compatible" with (41), that is, if

$$e_1 \geqq e_2 \geqq \cdots \geqq e_n \tag{50}$$

and $e_* > 0$. We, therefore, have established a result of the following sufficient condition.

LEMMA 4.1. *Suppose A is normal. Suppose A can be reduced by orthogonal transformation Q to the canonical form E (48) whose elements are arranged as in* (50). *Suppose Λ is a quasi diagonal in the form of* (40) *whose elements are arranged as in* (41). *Then*

1. *The orthogonal matrix Q is a local minimizer of F.*
2. *The local optimal value of F is given by $\frac{1}{2}\|\Lambda - E\|^2$.*

Remark. In the 3×3 numerical example above, we have $-4 = e_1 < e_2 = 1$. Thus (49) is positive only if $e_* > (e_2 - e_1)(\lambda_1 - \lambda_3)/\nu_* = 7.5$. Since $e_* = \pm 2$ in our example, we find that our descent flow X cannot converge to an E in the form of (47). In fact, it turns out that such an E is a local maximum for F.

In contrast to the preceding three theorems, it is rather surprising that when A has complex eigenvalues the differential equation (16) may have multiple limit points. This phenomenon can be observed numerically by starting with different initial values on the surface $M(\Lambda) := \{Q^T \Lambda Q \mid Q \in O(n)\}$. For instance, if we start with $X(0) = \Lambda^T \in M(\Lambda)$ for the above 3×3 example, the flow converges to another limit point

$$X = \begin{bmatrix} 13.442778205310 & -0.124823985983 & -6.168244962433 \\ -5.831716696280 & -2.460547718025 & -10.728214876180 \\ -2.013431726775 & 12.210156961630 & -1.982230487286 \end{bmatrix},$$

which is quite different from the one obtained earlier. The least squares distances from these two distinct limit points to A, nevertheless, are the same. We have experimented with many other numerical examples. It seems true that when A is normal and has complex eigenvalues, Problem A does not have a unique solution. Different least squares approximations to A may result in different optimal values of F. Problem A, therefore, has multiple local solutions.

At this point, it is worthwhile to look at Problem A from another aspect. The following general perturbation problem [15] is of significant importance in many areas.

PROBLEM B. Suppose we know exactly the eigenvalues of the matrix A and that A is perturbed to become $A + B$. How do the eigenvalues change?

Usually we are interested in finding bounds of the perturbed eigenvalues in terms of the perturbing matrix B. In application it is not uncommon to have a situation in which both the original matrix A and the perturbing matrix B are real and symmetric. In this case, and in the more general situation in which both A and $A + B$ are normal, a comprehensive bound, known as the Wielandt–Hoffman theorem (see [9], [10, p. 368], and [15, p. 104]), is available on the perturbation to all the eigenvalues.

THEOREM 4.2. *Let $A, B \in C^{n \times n}$. Assume that A and $A + B$ are both normal. Let $\mu_1, \cdots, \mu_n$ be the eigenvalues of A in some given order, and let $\lambda_1, \cdots, \lambda_n$ be the eigenvalues of $A + B$ in some order. Then there exists a permutation $\sigma(i)$ of the integers* 1, 2, $\cdots$, *n such that*

$$\sum_{i=1}^{n} |\lambda_{\sigma(i)} - \mu_i|^2 \leqq \|B\|^2. \tag{51}$$

In Problem A we have the situation that all the eigenvalues (the original ones and the perturbed ones) are known and that we want to minimize the norm of the perturbing matrix B.

What we have shown in Theorems 3.1 and 3.2 is that, in the real and symmetric case, the minimum of $\|B\|$ is attained if $A + B = Q^T \Lambda Q$ where columns of Q^T are orthogonal eigenvectors of A in a certain order. In this case, the equality in (51) holds. In other words, we have shown that the bound in (51) for eigenvalues is sharp. This is a reproof of the Wielandt–Hoffman theorem. We think our proof, being different from both the original proof of [9] and the one given in [15], is of interest in its own right.

When the matrix A is real and normal, we can see immediately that the proof given in [9] for Theorem 4.2 breaks down if the perturbed matrix $A + B$ is restricted to be only real and normal. Problem A, in which we try to minimize the right-hand side of the inequality (51), becomes an interesting but difficult question. In Lemma 4.1 we have proved that if eigenvalues of A and $A + B$ (both real and normal) are "compatible," then again the equality in (51) holds. Our numerical experiments seem to indicate, however, that generally the minimal $\|B\|$ may be far larger than any rearrangement of eigenvalues on the left-hand side of the inequality (51) if only real matrices are allowed in the perturbation. Taking the 3×3 example to demonstrate our point, we calculate $\|X - A\|^2 \approx 496.2$ in comparison with the eigenvalue variation

$$\min_{\sigma} \sum_{i=1}^{n} |\lambda_{\sigma(i)} - \mu_i|^2 = 461.$$

Case 6 (A is a general matrix). Given a quasi-diagonal matrix Λ as in (1), an arbitrary matrix $A \in R^{n\times n}$, and letting $X := Q^T \Lambda Q$, we have established that necessary conditions for $Q \in O(n)$ to be a local minimizer for Problem A are

$$XA^T + X^T A = AX^T + A^T X, \tag{52}$$

$$\langle [X, K], [A, K] \rangle \geqq 0 \quad \text{for every } K \in S(n)^{\perp}. \tag{53}$$

If the strict inequality holds in (53), then the above conditions are sufficient for $Q \in O(n)$ to be a strong local minimizer of Problem A.

Thus far, we are able to characterize an analytical solution of Problem A from (52) and (53) for the following cases:

1. All eigenvalues of Λ are real, and $A \in R^{n\times n}$ is arbitrary.
2. Λ has complex conjugate eigenvalues, and $A \in R^{n\times n}$ is symmetric.
3. Λ has complex conjugate eigenvalues, and $A \in R^{n\times n}$ is normal but not symmetric. (Indeed, only partial results are obtained for this case.)

For a general nonnormal matrix A, the analytic comprehension of solutions satisfying both (52) and (53) becomes a much harder problem.

We have pointed out (Case 3) that when $n = 2$, all orthogonal matrices $Q \in O(2)$ are stationary points and the corresponding X can only be either Λ or Λ^T. From here, we might be able to characterize some stationary points for higher-dimensional cases. For example, suppose Λ is given by (40). Suppose A can be reduced by orthogonal similarity to the matrix

$$E := \operatorname{diag}\left\{e_1, \cdots, \begin{bmatrix} e_k^{(11)} & e_k^{(12)} \\ e_k^{(21)} & e_k^{(22)} \end{bmatrix}, \cdots, e_n\right\}, \tag{54}$$

which is conformal with Λ except that $e_k^{(ij)}$, $1 \leqq i, j \leqq 2$ are arbitrary real numbers. Then we can show that (52) is satisfied. This, of course, is just one special type of stationary points.

Recently, the Wielandt–Hoffman theorem has been generalized to nondefective matrices [14], [16].

THEOREM 4.3. *Let $A, B \in C^{n\times n}$. Suppose both A and $A + B$ are nondefective, i.e., suppose there exist nonsingular matrices S and T such that*

$$S^{-1} A S = \operatorname{diag}\{\mu_1, \cdots, \mu_n\},$$

$$T^{-1}(A + B) T = \operatorname{diag}\{\lambda_1, \cdots, \lambda_n\}.$$

Then there exists a permutation $\sigma(i)$ *of integers* $1, 2, \cdots, n$ *such that*

$$\sum_{i=1}^{n} |\lambda_{\sigma(i)} - \mu_i|^2 \leqq (\kappa_2(S)\kappa_2(T))^2 \|B\|^2 \tag{55}$$

where $\kappa_2(S) := \|S\|_2 \|S^{-1}\|_2$ *is the condition number of* S *and* $\|\cdot\|_2$ *means 2-norm.*

In the context of our discussion, the matrix $A + B$ is required to be a real and normal matrix. In this case, clearly $\kappa_2(T) = 1$. Suppose the given matrix A is nondefective; then the inequality (55) becomes

$$\sum_{n=1}^{n} |\lambda_{\sigma(i)} - \mu_i|^2 \leqq (\kappa_2(S))^2 \|B\|^2. \tag{56}$$

The inequality (56) suggests that when A is a general nonnormal matrix, the minimum value of $\|X - A\|$ may be smaller than the so-called eigenvalue variation. That it indeed is the case can be seen from the 2×2 matrix considered in Case 3—Suppose $a_{21} = 0$, $a_{12} > 0$, $\nu > 0$. Then it holds that

$$\min_{\sigma} \sum_{i=1}^{2} |\lambda_{\sigma(i)} - \mu_i|^2 = (a_{11} - \lambda)^2 + (a_{22} - \lambda)^2 + 2\nu^2 \tag{57}$$

while

$$\min_{Q \in O(2)} \|Q^T \Lambda Q - A\|^2 = (a_{11} - \lambda)^2 + (a_{22} - \lambda)^2 + (a_{12} - \nu)^2 + \nu^2. \tag{58}$$

Obviously, the value in (58) is less than that in (57) if $a_{12} < 2\nu$. This observation is interesting when compared with the Wielandt–Hoffman theorem for normal matrices. In the latter case, the minimum value of $\|X - A\|$ is always bounded *below* by the eigenvalue variation.

Although closed forms of solutions of (52) and (53) generally are difficult to obtain, our approach offers an alternative way to solve Problem A. We note that the differential equation (16), derived from the projected gradient of the objective function F, is numerically traceable for an arbitrary matrix A. Thus, by following trajectories of (16), we may locate stationary solutions of the least squares problem numerically. Different starting points may lead to different stationary points. The asymptotic rate of convergence is expected to be similar to that of the usual steepest descent method. But the flow, by its definition, is guaranteed to converge regardless of the location of the starting point. Our numerical experience is that the flow usually reaches a stable equilibrium point within a reasonable interval of integration.

Acknowledgment. The author thanks an anonymous referee for bringing references [14] and [16] to his attention. The provision by the same referee of additional statistics that enlighten the numerical examples is also gratefully acknowledged.

REFERENCES

[1] V. I. ARNOLD, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Second Edition, Springer-Verlag, New York, 1988.

[2] M. T. CHU AND L. K. NORRIS, *Isospectral flows and abstract matrix factorizations*, SIAM J. Numer. Anal., 25 (1988), pp. 1383–1391.

[3] M. T. CHU AND K. R. DRIESSEL, *The projected gradient method for least squares matrix approximations with spectral constraints*, SIAM J. Numer. Anal., 27 (1990), 1050–1060.

[4] R. GABRIEL, *Matrizen mit maximaler Diagonale bien unitärer Similarität*, J. Reine Angew. Math., 307/308 (1979), pp. 31–52.
[5] F. R. GANTMACHER, *Matrix Theory*, Vol. 1, Chelsea, New York, 1959.
[6] P. E. GILL, W. MURRAY, AND M. H. WRIGHT, *Practical Optimization*, Academic Press, London, 1981.
[7] R. GRONE, C. R. JOHNSON, E. M. SA, AND H. WOLKOWICZ, *Normal matrices*, Linear Algebra Appl., 87 (1987), pp. 213–225.
[8] N. J. HIGHAM, *Matrix nearest problems and applications*, in Applications of Matrix Theory, M. J. C. Gover and S. Barnett, eds., Oxford University Press, Oxford, 1989, pp. 1–27.
[9] A. J. HOFFMAN AND H. WIELANDT, *The variation of the spectrum of a normal matrix*, Duke Math. J., 20 (1953), pp. 37–39.
[10] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, 1987.
[11] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, Second Edition, Academic Press, London, 1985.
[12] A. RUHE, *Closest normal matrix finally found!*, BIT, 27 (1987), pp. 585–598.
[13] L. F. SHAMPINE AND M. K. GORDON, *Computer Solution of Ordinary Differential Equations, The Initial Value Problem*, W. H. Freeman, San Francisco, 1975.
[14] J. G. SUN, *On the perturbation of the eigenvalues of a normal matrix*, Math. Numer. Sinica, 6 (1984), pp. 334–336.
[15] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.
[16] Z. Y. ZHANG, *On the perturbation of the eigenvalues of a nondefective matrix*, Math. Numer. Sinica, 8 (1986), pp. 106–108.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 1, pp. 128–145, January 1991

010

DIVIDE-AND-CONQUER SOLUTIONS OF LEAST-SQUARES PROBLEMS FOR MATRICES WITH DISPLACEMENT STRUCTURE*

J. CHUN† AND T. KAILATH‡

Abstract. A divide-and-conquer implementation of a generalized Schur algorithm enables (exact and) least-squares solutions of various block-Toeplitz or Toeplitz-block systems of equations with $O(\alpha^3 n \log^2 n)$ operations to be obtained, where the displacement rank α is a small constant (typically between two to four for scalar near-Toeplitz matrices) independent of the size of the matrices.

Key words. divide-and-conquer, least squares, displacement structure, fast convolution, Toeplitz, Schur complements, generalized Schur algorithm

AMS(MOS) subject classifications. primary 65F05, 65F30; secondary 15A06

1. Introduction. In recent years, there has been considerable research on fast algorithms for the solution of linear systems of equations with Toeplitz matrices. The Levinson and Schur algorithms allow solutions with $O(n^2)$ floating point operations (flops) for systems with $n \times n$ Toeplitz matrices.

In 1980 Brent, Gustavson, and Yun [5] described a scheme for obtaining a solution with $O(n \log^2 n)$ flops. This was based on two ideas—the use of the *Gohberg–Semencul formula* [11], [13], [17], [26] for the inverse of a Toeplitz matrix, and the use of divide-and-conquer (or doubling) techniques for computing (generators of) the Gohberg–Semencul formula.

Let $\mathbf{x}$ and $\mathbf{y}$ denote the first and last columns of $T^{-1} \in \mathbf{R}^{n\times n}$. Then if the first component of $\mathbf{x}$, say x_1, is nonzero, Gohberg and Semencul [13] showed that we could write

$$T^{-1} = \frac{1}{x_1}[L(\mathbf{x})L^T(\tilde{I}_n\mathbf{y}) - L(Z_n\mathbf{y})L^T(Z_n\tilde{I}_n\mathbf{x})], \qquad x_1 \neq 0,$$

where $\tilde{I}_n$ is the *reverse-identity matrix*, Z_n is the *shift matrix*,

$$\tilde{I}_n \equiv \begin{bmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{bmatrix}, \qquad Z_n \equiv \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix},$$

and $L(\mathbf{v})$ is a lower-triangular Toeplitz matrix with first column $\mathbf{v}$. The significance of the Gohberg–Semencul formula in the present application is that the product of a vector and a lower- or upper-triangular Toeplitz matrix is equivalent to the convolution of two vectors, which can be done using $O(n \log n)$ flops (see, e.g., [4]).

Brent, Gustavson, and Yun used a divide-and-conquer scheme for a certain Euclidean algorithm to factorize row-permuted Toeplitz matrices (i.e., Hankel matrices), and to

* Received by the editors December 5, 1988; accepted for publication (in revised form) November 6, 1989. This work was supported in part by the U.S. Army Research Office under contract DAAL03-86-K-0045, the Strategic Defense Initiative Organization/Innovative Science and Technology, managed by the Army Research Office under contract DAAL03-87-K-0033, and the National Science Foundation under grant MIP-21315-A2.

† Information Systems Laboratory, Stanford University, Stanford, California 94305. Present address, General Electric Research and Development, Room KWD 218, P.O. Box 8, Schenectady, New York 12301 (chun@rascals.stanford.edu).

‡ Information Systems Laboratory, Stanford University, Stanford, California 94305 (tk@isl.stanford.edu).

obtain the vectors $\{\mathbf{x}, \mathbf{y}\}$ of the Gohberg–Semencul formula with $O(n \log^2 n)$ flops. Later Bitmead and Anderson [3] and Morf [21] used another approach based on the displacement-rank properties of matrix Schur complements, to obtain similar results; while this approach allows for generalization to non-Toeplitz matrices, the hidden coefficient in their proposed $O(n \log^2 n)$ constructions turned out to be extremely large (see Sexton, Shensa, and Speiser [25]). Later Musicus [22], de Hoog [11], Ammar and Gragg [2] used a more direct approach based on a combination of the Schur and Levinson algorithms to obtain better coefficients; in particular, Ammar and Gragg made a detailed study and claimed an operation count of $8n \log^2 n$ flops. With this count, the new (called superfast in [2]) method for solving (exactly determined) Toeplitz systems is faster than the one based on the Levinson algorithm whenever $n > 256$. We should mention here that Schur-algorithm-based methods are natural in the context of transmission-line and layered-earth models, so it is not a surprise that similar techniques were also conceived in those fields (see Choate [7], McClary [20], Bruckstein and Kailath [6]). A good source for background on the Levinson and Schur algorithms, transmission line models, displacement representations as mentioned and used in the present paper may be [14].

The method we have taken in this paper is in the spirit of the generalized Schur algorithm (see, e.g., [8], [9]). Our algorithm can be applied to non-Toeplitz matrices, and is simpler than the methods of Bitmead and Anderson [3] or Morf [21]. Furthermore, we can readily handle matrices such as $(T^TT)^{-1}$ and $(T^TT)^{-1}T^T$, where T may be a near-Toeplitz matrix or a rectangular block-Toeplitz matrix, or a Toeplitz-block matrix; in particular, therefore, we can also obtain the *least-squares* solutions of overdetermined Toeplitz and near-Toeplitz systems with $O(n \log^2 n)$ flops. Our algorithm is closely related to the algorithm of Musicus [22]. However, our presentation is conceptually much simpler (especially for the non-Toeplitz cases treated in [22]) than previous approaches; in particular, we do not use the relationship between the Schur algorithm and Levinson algorithms needed in [2], [11], and [22].

An outline of our approach is the following. For a matrix E,

$$E = \begin{bmatrix} E_{1,1} & E_{1,2} \\ E_{2,1} & E_{2,2} \end{bmatrix}, \qquad E_{1,1}, \text{ nonsingular,} \tag{1}$$

the Schur complement of $E_{1,1}$ in E is

$$S \equiv E_{2,2} - E_{2,1} E_{1,1}^{-1} E_{1,2}.$$

Note that matrices such as

$$S_1 \equiv T^{-1}, \qquad S_2 \equiv (T^TT)^{-1}, \qquad S_3 \equiv (T^TT)^{-1}T^T \tag{2}$$

can be identified as the Schur complements of the northwest blocks in the following *extended matrices*:

$$E_1 = \begin{bmatrix} T & I \\ -I & O \end{bmatrix}, \qquad E_2 = \begin{bmatrix} T^TT & I \\ -I & O \end{bmatrix}, \qquad E_3 = \begin{bmatrix} T^TT & T^T \\ -I & O \end{bmatrix}. \tag{3}$$

Now the matrices E in (3) have the following (generalized) *displacement representation*, for suitably chosen matrixes $\{F^f, F^b\}$:

$$E = \sum_{i=1}^{\alpha} K(\mathbf{x}_i, F^f) K^T(\mathbf{y}_i, F^b),$$

where $K(\mathbf{x}_i, F^f)$ and $K(\mathbf{y}_i, F^b)$ are lower triangular matrices whose j columns are $(F^f)^{(j-1)}\mathbf{x}_i$ and $(F^b)^{(j-1)}\mathbf{y}_i$, respectively. The smallest possible number α is called the

displacement rank of E with respect to $\{F^f, F^b\}$. For an example, let T be an $m \times n$ scalar Toeplitz matrix, with $m \geqq n$. Then the matrix E_2 has displacement rank four with respect to $\{F, F\}$, where $F = \left[\begin{smallmatrix} Z_n & O \\ O & Z_n \end{smallmatrix}\right]$, and has a displacement representation [15],

$$\text{(4a)} \quad E_2 = \sum_{i=1}^{2} K(\mathbf{y}_i, F)K^T(\mathbf{x}_i, F) - \sum_{i=3}^{4} K(\mathbf{y}_i, F)K^T(\mathbf{x}_i, F), \qquad \mathbf{y}_i \equiv \begin{bmatrix} I_n & O \\ O & -I_n \end{bmatrix}\mathbf{x}_i.$$

If we define $\mathbf{x}_i^T \equiv [\mathbf{w}_i^T, \mathbf{v}_i^T]$, note that the matrix $K(\mathbf{x}_i, F)$ in (4a) has the form

$$\text{(4b)} \qquad \begin{bmatrix} L(\mathbf{w}_i) & O \\ L(\mathbf{v}_i) & O \end{bmatrix} \in \mathbf{R}^{2n \times 2n}, \qquad O \in \mathbf{R}^{n \times n},$$

where $L(\mathbf{w}_i)$ and $L(\mathbf{v}_i)$ are lower triangular Toeplitz matrices with first columns $\mathbf{w}_i$ and $\mathbf{v}_i$.

Given a displacement representation of E, we use a certain *generalized Schur algorithm* (see § 2) to successively compute displacement representations of the Schur complements of all the leading principal submatrices in E. For the above example, n steps of the generalized Schur algorithm will yield

$$\begin{bmatrix} O & O \\ O & (T^TT)^{-1} \end{bmatrix} = \sum_{i=1}^{2} K(\mathbf{u}_i, F)K^T(\mathbf{u}_i, F) - \sum_{i=3}^{4} K(\mathbf{u}_i, F)K^T(\mathbf{u}_i, F),$$

where the top n elements of $\mathbf{u}_i$ are zero. Therefore, if we denote the bottom n elements of $\mathbf{u}_i$ as $\mathbf{u}_{2,i}$, we can have the displacement representation

$$(T^TT)^{-1} = \sum_{i=1}^{2} L(\mathbf{u}_{2,i})L^T(\mathbf{u}_{2,i}) - \sum_{i=3}^{4} L(\mathbf{u}_{2,i})L^T(\mathbf{u}_{2,i}).$$

Now, the generalized Schur algorithm, which is a two-term polynomial recursion, can be implemented in a divide-and-conquer fashion with $O(\alpha^3 f(n) \log n)$ flops, where $f(n)$ denotes the number of operations for the multiplication of two polynomials. Therefore, if the multiplication of two polynomials is done again by divide-and-conquer, i.e., by using fast convolution algorithms, then the overall computation requires $O(\alpha^3 n \log^2 n)$ flops. Once we have a displacement representation of the desired Schur complement S, the matrix-vector multiplication, $S\mathbf{b}$, can be done with $O(\alpha n \log n)$ flops using fast convolutions. As an example, we can obtain the least squares solution for the Toeplitz system

$$T\mathbf{x} = \mathbf{b}, \qquad T \in \mathbf{R}^{m \times n}, \qquad m \geqq n$$

as follows:

(i) Form $T^T\mathbf{b}$ using two fast convolutions,

(ii) Obtain a displacement representation of $(T^TT)^{-1}$ using the divide-and-conquer version of the generalized Schur algorithm,

(iii) Form $(T^TT)^{-1}(T^T\mathbf{b})$ using eight fast convolutions.

If we had obtained the displacement representation of $(T^TT)^{-1}T^T$ directly (using E_3), then step (i) above would not be needed.

2. Generalized Schur algorithm. After a brief review of basic concepts and definitions, we shall describe the generalized Schur algorithm of references [8], [9], and [15], but in a polynomial form important for the divide-and-conquer implementations. We shall need to recall some definitions and basic properties.

Generators of matrices. Let F^f and F^b be nilpotent matrices. The matrix

$$\nabla_{(F^f,F^b)}A \equiv A - F^f A F^{bT}$$

is called the *displacement* of A with respect to the *displacement operators* $\{F^f, F^b\}$. Define the (F^f, F^b)-displacement rank of A as rank $[\nabla_{(F^f,F^b)}A]$. Any matrix pair $\{X, Y\}$ such that

$$(5) \qquad \nabla_{(F^f,F^b)}A = XY^T, \qquad X \equiv [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_\alpha], \qquad Y \equiv [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_\alpha]$$

is called a (*vector form*) *generator* of A with respect to $\{F^f, F^b\}$. The generator will be said to have *length* α. If the length α is equal to the displacement rank of A, we say that the generator is *minimal*. A generator such as $Y = X\Sigma$, where Σ is a diagonal matrix with 1 or -1 along the diagonal, is called a *symmetric generator*.

The following lemma [15], [16] establishes the connection between generators and displacement representations.

LEMMA. *Let E be an $m \times n$ matrix. If F^f and F^b are nilpotent, then the equation $\nabla_{(F^f,F^b)}E = \sum_1^\alpha \mathbf{x}_i\mathbf{y}_i^T$ has the unique solution $E = \sum_1^\alpha K(\mathbf{x}_i, F^f)K^T(\mathbf{y}_i, F^b)$, where $K(\mathbf{x}_i, F^f) \equiv [\mathbf{x}_i, F^f\mathbf{x}_i, \cdots, F^{f(n-1)}\mathbf{x}_i]$ and $K(\mathbf{y}_i, F^b) \equiv [\mathbf{y}_i, F^b\mathbf{y}_i, \cdots, F^{b(n-1)}\mathbf{y}_i]$.*

Choice of displacement operators. The generalized Schur algorithm operates with generators, and needs $O(\alpha mn)$ flops for sequential implementation and $O(\alpha^3 n \log^2 n)$ for divide-and-conquer implementation. Therefore, for a given matrix A, we should try to choose the displacement operators that give the smallest α. If the matrix A is an $n \times n$ Toeplitz matrix, the appropriate displacement operator F is Z_n, an $n \times n$ shift matrix. If A has some near-Toeplitz structure, then F would have forms such as

$$F = Z_n \oplus Z_m, \qquad F = \bigoplus_{i=1}^{n} Z_{n_i}, \qquad F = Z_n^\beta,$$

where $\oplus$ denotes the *direct sum*, $Z_n \oplus Z_m \equiv \left[\begin{smallmatrix} Z_n & O \\ O & Z_m \end{smallmatrix}\right]$, and $\oplus_{i=1}^n$ denotes the concatenated direct sum.

Example 1. Let $T = (t_{i-j})$ be an $m \times n$ pre- and post-windowed scalar Toeplitz matrix, i.e., $t_{i,j} = 0$ if $j > i$ or $i > m - n + j$ with $m \geqq n$. Then it is easy to check that the matrix $C = (c_{i-j}) \equiv T^T T$ is also an (unwindowed) Toeplitz matrix, and with respect to $\{Z_n \oplus Z_n, Z_n \oplus Z_m\}$, E_3 in (3) has a generator $\{X, Y\}$ of length two, where

$$\mathbf{x}_1 = [c_0, c_1, \cdots, c_{n-1}, -1, 0, \cdots, 0]^T / c_0^{1/2},$$

$$\mathbf{x}_2 = [0, c_1, \cdots, c_{n-1}, -1, 0, \cdots, 0]^T / c_0^{1/2},$$

$$\mathbf{y}_1 = [c_0, c_1, \cdots, c_{n-1}, t_0, t_1, \cdots t_{m-n}, 0, \cdots, 0]^T / c_0^{1/2},$$

$$\mathbf{y}_2 = -[0, c_1, \cdots, c_{n-1}, t_0, t_1, \cdots t_{m-n}, 0, \cdots, 0]^T / c_0^{1/2}. \qquad \square$$

Example 2. If T is a Toeplitz-block matrix, i.e.,

$$(6) \qquad T = \begin{bmatrix} T_{1,1} & T_{1,2} \cdot T_{1,N} \\ T_{2,1} & T_{2,2} \cdot T_{2,N} \\ \cdot & \cdot \cdot \cdot \\ T_{M,1} & T_{M,2} \cdot T_{M,N} \end{bmatrix} \in \mathbf{R}^{m \times n}, \qquad T_{i,j} = \text{scalar } m_i \times n_j \text{ Toeplitz matrix},$$

then for the matrices E in (3), we choose [9], [15] the following displacement operators:

$$E_1: F^f = \left[\bigoplus_{i=1}^{M} Z_{m_i}\right] \oplus F_1, \qquad F^b = \left[\bigoplus_{i=1}^{N} Z_{n_i}\right] \oplus F_1, \qquad m = n, \tag{7a}$$

$$E_2: F^f = \left[\bigoplus_{i=1}^{N} Z_{n_i}\right] \oplus F_1, \qquad F^b = \left[\bigoplus_{i=1}^{N} Z_{n_i}\right] \oplus F_1, \tag{7b}$$

$$E_3: F^f = \left[\bigoplus_{i=1}^{N} Z_{n_i}\right] \oplus F_1, \qquad F^b = \left[\bigoplus_{i=1}^{N} Z_{n_i}\right] \oplus \left[\bigoplus_{i=1}^{M} Z_{m_i}\right], \tag{7c}$$

where F_1 can be either Z_n or $\oplus_{i=1}^{N} Z_{n_i}$. However, for the divide-and-conquer implementation, we prefer to choose $\oplus_{i=1}^{N} Z_{n_i}$; see the remark in § 4.

Example 3. On the other hand, if the matrix T in (3) is a block-Toeplitz matrix with $\beta \times \beta$ blocks,

$$T = \begin{bmatrix} B_0 & B_{-1} & \cdot & B_{-N+1} \\ B_1 & B_0 & \cdot & B_{-N+2} \\ \cdot & \cdot & \cdot & \cdot \\ B_{M-1} & B_{M-2} & \cdot & B_{-N+M} \end{bmatrix} \in \mathbf{R}^{m \times n}, \quad B_k \in \mathbf{R}^{\beta \times \beta}, \quad m \equiv M\beta, \quad n \equiv N\beta, \tag{8}$$

then for the extended matrices E, we should choose [8], [9] the displacement operators

$$F^f = Z_n^{\beta} \oplus Z_n^{\beta}, \qquad F^b = Z_n^{\beta} \oplus Z_m^{\beta}, \tag{9}$$

where for E_1 we assumed that T is a square $n \times n$ matrix.

Generators of the above and other extended block-Toeplitz or Toeplitz-block matrices can be found in [8].

Polynomial form of generators. In general, the displacement operators F^f and F^b for both extended block-Toeplitz matrices and extended Toeplitz-block matrices have the form

$$F = \bigoplus_{i=1}^{N} Z_{n_i}^{\beta}, \qquad n \equiv \sum_{i=1}^{N} n_i. \tag{10}$$

We shall say that the displacement operator F in (10) has N *sections*. One of the key operations in generalized Schur algorithms is matrix-vector multiplication $F\mathbf{v}$, i.e., a *sectioned shift* operation. With the polynomial representation of vectors, the shift operations has a nice algebraic expression. For a given vector $\mathbf{v}$, let $v(z)$ denote the polynomial whose coefficient for the term z^i is the $(i+1)$st component of the vector, i.e.,

$$\mathbf{v} = [v_0, v_1, v_2, \cdots, v_{n-1}]^T \leftrightarrow v(z) = v_0 + v_1 z + v_2 z^2 + \cdots v_{n-1} z^{n-1}. \tag{11}$$

Then,

$$Z_n \mathbf{v} \equiv \mathbf{v}' = [0, v_0, v_1, \cdots, v_{n-2}]^T \leftrightarrow v(z) z \bmod z^n.$$

In general, for the matrix whose displacement operator is the F in (10), let us define integers $\{\delta_i\}$ by

$$\delta_i = \sum_{k=1}^{i} n_k, \qquad \delta_1 < \delta_2 < \cdots < \delta_N.$$

Let $v(z)$ and $\theta(z)$ be polynomials of degree less than or equal to $n-1$, and define the degree at most $(n_i - 1)$ polynomial, $v_i(z)$, by

$$v(z) = v_1(z) + z^{\delta_1} v_2(z) + z^{\delta_2} v_3(z) + \cdots + z^{\delta_{N-1}} v_{n_N}(z). \tag{12a}$$

Given two polynomials $v(z)$ and $\theta(z)$, and the displacement operator F in (10), the (*polynomial form*) *displacement operator* $\otimes_F$ is defined by the following operation:

$$(12b) \qquad v(z)\otimes_F\theta(z)\equiv r(z)\equiv r_1(z)+z^{\delta_1}r_2(z)+z^{\delta_2}r_3(z)+\cdots+z^{\delta_{N-1}}r_{n_N}(z),$$

where

$$(12c) \qquad r_i(z)\equiv v_i(z)\theta(z^{\beta}) \bmod z^{n_i},$$

i.e., $r_i(z)$ is the polynomial $v_i(z)\theta(z^{\beta})$ after chopping off the higher degree terms, so that $r_i(z)$ has the degree at most (n_i-1).

Let

$$X=[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_\alpha], \qquad Y=[\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_\alpha]$$

be a generator of a matrix A with respect to certain $\{F^f, F^b\}$, and let

$$\mathbf{x}_i\leftrightarrow x_i(z), \qquad \mathbf{y}_i\leftrightarrow y_i(w).$$

Then we call the pair of polynomial vectors $\{X(z), Y(w)\}$, where

$$X(z)\equiv[x_1(z),x_2(z),\cdots,x_\alpha(z)], \qquad Y(w)\equiv[y_1(w),y_2(w),\cdots,y_\alpha(w)],$$

a (*polynomial form*) *generator of* A, with respect to (*polynomial form*) *displacement operator* $\{\otimes_{F^f},\otimes_{F^b}\}$.

Example 1 (*continued*). The matrix E_3 in (3) has a generator $\{X(z), Y(w)\}$ with respect to $\{\otimes_{F^f},\otimes_{F^b}\}$, where $F^f=Z_n\oplus Z_n$, $F^b=Z_n\oplus Z_m$, and

$$x_1(z)=[c_0+c_1z+\cdots+c_{n-1}z^{n-1}-z^n]c_0^{-1/2},$$

$$x_2(z)=[c_1z+cz^2+\cdots+c_{n-1}z^{n-1}-z^n]c_0^{-1/2},$$

$$y_1(w)=[c_0+c_1w+\cdots+c_{n-1}w^{n-1}+t_0w^n+t_1w^{n+1}+\cdots+t_{m-n}w^m]c_0^{-1/2},$$

$$y_2(w)=-[c_1w+\cdots+c_{n-1}w^{n-1}+t_0w^n+t_1w^{n+1}+\cdots+t_{m-n}w^m]c_0^{-1/2}.$$

Also note that

$$x_1(z)\otimes_{F^f}z=[c_0z+c_1z^2+\cdots+c_{n-2}z^{n-1}-z^{n+1}]c_0^{-1/2},$$

$$y_1(w)\otimes_{F^b}w$$

$$=[c_0w+c_1w^2+\cdots+c_{n-2}w^{n-1}+t_0w^{n+1}+t_1w^{n+2}+\cdots+t_{m-n}w^{m+1}]c_0^{-1/2}. \quad \square$$

Next we note that for given vectors $\mathbf{a}$ and $\mathbf{b}$ such that $\mathbf{a}^T\mathbf{b}\neq 0$, we can always find [8] matrices Θ and Ψ such that

$$(13) \qquad \mathbf{a}^T\Theta=[a_1',0,0,\cdots,0], \qquad \mathbf{b}^T\Psi=[b_1',0,0,\cdots,0], \qquad \Theta\cdot\Psi^T=I,$$

and therefore, $\mathbf{a}^T\mathbf{b}=a_1'b_1'$. We define polynomial matrices $\Theta(z)$ and $\Psi(w)$ by

$$(14) \qquad \Theta(z)\equiv\Theta\begin{bmatrix} z & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}, \qquad \Psi(w)\equiv\Psi\begin{bmatrix} w & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}.$$

We also remark that if $\mathbf{a}=\mathbf{b}$, then we could choose $\Psi(w)=\Theta(w)$, and if $\mathbf{b}=\Sigma\mathbf{a}$, where $\Sigma\equiv I_p\oplus -I_q$, then $\Psi(w)=\Theta(w)\Sigma$, so that we only need to find, and post-multiply by, $\Theta(z)$.

Generalized Schur algorithm. Let a matrix E have a generator $\{X_0(z), Y_0(w)\}$ with respect to $\{\otimes_{F^f}, \otimes_{F^b}\}$, and define $E_{i,j}$ by

$$E = \begin{bmatrix} E_{1,1} & E_{1,2} \\ E_{2,1} & E_{2,2} \end{bmatrix} \in \mathbf{R}^{m \times n},$$

where $E_{1,1}$ is a $k \times k$ strongly nonsingular matrix, i.e., the one with all nonsingular leading submatrices. The k-step generalized Schur algorithm [8], [9], [15] presented below in polynomial form gives a generator of the matrix

$$\begin{bmatrix} O & O \\ O & S \end{bmatrix}, \qquad S \equiv E_{2,2} - E_{2,1} E_{1,1}^{-1} E_{1,2} \in \mathbf{R}^{(m-k)\times(n-k)},$$

with respect to $\{\otimes_{F^f}, \otimes_{F^b}\}$, or equivalently, a generator of S with respect to $\{\otimes_{\bar{F}^f}, \otimes_{\bar{F}^b}\}$, where $\bar{F}^f$ and $\bar{F}^b$ denote the trailing square submatrices of size $(m-k)$ and $(n-k)$ of F^f and F^b, respectively.

ALGORITHM (k-step generalized Schur algorithm).
Input: Generator of E, $\{X_0(z), Y_0(w)\}$; displacement operator $\{\otimes_{F^f}, \otimes_{F^b}\}$;
Number of steps k.
Output: Generator of S $\{X_k(z), Y_k(w)\}$
Procedure GeneralizedSchur
begin
for $i := 0$ to $k-1$ **do begin**
$\mathbf{a}^T := [z^{-i} X_i(z)]_{z=0}$;
$\mathbf{b}^T := [z^{-i} Y_i(z)]_{z=0}$;
Find $\Theta_i(z)$ and $\Psi_i(w)$ to transform $\mathbf{a}^T$ and $\mathbf{b}^T$ such as (13);
$X_{i+1}(z) = X_i(z) \otimes_{F_f} \Theta_i(z)$; $\quad Y_{i+1}(w) = Y_i(w) \otimes_{F_b} \Psi_i(w)$
end
return $\{X_k(z), Y_k(w)\}$
end

Remark. The polynomial vectors, $X_i(z)$ and $Y_i(w)$, have degrees $m-1$ and $n-1$, respectively, for all i. Each step eliminates the nonzero lowest degree term, and therefore the terms of $X_i(z)$ and $Y_i(w)$ whose degrees are less than z^i and w^i are zeros.

By applying the generalized Schur algorithm we can obtain generators, or equivalently displacement representations, for various interesting Schur complements.

3. Divide-and-conquer implementation. The (sequential) k-step generalized Schur algorithm in § 2 can also be implemented efficiently using the divide-and-conquer approach. We shall only explain how to find $X_k(z)$; essentially the same argument applies for $Y_k(w)$.

Let us define $\Theta_{p:q}(z)$ and $X_{p:q}(z)$ by

$$\Theta_{p:q}(z) \equiv \Theta_p(z)\Theta_{p+1}(z) \cdots \Theta_q(z),$$

$$X_{p:q}(z) \equiv X_{0:q}(z) \otimes_{F^f} \Theta_{0:p-1}(z), \qquad X_{0:q}(z) \equiv X_0(z) \bmod z^{q+1},$$

where $0 \leqq p \leqq q$. The polynomial matrix $\Theta_{p:q}(z)$ has a degree $q-p+1$. The polynomial vector $X_{p:q}(z)$ has degree q, and is obtained by dropping from $X_p(z)$ all terms of degree higher than z^q. Also note the useful properties,

$$[x(z) \otimes_F \theta_1(z)] \otimes_F \theta_2(z) = x(z) \otimes_F [\theta_1(z)\theta_2(z)],$$

$$[x_1(z) + x_2(z)] \otimes_F \theta(z) = [x_1(z) \otimes_F \theta(z)] + [x_2(z) \otimes_F \theta(z)].$$

These properties and the fact that $\Theta_{p:q}(z)$ is completely determined by $X_{p:q}(z)$ allow a divide-and-conquer implementation of the generalized Schur algorithm.

Given $X_{p:q}(z)$, we can compute $\Theta_{p:q}(z)$ as follows. If $p = q$, then we are successful, and compute $\Theta_{p:p}(z) = \Theta_p(z)$. Otherwise, we choose an "appropriate" (see § 4) *division point* r such that $p < r < q$, and try to solve the smaller subproblem of finding $\Theta_{p:r-1}(z)$, given $X_{p:r-1}(z)$. Once we know $\Theta_{p:r-1}(z)$, we can compute $X_{r:q}(z)$ by

$$X_{r:q}(z) = X_{0:q}(z) \otimes_{F^f} \Theta_{0:r-1}(z) = [X_{0:q}(z) \otimes_{F^f} \Theta_{0:p-1}(z)] \otimes_{F^f} \Theta_{p:r-1}(z) \tag{15a}$$

$$= X_{p:q}(z) \otimes_{F^f} \Theta_{p:r-1}(z). \tag{15b}$$

Now we again try to find $\Theta_{r:q}(z)$ given $X_{r:q}(z)$. After we obtain $\Theta_{r:q}(z)$, we can combine the two results, $\Theta_{p:r-1}(z)$ and $\Theta_{r:q}(z)$, by multiplication,

$$\Theta_{p:q}(z) = \Theta_{p:r-1}(z)\Theta_{r:q}(z). \tag{16}$$

Programming details of the above *recursive generalized Schur algorithm* are shown in the Appendix.

The previous recursive description can be visualized nonrecursively using *trees* (See Figs. 1 and 2). Each node in the tree is annotated with the *rules*: "find," "apply," and "combine,"

$f_{p:p}$: Find $\Theta_{p:p}(z)$,

$a_{p:q}$: $X_{r:q}(z) := X_{p:q}(z) \otimes_F \Theta_{p:r-1}(z)$,

$c_{p:q}$: $\Theta_{p:q}(z) := \Theta_{p:r-1}(z)\Theta_{r:q}(z)$.

We traverse the tree in *post-order* (i.e., follow the order labeled on each node of the tree), and evaluate the rules.

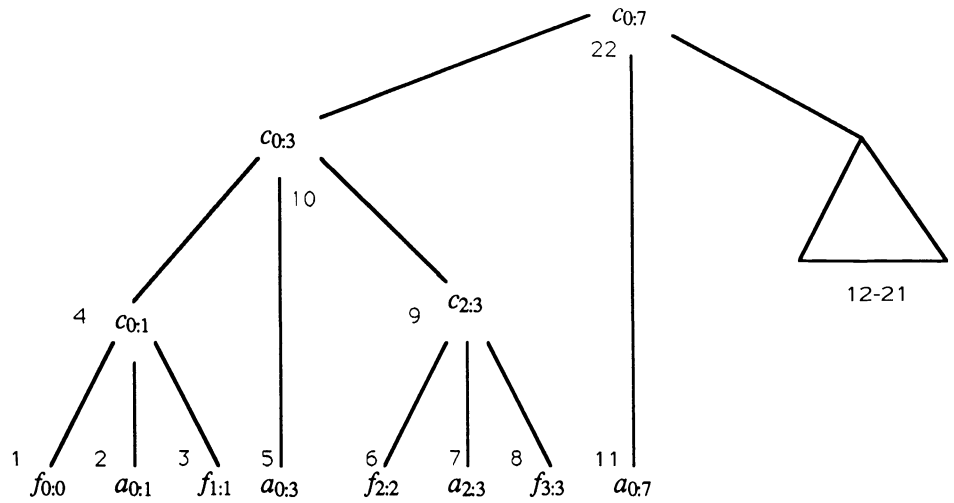


FIG. 1. *Sequence of computations for Example* 4.

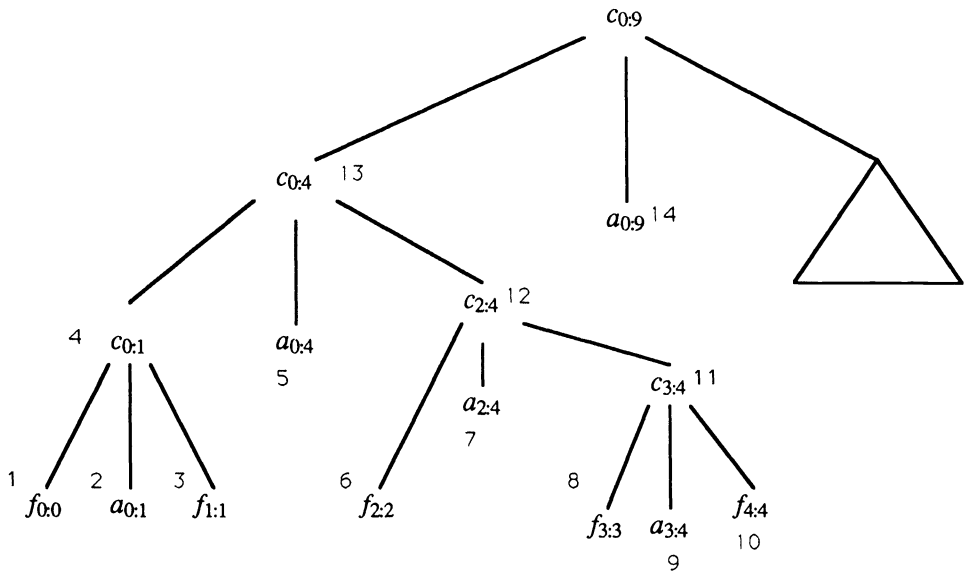


FIG. 2. *Sequence of computations for Example* 5.

Now, we shall consider two examples in detail.

Example 4. Pseudoinverse of pre- and post-windowed Toeplitz matrices. Consider the matrix E_3 in Example 1, where

$$T^TT=\begin{bmatrix}16&8&4&1\\8&16&8&4\\4&8&16&8\\1&4&8&16\end{bmatrix},\qquad T^T=\begin{bmatrix}3&2&1&1&-1&0&0&0\\0&3&2&1&1&-1&0&0\\0&0&3&2&1&1&-1&0\\0&0&0&3&2&1&1&-1\end{bmatrix}.$$

We would like to find a displacement representation of $(T^TT)^{-1}T^T$. This can be done by the four-step recursive generalized Schur algorithm. The input to the algorithm is a generator $\{X_0(z), Y_0(w)\}$ of

$$E_3=\begin{bmatrix}T^TT & T^T\\ -I & O\end{bmatrix},$$

with respect to $\{\otimes_{ff}, \otimes_{F^b}\}$, where $F^f = Z_n \oplus Z_n$, $F^b = Z_n \oplus Z_m$. The output $\{X_4(z), Y_4(w)\}$ is a generator of $(T^TT)^{-1}T^T$ with respect to $\{\otimes_{Z_n}, \otimes_{Z_m}\}$. The computational sequence is illustrated in Fig. 1, where it is assumed that the division points were chosen successively by two, one, and three.

[1] $f_{0:0}:\Theta_{0:0}(z)=\begin{bmatrix}1&0\\0&-1\end{bmatrix}\begin{bmatrix}z&\\&1\end{bmatrix}$ because $X_{0:0}(z)=[4,0]$.

[2] $a_{0:1}:X_{1:1}(z)=X_{0:1}(z)\otimes_{F^f}\Theta_{0:0}(z)=[4+2z,2z]\otimes_{F^f}\Theta_{0:0}(z)=[4z,-2z]$.

[3] $f_{1:1}:\Theta_{1:1}(z)=\frac{2}{\sqrt{3}}\cdot\begin{bmatrix}1&+\frac{1}{2}\\-\frac{1}{2}&-1\end{bmatrix}\begin{bmatrix}z&\\&1\end{bmatrix}$.

[4] $c_{0:1}:\Theta_{0:1}(z)=\Theta_{0:0}(z)\Theta_{1:1}(z)=\frac{2}{\sqrt{3}}\cdot\begin{bmatrix}z^2&-z/2\\-z/2&1\end{bmatrix}$.

[5] $a_{0:3}: X_{2:3}(z) = X_{0:3}(z) \circledast_{Ff} \Theta_{0:1}(z) = \frac{2}{\sqrt{3}} \cdot [3z^2 + 3z^3/2, -z^3/4].$

[6] $f_{2:2}: \Theta_{2:2}(z) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} z & \\ & 1 \end{bmatrix}$ because $X_{2:2}(z) = \frac{2}{\sqrt{3}} \cdot [3z^2, 0].$

[7] $a_{2:3}: X_{3:3}(z) = X_{2:3}(z) \circledast_{Ff} \Theta_{2:2}(z) = \frac{2}{\sqrt{3}} \cdot [3z^3, z^3/4].$

[8] $f_{3:3}: \Theta_{3:3}(z) = \frac{12}{\sqrt{143}} \cdot \begin{bmatrix} 1 & \frac{1}{12} \\ -\frac{1}{12} & -1 \end{bmatrix} \begin{bmatrix} z & \\ & 1 \end{bmatrix}.$

[9] $c_{2:3}: \Theta_{2:3}(z) = \Theta_{2:2}(z)\Theta_{3:3}(z) = \frac{12}{\sqrt{143}} \cdot \begin{bmatrix} z & z/12 \\ \frac{1}{12} & 1 \end{bmatrix} \begin{bmatrix} z & \\ & 1 \end{bmatrix}.$

[10] $c_{0:3}: \Theta_{0:3}(z) = \Theta_{0:1}(z)\Theta_{2:3}(z) = \frac{24}{\sqrt{3}\sqrt{143}} \cdot \begin{bmatrix} z^4 - z^2/24 & z^3/12 - z/12 \\ -z^3/12 + z/12 & -z^2/24 + 1 \end{bmatrix}.$

[11] $a_{0:7}: X_{4:7}(z) = [4 + 2z + z^2 + z^3/4 - z^4/4, 2z + z^2 + z^3/4 - z^4/4] \circledast_{Ff} \Theta_{0:3}(z)$

$$= [(4 + 2z + z^2 + z^3/4, 2z + z^2 + z^3/4) - z^4(\tfrac{1}{4}, \tfrac{1}{4})] \circledast_{Ff} \Theta_{0:3}(z)$$

$$= -z^4[(\tfrac{1}{4}, \tfrac{1}{4})\Theta_{0:3}(z) \bmod z^4]$$

$$= -\frac{6z^4}{\sqrt{3}\sqrt{143}}[z/12 - z^2/24 - z^3/2, 1 - z/2 - z^2/24 + z^3/12].$$

Because $T^T T$ is symmetric, $\Psi_{0:3}(w) = \Theta_{0:3}(w)\Sigma$, where $\Sigma = 1 \oplus -1$, and therefore,

$$Y_{4:13}(w) = [(4 + 2z + z^2 + z^3/4) + z^4(3/4 + z/2 + z^2/4 - z^4/4),$$
$$(2z + z^2 + z^3/4) + z^4(3/4 + z/2 + z^2/4 + z^3/4 - z^4/4)] \circledast_{Fb} \Theta_{0:3}(w)\Sigma$$
$$= \frac{z^4 6}{\sqrt{3}\sqrt{143}}[1/4z + z^2/24 - 3z^3/2 + 49z^4/24 + 11z^5/8 + 13z^6/24 + 3z^7/2,$$
$$-3 - z/2 + z^2/8 - 2z^3/3 + 11z^4/8 - 13/24z^5 - z^6/8 - z^7/12].$$

Therefore,

$$(T^T T)^{-1} T^T = \gamma^2 [L(\mathbf{x}_1)L^T(\mathbf{y}_1) + L(\mathbf{x}_2)L^T(\mathbf{y}_2)], \qquad \gamma = \frac{6}{\sqrt{3}\sqrt{143}},$$

where $L(\mathbf{x}_i)$ and $L(\mathbf{y}_i)$ are the lower triangular Toeplitz matrices whose first columns are $\mathbf{x}_i$ and $\mathbf{y}_i$, respectively, and

$$\mathbf{x}_1 = [0, -\tfrac{1}{12}, \tfrac{1}{24}, \tfrac{1}{2}]^T,$$
$$\mathbf{x}_2 = [-1, \tfrac{1}{2}, \tfrac{1}{24}, -\tfrac{1}{12}]^T,$$
$$\mathbf{y}_1 = [0, \tfrac{1}{4}, \tfrac{1}{24}, -\tfrac{3}{2}, \tfrac{49}{24}, \tfrac{11}{8}, \tfrac{13}{24}, \tfrac{3}{2}]^T,$$
$$\mathbf{y}_2 = [-3, -\tfrac{1}{2}, \tfrac{1}{8}, -\tfrac{2}{3}, \tfrac{11}{8}, -\tfrac{13}{24}, -\tfrac{1}{8}, \tfrac{1}{12}]^T.$$

Remark 1. For a symmetric generator of length two with $\beta = 1$, the 2×2 polynomial matrix $\Theta(z)$ in (14) can have the form (hyperbolic reflection)

$$\Theta_i(z) = \begin{bmatrix} ch_i z & sh_i \\ -sh_i z & -ch_i \end{bmatrix}, \qquad ch_i^2 - sh_i^2 = 1.$$

Let

$$\Theta_{p:q}(z) \equiv \Theta_p(z)\Theta_{p+1}(z) \cdots \Theta_q(z) \equiv \begin{bmatrix} \Theta_{1,1}(z) & \Theta_{1,2}(z) \\ \Theta_{2,1}(z) & \Theta_{2,2}(z) \end{bmatrix}.$$

Then, by induction, we can easily prove that

$$z^{q-p+1}\Theta_{1,1}(z^{-1}) = (-1)^{q-p+1}\Theta_{2,2}(z), \qquad z^{q-p+1}\Theta_{1,2}(z^{-1}) = (-1)^{q-p+1}\Theta_{2,1}(z).$$

Therefore, we need to compute and store only two entries of $\Theta_{p:q}(z)$.

Remark 2. For an unwindowed scalar Toeplitz matrix, the matrix E_2 in (3) has displacement rank four, whereas the matrix E_3 has displacement rank five. Therefore, when we solve Toeplitz least squares problems, it is more efficient to find a displacement representation of $(T^T T)^{-1}$ rather than of $(T^T T)^{-1} T^T$. With the notation in (4), the matrix E_2 for an unwindowed scalar Toeplitz matrix $T = (t_{i-j}) \in \mathbf{R}^{m \times n}$ $(m \geqq n)$ has a generator [15],

$$\mathbf{w}_1 = T^T \mathbf{t}_1 / \|\mathbf{t}_1\|, \qquad \mathbf{w}_2 = \mathbf{t}_2, \qquad \mathbf{w}_3 = Z_n Z_n^T \mathbf{w}_1, \qquad \mathbf{w}_4 = Z_n \mathbf{l},$$

$$\mathbf{t}_1 \equiv [t_0, t_1, \cdots, t_{m-1}]^T, \qquad \mathbf{t}_2 \equiv [0, t_{-1}, \cdots, t_{1-n}]^T, \qquad \mathbf{l} \equiv [t_{m-1}, \cdots, t_{m-n}]^T,$$

$$\mathbf{v}_1 = \mathbf{v}_3 = \mathbf{e}_1 / \|\mathbf{t}_1\|, \qquad \mathbf{v}_2 = \mathbf{v}_4 = 0,$$

where $\|\cdot\|$ denotes the Euclidean norm, and $\mathbf{e}_1$ is the vector with one in the first position, and zeros elsewhere.

Example 5. Displacement representation for the inverse of a Sylvester matrix. Let T denote the following Sylvester matrix,

$$T \equiv \begin{bmatrix} 2 & 0 & 0 & 1 & 0 \\ 1 & 2 & 0 & 2 & 1 \\ 3 & 1 & 2 & 1 & 2 \\ 0 & 3 & 1 & 1 & 1 \\ 0 & 0 & 3 & 0 & 1 \end{bmatrix} \tag{17}$$

and suppose that it is desired to obtain a displacement representation of T^{-1}. Then the appropriate extended matrix is

$$E_1 = \begin{bmatrix} T & I \\ -I & O \end{bmatrix}, \tag{18}$$

and it is easy to see that the following $\{X_0(z), Y_0(w)\}$ is a generator of E_1 with respect to $\{\otimes_{F^f}, \otimes_{F^b}\}$, where $F^f = Z_5 \oplus Z_5$, $F^b = Z_3 \oplus Z_2 \oplus Z_5$;

$$X_0(z) \equiv [x_1(z), x_2(z), x_3(z)], \qquad Y_0(w) \equiv [y_1(w), y_2(w), y_3(w)],$$

$$x_1(z) = 2 + z + 3z^2 - z^5, \qquad x_2(z) = 1 + 2z + z^2 + z^3 - z^8, \qquad x_3(z) = 1, \tag{19a}$$

$$y_1(w) = 1, \qquad y_2(w) = w^3, \qquad y_3(w) = w^5. \tag{19b}$$

Now the five-step recursive generalized Schur algorithm gives a desired generator of T^{-1}, with respect to $\{Z_5, Z_5\}$, and a possible computational sequence is shown in Fig. 2, where the division points are chosen successively as two, one, three, and four.

[1] $$f_{0:0}:\Theta_{0:0}(z)=\begin{bmatrix} z & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \Psi_{0:0}(w)=\begin{bmatrix} w & 0 & 0 \\ w/2 & 1 & 0 \\ w/2 & 0 & 1 \end{bmatrix}.$$

[2] $$a_{0:1}:X_{1:1}(z)=[2z, 3z/2, -z/2], \qquad Y_{1:1}(w)=[w, 0, 0].$$

[3] $$f_{1:1}:\Theta_{1:1}(z)=\begin{bmatrix} z & -\frac{3}{4} & -\frac{1}{4} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \Psi_{1:1}(w)=\begin{bmatrix} w & 0 & 0 \\ 3w/4 & 1 & 0 \\ -w/4 & 0 & 1 \end{bmatrix}.$$

[4] $$c_{0:1}:\Theta_{0:1}(z)=\begin{bmatrix} z^2 & -3z/4-1/2 & z/4-1/2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$\Psi_{0:1}(w)=\begin{bmatrix} w^2 & 0 & 0 \\ w^2/2+3w/4 & 1 & 0 \\ w^2/2-w/4 & 0 & 1 \end{bmatrix}.$$

[5] $$a_{0:4}:X_{2:4}(z)=[2z^2+z^3+3z^4, -5z^2/4-5z^3/4, -5z^2/4+3z^3/4],$$

$$\begin{aligned} Y_{2:4}(w) &= Y_{0:4}(w)\otimes_{F^b}\Psi_{0:1}(w) \\ &= [(1,0,0)\Psi_{0:1}(w) \bmod w^3]+w^3[(0,w,0)\Psi_{0:1}(w) \bmod w^2] \\ &= [w^2+3w^4/4, w^3, 0]. \end{aligned}$$

[6] $$f_{2:2}:\Theta_{2:2}(z)=\begin{bmatrix} z & \frac{5}{8} & \frac{5}{8} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \Psi_{2:2}(w)=\begin{bmatrix} w & 0 & 0 \\ -5w/8 & 1 & 0 \\ -5w/8 & 0 & 1 \end{bmatrix}.$$

[7] $$a_{2:4}:X_{3:4}(z)=[2z^3+z^4, -5z^3/8+15z^4/8, 11z^3/8+15z^4/8],$$

$$\begin{aligned} Y_{3:4}(w) &= Y_{2:4}(w)\otimes_{F^b}\Psi_{2:2}(w) \\ &= [(w^2,0,0)\Psi_{2:2}(w) \bmod w^3]+w^3[(3w/4,1,0)\Psi_{2:2}(w) \bmod w^2] \\ &= [-5w^4/8, w^3, 0]. \end{aligned}$$

[8] $$f_{3:3}:\Theta_{3:3}(z)=\begin{bmatrix} 0 & 1 & 0 \\ z & \frac{16}{5} & \frac{11}{5} \\ 0 & 0 & 1 \end{bmatrix}, \qquad \Psi_{3:3}(w)=\begin{bmatrix} -16w/5 & 1 & 0 \\ w & 0 & 0 \\ -11w/5 & 0 & 1 \end{bmatrix}.$$

[9] $$a_{3:4}:X_{4:4}(z)=[-5z^4/8, 7z^4, 6z^4], \qquad Y_{4:4}(w)=[w^4, -5w^4/8, 0].$$

[10] $$c_{4:4}:\Theta_{4:4}(z)=\begin{bmatrix} z/(2\sqrt{2}) & 28/(5\sqrt{2}) & \frac{6}{5} \\ -5z/(16\sqrt{2}) & 1/(2\sqrt{2}) & -\frac{3}{4} \\ 0 & 0 & 1 \end{bmatrix},$$

$$\Psi_{4:4}(w)=\begin{bmatrix} w(2\sqrt{2}) & 5/(16\sqrt{2}) & 0 \\ -28w/(5\sqrt{2}) & 1/(2\sqrt{2}) & 0 \\ -12\sqrt{2}w/5 & 0 & 1 \end{bmatrix}.$$

Operations [11]–[13] are obvious. After evaluating, $c_{3:4}$, $c_{2:4}$, and $c_{0:4}$, we obtain $\Theta_{0:4}(z)$ and $\Psi_{0:4}(w)$, and finally,

[14]
$$\begin{aligned} a_{0:9}: X_{0:9}(z) &= [x_1(z), x_2(z), x_3(z)] \circledast_{F^f} \Theta_{0:4}(z) \\ &= z^5[(-1, -z^3, 0) \circledast_{F^f} \Theta_{0:4}(z)] \\ &= z^5[(-1, -z^3, 0)\Theta_{0:4}(z) \bmod z^5] = z^5[u_1(z), u_2(z), u_3(z)], \end{aligned}$$

where

$$\begin{aligned} u_1(z) &= -z/(2\sqrt{2}) - z^2/(2\sqrt{2}) + z^3/\sqrt{2} + z^4/\sqrt{2}, \\ u_2(z) &= 4/(5\sqrt{2}) + 4z/\sqrt{2} + 16z^2/(5\sqrt{2}) - 28z^3/(5\sqrt{2}) - 28z^4/(5\sqrt{2}), \\ u_3(z) &= 2/5 + z/5 + 2z^2/5 + z^3/5 - 6z^4/5. \end{aligned}$$

$$\begin{aligned} Y_{0:9}(w) &= [y_1(w), y_2(w), y_3(w)] \circledast_{F^b} \Psi_{0:4}(w) \\ &= w^5[(0, 0, 1) \circledast_{F^b} \Psi_{0:4}(w)] = w^5[v_1(w), v_2(w), v_3(w)], \end{aligned}$$

where

$$\begin{aligned} v_1(w) &= -12\sqrt{2}w/5 + 12w^2/(5\sqrt{2}) + 12w^3/(5\sqrt{2}) - 12w^4/(5\sqrt{2}), \\ v_2(w) &= -w/\sqrt{2} + w^2/(2\sqrt{2}) + w^3/(2\sqrt{2}) - w^4/(2\sqrt{2}), \\ v_3(w) &= 1. \end{aligned}$$

Therefore,

$$T^{-1} = L(\mathbf{u}_1)L^T(\mathbf{v}_1) + L(\mathbf{u}_2)L^T(\mathbf{v}_2) + L(\mathbf{u}_3)L^T(\mathbf{v}_3),$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are the vectors whose jth component is the coefficient of z^{j-1} and w^{j-1} of $u_i(z)$ and $v_i(w)$, respectively. □

Remark 3. If we had chosen the displacement operator $F^f = Z_5 \oplus Z_3 \oplus Z_2$, $F^b = Z_3 \oplus Z_2 \oplus Z_5$ for the matrix T in (17) we would have the same generator (19) for E_1, but the obtained generator of T^{-1} would be the one with respect to $\{Z_3 \oplus Z_2, Z_5\}$ rather than with respect to $\{Z_5, Z_5\}$. The displacement ranks of T^{-1} with respect to both displacement operators are two, but the above procedure gives nonminimal generators of length three.

Remark 4. The following extended matrix:

(20)
$$\begin{bmatrix} T & \mathbf{b} \\ -I & 0 \end{bmatrix}, \quad T = \text{Sylvester matrix}$$

also has a displacement rank of three. We could as well obtain the solution $T^{-1}\mathbf{b}$ directly by applying the recursive generalized Schur algorithm to (20); the last column of X, where $\{X, \mathbf{y}\}$ is the computed generator of $T^{-1}\mathbf{b}$ with respect to $\{Z_n, 1\}$, can be shown to be the solution $T^{-1}\mathbf{b}$.

4. Polynomial products with fast convolutions. The product of two polynomials of degree d_1 and d_2 can be performed efficiently using $d \equiv d_1 + d_2 + 1$ point fast cyclic convolution algorithms [4]. A d-point fast cyclic convolution needs $O(d \log d)$ flops. Among others, Fast Fourier Transforms (FFTs) can be used for convolutions, and Ammar and Gragg [2] carefully examined the use of FFTs for a doubling algorithm for square Toeplitz systems of equations. We shall only consider the subtle complications that arise in the recursive generalized Schur algorithm in this paper.

The polynomial matrix-matrix product of (16) needs α^3 of $q-p$ point cyclic convolutions. The polynomial vector-matrix product of (15b) has α^2 of scalar polynomial products of the form, $x(z)\otimes_{F^f}\theta(z)$, where $x(z)$ is a polynomial with nonzero terms of $z^p, z^{p+1}, \cdots, z^q$. Let us assume that

$$0<\delta_1<\cdots<\delta_l\leqq p<\delta_{l+1}<\cdots<\delta_s\leqq r<\delta_{s+1}<\cdots<\delta_t\leqq q<\delta_{t+1}<\cdots<\delta_N.$$

Then

$$x'(z)\equiv x(z)\otimes_{F^f}\theta(z) \tag{21a}$$

$$=[z^{\delta_l}x_{l+1}(z)+z^{\delta_{l+1}}x_{l+2}(z)+\cdots+z^{\delta_s}x_{s+1}(z)$$

$$+\cdots+z^{\delta_t}x_{t+1}(z)]\otimes_{F^f}\theta(z) \tag{21b}$$

$$=[z^{\delta_l}x_{l+1}(z)+\cdots+z^{\delta_s-1}x_s(z)]\otimes_{F^f}\theta(z) \tag{22a}$$

$$+z^{\delta_s}[x_{s+1}(z)\theta(z^\beta) \bmod z^{n_s+1}] \tag{22b}$$

$$+z^{\delta_s+1}[x_{s+2}(z)\theta(z^\beta) \bmod z^{n_s+2}] \tag{22c}$$

$$\cdots$$

$$+z^{\delta_t}[x_{t+1}(z)\theta(z^\beta) \bmod z^{n_t+1}]. \tag{22d}$$

The terms in (22a) do not need to be computed because these terms will be summed to zeros after adding all the partial sums in the vector-matrix multiplication of (15b). Recall that $x_i(z)$ has degree n_i, and $\theta(z^\beta)$ has degree $\beta^{(q-p+1)}$. Therefore, the product $x_i(z)\theta(z^\beta)$ from (22b) to (22d) can be performed by

$2n_i+1$ point cyclic convolutions if degree $[\theta(z^\beta)]\geqq$ degree $[x_i(z)]$,

$n_i+\beta^{(q-p+1)}+1$ point cyclic convolutions if degree $[\theta(z^\beta)]<$ degree $[x_i(z)]$.

Remark 5. Note that two $d/2$ point convolutions take $cd\log(d/2)$ flops if one d point convolution takes $cd\log d$ flops. Therefore, the polynomial product (21) is more efficient for the displacement operator F^f with more sections, because such displacement operators break a long convolution into many smaller convolutions. Therefore, for a given matrix we prefer to choose a displacement operator with as many sections as possible, while keeping the displacement rank minimal. Also we remark that the first and last terms (22b) and (22d) need smaller point convolutions.

If the dimensions of the matrix are powers of two, then we can always choose the center division point $r=\lceil(p+q)/2\rceil$. This *balanced division* (or doubling) gives the least number of computations, in general. For this case, let $\eta\equiv p-q$, and $T(\eta)$ denote the number of computations for one recursion. Then

$$T(\eta)\leqq 2T(\eta/2)+W(\eta),\qquad W(\eta)\equiv O(\alpha^3\eta\log\eta),$$

and therefore, we can show [1] that the k-step recursion takes

$$T(k)\leqq O(\alpha^3 k\log^2 k).$$

However, in most cases the doubling is not possible, and for such circumstances, the desirable choice of r is such that $r-p$ and $q-r+1$ are highly composite numbers (so that fast convolution algorithms can be applied efficiently), as well as r is close to $(q-p)/2$ (so as to achieve balancing).

Matrix-vector products using displacement representation. The final step of finding solutions for linear equations is the matrix-vector multiplication $S\mathbf{b}$, given a displacement representation of $S \in \mathbf{R}^{m\times n}$,

$$S = \sum_{i=1}^{\alpha} K(\mathbf{x}_i, F^f) K^T(\mathbf{y}_i, F^b), \tag{23}$$

where the length α is a multiple of the block size β, $\alpha = \beta\delta$, say, and

$$F^f = \bigoplus_{i=1}^{M} Z^{\beta}_{m_i}, \quad F^b = \bigoplus_{i=1}^{N} Z^{\beta}_{n_i}, \quad m = \sum_{i=1}^{M} m_i, \quad n = \sum_{i=1}^{N} n_i.$$

The expression in (23) can be rewritten in the *block displacement form*

$$S = \sum_{i=1}^{\delta} K_\beta(X_i, F^f) K_\beta^T(Y_i, F^b), \qquad X_i \in \mathbf{R}^{m\times\beta}, \quad Y_i \in \mathbf{R}^{n\times\beta}, \tag{24}$$

where

$$K_\beta(X_i, F^f) = [X_i, F^f X_i, F^{f\,2} X_i, \cdots, F^{f[(m/\beta)-1]} X_i] \in \mathbf{R}^{m\times n}, \tag{25a}$$

$$K_\beta(Y_i, F^b) = [Y_i, F^b Y_i, F^{b2} Y_i, \cdots, F^{b[(n/\beta)-1]} Y_i] \in \mathbf{R}^{n\times n}. \tag{25b}$$

Furthermore, because F^f and F^b have M and N sections, respectively, (25a) and (25b) have the forms

$$K_\beta(X_i, F^f) = \begin{bmatrix} K_\beta(X_{1,i}, Z^{\beta}_{m_1}) & O \\ K_\beta(X_{2,i}, Z^{\beta}_{m_2}) & O \\ \cdot & \cdot \\ K_\beta(X_{M,i}, Z^{\beta}_{m_M}) & O \end{bmatrix}, \qquad K_\beta(Y_i, F^b) = \begin{bmatrix} K_\beta(Y_{1,i}, Z^{\beta}_{n_1}) & O \\ K_\beta(Y_{2,i}, Z^{\beta}_{n_2}) & O \\ \cdot & \cdot \\ K_\beta(Y_{N,i}, Z^{\beta}_{n_N}) & O \end{bmatrix},$$

where $K_\beta(X, Z^\beta)$ is the block lower triangular Toeplitz matrix with the first column block X. The matrix O denotes a null matrix of appropriate size such that $K_\beta(X_i, F^f)$ and $K_\beta(Y_i, F^b)$ are $m \times n$ and $n \times n$ matrices, respectively.

To see how to use convolutions for the product

$$K_\beta(X_i, F^f) K_\beta^T(Y_i, F^b)\mathbf{b}$$

it is enough to consider matrix-vector multiplications of the form $K_\beta(X, Z^\beta)\mathbf{b}$. Note that $K_\beta(X, Z^\beta)\mathbf{b}$ can be expressed as sum of β products of scalar lower triangular Toeplitz matrix and vectors. As an example,

$$\begin{bmatrix} a_0 & c_0 & & \\ a_1 & c_1 & & \\ a_2 & c_2 & a_0 & c_0 \\ a_3 & c_3 & a_1 & c_1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} a_0 & & & \\ a_1 & a_0 & & \\ a_2 & a_1 & a_0 & \\ a_3 & a_2 & a_1 & a_0 \end{bmatrix} \begin{bmatrix} b_0 \\ 0 \\ b_2 \\ 0 \end{bmatrix} + \begin{bmatrix} c_0 & & & \\ c_1 & c_0 & & \\ c_2 & c_1 & c_0 & \\ c_3 & c_2 & c_1 & c_0 \end{bmatrix} \begin{bmatrix} b_1 \\ 0 \\ b_3 \\ 0 \end{bmatrix}. \tag{26}$$

The multiplications in the right sides of (26) can be done by fast convolutions, and therefore, so can the multiplication $S\mathbf{b}$.

5. Concluding remarks. We have presented $O(\alpha^3 n \log^2 n)$ algorithms for the determination of exact and least squares solutions of linear systems with matrices having

(generalized) displacement rank α. Such algorithms for exact solutions have been studied by several authors, most recently by Ammar and Gragg [2] for Toeplitz systems. They also made a very close study of the implementation of the convolution operation in an attempt to obtain the smallest coefficient. Although we have not attempted so close an operation count for the more general algorithm in our paper, the hidden constant in the operation counts for solving Toeplitz least squares problems is quite high because $\alpha = 4$ for the matrices E_2 or E_3 (see (3)) with a full rectangular T. Also we conjecture that our algorithm suffers numerical stability problem when $E_{1,1}$ in (1) has a leading principal submatrix that is close to singular; nevertheless we might hope that numerical refinements devised for the Schur algorithm (see, e.g., Koltracht and Lancaster [18]) may be carried over to the divide-and-conquer framework as well.

We also mention that the fast algorithms for Hankel and close-to-Hankel matrices in [10] can be implemented with divide-and-conquer fashion using the spirit in this paper.

Appendix. We shall summarize the explanation in § 3 using a Pascal-like recursive procedure. First, note that the polynomial $\Theta_{p:q}(z)$ (and $\Psi_{p:q}(z)$) has $q - p + 2$ terms. The first column of $\Theta_{p:q}(z)$ has terms ranging from degree z to z^{q-p+1}, and the other columns have terms from 1 to z^{q-p}. Hence, by shifting the first column by one position, we can store $\Theta_{p:q}(z)$ and $\Psi_{p:q}(z)$ in the array "Poly" from p to q slots inclusive:

```
Poly: array [1..α, 1..α, 0..MAX-1] of record
        θ: coefficients;
        ψ: coefficients
      end;
```

The computation of $\Theta_{p:q}(z)$ is sequential, i.e., once we compute $\Theta_{p:q}(z)$, we do not need to keep $\Theta_{p:r-1}(z)$, and therefore, the array "Poly" can be kept as a single global variable.

The polynomial vector $X_{p:q}(z)$ has $q - p + 1$ terms, and therefore, can be stored in an array type GENERATORS:

```
type
    GENERATORS = array [1..α, 0..MAX-1] of record
        x: coefficient;
        y: coefficient
    end
```

However, $X_{p:q}(z)$ cannot be kept as a global variable, and local copies should be maintained until we compute $X_{r:q}(z)$.

Now we can describe the recursive generalized Schur algorithm as follows.

```
ALGORITHM (recursive k-step generalized Schur algorithm).
Input: Generator of E, {X_0(z), Y_0(w)}; displacement operator {⊗_{F^f}, ⊗_{F^b}};
       Number of steps, k.
Output: Generator of S, {X_k(z), Y_k(w)};
procedure RecursiveSchur
    var
        G, LowerG: GENERATORS;
    begin
        Find(0, k-1, G);
        Apply(0, k, n, G, LowerG);
        return(LowerG)
    end
```

The procedure Find (p, q, G) computes $\Theta_{p:q}(z)$, and $\Psi_{p:q}(w)$ given $\{X_{p:q}(z), Y_{p:q}(w)\}$, and the procedure Apply $(p, r, q, G, \text{Lower}G)$ returns Lower$G$ = $\{X_{r:q}(z), Y_{r:q}(w)\}$ given $G = \{X_{p:q}(z), Y_{p:q}(w)\}$

```
procedure Find(p, q: index; G: GENERATORS);
    var
        r: index;
        G, LowerG: GENERATORS;
    begin
        if p = q then begin
            Compute Θ_{p:q}(z) and Ψ_{p:q}(w);
            return
        end
    r := appropriate integer close to ⌈(p+q)/2⌉;
    Find(p. r-1, G);
    Apply(p, r, q, G, LowerG);
    Find(r, q, LowerG);
    (* Use fast convolution for polynomial products *)
    Θ_{p:q}(z) := Θ_{p:r-1}(z)Θ_{r,q}(z);
    Ψ_{p:q}(w) := Ψ_{p:r-1}(w)Ψ_{r,q}(w)
end

procedure Apply (p, r, q: index; G: GENERATORS; var LowerG: GENERATORS);
    begin
        (* Use fast convolution for polynomial products *)
        X_{r:q}(z) := X_{p:q}(z)⊛_{F^f} Θ_{p:r-1}(z);
        Y_{r:q}(w) := Y_{p:q}(w)⊛_{F^b} Ψ_{p:r-1}(w);
        LowerG := {X_{r:q}(z), Y_{r:q}(w)}
        Free the storage of {X_{p:q}(z), Y_{p:q}(w)};
        return (LowerG);
    end
```

REFERENCES

[1] A. AHO, J. HOPCROFT, AND J. ULLMAN, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974, p. 305.

[2] G. AMMAR AND W. GRAGG, *Superfast solution of real positive definite Toeplitz systems*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 61–76.

[3] G. BITMEAD AND B. D. O. ANDERSON, *Asymptotically fast solution of Toeplitz and related systems of linear equations*, Linear Algebra Appl., 34 (1980), pp. 103–116.

[4] R. BLAHUT, *Fast Algorithms for Digital Signal Processing*, Addison-Wesley, Reading, MA, 1985.

[5] R. BRENT, F. GUSTAVSON, AND D. YUN, *Fast solution of Toeplitz systems of equations and computation of Padé approximants*, J. Algorithms, 1 (1980), pp. 259–295.

[6] A. BRUCKSTEIN AND T. KAILATH, *Doing inverse scattering the fast(est) way*, Tech. Report, Department of Electrical Engineering, Stanford University, Stanford, CA, July 1985.

[7] W. CHOATE, *A fast algorithm for normal incidence seismograms*, Geophysics, 47 (1982), pp. 196–202.

[8] J. CHUN, *Fast array algorithms for structured matrices*, Ph.D. thesis, Department of Electrical Engineering, Stanford University, Stanford, CA, June 1989.

[9] J. CHUN AND T. KAILATH, *Generalized displacement structure for block-Toeplitz, Toeplitz-block and Toeplitz-derived matrices*, NATO Conference on Signal Processing, Leuven, Belgium, August, 1988.

[10] ———, *Displacement structure for Hankel, Vandermonde and related matrices*, IMA Conference, Minneapolis, MN, June, 1988.

[11] F. DE HOOG, *A new algorithm for solving Toeplitz systems of equations*, Linear Algebra Appl., 88/89 (1987), pp. 122–138.

[12] I. GOHBERG AND I. FEL'DMAN, *Convolution Equations and Projection Methods for Their Solutions*, Trans. Math. Monographs, Vol. 41, American Mathematical Society, Providence, RI, 1974.

[13] I. GOHBERG AND A. SEMENCUL, *On the inversion of finite Toeplitz matrices and their continuous analogs*, Mat. Issled., 2 (1972), pp. 201–233.

[14] T. KAILATH, *Signal processing applications of some moment problems*, Proc. Sympos. Appl. Math., 37 (1987), pp. 71–109.

[15] T. KAILATH AND J. CHUN, *Generalized Gohberg–Semencul formulas for matrix inversion*, in the Gohberg Anniversary Collection, Vols. I and II, H. Dym, S. Goldberg, M. Kaashoek, P. Lancaster, eds., Birkhäuser-Verlag, Basel, Switzerland, 1989.

[16] T. KAILATH, S. KUNG, AND M. MORF, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407; see also Bull. Amer. Math. Soc., 1 (1979), pp. 769–773.

[17] T. KAILATH, A. VIEIRA, AND M. MORF, *Inverses of Toeplitz operators, innovations, and orthogonal polynomial*, SIAM Rev., 20 (1978), pp. 106–119.

[18] I. KOLTRACHT AND P. LANCASTER, *Threshold algorithms for the prediction of reflection coefficients in a layered medium*, Geophysics, 53 (1988), pp. 908–919.

[19] H. LEV-ARI AND T. KAILATH, *Triangular Factorization of Structured Hermitian Matrices*, Operator Theory, Advances and Applications, Vol. 18, Birkhäuser, Boston, 1986, pp. 301–324.

[20] W. MCCLARY, *Fast seismic inversion*, Geophysics, 48 (1983), pp. 1371–1372.

[21] M. MORF, *Doubling algorithms for Toeplitz and related equations*, in Proc. IEEE Internat. Conference on Acoustics, Speech and Signal Processing, Denver, CO, 1980, pp. 954–959.

[22] B. MUSICUS, *Levinson and fast Cholesky algorithms for Toeplitz and almost Toeplitz matrices*, Res. Report, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, 1981.

[23] I. SCHUR, *Über Potenzreihen, die im Innern des Einheitskreises beschrankt sind*, J. Reine Angew. Math., 147 (1917), pp. 205–232.

[24] ———, *On Power Series Which Are Bounded in the Interior of the Unit Circle*. 1, Operator Theory, Advances and Applications, Vol. 18, Birkhäuser, Boston, (1986), pp. 31–60. (English translation of [23].)

[25] H. SEXTON, M. SHENSA, AND J. SPEISER, *Remarks on a displacement-rank inversion method for Toeplitz systems*, Linear Algebra Appl., 45 (1982), pp. 127–130.

[26] W. TRENCH, *An algorithm for inversion of finite Toeplitz matrices*, J. Soc. Indust. Appl. Math., 12 (1964), pp. 515–522.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 1, pp. 146–159, January 1991

011

INERTIA, NUMERICAL RANGE, AND ZEROS OF QUADRATIC FORMS FOR MATRIX PENCILS*

NAM-KIU TSING† AND FRANK UHLIG‡

Abstract. Definite, semidefinite, and indefinite Hermitian and symmetric matrix pencils $P(A, B)$ are classified by their $l_{\mathbf{C}}$ and $l_{\mathbf{R}}$ numbers where $l_{\mathbf{F}} = \dim \text{span}\{X \in \mathbf{F}^n: x^*Ax = x^*Bx = 0\}$. Using ideas from numerical range theory, it is proved for $\mathbf{F} = \mathbf{C}$ that $P(A, B)$ is a definite pencil if and only if $l_{\mathbf{C}} = 0$, $P(A, B)$ is an indefinite pencil if and only if $l_{\mathbf{C}} = n$, while $P(A, B)$ is a semidefinite pencil if and only if $0 < l_{\mathbf{C}} < n$. In contrast, for $\mathbf{F} = \mathbf{R}$ the $l_{\mathbf{R}}$ number for indefinite pencils can be as low as $n - 2$. In the cases for $\mathbf{F} = \mathbf{R}$ with $l_{\mathbf{R}} = n - 2$ or $n - 1$, the Kronecker canonical form theory is used to describe sets of generators for indefinite and semidefinite pencils $P(A, B)$.

Key words. matrix pencil, inertia, numerical range, Kronecker form

AMS(MOS) subject classifications. 15A48, 15A42, 15A57, 15A63, 15A60

1. Introduction. Let H_n be the set of all $n \times n$ Hermitian matrices, and let S_n be the set of all $n \times n$ real symmetric matrices. For any pair of A, B in H_n or S_n, the *pencil generated by A and B* is the set

$$P(A,B) = \{aA + bB: a, b \in \mathbf{R}\}.$$

$P(A, B)$ is called a *definite pencil* (d.*pencil*) if it contains a definite matrix; it is a *semidefinite pencil* (s.d.*pencil*) if it contains no definite matrix but contains a nonzero semidefinite matrix; it is an *indefinite pencil* (i.*pencil*) if all its nonzero elements are indefinite. In particular, if $A = B = 0$, then $P(A, B) = \{0\}$ is an i.pencil.

Let X^* denote the conjugate transpose of X if X is a complex vector or matrix, and the transpose of X, i.e., X^t, if X is a real vector or matrix. Let $\mathbf{F}$ stand for either $\mathbf{C}$ or $\mathbf{R}$. For any $A, B \in H_n$, or S_n we define

$$l_{\mathbf{F}}(A,B) = \dim(\text{span}\{x \in \mathbf{F}^n: x^*Ax = x^*Bx = 0\}),$$

and the $\mathbf{F}$-numerical range of A and B by

$$W_{\mathbf{F}}(A,B) = \{(x^*Ax, x^*Bx): x \in \mathbf{F}^n, x^*x = 1\}.$$

The number $l_{\mathbf{F}}(A, B)$ may be regarded as a measure of the "size" of the set of vectors $x \in \mathbf{F}^n$ which are annihilated simultaneously by the Hermitian forms A and B. As the unit sphere $\{x \in \mathbf{F}^n: x^*x = 1\}$ is compact, it follows that $W_{\mathbf{F}}(A, B)$ is a compact subset in $\mathbf{R}^2$.

In this paper we want to explore the relationship between the inertia of a matrix pencil $P(A, B)$, the associated numbers $l_{\mathbf{C}}(A, B)$ or $l_{\mathbf{R}}(A, B)$, and properties of the field of values $W_{\mathbf{F}}(A, B) \subset \mathbf{R}^2$. Definite matrix pencils have been studied for over 50 years. The first classification in 1936/1937 is due to Finsler [Fi]; we refer the reader to the survey [Uhe] for the history of this subject. The field of values or numerical range of a matrix has been studied for 70 years since [To] and [Ha], while the l-numbers were originally introduced in 1973 in [Uhb]–[Uhd].

* Received by the editors November 21, 1988; accepted for publication (in revised form) November 28, 1989.

† Systems Research Center, University of Maryland, College Park, Maryland 20742 (tsing@cacse.src.umd.edu).

‡ Department of Algebra, Combinatorics and Analysis, Auburn University, Auburn, Alabama 36849-5307 (fuhlig@auducvax.bitnet).

In § 2 we shall first obtain some auxiliary results, and then deal with the case $\mathbf{F} = \mathbf{C}$ completely to obtain a very clear result (Theorem 2.4) for the inertia of Hermitian matrix pencils and their $l_{\mathbf{C}}$ numbers: definite pencils can only have $l_{\mathbf{C}} = 0$, semidefinite pencils can have any $l_{\mathbf{C}}$ number greater than zero and less than n, while indefinite pencils have $l_{\mathbf{C}} = n$.

Section 3 deals with the case $\mathbf{F} = \mathbf{R}$ and $A, B \in S_n$. The results here become more complicated due to the fact (Theorem 3.2) that when $n \geqq 3$, indefinite real symmetric pencils can have $l_{\mathbf{R}}$ numbers between $n-2$ and n, while the $l_{\mathbf{R}}$ numbers of semidefinite pencils are bounded by 1 and $n-1$ as in the Hermitian case. The "overlapping region," $l_{\mathbf{R}} = n-2, n-1$, for s.d. or i.pencils, will be investigated in § 4 where we will use the Kronecker pair form to determine sets of generators for the four overlapping cases if A, $B \in S_n$.

2. Some auxiliary results and the complex case. We shall first develop some of the properties of numerical ranges for mixed base fields in Lemma 2.1, and relate the position of $(0, 0) \in \mathbf{R}^2$ with regard to $W_{\mathbf{F}}$ to the inertia of the pencil as well.

LEMMA 2.1. *Let $A, B \in H_n$. Then*

(a) $W_{\mathbf{C}}(A, B)$ *is convex*;

(b) $W_{\mathbf{R}}(A, B)$ *is convex if* $n \neq 2$;

(c) $W_{\mathbf{R}}(A, B)$ *is a (possibly degenerate) ellipse in* $\mathbf{R}^2$ *if* $n = 2$.

Proof. Part (a) of the lemma follows from the well-known Toeplitz–Hausdorff theorem [Ha], [To], which asserts that the set $\{x^*(A + iB)x\colon x \in \mathbf{C}^n, x^*x = 1\}$ is a convex subset of $\mathbf{C}$. For $A, B \in S_n$, part (b) was proved by Brickman [Br]. In general, if A is Hermitian and $x \in \mathbf{R}^n$, then

$$x^tAx = \{x^tAx)^t = x^tA^tx = x^t\bar{A}x,$$

where $\bar{A}$ is the complex conjugate of A. Note that

$$\tilde{A} = (A + \bar{A})/2$$

is real symmetric and that

$$W_{\mathbf{R}}(A, B) = W_{\mathbf{R}}(\tilde{A}, \tilde{B}).$$

Hence (b) is true for Hermitian A and B also.

If $n = 2$, by letting $x = \cos\theta e_1 + \sin\theta e_2$, where $\theta \in \mathbf{R}$ and $\{e_1, e_2\}$ is the standard basis for $\mathbf{R}^2$, we see that

$$W_{\mathbf{R}}(A, B) = \{\cos 2\theta(a_1, a_2) + \sin 2\theta(b_1, b_2) + (c_1, c_2)\colon \theta \in \mathbf{R}\}, \quad \text{where}$$

$$(a_1, a_2) = [(e_1^tAe_1, e_1^tBe_1) - (e_2^tAe_2, e_2^tBe_2)]/2,$$

$$(b_1, b_2) = [(e_1^tAe_2, e_1^tBe_2) + (e_2^tAe_1, e_2^tBe_1)]/2,$$

$$(c_1, c_2) = [(e_1^tAe_1, e_1^tBe_1) + (e_2^tAe_2, e_2^tBe_2)]/2.$$

Hence $W_{\mathbf{R}}(A, B)$ is an ellipse in $\mathbf{R}^2$. □

Let $A, B \in H_n$ where $n \neq 2$ if $\mathbf{F} = \mathbf{R}$. By Lemma 2.1, $W_{\mathbf{F}}(A, B)$ is either a convex set with nonempty interior, or a line segment, or a point in $\mathbf{R}^2$. If $W_{\mathbf{F}}(A, B)$ has nonempty interior, we use $\partial W_{\mathbf{F}}(A, B)$ to denote its topological boundary. If $W_{\mathbf{F}}(A, B)$ is a line segment, its two endpoints will form the set $\partial W_{\mathbf{F}}(A, B)$. If $W_{\mathbf{F}}(A, B)$ is a single point, we define $\partial W_{\mathbf{F}}(A, B)$ to be the empty set. In all cases, we define

$$\text{int } W_{\mathbf{F}}(A, B) = W_{\mathbf{F}}(A, B) \backslash \partial W_{\mathbf{F}}(A, B).$$

The following theorem relates the inertia of the pencil to the position of $(0, 0) \in \mathbf{R}^2$ with respect to $W_{\mathbf{F}}$.

THEOREM 2.2. *Let $A, B \in H_n$ if $F = \mathbf{C}$, and $A, B \in S_n$ with $n \neq 2$ if $\mathbf{F} = \mathbf{R}$. Then*

(a) *$P(A, B)$ is a* d.*pencil if and only if* $(0, 0) \notin W_{\mathbf{F}}(A, B)$;

(b) *$P(A, B)$ is a* s.d.*pencil if and only if* $(0, 0) \in \partial W_{\mathbf{F}}(A, B)$;

(c) *$P(A, B)$ is an* i.*pencil if and only if* $(0, 0) \in \operatorname{int} W_{\mathbf{F}}(A, B)$.

Proof. We can use a Hahn–Banach style argument which was first introduced to this problem by Taussky [Ta]: By (a) and (b) of Lemma 2.1, $W_{\mathbf{F}}(A, B)$ is convex. Therefore, if $(0, 0) \notin W_{\mathbf{F}}(A, B)$, there is a straight line in $\mathbf{R}^2$ with equation $ax_1 + bx_2 = c$ where $c > 0$, which separates $(0, 0)$ and $W_{\mathbf{F}}(A, B)$. Consequently, we have

$$x^*(aA + bB)x = ax^*Ax + bx^*Bx > c > 0$$

for all $x \in \mathbf{F}^n$ with $x^*x = 1$. This means $(aA + bB)$ is positive definite, and hence $P(A, B)$ is a d.pencil. Conversely, if $P(A, B)$ is a d.pencil, we may reverse the argument and get $(0, 0) \notin W_{\mathbf{F}}(A, B)$.

If $(0, 0) \in \partial W_{\mathbf{F}}(A, B)$, then there exists a straight line in $\mathbf{R}^2$ with equation $ax_1 + bx_2 = 0$, such that $W_{\mathbf{F}}(A, B)$ is on one side of the line and $W_{\mathbf{F}}(A, B)$ is not entirely contained in the line (because we have assumed that $W_{\mathbf{F}}(A, B)$ cannot be a single point in $\mathbf{R}^2$ in this case). Hence we may assume that

$$x^*(aA + bB)x \geqq 0 \quad \text{for all } x \in \mathbf{F}^n \text{ with } x^*x = 1. \tag{1}$$

Since $P(A, B)$ is not a d.pencil (because $(0, 0) \in W_{\mathbf{F}}(A, B)$), $(aA + bB)$ is nonzero positive semidefinite by (1). Therefore $P(A, B)$ is a s.d.pencil. By reversing the argument, we get the converse.

Since (a) and (b) hold, by the principle of exhaustion, (c) must hold also. □

If $A, B \in H_2$, then $W_{\mathbf{R}}(A, B)$ is a (possibly degenerate) ellipse in $\mathbf{R}^2$ by Lemma 2.1(c). For any subset S of R^2, denote the *convex hull* of S by conv S. Then conv $W_{\mathbf{R}}(A, B)$ is a (possibly degenerate) elliptical disc. We define ∂ conv $W_{\mathbf{R}}(A, B)$ and int conv $W_{\mathbf{R}}(A, B)$ for the set conv $W_{\mathbf{R}}(A, B)$ in the same manner as for the set $W_{\mathbf{F}}(A, B)$ above. Using Taussky's idea again, we have Theorem 2.3.

THEOREM 2.3. *Let $A, B \in S_2$. Then*

(a) *$P(A, B)$ is a* d.*pencil if and only if* $(0, 0) \notin \operatorname{conv} W_{\mathbf{R}}(A, B)$;

(b) *$P(A, B)$ is a* s.d.*pencil if and only if* $(0, 0) \in \partial \operatorname{conv} W_{\mathbf{R}}(A, B)$;

(c) *$P(A, B)$ is an* i.*pencil if and only if* $(0, 0) \in \operatorname{int} \operatorname{conv} W_{\mathbf{R}}(A, B)$.

Now we use the value $l_{\mathbf{C}}(A, B)$ to characterize the pencil $P(A, B)$.

THEOREM 2.4. *Let $A, B \in H_n$. Then*

(a) *$P(A, B)$ is a* d.*pencil if and only if* $l_{\mathbf{C}}(A, B) = 0$;

(b) *$P(A, B)$ is a* s.d.*pencil if and only if* $0 < l_{\mathbf{C}}(A, B) < n$;

(c) *$P(A, B)$ is an* i.*pencil if and only if* $l_{\mathbf{C}}(A, B) = n$.

Proof. (a) This follows from the definition of $l_{\mathbf{C}}(A, B)$ and Theorem 2.2.

(c) By Theorem 2.2, $P(A, B)$ is an i.pencil if and only if $(0, 0) \in \operatorname{int} W_{\mathbf{C}}(A, B)$. Suppose $W_{\mathbf{C}}(A, B)$ is a single point. Then $(0, 0) \in \operatorname{int} W_{\mathbf{C}}(A, B) = W_{\mathbf{C}}(A, B)$ if and only if $A = B = 0$ and hence $l_{\mathbf{C}}(A, B) = n$. If $W_{\mathbf{C}}(A, B)$ is a line segment, then $(0, 0) \in \operatorname{int} W_{\mathbf{C}}(A, B)$ if and only if $(0, 0)$ is not an extreme point of $W_{\mathbf{C}}(A, B)$ and $W_{\mathbf{C}}(A, B) \subset L$ where L is the supporting line of $W_{\mathbf{C}}(A, B)$ at $(0, 0)$. If $W_{\mathbf{C}}(A, B)$ has nonempty interior, then $(0, 0) \in \operatorname{int} W_{\mathbf{C}}(A, B)$ if and only if $(0, 0)$ is an interior point of $W_{\mathbf{C}}(A, B)$. In the above two situations, we can apply a result of Embry [Em, Thm. 1] to show that the conditions are equivalent to $l_{\mathbf{C}}(A, B) = n$.

Since (a) and (c) hold, by the principle of exhaustion, (b) must hold also. □

3. The real case. As shown in Theorem 2.4, the number $l_{\mathbf{C}}(A, B)$ can be used as an indicator for the definiteness of the Hermitian pencil $P(A, B)$. In this section, we consider real symmetric pencils and the number $l_{\mathbf{R}}(A, B)$ instead. Clearly,

$$0 \leqq l_{\mathbf{R}}(A,B) \leqq l_{\mathbf{C}}(A,B).$$

We shall deal with the two-dimensional case first.

THEOREM 3.1. *Let* $A, B \in S_2$. *Then*

(a) $l_{\mathbf{R}}(A, B) = 0$ *if* $P(A, B)$ *is a* d.*pencil*;

(b) $l_{\mathbf{R}}(A, B) = 1$ *if and only if* $P(A, B)$ *is a* s.d.*pencil*;

(c) $l_{\mathbf{R}}(A, B) = 0$ *or* 2 *if* $P(A, B)$ *is an* i.*pencil*.

Proof. (a) If $P(A, B)$ is a d.pencil then $l_{\mathbf{C}}(A, B) = 0$ by Theorem 2.4. Hence $l_{\mathbf{R}}(A, B) = 0$.

(c) Suppose $P(A, B)$ is an i.pencil. Then by Theorem 2.3,

$$(0,0) \in \operatorname{int} \operatorname{conv} W_{\mathbf{R}}(A,B).$$

If $W_{\mathbf{R}}(A, B)$ is a single point, then clearly the above will imply $(0, 0) = W_{\mathbf{R}}(A, B)$ and hence $l_{\mathbf{R}}(A, B) = 2$. If $W_{\mathbf{R}}(A, B)$ is a line segment then, since $(0, 0) \in W_{\mathbf{R}}(A, B)$, $W_{\mathbf{R}}(A, B)$ is contained in some straight line in $\mathbf{R}^2$ with equation $ax + by = 0$. Therefore $x^T(aA + bB)x = 0$ for all $x \in \mathbf{R}^n$ with $x^Tx = 1$. It then follows that $aA + bB = 0$ and hence A and B are linearly dependent. We may therefore assume $B = 0$ and $A \neq 0$. As A is indefinite, let $-a$, b $(a, b > 0)$ be the eigenvalues of A, and let $u_1, u_2 \in \mathbf{R}^2$ be the corresponding orthonormal eigenvectors. Then $x^tAx = 0$ where

$$x = b^{-1/2}u_1 \pm a^{-1/2}u_2.$$

Hence $l_{\mathbf{R}}(A, B) = 2$. If $W_{\mathbf{R}}(A, B)$ is a nondegenerate ellipse, then $(0, 0) \notin W_{\mathbf{R}}(A, B)$. Hence $l_{\mathbf{R}}(A, B) = 0$.

(b) Suppose $P(A, B)$ is a s.d.pencil. Then $(0, 0) \in \partial \operatorname{conv} W_{\mathbf{R}}(A, B) \subset W_{\mathbf{R}}(A, B)$ and hence $l_{\mathbf{R}}(A, b) \geqq 1$. Note that $W_{\mathbf{R}}(A, B)$ cannot be a single point in this case. Therefore there is a supporting line L, with equations $ax + by = 0$, to $W_{\mathbf{R}}(A, B)$ at $(0, 0)$, and $W_{\mathbf{R}}(A, B) \not\subset L$. As a result, $aA + bB$ is a semidefinite matrix with zero as an eigenvalue of multiplicity one. Therefore

$$1 \leqq l_{\mathbf{R}}(A,B) \leqq \dim\,(\operatorname{span}\,\{x \in \mathbf{R}^n : x^t(aA+bB)x = 0\}) = 1.$$

Conversely, if $l_{\mathbf{R}}(A, B) = 1$, then by (a) and (c) and the exhaustion principle, $P(A, B)$ must be a s.d.pencil. □

THEOREM 3.2. *Let* $A, B \in S_n$, *where* $n \neq 2$. *Then*

(a) $l_{\mathbf{R}}(A, B) = 0$ *if and only if* $P(A, B)$ *is a* d.*pencil*;

(b) $0 < l_{\mathbf{R}}(A, B) < n$ *if* $P(A, B)$ *is a* s.d.*pencil*;

(c) $\max\{1, n - 2\} \leqq l_{\mathbf{R}}(A, B) \leqq n$ *if* $P(A, B)$ *is an* i.*pencil*.

Proof. It is obvious that the theorem is true if $n = 1$. Hence in view of Theorem 3.1 we consider only the case $n \geqq 3$.

Part (a) follows from Theorem 2.2(a) and the definition of $l_{\mathbf{R}}$.

(b) Suppose $P(A, B)$ is a s.d.pencil. Then $(0, 0) \in \partial W_{\mathbf{R}}(A, B) \subset W_{\mathbf{R}}(A, B)$ by Theorem 2.2, and hence $1 \leqq l_{\mathbf{R}}(A, B)$. As $W_{\mathbf{R}}(A, B)$ cannot be a single point in this case, we may follow the argument in the proof of Theorem 3.1(b) to conclude that, for some $a, b, \in \mathbf{R}$, $(aA + bB)$ is a nonzero semidefinite matrix. Hence

$$1 \leqq l_{\mathbf{R}}(A,B) \leqq \dim\,(\operatorname{span}\,\{x \in \mathbf{R}^n : x^t(aA+bB)x = 0\}) \leqq n-1.$$

(c) Now suppose $P(A, B)$ is an i.pencil. By Theorem 2.2(c) and Lemma 2.1(b), we have

$$(0,0)\in\operatorname{int} W_{\mathbf{R}}(A,B)\subset W_{\mathbf{R}}(A,B).$$

Therefore $l_{\mathbf{R}}(A, B) \geqq 1$. Hence (c) holds if $n = 3$. Suppose $n \geqq 4$. If $W_{\mathbf{R}}(A, B)$ is a single point, then $l_{\mathbf{R}}(A, B) = n$. Hence we may assume $W_{\mathbf{R}}(A, B)$ is not a single point, and we want to show that $l_{\mathbf{R}}(A, B) \geqq n - 2$. In fact, if $(0, 0) \in \operatorname{int} W_{\mathbf{R}}(A, B)$ and $l_{\mathbf{R}}(A, B) = m \leqq n - 3$, let $u_1, \cdots, u_m$ be linearly independent vectors in $\mathbf{R}^n$ such that

$$u_j^t A u_j = u_j^t B u_j = 0 \quad \text{for } j = 1, \cdots, m.$$

Let $V = \{u_1, \cdots, u_m\}^{\perp}$, where the orthogonal complement is taken in $\mathbf{R}^n$. Then $\dim V \geqq 3$, and $u \in V^{\perp}$ whenever $u \in \mathbf{R}^n$ satisfies

$$u^t A u = u^t B u = 0.$$

It follows that $W_{\mathbf{R}}|_V(A, B)$, i.e., the $\mathbf{R}$-numerical range of A and B when restricted on the subspace V, which is defined by $W_{\mathbf{R}}|_V(A, B) = \{(x^t A x, x^t B x): x \in V, x^t x = 1\}$, does not contain $(0, 0)$. Let v, v_1, v_2 be three linearly independent unit vectors in V. Denote the point $(v^t A v, v^t B v)$ in $W_{\mathbf{R}}|_V(A, B)$ by η. Then $\eta \neq (0, 0)$. Since $(0, 0) \in \operatorname{int} W_{\mathbf{R}}(A, B)$ and $W_{\mathbf{R}}(A, B)$ is convex, there exists a chord $[\eta_1, \eta_2]$ (in $\mathbf{R}^2$) of $W_{\mathbf{R}}(A, B)$ which passes through $(0, 0)$ and η, such that $(0, 0)$ lies between η and η_1, and $\eta_1 \neq (0, 0)$. Let w be a unit vector in $\mathbf{R}^n$ such that $(w^t A w, w^T B w) = \eta_1$. Clearly, $w \notin V$ (otherwise $W_{\mathbf{R}}|_V(A, B)$ contains both η and η_1, and hence $(0, 0)$ also, as $W_{\mathbf{R}}|_V(A, B)$ is convex for $\dim V \geqq 3$). Let $U = \operatorname{span}\{v, w\}$. Then $\dim U = 2$, and by Lemma 2.1(c), $W_{\mathbf{R}}|_U(A, B)$ is a (possibly degenerate) ellipse in $\mathbf{R}^2$ which contains the two distinct points η and η_1. We consider two cases.

Case 1. If $W_{\mathbf{R}}|_U(A, B)$ degenerates into a line segment, then it must contain $(0, 0)$ (which lies between η and η_1). Since $\dim U = 2$, and $(0, 0) \in \operatorname{int} W_{\mathbf{R}}|_U(A, B) \subset W_{\mathbf{R}}|_U(A, B)$ in this case, by Theorem 2.2(c) and Theorem 3.1(c), there exist two linearly independent vectors w_1 and w_2 in U, such that

$$w_i^t A w_i = w_i^t B w_i = 0 \quad \text{for } i = 1, 2.$$

Thus $w_1, w_2 \in V^{\perp}$. Since $v \in V$, w_1 and w_2 are orthogonal to v. But this is impossible, as w_1, w_2, and v are in U and $\dim U = 2$.

Case 2. Suppose $W_{\mathbf{R}}|_U(A, B)$ is a nondegenerate ellipse. Then $(0, 0) \notin W_{\mathbf{R}}|_U(A, B)$. Let $i = 1$ or 2 and define $U_i = \operatorname{span}\{v_i, v, w\}$. Note that $\dim U_i = 3$, and hence $W_{\mathbf{R}}|_{U_i}(A, B)$ is convex and contains $(0, 0)$. It follows that there exists a nonzero vector w_i in U_i such that

$$w_i^t A w_i = w_i^t B w_i = 0.$$

The vectors w_1 and w_2 must be linearly independent (otherwise $w_1 \in U_1 \cap U_2 = \operatorname{span}\{v, w\} = U$, and hence $(0, 0) \in W_R|_U(A, B)$, a contradiction). However, this is impossible, since

$$w_1, w_2 \in \operatorname{span}\{v_1, v_2, v, w\}$$

and

$$w_1, w_2 \in V^{\perp} \subset \{v_1, v_2, v\}^{\perp}.$$

Hence we must have $l_{\mathbf{R}}(A, B) \geqq n - 2$ if $P(A, B)$ is an i.pencil. □

From the above, we see that if $A, B \in S_n$ where $n = 2$, and $l_{\mathbf{R}}(A, B) = 0$, then $P(A, B)$ can either be a d.pencil or an i.pencil. Also, when $n \geqq 3$ and $n - 2 \leqq l_{\mathbf{R}}(A, B) \leqq n - 1$, then $P(A, B)$ can either be a s.d.pencil or an i.pencil. In such cases, we need further information to determine the inertia of $P(A, B)$. If for any pair of $A, B \in S_n$, we are in one of the "overlapping cases" for which the number $l_{\mathbf{R}}(A, B)$ alone does not determine the inertia of the pencil $P(A, B)$, we will find a set of sparse generators C, D for $P(A, B)$ in § 4. There we are interested in finding some canonical forms, up to congruences, for the generators of those pencils whose $l_{\mathbf{R}}$ number alone cannot determine their inertia.

4. The canonical pair form approach. Originally, classifying definite, semidefinite, and indefinite real symmetric matrix pencils via their $l_{\mathbf{R}}$ numbers was done for nonsingular pairs A and B (A nonsingular) by use of the real canonical pair form (see, e.g., Uhlig [Uha]) in [Uhb]–[Uhd]. Here we shall use the Kronecker canonical pair form as described in Gantmacher [Ga] or Thompson [Th], [Th1], for example, for not necessarily nonsingular pairs A and B. Our proofs will rely on our original work in [Uhb]–[Uhd]. The numerical computation of the Kronecker form has recently become important in numerical methods for linear control theory (see, e.g., Van Dooren [VDo]).

Kronecker canonical form for a pair of real symmetric matrices A, B: A pair of real symmetric matrices A and B is simultaneously congruent over $\mathbf{R}$ to a direct sum of three (possibly void) parts:

The regular part:

$$\operatorname{diag}(\varepsilon_i E_i) \quad \text{and} \quad \operatorname{diag}(\varepsilon_i E_i J_i) \quad \text{where } \varepsilon_i = \pm 1,$$

$$E_i = \begin{pmatrix} 0 & \cdots & 0 & 1 \\ \vdots & & \cdot & 0 \\ 0 & \cdot & & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}$$

and J_i is a real or complex real Jordan block;

The E part:

$$\operatorname{diag}\left(\varepsilon_j \begin{pmatrix} 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & & & \cdot & 1 \\ \cdot & & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & & \vdots \\ 0 & 1 & 0 & 0 & \end{pmatrix} \right)$$

with $\varepsilon_j = \pm 1$ and

$$\operatorname{diag}\left(\varepsilon_j \begin{pmatrix} 0 & \cdot & \cdot & 0 & 1 \\ \vdots & & & \cdot & 0 \\ 0 & \cdot & & & \vdots \\ 1 & 0 & \cdot & \cdot & 0 \end{pmatrix} \right)$$

The L part:

$$\operatorname{diag}(L_{1,2k_m+1}) \quad \text{and} \quad \operatorname{diag}(L^{\lambda_m, 2k_m+1}),$$

where

$$L_{1,2k+1} = \left(\begin{array}{c|c} 0_{k+1,k+1} & \begin{matrix} 0 & 0 \\ 1 & \ddots \\ & \ddots & 0 \\ & & 0 & 1 \end{matrix} \\ \hline \begin{matrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ 0 & & 0 & 1 \end{matrix} & 0_{k,k} \end{array}\right)$$

and

$$L^{\lambda,2k+1} = \left(\begin{array}{c|c} 0_{k+1,k+1} & \begin{matrix} \lambda & 0 \\ 0 & \ddots \\ & \ddots & \lambda \\ & & 0 & 0 \end{matrix} \\ \hline \begin{matrix} \lambda & 0 & & 0 \\ & \ddots & \ddots & \\ 0 & & \lambda & 0 \end{matrix} & 0_{k,k} \end{array}\right) \quad \text{for } \lambda \in \mathbf{R},$$

while for $k = 0$: $L_1 = L^\lambda = (0)$.

Note that all L and E blocks are square. Complete proofs can be found in Gantmacher [Ga, Vol. II, p. 44] and Thompson [Th, pp. 4, 18, 24, 30] or [THl, § 2]. The aim of this section is to provide a complete description of the "overlapping cases" for the critical $l_{\mathbf{R}}$ numbers $n - 1$ and $n - 2$ in terms of generators and the finest simultaneous block diagonal structure for A and B. Note that if the pairs A, B and C, D are congruent, then, because of the invertibility of the congruence transformation, $l_{\mathbf{R}}(A, B) = l_{\mathbf{R}}(C, D)$. What complicates matters for S_n is the fact that the analogue of Theorem 2.4 for S_n is true only for linearly dependent pencils $P(A, B)$.

THEOREM 4.1. *Let $A, B \in S_n$ be linearly dependent. Then*

(a) *$P(A, B)$ is a* d.*pencil if and only if $l_{\mathbf{R}}(A, B) = 0$;*

(b) *$P(A, B)$ is a* s.d.*pencil if and only if $0 < l_{\mathbf{R}}(A, B) < n$;*

(c) *$P(A, B)$ is an* i.*pencil if and only if $l_{\mathbf{R}}(A, B) = n$.*

To simplify notation we define the quadratic hypersurface

$$Q_A = \{x \in \mathbf{R}^n : x^t A x = 0\} \quad \text{for } A \in S_n.$$

Proof. If A and $B \in S_n$ are linearly dependent, we can assume without loss of generality that $P(A, B) = \{\alpha A : \alpha \in \mathbf{R}\}$, and hence $l_{\mathbf{R}}(A, B) = \dim \text{span } Q_A$. To compute this dimension we consider A diagonalized by a real congruence $D = X^t A X$ and use Lemma 1 of [Uhc, p. 545] again so that (a) and (b) follow immediately. In (c) we can, without loss of generality, assume that for $D = \text{diag}(d_i)$ we have $d_1 d_2 < 0$. Then $|d_2|^{1/2} e_1 \pm |d_1|^{1/2} e_2 \in Q_A$ and if $d_k = 0$ then $e_k \in Q_A$, while in case $d_1 k_k > 0$ we have $|d_k|^{1/2} e_2 + |d_2|^{1/2} e_k \in Q_A$, and in case $d_1 d_k < 0$ we have $|d_k|^{1/2} e_1 + |d_1|^{1/2} e_k \in Q_A$ for $k \geqq 2$. Hence (c) holds. □

We note in passing that linearly dependent pairs A and B will not have any E or L part in their Kronecker normal form. Next we give some remarks on the $l_{\mathbf{R}}$ numbers of E parts and L parts.

Remark 4.2. (a) $l_{\mathbf{R}}(L_1, L^\lambda) = \dim L_1 = \dim L^\lambda = 2k + 1$,

(b)

$$l_{\mathbf{R}}(E_{(0)}, E_{(1)}) = \begin{cases} \dim E_{(0)} & \text{if } \dim E_{(0)} \geqq 4, \\ \dim E_{(0)} - 1 & \text{if } \dim E_{(0)} < 4, \end{cases}$$

where

$$E_{(0)} = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & 1 \\ \vdots & \iddots & \\ 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad E_{(1)} = \begin{pmatrix} 0 \cdots 0 & 1 \\ \vdots & \iddots & 0 \\ 0 & \iddots & \vdots \\ 1 & 0 \cdots 0 & \end{pmatrix}.$$

Proof. (a) For all $1 \leqq j \leqq k+1$ and $k+2 \leqq k \leqq 2k+1$, we have $e_j^t L_1 e_j = 0 = e_j^t L^\lambda e_j$, where e_j is the jth unit vector.

(b) Let $\dim E_{(0)} = \dim E_{(1)} = j$. If $j = 1$, then $E_{(0)} = (0)$, $E_{(1)} = (1)$, and so $\dim Q_{E_{(0)}} \cap Q_{E_{(1)}} = 0$. If $j = 2$, then $Q_{E_{(0)}} \cap Q_{E_{(1)}} = \operatorname{span}\{e_1\}$, while for $j = 3$, $Q_{E_{(0)}} \cap Q_{E_{(1)}} = \operatorname{span}\{e_1, e_3\}$. For $j > 4$, we can use Theorem 1(i) of [Uhc, p. 544] to obtain $l_{\mathbf{R}}(E_{(0)}, E_{(1)}) = \dim E_{(0)}$. □

In view of Theorem 4.1 we will henceforth only consider matrix pencils in S_n with linearly independent generators A, B.

First we deal with $n = 2$.

THEOREM 4.3. *Suppose $A, B \in S_2$ are linearly independent matrices with $l_{\mathbf{R}}(A, B) = 0$. Then*

(a) *$P(A, B)$ is a* d.*pencil if and only if $P(A, B)$ is congruent to $P(C, D)$ where*

$$C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \textit{and} \quad D = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix};$$

(b) *$P(A, B)$ is an* i.*pencil if and only if $P(A, B)$ is congruent to $P(C, D)$ where*

$$C = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \textit{and} \quad D = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Here we call two pencils $P(A, B)$ and $P(C, A)$ *congruent* if there is a nonsingular matrix $S \in \mathbf{R}^{nn}$ with $S^t(P(A, B))S = P(C, D)$.

Proof. (a) Let $P(A, B)$ be a d.pencil. Then $P(A, B)$ contains a definite matrix. We may let this definite matrix, after a suitable congruence transform if necessary, be the identity matrix I. Let $E \in P(A, B)$ be linearly independent of I. After subtracting a scalar multiple of I, we may assume E to be indefinite. Let $X \in \mathbf{R}^{2\times 2}$ be orthogonal such that

$$X^t E X = \begin{pmatrix} 0 & a \\ a & 0 \end{pmatrix}$$

for some $a > 0$. Then $P(A, B)$ is congruent to $P(I, D)$ where $D = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. The converse can be verified directly.

(b) Let $P(A, B)$ be an i.pencil. Choose any indefinite matrix $D \in P(A, B)$. After a suitable congruence we may assume D in the form

$$D = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Then choose any $E \in P(A, B)$ which is linearly independent of D. After subtracting a scalar multiple of D, we may assume E in the form

$$E = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}.$$

Since E is indefinite, we can assume $a > 0 > b$. With a suitable congruence transform, we see that $P(A, B)$ is congruent to $P(C, D)$ where

$$C = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The converse clearly holds as well. □

Next we deal with s.d.pencils, $n \geqq 3$ and large $l_{\mathbf{R}}$ numbers.

THEOREM 4.4. *$P(A, B)$ is an* s.d.*pencil with $l_{\mathbf{R}}(A, B) = n - 1$, $n \geqq 3$ where A, B are linearly independent if and only if A and B or B and A are simultaneously congruent to*:

(a) *Cases $(D) + 0$ and $(E) + 0$, where the (D) type block has size greater than or equal to four and the (E) type block has size greater than or equal to three (here (D) and (E) stand for the block structure described in the main theorem of* [Uhb, pp. 537, 538]); *or*

(b) $\operatorname{diag}(1, -1, \pm 1, \cdots, \pm 1, 0, 0, \cdots, 0)$, *and*

$\operatorname{diag}(\lambda, -\lambda, \pm\lambda, \cdots, \pm\lambda, \pm 1, 0, \cdots, 0)$ for $\lambda \in \mathbf{R}$; *or*

(c) $\operatorname{diag}\left(\pm\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, 0, \cdots, 0\right)_g$ *and*

$\operatorname{diag}\left(\pm\begin{pmatrix} 0 & \lambda \\ \lambda & 1 \end{pmatrix}, 0, \cdots, 0\right)$ *for* $\lambda \in \mathbf{R}$.

Proof. The Kronecker canonical form of an s.d.pencil can contain E parts and L parts only of block sizes less than or equal to two, for otherwise the pencil would be indefinite.

Let $l_{\mathbf{R}} = l_{\mathbf{R}}(A, B)$ and assume that the overall dimension of the regular part is $m \leqq n$. Much of our work has already been prepared in [Uhb] for nonsingular pencils $P(A, B)$, where the first matrix A is nonsingular. To be able to use this, we need to "symmetrize" the statements about the generator A and B by allowing "A and B or B and A" to be congruent to specific generators as indicated in this and the following theorems. With this understanding here, the regular part can be (1) indefinite with $m \geqq 2$, or (2) definite, or (3) semidefinite, or (4) void.

Case (1). There is only one possibility for an s.d.pencil to have an indefinite regular part: $\operatorname{diag}(1, -1, \pm 1, \cdots, \pm 1)$ and $\operatorname{diag}(\lambda, -\lambda, \pm\lambda, \cdots, \pm\lambda)$ so that the regular part of $\lambda A - B$ is congruent to $0_{m,m}$. In this case the E part cannot have a two-dimensional block since $\lambda\left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right) - \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ is indefinite. If the E part contains more than one one-dimensional block we would drop more than one l number, contradicting $l_{\mathbf{R}}(A, B) = n - 1$. Since the L part must be empty or made up of one-dimensional blocks we clearly have case (b).

Case (2). If the regular part is definite, then independently of its dimension m, we have $l_{\text{reg}} = 0$. Since overall $l_{\mathbf{R}}(A, B) = n - 1$, we must have $m = 1$, so the regular part for A and B is congruent to (± 1), $(\pm\lambda)$. The E part must be empty for we cannot drop any more l numbers. So A and B are simultaneously congruent to $\operatorname{diag}(\pm 1, 0, \cdots, 0)$ and $\operatorname{diag}(\pm\lambda, 0, \cdots, 0)$, contradicting the assumed linear independence of A and B.

Case (3). (a) The regular part for A and B is semidefinite of size $m \geqq 3$: Using the main theorem of [Uhb], we get that $1 \leqq l_{\text{reg}} \leqq m - 1$. So $l_{\text{reg}} = m - 1$ if we want $l_{\mathbf{R}}(A, B) = n - 1$. That makes cases (D) or (E) from the main theorem in [Uhb] augmented by arbitrarily many one-dimensional L blocks or case (a).

(b) The regular part for A and B is semidefinite of size $m = 2$: Using Theorem 3 of [Uhc p. 557] we can see that the regular parts in the Kronecker form for the pair A, B must be $\pm\left(\begin{smallmatrix}0 & 1\\ 1 & 0\end{smallmatrix}\right)$ and $\pm\left(\begin{smallmatrix}0 & \lambda\\ \lambda & 1\end{smallmatrix}\right)$ with $l_{\text{reg}} = 1$ or case (c). Note that $m = 1$ would not yield a semidefinite regular part.

Case (4). If a semidefinite pencil has no regular part and $l_{\mathbf{R}}(A, B) = n - 1$, then the L-part can have arbitrarily many one-dimensional L blocks and the E part can consist of either (aa) one two-dimensional E block, or (bb) one one-dimensional E block.

In the case of (aa) we have case (c) in the theorem, for $\lambda = 0$ and for the pair A and B, while in the case of (bb), A and B are linearly dependent matrices. The converse is obvious. □

THEOREM 4.5. *$P(A, B)$ is an* s.d.*pencil with $l_{\mathbf{R}}(A, B) = n - 2$, $n \geqq 3$ and A and B linearly independent if and only if A and B or B and A are simultaneously congruent to*

(a) $\operatorname{diag}\left(\varepsilon_1\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \varepsilon_3, 0, \cdots, 0\right)$, *and*

$\operatorname{diag}\left(\varepsilon_1\begin{pmatrix}0 & \lambda\\ \lambda & 1\end{pmatrix}, \varepsilon_3\mu_3, 0, \cdots, 0\right)$,

where $\varepsilon_1\varepsilon_3(\mu_3 - \lambda) > 0$, $\varepsilon_i = \pm 1$, $\lambda, \mu_3 \in \mathbf{R}$; *or*

(b) $\operatorname{diag}\left(\varepsilon\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \varepsilon\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, 0, \cdots, 0\right)$, *and*

$\operatorname{diag}\left(\varepsilon\begin{pmatrix}0 & \lambda\\ \lambda & 1\end{pmatrix}, \varepsilon\begin{pmatrix}0 & \lambda\\ \lambda & 1\end{pmatrix}, 0, \cdots, 0\right)$ *where* $\varepsilon = \pm 1, \lambda \in \mathbf{R}$; *or*

(c) $\operatorname{diag}\left(\varepsilon_1\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \varepsilon_1\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \varepsilon_5, \cdots, \varepsilon_m, 0, \cdots, 0\right)$,

$\operatorname{diag}\left(\varepsilon_1\begin{pmatrix}0 & \lambda\\ \lambda & 1\end{pmatrix}, \varepsilon_1\begin{pmatrix}0 & \lambda\\ \lambda & 1\end{pmatrix}, \varepsilon_5\lambda, \cdots, \varepsilon_m\lambda, 0, \cdots, 0\right)$,

where $\varepsilon_i = \pm 1$, $\lambda \in \mathbf{R}$, *and* $\varepsilon_i\varepsilon_j = -1$ *at least once for* $i, j \geqq 5$ *and* $m \geqq 6$; *or*

(d) $\operatorname{diag}(1, \cdots \cdots, 1, -1, \cdots, -1)$, *where* 1 *appears* k *fold and* -1 *appears* $s = n - k$ *fold on the diagonal, and*

(d1) $\operatorname{diag}(\lambda, \cdots, \lambda, \mu, \kappa, -\lambda, \cdots, -\lambda)$ *with either* $\mu, \kappa < \lambda$ *or* $\mu, \kappa > \lambda$, *where* λ *appears* $k - 2$ *fold, while* $-\lambda$ *appears* s *fold on the diagonal and* $k \geqq 3$, $s \geqq 1$; *or*

(d2) $\operatorname{diag}(\lambda, \cdots, \lambda, \mu, -\lambda, \cdots, -\lambda, \kappa)$ *with either* $\mu < \lambda$ *and* $\kappa < -\lambda$ *or* $\mu > \lambda$ *and* $\kappa > -\lambda$. *Here* λ *appears* $k - 1$ *fold and* $-\lambda$ *appears* $s - 1$ *fold and* $k \geqq 2$, $s \geqq 2$; *or*

(d3) $\operatorname{diag}(\lambda, \cdots, \lambda, -\lambda, \cdots, -\lambda, \mu, \kappa)$, *where either* $\mu, \kappa > \lambda$ *or* $\mu, \kappa < \lambda$ *with* λ *appearing* k *fold and* $-\lambda$ $s - 2$ *fold and* $k \geqq 1$, $s \geqq 3$; *or*

(e) $\operatorname{diag}(\varepsilon_1, \varepsilon_2, 0, \cdots, 0)$, *and*
$\operatorname{diag}(\varepsilon_1\lambda, \varepsilon_2\mu, 0, \cdots, 0)$ *for*
$\varepsilon_i = \pm 1$, $\lambda, \mu \in \mathbf{R}$ *with* $\lambda \neq \mu$; *or*

(f) $\operatorname{diag}\left(\varepsilon_1\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \varepsilon_3, \cdots, \varepsilon_n\right)$, *and*

$$\operatorname{diag}\left(\varepsilon_1\begin{pmatrix}0 & \lambda\\ \lambda & 1\end{pmatrix}, \varepsilon_3\mu, \varepsilon_4\lambda, \cdots, \varepsilon_n\lambda\right),$$

where $\varepsilon_i = \pm 1$, $\lambda, \mu \in \mathbf{R}$, $\varepsilon_1\varepsilon_3(\lambda - \mu) > 0$; *or*

(g) $$\operatorname{diag}\left(\varepsilon_1\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \varepsilon_3, \cdots, \varepsilon_k, 0, \cdots, 0\right), \quad \textit{and}$$

$$\operatorname{diag}\left(\varepsilon_1\begin{pmatrix}0 & \lambda\\ \lambda & 1\end{pmatrix}, \varepsilon_3\lambda, \cdots, \varepsilon_k\lambda, \delta, 0, \cdots, 0\right),$$

where $\varepsilon_i = \pm 1$, $k \geqq 4$, $\varepsilon_i\varepsilon_j = -1$ *at least once for* $i, j \geqq 3$, $k \geqq 4$ *and* $\varepsilon_1\delta = 1$; *or*

(h) $\operatorname{diag}(\varepsilon_1, \cdots, \varepsilon_{m-1}, \varepsilon_m, 0, \cdots \cdots, 0)$, *and*
$\operatorname{diag}(\varepsilon_1\lambda, \cdots, \varepsilon_{m-1}\lambda, \varepsilon_m\mu, \varepsilon_{m+1}, 0, \cdots, 0)$,
where $\varepsilon_i = \pm 1$, $\varepsilon_i\varepsilon_j = -1$ *at least once for* $i, j \leqq m$, $m \geqq 3$, $\lambda \neq \mu$ *and* $\varepsilon_m(\lambda - \mu)\varepsilon_{m+1} \leqq 0$; *or*

(i) $\operatorname{diag}(\varepsilon_1, 0, \cdots, 0)$, *and*
$\operatorname{diag}(\varepsilon_1\lambda, \varepsilon_2, 0, \cdots, 0)$ *where* $\varepsilon_i = \pm 1$, $\lambda \in \mathbf{R}$; *or*

(j) $\operatorname{diag}(\varepsilon_1, \cdots, \varepsilon_m, 0, \cdots \cdots \cdots, 0)$, *and*
$\operatorname{diag}(\varepsilon_1\lambda, \cdots, \varepsilon_m\lambda, \varepsilon_{m+1}, \varepsilon_{m+2}, 0, \cdots, 0)$
with $\varepsilon_i = \pm 1$, $\varepsilon_i\varepsilon_j = -1$ *at least once for* $i, j \leqq m$, $m \geqq 2$ *and* $\varepsilon_{m+1}\varepsilon_{m+2} = 1$; *or*

(k) $$\operatorname{diag}\left(\varepsilon_1\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, 0, \cdots, 0\right), \quad \textit{and}$$

$$\operatorname{diag}\left(\varepsilon_1\begin{pmatrix}0 & \lambda\\ \lambda & 1\end{pmatrix}, \varepsilon_2, 0, \cdots, 0\right) \quad \textit{where } \varepsilon_i = \pm 1, \varepsilon_1\varepsilon_2 = 1, \lambda \in \mathbf{R}; \quad \textit{or}$$

(l) $$\operatorname{diag}\left(\varepsilon_1\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}, 0, \cdots, 0\right), \quad \textit{and}$$

$$\operatorname{diag}\left(\varepsilon_1\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \varepsilon_2, 0, \cdots, 0\right) \quad \textit{where } \varepsilon_i = \pm 1.$$

Proof. As we remarked earlier, a semidefinite pencil cannot have any L blocks of large dimension or it would be indefinite. Similarly, the E part must have only one- or two-dimensional blocks.

If $l_{\text{reg}} = m - 2$, then the E part must be void and the L part must contain one-dimensional blocks only. For $l_{\text{reg}} = m - 2$ we conclude from [Uhc, p. 545] for a s.d.pencil that we must have cases (viia) or (viib) or (ix) only. For (viia), we must have $k = m - 2$ where $2k \leqq m$, so $k = 1$ or 2 are the only possibilities. If $k = 1$, then $m = 3$ and case (a). If $k = 2$, then $m = 4$ and case (b). If (viib) holds with $k = 2$, then $r = m - 4$ makes $l_{\text{reg}} = k + r = m - 2$ or case (c). If (viib) holds with $k = 2$ and $r = m - 2$, then $l_{\text{reg}} = m - 1$ and we must drop one l number in the E part or case (g). If (viib) holds with $k = 1$ and $r = m - 3$, then $l_{\text{reg}} = m - 2$ and we have case (f).

For (ix) we use Theorem 1 of [Uhb, p. 538] for $m > 2$: A and B must be simultaneously congruent to $\operatorname{diag}(a_i)$, $\operatorname{diag}(b_i)$ where

$$\max_{a_i > 0} \frac{b_i}{a_i} = \min_{a_i < 0} \frac{b_i}{a_i} \quad \text{or} \quad \min_{a_i > 0} \frac{b_i}{a_i} = \max_{a_i < 0} \frac{b_i}{a_i}.$$

(Note the obvious misprinting of formula (i) of [Uhb, p. 538].) So without loss of generality we can assume that A and B are simultaneously congruent to diag $(1, \cdots, 1, -1, \cdots, -1)$ and diag $(\lambda_1, \cdots, \lambda_k, \mu_1, \cdots, \mu_s)$ where $k, s \geqq 1$ and $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_k$, $\mu_1 \geqq \cdots \geqq \mu_s$. If $\max_{a_i>0} b_i/a_i = \min_{a_i<0} b_i/a_i$, then $\lambda_1 = -\mu_1$, while $\lambda_k = -\mu_s$ in the alternate case. In the former case $-\lambda_1 A + B$ = diag $(0, \lambda_2 - \lambda_1, \cdots, \lambda_k - \lambda_1, 0, \mu_2 - \mu_1, \cdots, \mu_s - \mu_1)$ is a nonpositive diagonal matrix, while in the latter case $\lambda_k A - B$ = diag $(\lambda_k - \lambda_1, \cdots, \lambda_2 - \lambda_1, 0, \mu_s - \mu_1, \cdots, \mu_2 - \mu_1, 0)$ is nonpositive as well. But $l_{\mathbf{R}}(A, B) = n - 2$, so only two nonzero entries $\lambda_{..} - \lambda_1$, and/or $\mu_{..} - \mu_1$ are possible, giving case (d) with its three possibilities. If $m = 2$ and (ix), the regular parts of A and B are simultaneously congruent to diag $(\varepsilon_1, \varepsilon_2)$ and diag $(\varepsilon_1\lambda, \varepsilon_2\mu)$. Then $(\lambda A - B)_{\text{reg}}$ = diag $(0, \varepsilon_2(\lambda - \mu))$ is semidefinite if $\lambda \neq \mu$, in which case $l_{\text{reg}}(A, B) = 0$ or case (e).

Note that cases (a), $\cdots$, (f) all deal with $l_{\text{reg}} = m - 2$. If $l_{\text{reg}} = m - 1$, then we can use the cases from Theorem 4.4 for s.d.pencils with $l_{\mathbf{R}} = m - 1$ if we add one more one-dimensional E block to A and B. The previous case (a) involving (D) makes case (g) now; while old (a) involving (E) makes case (h) here. Case (i) derives from the rejected possibility (2) in the previous proof: We can add one one-dimensional E block to the pencil, which then becomes linearly independent, hence case (i). The remaining cases (j), (k), and (l) come from the cases (b), (c), and (d) of the previous theorem. Finally, note that the old case (bb) does not yield a new case here either, since adding one more one-dimensional E-block would keep A and B linearly dependent. The converse is again obvious. □

Finally, we will classify i-pencils with $l_{\mathbf{R}}$ numbers, smaller than n.

THEOREM 4.6. *$P(A, B)$ is an* i-*pencil with $l_{\mathbf{R}}(A, B) = n - 1$, $n \geqq 3$, and A, B linearly independent if and only if one of the following* 14 *cases holds for the regular, E and L parts of the Kronecker canonical pair form of A and B or B and A, where we set* dim (*regular part*) $= m \leqq n$. (*Note that the cases* (A), $\cdots$, (*E*) *in part* (d) (*Table* 1) *below refer to the cases mentioned in the main theorem in* [Uhb, p. 538].)

Note that there are 14 possible cases here: (a), (a1), (b), (c), (c1), (d1), (d2), (d20), (d3), (d30), (d4), (d40), (e), (e1).

Proof. If $m = 1$, then $l_{\text{reg}} = 0$ so that E and L parts of A, B cannot drop any l-number. Hence the E blocks must have dimension greater than or equal to four by Remark 4.2. To ensure an i-pencil, should no E block occur, then at least one L block must be of size greater than or equal to three. If $m \geqq 2$ and $l_{\text{reg}} = m$, then exactly one E block must be of size less than or equal to three to drop $l_{\mathbf{R}}$ to $n - 1$. If $l_{\text{reg}} = m$, then the regular part of the pencil is indefinite according to the main theorem of [Uhb, p. 537]. If $m = 2$ with $l_{\text{reg}} = 1$, then by Theorem 3 of [Uhc, p. 557], $\pm(A, B)_{\text{reg}}$ or $\pm(B, A)_{\text{reg}}$ is congruent to $\left(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & \lambda \\ \lambda & 1 \end{pmatrix}\right)$ and the E and L parts cannot drop another l number and must be indefinite. If $m \geqq 3$ and $l_{\text{reg}} = m - 1$, then the E and L parts must have full l numbers and if the E part should be void and the regular part semidefinite (in cases (D) and (E)), then at least one L block must have size greater than or equal to three to ensure an i-pencil. Finally cases (e) and (el) are obvious if the regular part of (A, B) is void. □

Note that this theorem describes the finest simultaneous block structure of A and B completely (as did the previous two theorems for s.d.pencils) except in case (b) when $l_{\text{reg}} = m$. While Theorem 1 and 2 of [Uhc, pp. 544, 545] described some i-pencils with $l = n$ in (i), $\cdots$, (viii), we did not attempt to describe all such pencils then nor are we able to do so now.

THEOREM 4.7. *$P(A, B)$ is an* i-*pencil with $l_{\mathbf{R}}(A, B) = n - 2$, $n \geqq 3$ and A and B linearly independent if and only if*

TABLE 1

Regular part	E part	L part
(a) $m = 1$, $l_{\text{reg}} = 0$	All blocks of size $\geqq 4$ or void	Any size blocks or void
	(1) **But** if E part void, then one L block of size $\geqq 3$	
(b) $m \geqq 2$, $l_{\text{reg}} = m$	Exactly one block of size $\leqq 3$, possibly more of size $\geqq 4$	Any size blocks or void
(c) $m = 2$, $l_{\text{reg}} = 1$, and $\pm(S, T)_{\text{reg}}$ or $\pm(T, S)_{\text{reg}}$ is congruent to $\left(\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \begin{pmatrix}0 & \lambda\\ \lambda & 1\end{pmatrix}\right)$	All blocks of size ≥ 4 or void	Any size blocks or void
	(1) **But** if E part void, then one L block of size $\geqq 3$	
(d) $m \geqq 3$, $l_{\text{reg}} = m - 1$		
(1) Case (A), or	All blocks of size $\geqq 4$ or void	Any size blocks or void
(2) Case (B), or	All blocks of size $\geqq 4$ or void	Any size blocks or void
(3) Case (D), or	All blocks of size $\geqq 4$ or void	Any size blocks or void
(4) Case (E),	All blocks of size $\geqq 4$ or void	Any size blocks or void
	(0) **But** if E part void in case (2), (3), or (4), then at least one L block of size $\geqq 3$.	
(e) $m = 0$, regular part void	Exactly one block of size $\leqq 3$, possibly more of sizes $\geqq 4$	Any size blocks or void
	(1) **But** if there is a one-dimensional E block, then there must be an E-block of size $\geqq 4$ or an L block of size $\geqq 3$.	

(a) *Any of the cases of an* i *pencil with* $l = n - 1$ *from Theorem* 4.4 *holds* <u>*except*</u> *that the E part must now contain one extra block of size less than or equal to three; or*

(b) *Any of the cases of an* s.d.*pencil and* $l = n - 2$ *from Theorem* 4.3 *holds* <u>*except*</u> *that there must be an additional E block of size greater than or equal to four or an additional L block size greater than or equal to three*; *or*

(c) *If the regular part is void, then there must be exactly two E blocks of sizes less than or equal to three. If both are one-dimensional or one is one-dimensional, the second two-dimensional, then there must be an E block of size greater than or equal to four or an L block of size greater than or equal to three. If both small E blocks are two-dimensional, then they must carry opposite signs or there must be an E block of size greater than or equal to four or an L block of size greater than or equal to three.*

The proof is obvious since for regular i-*pencils* $l \leqq n - 1$. Note that the extra conditions for small E blocks ensure an i-*pencil* of dimension two.

REFERENCES

[Br] L. BRICKMAN, *On the field of values of a matrix*, Proc. Amer. Math. Soc., 12 (1961), pp. 61–66.

[Em] M. R. EMBRY, *The numerical range of an operator*, Pacific J. Math., 32 (1970), pp. 647–650.

[Fi] P. FINSLER, *Uber das Vorkommen definiter und semidefiniter Formen in Scharen quadratischer Formen*, Comment. Math. Helv., 9 (1936/1937), pp. 188–192.

[Ga] F. R. GANTMACHER, *The Theory of Matrices*, Vols. 1, 2, Chelsea, New York, 1977.
[Ha] F. HAUSDORFF, *Der Wertevorrat einer Bilinearform*, Math. Z., 3 (1919), pp. 314–316.
[Ta] O. TAUSSKY, *Positive-definite matrices*, in Inequalities, O. Shisha, ed., Academic Press, New York, 1967, pp. 309–319.
[Th] R. C. THOMPSON, *Pencils of complex and real symmetric and skew matrices*, manuscript, University of California, Santa Barbara, CA, circa 1976.
[Thl] ———, *The characteristic polynomial of a principal subpencil of a Hermitian matrix pencil*, Linear Algebra Appl., 14 (1976), pp. 135–177.
[To] O. TOEPLITZ, *Das algebraische Analogon zu einem Satz von Feher*, Math. Z., 2 (1918), pp. 187–197.
[Uha] F. UHLIG, *A canonical form for a pair of real symmetric matrices that generate a nonsingular pencil*, Linear Algebra Appl., 14 (1976), pp. 189–209.
[Uhb] ———, *The number of vectors jointly annihilated by two real quadratic forms determines the inertia of matrices in the associated pencils*, Pacific J. Math., 49 (1973), pp. 537–542.
[Uhc] ———, *On the maximal number of linearly independent real vectors annihilated simultaneously by two real quadratic forms*, Pacific J. Math., 49 (1973), pp. 543–560.
[Uhd] ———, *Definite and semidefinite matrices in a real symmetric matrix pencil*, Pacific J. Math., 49 (1973), pp. 561–568.
[Uhe] ———, *A recurring theorem about pairs of quadratic forms and extensions: A survey*, Linear Algebra Appl., 25 (1979), pp. 219–237.
[VDo] P. VAN DOOREN, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27 (1979), pp. 103–140.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 1, pp. 160–165, January 1991

012

EXPONENTIAL NONNEGATIVITY ON THE ICE CREAM CONE*

RONALD J. STERN† AND HENRY WOLKOWICZ‡

Abstract. Let K_n denote the n-dimensional ice cream cone. This paper investigates the structure of those matrices A such that $e^{tA}K_n \subset K_n$ for all $t \geqq 0$. The characterizations extend to general ellipsoidal cones.

Key words. ice cream cone, ellipsoidal cone, matrices, exponential nonnegativity, copositivity, spectrum

AMS(MOS) subject classification. 15A48

1. Introduction. A set $C \subset R^n$ is a *cone* provided that $\alpha C \subset C$ for all $\alpha \geqq 0$. We call a cone C *proper* provided that it is closed, convex, possesses nonempty interior, and is pointed ($C \cap \{-C\} = \{0\}$). Given a proper cone $C \subset R^n$, we denote by $p(C)$ the set of matrices $A \in R^{n,n}$ which are *exponentially nonnegative* on C; that is, $e^{tA}C \subset C$ for all $t \geqq 0$, where $e^{tA} = \sum_{j=0}^{\infty} (tA)^j/j!$ is the familiar matrix exponential. Hence $p(C)$ is the set of matrices A such that for an arbitrary start point $x(0) \in C$, the solution $x(t) = e^{tA}x(0)$ of the linear differential equation $\dot{x}(t) = Ax(t)$ remains in C for all future time.

The purpose of this paper is to investigate the structure of the set of matrices $p(K_n)$, where

$$K_n = \left\{ x \in R^n : \sum_{i=1}^{n-1} x_i^2 \leqq x_n^2,\ x_n \geqq 0 \right\}$$

is the *n-dimensional ice cream cone.* It will be seen that our results can be extended to general ellipsoidal cones.

In the following section, we review some required technical material on ellipsoidal cones. Then, in § 3, the main results are presented. A key result which we employ is a lemma on copositivity for the ice cream cone K_n due to Loewy and Schneider [3]. To a certain extent our results complement some of those in [3], which provided characterizations of those matrices which leave K_n invariant.

2. Ellipsoidal cones. Let $Q \subset R^{n,n}$ be a symmetric nonsingular matrix, with a single negative eigenvalue λ_n. Therefore Q has *inertia* $(n-1, 0, 1)$, where by inertia we mean the triple (P, Z, N), indicating the number of positive, zero, and negative eigenvalues, respectively. Let u_n be a unit eigenvector of Q corresponding to λ_n. With Q we associate two *ellipsoidal cones*; these are

$$K = K(Q, u_n) = \{ x \in R^n : x^tQx \leqq 0,\ x^tu_n \geqq 0 \} \tag{2.1}$$

and $-K = K(Q, -u_n)$. In the sequel we will employ the fact that at each $0 \neq x \in \partial K = \{ x \in K : x^tQx = 0 \}$, the vector Qx is an outward pointing normal at x (where ∂ denotes boundary).

* Received by the editors July 14, 1989; accepted for publication (in revised form) November 26, 1989.

† Department of Mathematics and Statistics, Concordia University, Montreal, Quebec, Canada H4B 1R6 (stern@conu2.bitnet). The research of this author was supported by Natural Sciences and Engineering Research Council of Canada grant A4641.

‡ Department of Combinatorics and Optimization, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 (hwolkocz@orion.waterloo.edu). The research of this author was supported by Natural Sciences and Engineering Research Council of Canada grant A9161.

Clearly, K_n is an ellipsoidal cone with

$$Q = Q_n := \begin{pmatrix} I_{n-1} & 0 \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad u_n = e_n,$$

where I_{n-1} denotes the $(n-1) \times (n-1)$ identity matrix. Also, we denote the kth unit vector by e_k.

We shall require the following lemma from [5], which says that in formula (2.1) we may replace the eigenvector u_n with vectors v satisfying certain requirements (which are met by u_n itself).

LEMMA 2.2. *Suppose that K is as above and assume that $v \in R^n$ satisfies*

$$\{v\}^{\perp} \cap \{K \cup \{-K\}\} = \{0\} \tag{2.3}$$

and

$$v^t u_n \geqq 0. \tag{2.4}$$

Then

$$K = \{x \in R^n : x^t Q x \leqq 0,\ x^t v \geqq 0\}. \tag{2.5}$$

Remark 2.6. In view of the fact that the orthogonal complement $\{u_n\}^{\perp}$ is a hyperplane which supports the proper cones K and $-K$ only at the origin, it follows from the preceding lemma that if v is a vector whose distance from u_n is sufficiently small, then (2.5) holds.

For Q as above, let the spectrum be $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ where $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_{n-1} > 0 > \lambda_n$, and let the orthogonal diagonalization of Q be given by $U^t Q U = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n)$. The following lemma will also prove to be useful. Its proof, which employs Sylvester's theorem, may be found in [5].

LEMMA 2.7. *K is an ellipsoidal cone as in* (2.1) *if and only if $K = TK_n$ for some nonsingular $T \in R^{n,n}$.*

In particular, for a given ellipsoidal cone $K = K(Q, u_n)$, we have $K = TK_n$ for $T = UD$, where D is the diagonal matrix with entries $d_{ii} = |\lambda_i|^{-1/2}$, $i = 1, 2, \cdots, n$, and then $Q = (T^{-1})^t Q_n T^{-1}$. Conversely, for a given nonsingular $T \in R^{n,n}$, the matrix $(T^{-1})^t Q_n T^{-1}$ has inertia $(n-1, 0, 1)$ and $TK_n = K((T^{-1})^t Q_n T, (T^{-1})^t e_n)$.

3. Main results. To begin, we require the following lemma, in which $\langle \cdot, \cdot \rangle$ denotes the standard inner product on R^n.

LEMMA 3.1. *Let K be an ellipsoidal cone as in* (2.1). *Then*

$$p(K) = \{A \in R^{n,n} : \langle Ax, Qx \rangle \leqq 0 \text{ for all } x \in \partial K\}. \tag{3.2}$$

Proof. Since Qx is the unique outward pointing normal vector (up to scalar multiples) to K at any nonzero $x \in \partial K$, then the condition that $\langle Ax, Qx \rangle \leqq 0$, for all such x, is, in the terminology of Schneider and Vidyasagar [4], *cross-positivity* of A on K, which was shown in [4] to be equivalent to exponential nonnegativity. □

We now turn our attention to the problem of characterizing $p(K_n)$. We will make use of the following copositivity result from Loewy and Schneider [3].

LEMMA 3.3 [3, Lemma 2.2]. *Let $W \in R^{n,n}$ be symmetric. Then there exists $\mu \geqq 0$ such that $W - \mu Q_n$ is negative semidefinite if and only if*

$$x \in K_n \Rightarrow x^t W x \leqq 0. \tag{3.4}$$

Our main characterization of $p(K_n)$ is given next.

THEOREM 3.5. *A necessary and sufficient condition for $A \in p(K_n)$ is that there exists $\xi \in R$ such that*

$$Q_n A + A^t Q_n - \xi Q_n \leqq 0, \tag{3.6}$$

where "$\leqq$" means negative semidefinite.

Proof. Let us denote

$$W(Q_n, A) := Q_n A + A^t Q_n.$$

Upon symmetrizing the quadratic form $\langle Ax, Qx \rangle$, it follows that $A \in p(K_n)$ if and only if

$$x \in \partial K_n \Rightarrow x^t W(Q_n, A) x \leqq 0. \tag{3.7}$$

Since $x^t Q_n x = 0$ for all $x \in \partial K_n$, we have that (3.7) is equivalent to

$$x \in \partial K_n \Rightarrow x W(Q_n, A + \gamma I) x \leqq 0 \tag{3.8}$$

for any given $\gamma \in R$. Since

$$W(Q_n, A + \gamma I) = W(Q_n, A) + 2\gamma Q_n, \tag{3.9}$$

we may choose γ large enough to ensure that $W(Q_n, A + \gamma I)$ has inertia $(n - 1, 0, 1)$. For such γ, consider the ellipsoidal cone

$$C(\gamma) := \{ x \in R^n : x^t W(Q_n, A + \gamma I) x \leqq 0,\ x^t u_n(\gamma) \geqq 0 \},$$

where $u_n(\gamma)$ is a unit eigenvector of $W(Q_n, A + \gamma I)$ corresponding to its only negative eigenvalue. Since γ may be chosen so large that $u_n(\gamma)$ approximates e_n to any prescribed tolerance, Remark 2.6 tells us that for sufficiently large γ we have

$$C(\gamma) = \{ x \in R^n : x^t W(Q_n, A + \gamma I) x \leqq 0,\ x^t e_n \geqq 0 \}. \tag{3.10}$$

Hence (3.8) implies that $A \in p(K_n)$ if and only if for all γ sufficiently large we have

$$\partial K_n \subset C(\gamma). \tag{3.11}$$

Since $C(\gamma)$ is an ellipsoidal and therefore convex cone for large γ, it follows that for such γ, (3.11) is equivalent to

$$K_n \subset C(\gamma). \tag{3.12}$$

Therefore, Lemma 3.3 implies that $A \in p(K_n)$ if and only if for each sufficiently large γ there exists $\mu_\gamma \geqq 0$ such that

$$W(Q_n, A + \gamma Q) - \mu_\gamma Q_n \leqq 0. \tag{3.13}$$

Since

$$W(Q_n, A + \gamma I) - \mu_\gamma Q_n = W(Q_n, A) + (2\gamma - \mu_\gamma) Q_n, \tag{3.14}$$

the theorem is proven. □

In what follows, we shall partition A as

$$A = \left(\begin{array}{c|c} A_1 & c \\ \hline d^t & a_{nn} \end{array} \right),$$

where A_1 denotes the leading $(n-1)\times(n-1)$ principal submatrix of A. Then

$$W(Q_n, A) = \left(\begin{array}{c|c} A_1 + A_1^t & g \\ \hline g^t & -2a_{nn} \end{array}\right), \tag{3.15}$$

where

$$g := c - d,$$

and therefore

$$W(Q_n, A) - \xi Q_n = \left(\begin{array}{c|c} A_1 + A_1^t - \xi I_{n-1} & g \\ \hline g^t & \xi - 2a_{nn} \end{array}\right). \tag{3.16}$$

We have the following corollary to Theorem 3.5. It provides sufficient conditions for membership and nonmembership in $p(K_n)$.

COROLLARY 3.17. *Let $A \in R^{n,n}$. Then the following hold*:

$$\max_{1 \le i \le n-1} \left\{ 2a_{ii} + |g_i| + \sum_{i \ne j = 1}^{n-1} |a_{ij} + a_{ji}| \right\} \le 2a_{nn} - \sum_{i=1}^{n-1} |g_i| \Rightarrow A \in p(K_n), \tag{3.18}$$

$$\max_{1 \le i \le n-1} \left\{ 2a_{ii} - |g_i| - \sum_{i \ne j = 1}^{n-1} |a_{ij} + a_{ji}| \right\} > 2a_{nn} + \sum_{i=1}^{n-1} |g_i| \Rightarrow A \notin p(K_n). \tag{3.19}$$

Proof. Theorem 3.5 implies that $A \in p(K_n)$ if and only if there exists $\xi \in R$ such that the (symmetric) matrix $W(Q_n, A) - \xi Q_n$ has no positive eigenvalues. A straightforward application of Gershgorin's theorem then yields (3.18) and (3.19). □

A different sufficient condition for $A \in p(K_n)$ is provided in the following result. We shall denote the euclidean norm by $\|\cdot\|$, and the largest eigenvalue of a symmetric matrix M by $\lambda_1(M)$.

THEOREM 3.20. *A sufficient condition for $A \in p(K_n)$ is*

$$\lambda_1(A_1 + A_1^t) \le 2(a_{nn} - \|g\|). \tag{3.21}$$

Proof. Let us write

$$W(Q_n, A) - \xi Q_n = U(\xi) + V,$$

where

$$U(\xi) = \left(\begin{array}{c|c} A_1 + A_1^t - \xi I_{n-1} & 0 \\ \hline 0 & \xi - 2a_{nn} \end{array}\right) \quad \text{and} \quad V = \left(\begin{array}{c|c} 0 & g \\ \hline g^t & 0 \end{array}\right).$$

Then, since $U(\xi)$ and V are symmetric, we have

$$\lambda_1(U(\xi) + V) \le \lambda_1(U(\xi)) + \lambda_1(V). \tag{3.22}$$

(See, e.g., Wilkinson [6, p. 101].) Therefore, in view of Theorem 3.5, a sufficient condition for $A \in p(K_n)$ is the existence of $\xi \in R$ such that

$$\lambda_1(U(\xi)) + \lambda_1(V) \le 0. \tag{3.23}$$

Since $\lambda_1(V) = \|g\|$, the existence of such a ξ is readily seen to be guaranteed by (3.21). □

It is not difficult to construct examples where the sufficient condition (3.21) holds, but (3.18) fails. The reverse may occur as well, as is evidenced by the matrix

$$A = \begin{pmatrix} -2 & 0 & 2 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

The next result provides a general necessary condition for $A \in p(K_n)$.

THEOREM 3.24. *Let $A \in p(K_n)$. Then*

$$\lambda_1(A_1 + A_1^t) \leqq 2a_{nn}. \tag{3.25}$$

Proof. Theorem 3.5 tells us that if $A \in p(K_n)$, then there exists a real number ξ such that all the spectrum of $W(Q_n, A) - \xi Q_n$ is nonpositive, which implies that each principal submatrix has nonpositive spectrum as well. Applying this fact to the principal submatrices $A_1 + A_1^t - \xi I_{n-1}$ and $\xi - 2a_{nn}$ readily yields (3.25). □

Theorems 3.20 and 3.24 immediately yield the following complete characterization of $p(K_n)$ for matrices satisfying a certain "partial symmetry" condition.

COROLLARY 3.26. *Let $A \in R^{n,n}$ be such that $a_{in} = a_{ni}$ for all $1 \leqq i \leqq n-1$ (i.e., $g = 0$). Then (3.25) is necessary and sufficient for $A \in p(K_n)$.*

Another general necessary condition is given next.

THEOREM 3.27. *Assume that $A \in p(K_n)$. Let $\{\mu_1, \mu_2, \cdots, \mu_k\}$ be any set of eigenvalues of A (not necessarily distinct), and let $\{x_1, x_2, \cdots, x_k\}$ be a corresponding set of eigenvectors. Consider the (possibly empty) index sets*

$$I_+ = \{i : x_i^* Q_n x_i > 0\} \quad \text{and} \quad I_- = \{i : x_i^* Q_n x_i < 0\}.$$

Then

$$\inf\{\operatorname{Re}\mu_i : i \in I_-\} \geqq \sup\{\operatorname{Re}\mu_i : i \in I_+\} \tag{3.28}$$

(where $\sup(\varnothing) = -\infty$ and $\inf(\varnothing) = \infty$, $\varnothing$ denoting the empty set).

Proof. Since $A \in p(K_n)$, there exists $\xi \in R$ such that

$$H(\xi) := Q_n A + A^t Q_n - \xi Q_n \leqq 0. \tag{3.29}$$

Then

$$x_i^* H(\xi) x_i = 2x_i^* Q_n x_i (\operatorname{Re}\mu_i - \xi) \leqq 0 \quad \text{for all } i = 1, 2, \cdots, k. \tag{3.30}$$

Hence $\xi \geqq \operatorname{Re}\mu_i$ for all $i \in I_+$ and $\xi \leqq \operatorname{Re}\mu_i$ for all $i \in I_-$, yielding (3.28). □

Our final result provides a characterization of the set of matrices

$$p(\partial K_n) := \{A \in R^{n,n} : e^{tA}(\partial K_n) \subset \partial K_n \text{ for all } t \geqq 0\}.$$

Hence $p(\partial K_n)$ is the set of matrices A such that solutions of the linear differential equation $\dot{x}(t) = Ax(t)$ with $x(0) \in \partial K_n$ remain in ∂K_n for all $t \geqq 0$.

THEOREM 3.31. *A necessary and sufficient condition for $A \in p(\partial K_n)$ is that $A = B + \alpha I$, where $\alpha \in R$ and*

$$B = \left(\begin{array}{c|c} B_1 & b \\ \hline b^t & 0 \end{array}\right)$$

with B_i being an $(n-1) \times (n-1)$ skew-symmetric matrix.

Proof. The matrix $A \in p(\partial K_n)$ if and only if the vector field Ax is tangent to the locally smooth surface $\partial K_n / \{0\}$; that is,

$$\langle Ax, Qx \rangle = 0 \quad \text{for all } x \in \partial K_n. \tag{3.32}$$

This is equivalent to $A \in p(K_n)$ and $-A \in p(K_n)$. Hence in view of Theorem 3.5, (3.32) is equivalent to the existence of real numbers ξ_1 and ξ_2 such that

$$W(Q_n, A) - \xi_1 Q_n \leqq 0 \quad \text{and} \quad W(Q_n, -A) - \xi_2 Q_n \leqq 0. \tag{3.33}$$

But (3.33) implies that $\xi_1 = -\xi_2$ and $W(Q_n, A) = \xi_1 Q_n$. In view of (3.15), the conclusion of the theorem follows. □

We conclude with some remarks.

Remark 3.34. (i) The proof of Theorem 3.31 shows that $p(\partial K_n)$ is the maximal subspace of the closed convex cone $p(K_n) \in R^{n,n}$. The theorem implies that $\dim (p(\partial K_n)) = (n^2 - n + 2)/2$.

(ii) It is interesting to note that if A satisfies either of the sufficient conditions (3.18) or (3.21), or if A is of the form specified in Theorem 3.31, then A must satisfy the conditions of Elsner [1] for the existence of a proper cone K such that $A \in p(K)$; namely, that the *spectral abscissa*

$$\lambda(A) := \max \{ \operatorname{Re} \lambda : \lambda \text{ is an eigenvalue of } A \}$$

is an eigenvalue of A and no eigenvalue λ of A with $\operatorname{Re} \lambda = \lambda(A)$ can have degree exceeding that of $\lambda(A)$. (By the degree of an eigenvalue, we mean its degree in the minimal polynomial.)

(iii) Our results can be extended to general ellipsoidal cones by applying Lemma 2.7. In particular, let $K = K(Q, u_n)$ be a given ellipsoidal cone, and let T be a nonsingular matrix such that $K = TK_n$. (One such T is provided by Lemma 2.7.) Then $A \in p(K)$ if and only if $T^{-1}AT \in p(K_n)$, and likewise, $A \in p(\partial K)$ if and only if $T^{-1}AT \in p(\partial K_n)$.

(iv) In view of (3.7), $A \in p(K_n)$ if and only if $x^t W(Q_n, A)x \leqq 0$ for all $x \in R^n$ such that $x_n = 1$ and $\sum_{i=1}^{n-1} x_i^2 = 1$. Hence a necessary and sufficient condition for $A \in p(K_n)$ is

$$\max \{ y^t(A_1 + A_1^t)y + 2y^t(c - d) : \|y\| = 1 \} \leqq 0. \tag{3.35}$$

A numerical method for obtaining the maximum in (3.35) may be found, e.g., in Fletcher [2]. Thus we can computationally check whether $A \in p(K_n)$ in cases where our necessary conditions are met, but sufficiency is not.

Acknowledgments. We are indebted to the referees and A. Berman for detecting errors in earlier versions of this work.

REFERENCES

[1] L. ELSNER, *Monotonie und Randspektrum bei Vollstetigen Operatoren*, Arch. Rational Mech. Anal., 36 (1970), pp. 356–365.

[2] R. FLETCHER, *Practical Methods of Optimization*, Second Edition, John Wiley, New York, 1987.

[3] R. LOEWY AND H. SCHNEIDER, *Positive operators on the n-dimensional ice-cream cone*, J. Math. Anal. Appl., 49 (1975), pp. 375–392.

[4] H. SCHNEIDER AND M. VIDYASAGAR, *Cross-positive matrices*, SIAM J. Numer. Anal., 7 (1970), pp. 508–519.

[5] R. STERN AND H. WOLKOWICZ, *Invariant ellipsoidal cones*, Res. Report CORR 88-53, University of Waterloo, Waterloo, Ontario, Canada, 1988.

[6] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 1, pp. 166–171, January 1991

013

NONNEGATIVE IDEMPOTENT MATRICES AND THEIR GRAPHS*

MORDECHAI LEWIN†

Abstract. A graph-theoretic characterization of nonnegative matrices having idempotent pattern is given. Given a directed graph, it is either decided that its adjacency matrix has no idempotent pattern, or else a nonnegative, idempotent matrix whose graph is the given graph is supplied.

Key words. nonnegative matrix, idempotent, pattern, directed graph, clique, source, sink

AMS(MOS) subject classifications. 15A18, 15A47, 05C50

1. Introduction. A square matrix A is called *idempotent* if $A^2 = A$. In [2] Flor establishes the structure of nonnegative idempotent matrices. In particular, he shows the following proposition.

PROPOSITION [2, Thm. 2]. *Let A be a nonnegative idempotent matrix. Then there exists a permutation matrix P such that*

$$P^{-1}AP = \begin{bmatrix} J & JT & 0 & 0 \\ 0 & 0 & 0 & 0 \\ SJ & SJT & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

where J is the direct sum of nonnegative idempotent matrices of rank one. Conversely, while S and T are arbitrary matrices of appropriate sizes, every matrix of the above-mentioned form is idempotent.

Let $A = (a_{ij})$ be a given square matrix. Let $G(A)$ be the directed graph associated with A such that the order of the graph is the order of the matrix with $a_{ij} \neq 0$ if $(i, j) \in E(G)$ where $E(G) = E$ is the set of edges of G. A nonnegative matrix A is *positive* if all its elements are positive.

A nonnegative matrix A is of *idempotent pattern* if there exists a nonnegative idempotent matrix having the same zero-pattern as A. Let $\mathfrak{I}$ be the set of nonnegative idempotent matrices and let $\mathfrak{N}$ be the set of all matrices of idempotent pattern.

The purpose of this paper is to characterize this idempotent pattern family graph theoretically. This characterization enables us to decide whether an arbitrarily given matrix A is in $\mathfrak{N}$ or not, simply by observing the graph of the matrix. In case of an affirmative answer, we suggest a construction of a nonnegative idempotent matrix having the same zero pattern as A.

2. Definitions. Let $A = (a_{ij})$ be a square matrix and let G be its directed graph. Let $V = \{1, 2, \cdots, n\}$ be the set of vertices of G, n being the order of the matrix. An (i, j)-*walk* in G is a sequence of directed edges from i to j of the form $(i, v_1), (v_1, v_2), \cdots, (v_k, j)$. We shall also use the form $(i, v_1, v_2, \cdots, v_k, j)$. An (i, j)-*path* is a walk in which no edge appears more than once. A k-(i, j)-*walk*(*path*) in G is a walk(path) of length k (number of edges from vertex i to vertex j). A *clique* in a (directed) graph is a maximal complete (directed) subgraph. A *proper clique* is a clique whose set of edges is nonvoid. A clique of order one is a *loop*. A *null-clique* is a nonisolated vertex not belonging to any proper clique. A subgraph of G is *clique-free* if none of its edges belongs to a proper clique. A *source* (*sink*) is a vertex with positive outgoing (incoming) degree and zero

* Received by the editors March 22, 1989; accepted for publication (in revised form) November 26, 1989. This research was supported by the Technion V.P.R. Fund and the Loewengard Research Fund.

† Department of Mathematics, Technion–Israel Institute of Technology, Haifa 32000, Israel (MAR32AA@TECHNION. BITNET).

incoming (outgoing) degree. Let *So* (*Si*) stand for the set of sources (sinks) of a given graph. A graph is *transitive* if any two of its edges (i, j), (j, k) imply that (i, k) is in the graph.

A directed graph is *strongly connected* if there is a walk from every vertex to every other vertex in the graph.

A matrix A is *reducible* if there is a permutation matrix P such that

$$P^{-1}AP=\begin{bmatrix} X & 0 \\ Y & Z \end{bmatrix}$$

with X and Z both square blocks. Otherwise, the matrix is called *irreducible*. It is now common knowledge that A is irreducible if and only if $G(A)$ is strongly connected (see, for example, [4]).

3. A necessary condition. Let $\bar{A}$ be a matrix in $\mathfrak{N}$. Then there exists a matrix A in $\mathfrak{I}$ having the same pattern as $\bar{A}$. Considering $G(A)$, it follows that $(i, j) \in E(G(A))$ if and only if for some k, $1 \leqq k \leqq n$ and (i, k) and (k, j) are both in E. An immediate consequence of the idempotence of A is that $G(A)$ is transitive. Now let A be irreducible. Then $G(A)$ is strongly connected. Since it is also transitive, it is necessarily a complete directed graph, so that A is positive. We should also bear in mind that in a transitive directed graph, distinct proper cliques are disjoint. We thus have obtained the following lemma.

LEMMA 1. *A nonnegative idempotent matrix is either positive or reducible.*

Let $A \in \mathfrak{I}$. Without loss of generality, we may assume $G(A)$ to be connected, since each connected component of $G(A)$ is the graph of an idempotent matrix of a smaller order. We have the following lemma.

LEMMA 2. *Let $A \in \mathfrak{I}$. Let further (x_0, x_1, x_2, x_3) be a 3-walk in $G(A)$. Then x_1 and x_2 belong to the same proper clique in G.*

Proof. Let U_1 be the subgraph of $G(A)$ spanned on x_1 and all its incoming vertices and let U_2 be spanned on the rest of the vertices. Then $G(A)$ consists of the disjoint union of U_1 and U_2 (the latter may be empty) and possibly some edges from U_1 to U_2. Let k be the order of U_2. Number its vertices from 1 to k and those of U_1 from $k+1$ to n. Then A already assumes the form

$$\begin{pmatrix} X & 0 \\ Y & Z \end{pmatrix}.$$

If x_2 is in U_1, then, by the transitivity condition, x_2 and x_1 belong to the same proper clique. We may therefore assume x_2 in U_2. Then x_3 is also in U_2. The edges of U_2 and U_1 represent the nonzero entries of X and Z, respectively; the edges from U_1 to U_2 represent the nonzero entries of Y. We now have

$$A^2=\begin{bmatrix} X^2 & 0 \\ YX+ZY & Z^2 \end{bmatrix}=\begin{bmatrix} X & 0 \\ YX+ZY & Z \end{bmatrix}.$$

It follows that both blocks X and Z are idempotent. We also get $YX + ZY = Y$. Multiplying both sides from the left by Z we get $ZYX + ZY = ZY$ and hence

$$ZYX=0. \tag{1}$$

The graph theoretic interpretation of (1) is that there is no 3-walk starting in U_1 and ending in U_2. This contradiction implies Lemma 2.

COROLLARY 1. *Let $A \in \mathfrak{N}$. Then every null-clique of $G(A)$ is either a source or a sink.*

Proof. Let x be a null-clique in $G(A)$. Suppose x is neither a source nor a sink. Since x is nonisolated by definition, it has an incoming edge (y, x) and an outgoing edge (x, z) with x, y, z all distinct. Since A is of idempotent pattern, there is a 2-walk from y to x in $G(A)$. Let it be (y, u, x) for some u. Then (y, u, x, z) is a 3-walk in $G(A)$ and so, by Lemma 2, u and x belong to the same proper clique in G, an obvious contradiction.

Let $A \in \mathfrak{N}$, and let x be a vertex and C a proper clique of $G(A)$. Then, because of the transitivity of G, x is adjacent to all the vertices of C or to none. A *contraction* of G is a graph derived from G, whose vertices are the cliques of G. The edges in G whose vertices belong to distinct cliques are edges from a source to a proper clique, from a proper clique to a sink, or from a source to a sink. The sets So and Si may be void (one or both), but the set C of proper cliques may not be empty, unless $A = 0$. We then obtain the following corollary.

COROLLARY 2. *Let $A \in \mathfrak{N}$. Then the longest path in the contraction of G is of length at most two.*

A transitive graph consisting of a disjoint union of proper cliques, with sources or sinks attached to some of the proper cliques, and possibly some isolated vertices, will henceforth be termed an *admissible graph*.

Let G be an admissible graph. Order the cliques in ascending order: first the sinks, then the proper cliques, and finally the sources. Now order the vertices of the proper cliques lexicographically. This induces a numbering of the vertices of G which we shall refer to as *admissible numbering*.

COROLLARY 3. *Distinct proper cliques are completely disjoint.*

4. Sufficiency. We now state Lemma 3.

LEMMA 3. *Let G be an admissible graph. Then there exists a nonnegative, idempotent matrix A such that $G(A)$ is its graph.*

Proof. Let $C_1, C_2, \cdots, C_k$ be the proper cliques of G and let $x_1, x_2, \cdots, x_n$ be an admissible numbering of G. For each C_i there exists a positive idempotent matrix A_i. Let A_0 be the direct sum of the A_i. Consider A_0 as a principal submatrix of A. Let x be a sink in G adjacent to some of the cliques $C_1, C_2, \cdots$. By Perron's theorem [3] there exists a positive eigenvector for each A_i in A_0. Let $\mathbf{v}_{ix}$ be an eigenvector of A_i and let its entries correspond to the edges from C_i to x. Consider the entry a_{jx} of A. We have the following cases.

Case 1. $(j, x) \notin G$. Then put $a_{jx} = 0$.

Case 2. $(j, x) \in G$. We distinguish two subcases.

Subcase 2.1. $j \in C_i$ for some i. Then let a_{jx} be an entry of the corresponding eigenvector $\mathbf{v}_{ix}$ of A_i.

Subcase 2.2. The vertex j is a source. Then leave the value for a_{jx} open for the time being.

After having dealt with all the sinks, let us turn to the sources. Let y denote such a source. Repeat the same argument for the eigenvectors $\mathbf{u}^t_{yk}$ of A^t_i and consider the cases $(y, k) \notin G$ and $(y, k) \in G$, with k belonging to some proper clique, leaving open the case where k is a sink.

Case 1 and subcase 2.1 supply us with q-dimensional eigenvectors of A_0, where q is the order of A_0. To each sink (source) there corresponds a vertical (horizontal) concatenation of eigenvectors of the A_i (A^t_i), and maybe zeros.

We may now conclude subcase 2.2. Let $(y, x) \in G$, y a source, x a sink. Let further $\mathbf{v}_x$ and $\mathbf{u}^t_y$ correspond to the appropriate eigenvectors of A_0 (A^t_0). Put $a_{yx} = \langle \mathbf{u}^t_y, \mathbf{v}_x \rangle$, the standard inner product.

Finally, put $a_{xj} = a_{ky} = 0$ for all x in *Si*, y in *So*, and all j and k. This yields a zero block of order $r \times n$ above and a zero block of order $n \times s$ on the right-hand side of A where $r = |Si|$, $s = |So|$ ($|S|$ meaning the number of elements of S). We thus obtain

a matrix of the form

$$A = \begin{bmatrix} 0 & 0 & 0 \\ V_x & A_0 & 0 \\ V_y V_x & V_y & 0 \end{bmatrix}$$

where V_x is a $q \times r$ block, whose columns are all eigenvectors of A_0 and where V_y is an $s \times q$ block whose rows are all transposes of eigenvectors of A_0^t. The block-scheme of A is thus

$$\begin{bmatrix} r\times r & r\times q & r\times s \\ q\times r & q\times q & q\times s \\ s\times r & s\times q & s\times s \end{bmatrix}.$$

It is quite clear that a matrix so described is idempotent.

This completes the proof of Lemma 3.

As an example we present a graph of order 10 by means of its contraction (Fig. 1) and then a matrix which is nonnegative and idempotent. Note that all the parameters are arbitrary and independent of each other.

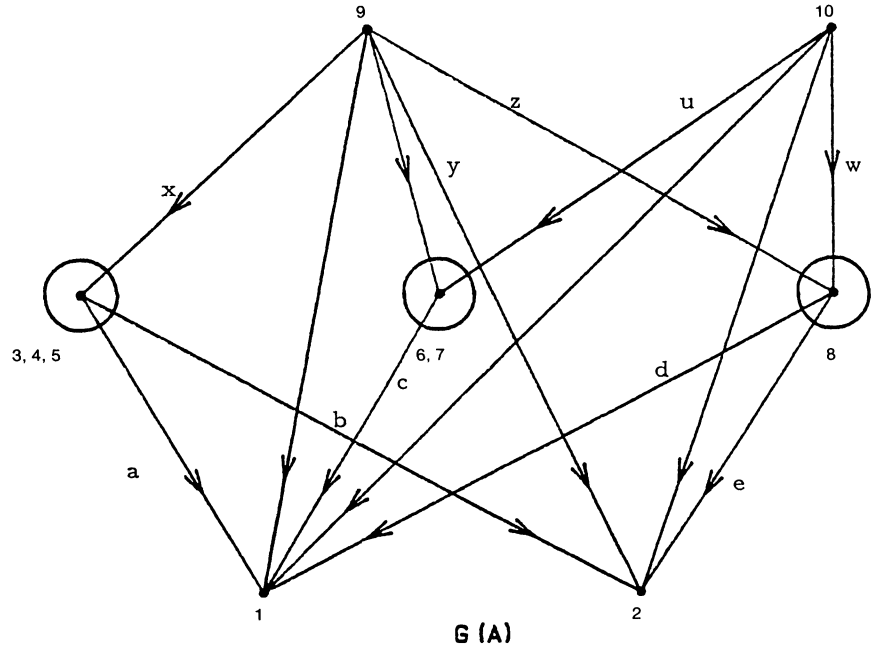


FIG. 1

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ a & b & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ a & b & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ a & b & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ c & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ c & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ d & e & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ a_{91} & a_{92} & x & x & x & y & y & z & 0 & 0 \\ a_{10,1} & a_{10,2} & 0 & 0 & 0 & u & u & w & 0 & 0 \end{bmatrix}$$

with $a_{91} = 3ax + 2cy + dz$

$a_{92} = 3bx + ez$

$a_{10,1} = 2cu + dw$

$a_{10,2} = ew$

Now put $A_0 = J$, $V_x = JT$, $V_y = SJ$ with arbitrary T and S. Considering that $V_x = A_0V_x = JT$, $V_y = V_yA_0 = SJ$ we have $V_yV_x = SJT$. By interchanging the first and second rows and columns of blocks we arrive at Flor's matrix [2, Thm. 2]. (A fourth row and column of blocks will appear if there are isolated vertices in $G(A)$.) V_x is necessarily composed of eigenvectors of J, so is V_y (from the left). We thus obtain Flor's result.

Since J is a direct sum of irreducible, idempotent blocks, the latter are all positive and hence, once again by Perron's theorem, each has a simple eigenvalue 1, so that the rank of each positive block is 1.

Note that *nonnegative* in Flor's result may be replaced by *positive*.

5. Conclusion. We are now in a position to state our Main Theorem.

THEOREM 1. *A nonnegative square matrix is of idempotent pattern if and only if its graph is admissible.*

A matrix $A = (a_{ij})$ is of *symmetric* pattern if $a_{ij} = 0$ implies $a_{ji} = 0$ for arbitrary i, j. We have the following corollary.

COROLLARY 3. *A nonnegative matrix of symmetric pattern is of idempotent pattern if and only if it is permutationally equivalent to a direct sum of positive (and possibly one zero) blocks. (Compare with* [1, (3.4)].)

Proof. A necessary condition for idempotency of a matrix A is that its graph be admissible. But because of symmetricity, G has neither sources nor sinks, so that G is a disjoint union of cliques and possibly isolated vertices. The converse is clear. This proves the corollary.

Here is a characterization of positive, symmetric, idempotent matrices. Let $\alpha_1 = 1$ and let $\alpha_2, \alpha_3, \cdots, \alpha_n$ be arbitrary positive numbers. We then have the following theorem.

THEOREM 2. *A positive, symmetric matrix $A = (a_{ij})$ is idempotent if and only if it has the form $A = aB$, $B = (b_{ij})$, $b_{ij} = \alpha_i\alpha_j$, and*

$$a = \left(\sum_{i=1}^{n} \alpha_i^2 \right)^{-1}.$$

Proof. Let A be as described. Put $A^2 = (c_{ij})$, with

$$c_{ij} = \sum_{k=1}^{n} a_{ik}a_{kj} = \sum_{k=1}^{n} a_{ik}a_{jk} = a^2 \sum_{k=1}^{n} \alpha_i\alpha_j\alpha_k^2 = a^2\alpha_i\alpha_j \sum_{k=1}^{n} \alpha_k^2 = a\alpha_i\alpha_j = ab_{ij} = a_{ij},$$

so that A is idempotent.

Now let A be positive and idempotent. Since A is positive, its rank equals 1. This means that each row is a positive multiple of the first row. Let the ith row, $r_i = \lambda_i r_1$, with $\lambda_1 = 1$. Then $a_{ij} = \lambda_i\lambda_j a_{11}$. Put $a_{11} = u$. We may normalize A by extracting a from the matrix and writing $A = aB$. Then $b_{ij} = \lambda_i\lambda_j$. Since $a = c_{11} = a^2 \sum_{i=1}^{n} \lambda_i^2$, we get $a = (\sum_{i=1}^{n} \lambda_i^2)^{-1}$. This proves the theorem (see also [1, Cor. 3.5]).

Let A be nonnegative, stochastic, and idempotent. Then clearly $G(A)$ has no sinks. A matrix A is of *stochastic pattern* if there exists a stochastic matrix having the same zero-pattern as A. We have Corollary 4.

COROLLARY 4. *A matrix is of stochastic and idempotent pattern if and only if its graph is admissible and has no sinks.*

Proof. Let A be of stochastic, idempotent pattern. Let A_0 be the stochastic, idempotent representative. By what we just showed, $G(A_0)$ is admissible and has no sinks. Now let G be an admissible graph without sinks. For every clique of G we may introduce stochasticity conditions. Every source contributes a row vector which is the transpose of an eigenvector of A_0 where A_0 is constructed from the disjoint union of the cliques.

Therefore every such row vector may be normalized by multiplying each element of that row vector by its row sum. We thus get a stochastic matrix which is still idempotent. This proves the corollary.

COROLLARY 5. *A matrix A is of doubly stochastic and idempotent pattern if and only if its graph is a disjoint union of proper cliques.*

Proof. Clearly $G(A)$ may have neither sinks nor sources.

COROLLARY 5′. *A doubly stochastic matrix is of idempotent pattern if and only if it is permutationally equivalent to a direct sum of positive square blocks.*

Remark. It has already been mentioned in [1] that the only positive doubly stochastic idempotent n-square matrix is the one whose entries are all n^{-1}.

REFERENCES

[1] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[2] P. FLOR, *On groups of nonnegative matrices*, Compositio Math., 21 (1969), pp. 376–382.

[3] O. PERRON, *Zur Theorie der Matrizen*, Math. Ann., 64 (1907), pp. 248–263.

[4] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 1, pp. 172–194, January 1991

014

THE RESTRICTED SINGULAR VALUE DECOMPOSITION OF MATRIX TRIPLETS*

HONGYUAN ZHA†

Abstract. In this paper the concept of *restricted singular values* of matrix triplets is introduced. A decomposition theorem concerning the general matrix triplet (A, B, C), where $A \in \mathcal{C}^{m\times n}$, $B \in \mathcal{C}^{m\times p}$, and $C \in \mathcal{C}^{q\times n}$, which is called *the restricted singular value decomposition* (RSVD), is proposed. This result generalizes the well-known singular value decomposition, the generalized singular value decomposition, and the recently proposed product-induced singular value decomposition. Connection of restricted singular values with the problem of determination of matrix rank under restricted perturbation is also discussed.

Key words. matrix rank, singular values, generalized singular values, product-induced singular values, restricted singular values, matrix decompositions

AMS(MOS) subject classifications. 15A09, 15A12, 15A23, 65F20

1. Introduction.

Rank determination of matrices is an important problem in numerical linear algebra [7]. In applications, the matrix A_0, the rank of which is to be determined, is always contaminated with errors, i.e., instead of knowing A_0 exactly we only have $A = A_0 + E$, an approximation of A_0, where E represents the error or perturbation matrix. The rank determination problem is how to estimate the rank of A_0, if A and some information of E are available. Usually only an upper bound on certain norms of E, e.g., 2-norm, is assumed to be known. In this case *the singular value decomposition* (SVD) is a useful tool for solving the problem [4], [7].

In many situations, however, more information about the error matrix E than the simple upper bound of its 2-norm is available, e.g., E has some special structure or, in other words, is restricted to a special class of matrices. SVD-based methods in these situations are likely to lead to conservative rank estimations.

In order to illustrate the situation, we give the following simple example. Consider the matrix

$$A_0 = \begin{pmatrix} 0 & 1 \\ a_2 & a_1 \end{pmatrix}.$$

If we assume that A_0 results from the second-order ordinary differential equation

$$\frac{d^2x}{dt^2} - a_2\frac{dx}{dt} - a_1 x = f,$$

then only a_1 and a_2 are subject to errors, and the "0" and "1" entries in A_0 are exact. Hence the error matrix E can only be of the following three forms:

(i) Only a_2 is changeable:

$$E = \begin{pmatrix} 0 & 0 \\ e_{21} & 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} e_{21}(1, 0).$$

* Received by the editors November 6, 1987; accepted for publication (in revised form) November 28, 1989. This paper was finished while the author was a visitor at Konrad-Zuse-Zentrum für Informationstechnik Berlin, Federal Republic of Germany ; Part of the work was done while the author was a graduate student with the group of Professor Jiang at Institute of Mathematics, Fudan University, Shanghai, People's Republic of China

† Scientific Computing and Computational Mathematics, MJH 460, Stanford University, Stanford, California 94305-2140 (zha@patience.stanford.edu).

(ii) Only a_1 is changeable:

$$E = \begin{pmatrix} 0 & 0 \\ 0 & e_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} e_{22}(0,1).$$

(iii) Both a_1 and a_2 are changeable:

$$E = \begin{pmatrix} 0 & 0 \\ e_{21} & e_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} (e_{21}, e_{22}).$$

Observe that any E of the form in (ii) cannot change the rank of the original matrix A_0, while SVD-based methods cannot lead to such a conclusion.

In this paper we consider the error matrix E which is restricted to a special class of matrices, i.e., $E = BDC$, where B and C are known matrices, and D is an arbitrary matrix with an upper bound on its 2-norm. In §2 we introduce the concept of *restricted singular values* (RSVs) for the restricted error matrix $E = BDC$ and discuss the problem of rank determination of matrices under the perturbation of this special class of error matrices. In §3 we consider two special cases of RSVs, i.e., singular values (SVs) and generalized singular values (GSVs). In §4 we derive the main result of this paper, which we call *the restricted singular value decomposition* (RSVD) of matrix triplets concerning the simultaneous reduction of three matrices into quasi-diagonal form. Section 5 summarizes the paper and gives some comments concerning the further research on the subject of RSVD. Although only 2-norm is used in this paper, we note that the results of this paper can be extended to the case of unitarily invariant norms [5].

Notation. In this paper, only the complex matrices are considered, while the case of real matrices can be similarly considered. Throughout the paper $\mathcal{C}^{m\times n}$ denotes the set of all $m \times n$ complex matrices. The matrix A^H is the complex conjugate transpose of A, $\|\cdot\|$ and $\|\cdot\|_F$ are the 2-norm and Frobenius norm, respectively. I_s represents the identity matrix of order s; O with different subscripts and superscripts (e.g., $O_A^{(1)}$) denotes zero matrices of different dimensions. Sometimes we just use I and O to denote an identity matrix or a zero matrix of different dimensions when their dimensions are clear from the context.

Note. Originally we used the name "Structured Singular Values" for the concept introduced in this paper. Some people, especially B. De Moor, G. Golub, and S. Van Huffel,[1] brought to our attention that the name had been used in control theory under a different setting. Therefore we adopt here the name "Restricted Singular Values," which was suggested by B. De Moor and G. Golub.

2. Restricted singular values and rank determination of matrices. Let $A \in \mathcal{C}^{m\times n}$ and the error matrix be of the form $E = BDC$, where $B \in \mathcal{C}^{m\times p}$, $D \in \mathcal{C}^{p\times q}$, and $C \in \mathcal{C}^{q\times n}$.

DEFINITION 2.1. The restricted singular values (RSVs) of the matrix triplet (A, B, C) are defined as follows:

$$(2.1)\quad \sigma_k(A,B,C) = \min_{D\in\mathcal{C}^{p\times q}} \{\|D\|_2 \mid \operatorname{rank}(A+BDC) \le k-1\}, \qquad k = 1,\cdots,n.$$

Before we proceed, some remarks are in order concerning the above definition.

Remark 2.1. If for some k $(1 \le k \le n)$ there is no $D \in \mathcal{C}^{p\times q}$ such that $\operatorname{rank}(A+BDC) \le k-1$, then $\sigma_k(A,B,C)$ is defined to be ∞.

[1] Private communications, March 1989.

Remark 2.2. For notational convenience, we define $\sigma_k(A,B,C) = 0$, for $k = n - \min(m,n), \cdots, n$.

Remark 2.3. It can be readily verified that the RSVs are arranged in nondecreasing order, i.e.,

$$\sigma_k(A,B,C) \geq \sigma_{k+1}(A,B,C), \qquad k = 1, \cdots, n-1. \tag{2.2}$$

Considering the example in the above section, we distinguish three cases correspondingly. In the notation of the above definition we have

$$\text{(i)} \quad A = \begin{pmatrix} 0 & 1 \\ a_2 & a_1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad C = (1,0),$$

$$\sigma_1(A,B,C) = \infty, \qquad \sigma_2(A,B,C) = |a_2|.$$

$$\text{(ii)} \quad A = \begin{pmatrix} 0 & 1 \\ a_2 & a_1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad C = (0,1),$$

$$\begin{aligned} &\sigma_1(A,B,C) = \infty, \quad \sigma_2(A,B,C) = \infty \quad \text{if } a_2 \neq 0, \\ &\sigma_1(A,B,C) = \infty, \quad \sigma_2(A,B,C) = 0 \quad \text{if } a_2 = 0. \end{aligned}$$

$$\text{(iii)} \quad A = \begin{pmatrix} 0 & 1 \\ a_2 & a_1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\sigma_1(A,B,C) = \infty, \qquad \sigma_2(A,B,C) = |a_2|.$$

This is an example of matrices of low orders, and we can find the RSVs by direct computation. For matrices of higher orders, we need a decomposition theorem, which will be the subject of §4. We now briefly discuss the connection of RSVs and rank determination of matrices. The problem is to estimate the rank of

$$A_0 = A + BDC$$

where (A,B,C) is known, and in addition $\| D \|_2 \leq \varepsilon$.

Assume further that the following inequalities for ε hold:

$$\sigma_1(A,B,C) \geq \cdots \geq \sigma_k(A,B,C) > \varepsilon \geq \sigma_{k+1}(A,B,C) \geq \cdots \geq \sigma_n(A,B,C);$$

then the best possible estimation of the rank of A_0 is k, in the sense that there exists a matrix D_0, satisfying $\| D_0 \|_2 \leq \varepsilon$ such that

$$\text{rank } (A + BD_0C) = k$$

but there exists no D satisfying $\| D \|_2 \leq \varepsilon$ such that

$$\text{rank } (A + BDC) < k.$$

Such strategy of estimation is also used in the determination of numerical rank [4], [7].

3. Singular values and generalized singular values. In this section we discuss two special cases of RSVs, i.e.,

$$\begin{array}{lll} (1) & B = I_m & \text{and} \quad C = I_n, \\ (2) & B = I_m & \text{or} \quad C = I_n. \end{array}$$

We will show that the RSVs of the matrix triplet (A, B, C) corresponding to these two special cases are just the well-known *singular values* (SVs) and *generalized singular values* (GSVs), respectively [4], [6].

3.1. Singular values of a complex matrix. We first cite the following result.

THEOREM 3.1 ([4], [5]). *Let the* SVs *of* A *be*

$$\sigma_1 \geq \cdots \geq \sigma_n \geq 0; \tag{3.1}$$

then

$$\sigma_k = \min_{E \in \mathcal{C}^{m \times n}} \{ \| E \|_2 \mid \text{rank}\,(A + E) \leq k - 1 \}, \qquad k = 1, \cdots, n, \tag{3.2}$$

and there exists a matrix E_k *satisfying* $\|E_k\|_2 = \sigma_k$ *such that*

$$\text{rank}\,(A + E_k) = k - 1, \qquad i = 1, \cdots, n.$$

We note that Remark 2.2 is also applicable here, i.e., we simply define $\sigma_k = 0$ for $k = n - \min(m, n), \cdots, n$. Using the notation of Definition 2.1 we can rewrite Theorem 3.1 as Corollary 3.1.

COROLLARY 3.1.

$$\sigma_k(A, I_m, I_n) = \sigma_k, \qquad k = 1, \cdots, n. \tag{3.3}$$

It is also easy to establish the following inequalities.

COROLLARY 3.2. *Assume that* $B \neq 0$ *and* $C \neq 0$*; then*

$$\sigma_k \leq \| B \|_2 \| C \|_2 \sigma_k(A, B, C), \qquad k = 1, \cdots, n. \tag{3.4}$$

Proof. Let $D_k \in \mathcal{C}^{p \times q}$ satisfy $\|D_k\|_2 = \sigma_k(A, B, C,)$ (see Theorem 4.2) and

$$\text{rank}\,(A + BD_kC) = k - 1;$$

then from the above theorem, it follows that

$$\sigma_k \leq \|BD_kC\|_2 \leq \| B \|_2 \| C \|_2 \sigma_k(A, B, C)$$

which proves the corollary. □

3.2. Generalized singular values. We only consider the case $B = I_m$ and C is a general complex matrix. The error matrix is now $E = DC$. The dual case that B is a general matrix and $C = I_n$ can be discussed similarly.

The concept of GSVs of matrix pencils was introduced by Van Loan [8] (where he used the term B-singular values). Paige and Saunders provided a slight generalization of Van Loan's result in order to treat all the possible cases [6]. Since GSVs have many applications in numerical linear algebra problems and thus are of their own

interests, here we give an alternative derivation of the so-called generalized singular value decomposition (GSVD) of matrix pairs, in which the two matrices have the same number of columns. Our approach here is different from those in [6] and [8].

THEOREM 3.2 ([6], [8]). *Let $A \in \mathcal{C}^{m\times n}$ and $C \in \mathcal{C}^{q\times n}$; then there exist unitary matrices U and V and nonsingular matrix Q such that*

$$UAQ = \begin{matrix} k & n-k \end{matrix} \begin{pmatrix} \Sigma_A, & O \end{pmatrix}, \qquad VCQ = \begin{matrix} k & n-k \end{matrix} \begin{pmatrix} \Sigma_C, & O \end{pmatrix}, \tag{3.5}$$

$$\Sigma_A = \begin{pmatrix} I_r & & \\ & S_A & \\ & & O_A \end{pmatrix}, \qquad \Sigma_C = \begin{pmatrix} O_C & & \\ & S_C & \\ & & I_{k-r-s} \end{pmatrix} \tag{3.6}$$

where

$$S_A = \operatorname{diag}(\alpha_{r+1}, \cdots, \alpha_{r+s}), \qquad S_C = \operatorname{diag}(\beta_{r+1}, \cdots, \beta_{r+s})$$

and

$$\begin{aligned} 1 > \alpha_{r+1} \geq \cdots \geq \alpha_{r+s} > 0, \qquad & 0 < \beta_{r+1} \leq \cdots \leq \beta_{r+s} < 1, \\ \alpha_i^2 + \beta_i^2 = 1, \qquad & i = r+1, \cdots, r+s. \end{aligned} \tag{3.7}$$

The integer indices can be expressed as follows:

$$k = \operatorname{rank}\begin{pmatrix} A \\ C \end{pmatrix}, \qquad r = \operatorname{rank}\begin{pmatrix} A \\ C \end{pmatrix} - \operatorname{rank}(C),$$

and

$$s = \operatorname{rank}(A) + \operatorname{rank}(C) - \operatorname{rank}\begin{pmatrix} A \\ C \end{pmatrix}.$$

Proof. The proof is constructive and consists of four steps. The transformations of each step are of the following form:

$$A^{(k+1)} = U^{(k)} A^{(k)} Q^{(k)}, \qquad C^{(k+1)} = V^{(k)} C^{(k)} Q^{(k)}$$

where $U^{(k)}$ and $V^{(k)}$ are unitary matrices and $Q^{(k)}$ nonsingular. In each step we only specify the $U^{(k)}$, $V^{(k)}$, and $Q^{(k)}$ and the resulting matrices $A^{(k+1)}$ and $C^{(k+1)}$. Set $A^{(1)} = A$ and $C^{(1)} = C$.

Step 1. Let the SVD of the matrix C be $U_1 C V_1 = \operatorname{diag}(O, \Sigma_C^{(1)})$, where $\Sigma_C^{(1)} = \operatorname{diag}(s_1, \cdots, s_t)$ and $s_1 \geq, \cdots, \geq s_t > 0$. Set

$$\begin{aligned} U^{(1)} &= I, \qquad V^{(1)} = U_1, \\ Q^{(1)} &= V_1 \operatorname{diag}(I, \Sigma_C^{-1}); \end{aligned}$$

then

$$A^{(2)} = \begin{matrix} n-t & t \end{matrix} \begin{pmatrix} A_1^{(2)}, & A_2^{(2)} \end{pmatrix},$$

$$C^{(2)} = \begin{pmatrix} O & O \\ O & I_t \end{pmatrix}.$$

Step 2. Let the SVD of the matrix $A_1^{(2)}$ be $U_2 A_1^{(2)} V_2 = \operatorname{diag}(\Sigma_A^{(2)}, O)$, where $\Sigma_A^{(2)} = \operatorname{diag}(t_1, \cdots, t_r)$ and $t_1 \geq, \cdots, \geq t_r > 0$. Set

$$U^{(2)} = U_2, \qquad V^{(2)} = I,$$
$$Q^{(2)} = \operatorname{diag}(V_2, I) \operatorname{diag}((\Sigma_A^{(2)})^{-1}, I);$$

then

$$A^{(3)} = \begin{array}{c} \\ r \\ m-r \end{array} \begin{array}{c} \begin{array}{ccc} r & n-r-t & t \end{array} \\ \begin{pmatrix} I_r & O & A_{13}^{(3)} \\ O & O & A_{23}^{(3)} \end{pmatrix} \end{array}, \qquad C^{(3)} = C^{(2)}.$$

Step 3. Let the SVD of the matrix $A_{23}^{(3)}$ be $U_3 A_{23}^{(3)} V_3 = \operatorname{diag}(\Sigma_A^{(3)}, O)$, where $\Sigma_A^{(3)} = \operatorname{diag}(w_1, \cdots, w_s)$ and $w_1 \geq, \cdots, \geq w_s > 0$. Let $\alpha_i = w_i(1 + w_i^2)^{-1/2}$ and $\beta_i = (1 + w_i^2)^{-1/2}$, $i = r+1, \cdots, r+s$, and $S_A = \operatorname{diag}(\alpha_{r+1}, \cdots, \alpha_{r+s})$, $S_C = \operatorname{diag}(\beta_{r+1}, \cdots, \beta_{r+s})$. It is easy to check that α_i, β_i $(i = r+1, \cdots, r+s)$ satisfy (3.7). Set

$$U^{(3)} = \operatorname{diag}(I, U_3), \qquad V^{(3)} = \operatorname{diag}(I, V_3^H),$$
$$Q^{(3)} = \begin{pmatrix} I & -A_{13}^{(3)} \\ O & I \end{pmatrix} \operatorname{diag}(I, V_3) \operatorname{diag}(I, S_C, I);$$

then

$$A^{(4)} = \begin{array}{c} \\ r \\ s \\ m-r-s \end{array} \begin{array}{c} \begin{array}{cccc} r & n-r-t & s & t-s \end{array} \\ \begin{pmatrix} I_r & O & O & O \\ O & O & S_A & O \\ O & O & O & O \end{pmatrix} \end{array},$$

$$C^{(4)} = \begin{array}{c} \\ n-k+r \\ s \\ k-r-s \end{array} \begin{array}{c} \begin{array}{ccc} n-t & s & t-s \end{array} \\ \begin{pmatrix} O & O & O \\ O & S_C & O \\ O & O & I_{k-r-s} \end{pmatrix} \end{array}.$$

Step 4. After suitable permutations P_1 and P_2 and set $k = t + r$ we obtain

$$A^{(5)} = A^{(4)} P_1 = \left(\begin{array}{ccc|c} I_r & & & \\ & S_A & & O \\ & & O_A & \end{array}\right),$$
$$C^{(5)} = P_2 C^{(4)} P_1 = \left(\begin{array}{ccc|c} O_C & & & \\ & S_C & & O \\ & & I_{k-r-s} & \end{array}\right),$$

thus we have obtained the required quasi-diagonal form. It is easy to verify that

$$\operatorname{rank}(A) = r + s, \qquad \operatorname{rank}(C) = k - r, \qquad \operatorname{rank}\begin{pmatrix} A \\ C \end{pmatrix} = k,$$

which complete the proof. □

According to [6], corresponding to each column in (3.5) is ascribed a generalized singular pair (α_i, β_i). Following (3.6) we take for the first k of those as

$$\alpha_i = 1, \quad \beta_i = 0, \quad i = 1, \cdots, r, \tag{3.8}$$

$$\alpha_i, \beta_i \quad \text{as in } S_A \text{ and } S_B \quad i = r+1, \cdots, r+s, \tag{3.9}$$

$$\alpha_i = 0, \quad \beta_i = 1, \quad i = r+s+1, \cdots, k, \tag{3.10}$$

and call them the nontrivial generalized singular pairs of (A, C); α_i/β_i, $i = 1, \cdots, k$, are called the nontrivial generalized singular values of (A, C). The other $n-k$ pairs corresponding to the zero columns in (3.5) are called trivial generalized singular pairs of (A, C), and no particular numbers are assigned to them.

The following result gives a new characterization of the GSVs of a general matrix pencil and states that GSV's are a special case of RSVs.

THEOREM 3.3. *Using the notation of Definition* 2.1 *and Theorem* 3.4 *we have the following results:*

(1)

$$\sigma_i(A, I_m, C) = \frac{\alpha_i}{\beta_i}, \qquad i = 1, \cdots, k, \tag{3.11}$$

and

$$\sigma_i(A, I_m, C) = 0, \qquad i = k+1, \cdots, n. \tag{3.12}$$

(2) *Let* $l = \text{rank} \begin{pmatrix} A \\ C \end{pmatrix} - \text{rank}\,(C)$ *and* $u = \min\left(m, \text{rank} \begin{pmatrix} A \\ C \end{pmatrix}\right) rm$ *; then for all* $D \in \mathcal{C}^{m\times q}$

$$l \le \text{rank}\,(A + DC) \le u \tag{3.13}$$

and for all integers k *satisfying* $l \le k \le n$, *there exists matrix* $D_k \in \mathcal{C}^{m\times q}$ *such that*

$$\text{rank}\,(A + D_k C) = k.$$

Proof. (1) Let the GSVD of (A, C) be as in Theorem 3.2. For arbitrary $D \in \mathcal{C}^{m\times q}$, let $UDV^H = (D_{ij})_{i,j=1}^3$ be partitioned conformally with the partitionings of Σ_A and Σ_C; then

$$\begin{aligned}
&\text{rank}\,(A + DC) \\
&= \text{rank}\,(UAQ + UDV^H VCQ) \\
&= \text{rank}\left\{\begin{pmatrix} I_r & D_{12}S_C & D_{13} & O \\ O & S_A + D_{22}S_C & D_{23} & O \\ O & D_{32}S_C & D_{33} & O \end{pmatrix}\right\} \\
&= r + \text{rank}\left(\begin{pmatrix} S_A S_C^{-1} & O \\ O & O \end{pmatrix} + \begin{pmatrix} D_{22} & D_{23} \\ D_{32} & D_{33} \end{pmatrix}\right).
\end{aligned}$$

The result follows from Theorem 3.1.

(2) As is proved in Theorem 3.4, we have

$$k = \text{rank} \begin{pmatrix} A \\ C \end{pmatrix}, \quad r = \text{rank} \begin{pmatrix} A \\ C \end{pmatrix} - \text{rank}\,(C).$$

The proof of this part can be easily derived from the above expressions and Theorem 3.1. □

In the following we discuss the problem of uniqueness of GSVD. From the GSVD in Theorem 3.2, let

$$U_iAQ_i = (\Sigma_A, O), \qquad V_iCQ_i = (\Sigma_C, O), \quad (i = 1, 2)$$

be two GSVDs of A and C; then

$$(U_2U_1^H)(\Sigma_A, O) = (\Sigma_A, O)(Q_2^{-1}Q_1), \tag{3.14}$$

$$(V_2V_1^H)(\Sigma_C, O) = (\Sigma_C, O)(Q_2^{-1}Q_1). \tag{3.15}$$

Let

$$U_2U_1^H = (U_{ij})_{i,j=1}^3, \qquad V_2V_1^H(V_{ij})_{i,j=1}^3,$$

and

$$Q_2^{-1}Q_1 = (Q_{ij})_{i,j=1}^4$$

be block matrices partitioned conformally with the partitions of Σ_A and Σ_C. Equation (3.14) gives the following identity:

$$\begin{pmatrix} U_{11} & U_{12}S_A & O & O \\ U_{21} & U_{22}S_A & O & O \\ U_{31} & U_{32}S_A & O & O \end{pmatrix} = \begin{pmatrix} Q_{11} & Q_{12} & Q_{13} & Q_{14} \\ S_AQ_{21} & S_AQ_{22} & S_AQ_{23} & S_AQ_{24} \\ O & O & O & O \end{pmatrix},$$

which yields

$$\begin{gathered} U_{31} = O, \quad U_{32} = O, \quad Q_{13} = O, \quad Q_{14} = O, \quad Q_{23} = O, \quad Q_{24} = O, \\ U_{11} = Q_{11}, \quad U_{12}S_A = Q_{12}, \quad U_{21} = S_AQ_{21}, \quad U_{22}S_A = S_AQ_{22}. \end{gathered}$$

From the fact that $U_2U_1^H$ is unitary, it follows that $U_{13} = O$, $U_{23} = O$. Similarly, equation (3.15) results in the following identity:

$$\begin{pmatrix} O & V_{12}S_C & V_{13} & O \\ O & V_{22}S_C & V_{23} & O \\ O & V_{32}S_C & V_{33} & O \end{pmatrix} = \begin{pmatrix} O & O & O & O \\ S_CQ_{21} & S_CQ_{22} & S_CQ_{23} & S_CQ_{24} \\ Q_{31} & Q_{32} & Q_{33} & Q_{34} \end{pmatrix},$$

which yields

$$\begin{gathered} V_{12} = O, \quad V_{13} = O, \quad Q_{21} = O, \quad Q_{31} = O, \quad Q_{34} = O, \\ V_{22}S_C = S_CQ_{22}, \quad V_{23} = S_CQ_{23}, \quad V_{32}S_C = Q_{32}, \quad Q_{33} = V_{33}. \end{gathered}$$

From the fact that $V_2V_1^H$ is unitary, it follows that $V_{21} = O, V_{31} = O$. Furthermore, since $U_{21} = S_AQ_{21} = O$, hence $U_{12} = O$ and $Q_{12} = O$. Since $V_{32} = Q_{32}S_C^{-1} = O$, hence $V_{23} = O$ and $Q_{23} = O$. From $P_{22} = S_A^{-1}U_{22}S_A$ and $P_{22} = S_C^{-1}V_{22}S_C$, we obtain

$$(S_AS_C^{-1})V_{22} = U_{22}(S_AS_C^{-1}).$$

Let $\sigma_i = \alpha_{i+r}/\beta_{i+r}, i = 1, \cdots, s$ and $\Sigma := S_AS_C^{-1} = \text{ diag } (\sigma_{i_1}I_{s_1}, \cdots, \sigma_{i_l}I_{s_l})$, where $\sigma_{i_1} > \cdots > \sigma_{i_l}$ and $\sum_{t=1}^{l} s_t = s$. Since $\alpha_{i+r}^2 + \beta_{i+r}^2 = 1$, $i = 1, \cdots, s$, hence S_A and S_C have the same partitioning as that of Σ, i.e.,

$$S_A = \text{ diag } (\alpha_{i_1}I_{s_1}, \cdots, \alpha_{i_l}I_{s_l}), \qquad S_C = \text{ diag } (\beta_{i_1}I_{s_1}, \cdots, \beta_{i_l}I_{s_l}).$$

From $\Sigma V_{22} = U_{22}\Sigma$, we can verify

$$\Sigma^2 V_{22} = V_{22}\Sigma^2, \qquad \Sigma^2 U_{22} = U_{22}\Sigma^2,$$

therefore

$$U_{22} = V_{22} = \text{ diag } (\tilde{U}_1, \cdots, \tilde{U}_l),$$

where $\tilde{U}_i$, $(i = 1, \cdots, l)$ is unitary matrix of order s_i.

Summarizing the above, we obtain

$$Q_1 = Q_2 \left(\begin{array}{ccc|c} U_{11} & & & \\ & U_{22} & & O \\ & & V_{33} & \\ Q_{41} & Q_{42} & Q_{43} & Q_{44} \end{array} \right), \qquad U_1^H = U_2^H \text{ diag } (U_{11}, U_{22}, U_{33})$$

and

$$V_1^H = V_2^H \text{ diag } (V_{11}, U_{22}, V_{33}) \tag{3.16}$$

where $U_{11}, U_{22}, U_{33}, V_{11}, V_{33}$ are unitary; the matrix Q_{44} is nonsingular, and $U_{22} = \text{diag } (\tilde{U}_1, \cdots, \tilde{U}_l)$.

As pointed out in [6], the GSVs of (A, C) are just the SVs of AC^{-1}, if C is nonsingular. In the following we further discuss the case in which C is a general matrix.

COROLLARY 3.3. *We use the notation of Theorem* 3.2 *and let*

$$C_A^+ = Q \begin{pmatrix} O_C^H & & \\ & S_C^{-1} & \\ & & I \end{pmatrix} V .$$

If rank $(A^H, C^H)^H = n$, *then* C_A^+ *is uniquely defined and the* SV*s of* AC_A^+ *contain the finite* GSV*s of* (A, C).

Proof. Since rank $(A^H, C^H)^H = n$, any two sets of transformations in Theorem 3.2 satisfy the following relations:

$$Q_1 = Q_2 \text{ diag } (U_{11}, U_{22}, V_{33}), \quad U_1^H = U_2^H \text{ diag } (U_{11}, U_{22}, U_{33}),$$

and $V_1^H = V_2^H \text{ diag } (V_{11}, U_{22}, V_{33})$, hence

$$\begin{aligned} & Q_1 \begin{pmatrix} O_C^H & & \\ & S_C^{-1} & \\ & & I \end{pmatrix} V_1 \\ = \; & Q_2 \begin{pmatrix} U_{11} & & \\ & U_{22} & \\ & & V_{33} \end{pmatrix} \begin{pmatrix} O_C^H & & \\ & S_C^{-1} & \\ & & I \end{pmatrix} \begin{pmatrix} V_{11}^H & & \\ & U_{22}^H & \\ & & V_{33}^H \end{pmatrix} V_2 \\ = \; & Q_2 \begin{pmatrix} O_C^H & & \\ & S_C^{-1} & \\ & & I \end{pmatrix} V_1. \end{aligned}$$

Therefore we have proved that C_A^+ is well defined. Furthermore, observe that

$$UAC_A^+V^H = \text{ diag } (O, S_A S_C^{-1}, O)$$

and only the infinite GSVs of (A, C) are changed to zero SVs of AC_A^+; the other GSVs are preserved in AC_A^+. □

In the following we discuss some properties of C_A^+. It is easy to check that C_A^+ satisfies the following equations:

$$CC_A^+C = C, \tag{3.17}$$

$$C_A^+CC_A^+ = C_A^+, \tag{3.18}$$

$$(CC_A^+)^H = CC_A^+ . \tag{3.19}$$

Therefore in the notations of [1], C_A^+ is a $\{1, 2, 3\}$-inverse of C. It will be interesting to know how we can uniquely characterize C_A^+ in the class of a $\{1, 2, 3\}$-inverse of C. The following theorem answers this question under the assumption that rank $(A^H, C^H)^H = n$.

THEOREM 3.4. *If $(A^H, C^H)^H$ is of full column rank, then C_A^+ is the unique solution of the following constrained minimization problem:*

$$\min_{X \in C^{n \times q}} \|AX\|_F, \tag{3.20}$$

subject to

$$CXC = C, \tag{3.21}$$

$$XCX = X, \tag{3.22}$$

$$(CX)^H = CX. \tag{3.23}$$

The minimum value is $\sqrt{\sum_{i=r+1}^{r+s} (\alpha_i/\beta_i)^2}$.

Proof. Let C have the decomposition as in (3.5):

$$VCQ = \begin{array}{c} \begin{matrix} k & n-k \end{matrix} \\ \begin{pmatrix} \Sigma_C, & O \end{pmatrix} \end{array}.$$

Since rank $(A^H, C^H)^H = n$, so $k = n$ and $C = V^H \Sigma_C Q^{-1}$. Partition $Q^{-1}XV^H = (X_{ij})_{i,j=1}^3$ conformally with the partitionings of Σ_A and Σ_C. We can verify that X should be of the following form:

$$X = Q \begin{pmatrix} O & X_{12} & X_{13} \\ O & S_C^{-1} & O \\ O & O & I_{n-r-s} \end{pmatrix} V$$

in order to satisfy (3.21)–(3.23).

Since

$$\begin{aligned} & \|AX\|_F^2 \\ = \; & \|UAQ\, Q^{-1}XV^H\|_F^2 \\ = \; & \left\| \begin{pmatrix} I_r & & \\ & S_A & \\ & & O_A \end{pmatrix} \begin{pmatrix} O & X_{12} & X_{13} \\ O & S_C^{-1} & O \\ O & O & I_{n-r-s} \end{pmatrix} \right\|_F^2 \\ = \; & \|(X_{12}, X_{13})\|_F^2 + \|S_A S_C^{-1}\|_F^2 \\ \geq \; & \|S_A\, S_C^{-1}\|_F^2 \\ = \; & \sum_{i=r+1}^{r+s} (\frac{\alpha_i}{\beta_i})^2. \end{aligned}$$

The equality is satisfied if and only if $X_{12} = O$ and $X_{13} = O$, i.e., $X = C_A^+$. □

Remark 3.1. Along the lines of the proof of Theorem 3.4 we can also verify that C_A^+ is the unique solution of the following constrained minimization problem:

$$\min_{X \in \mathbb{C}^{q \times n}} \|AX\|_F,$$

subject to

$$\begin{array}{ll} (1) & CXC = C, \\ (2) & (CX)^H = CX. \end{array}$$

Remark 3.2. Exchanging the rolls of A and C in (3.20) and (3.14), we can also show that

$$A_C^- := Q \begin{pmatrix} I & & \\ & S_A^{-1} & \\ & & O \end{pmatrix} U$$

is the unique solution of the corresponding minimization problem. Another way of uniquely characterizing C_A^+ is to generalize the Moore-Penrose conditions.

THEOREM 3.5. *If $(A^H, C^H)^H$ has full column rank, then C_A^+ is the unique solution of the following four equations:*

$$CXC = C, \tag{3.24}$$

$$XCX = X, \tag{3.25}$$

$$(CX)^H = CX, \tag{3.26}$$

$$(A^H AXC)^H = A^H AXC. \tag{3.27}$$

Proof. As in the proof of Theorem 3.4, X should be of the following form:

$$X = Q \begin{pmatrix} O & X_{12} & X_{13} \\ O & S_C^{-1} & O \\ O & O & O \end{pmatrix} V$$

in order to satisfy (3.24)–(3.26). Since

$$\begin{aligned} A^H AXC &= Q^{-H} \begin{pmatrix} I & & \\ & S_A & \\ & & O \end{pmatrix} UU^H \begin{pmatrix} I & & \\ & S_A & \\ & & O \end{pmatrix} Q^{-1} Q \\ &\quad \cdot \begin{pmatrix} O & X_{12} & X_{13} \\ O & S_C^{-1} & O \\ O & O & I \end{pmatrix} VV^T \begin{pmatrix} O & & \\ & S_C^{-1} & \\ & & I \end{pmatrix} Q^{-1}, \\ &= Q^{-H} \begin{pmatrix} O & X_{12} & X_{13} \\ O & S_A^2 S_C^{-2} & O \\ O & O & O \end{pmatrix} Q^{-1}, \end{aligned}$$

therefore $(A^H AXC)^H = A^H AXC$ if and only if $X_{12} = O$ and $X_{13} = O$, i.e., $X = C_A^+$. □

4. The restricted singular value decomposition. In this section, B and C are assumed to be general matrices. The key observation is the following.

LEMMA 4.1. *Let $P \in \mathcal{C}^{m\times m}$ and $Q \in \mathcal{C}^{n\times n}$ be nonsingular matrices, and let $U \in \mathcal{C}^{p\times p}$ and $V \in \mathcal{C}^{q\times q}$ be unitary matrices; then*

$$\sigma_k(PAQ, PBU, VCQ) = \sigma_k(A, B, C), \qquad k = 1, \cdots, n. \tag{4.1}$$

This lemma specifies a class of transformations which preserves the RSVs of a matrix triplet.

THEOREM 4.1. *Let $A \in \mathcal{C}^{m\times n}$, $B \in \mathcal{C}^{m\times p}$, and $C \in \mathcal{C}^{q\times n}$; then there exist nonsingular matrices $P \in \mathcal{C}^{m\times m}$ and $Q \in \mathcal{C}^{n\times n}$, unitary matrices $U \in \mathcal{C}^{p\times p}$, and $V \in \mathcal{C}^{q\times q}$ such that*

$$PAQ = \begin{array}{c} \\ m-t_1 \\ t_2 \end{array} \begin{array}{c} \begin{array}{cc} n-t_1 & t_1 \end{array} \\ \begin{pmatrix} \Sigma_A & \\ & O_A^{(2)} \end{pmatrix} \end{array}, \tag{4.2}$$

$$PBU = \begin{array}{c} \\ t_2 \end{array} \begin{array}{c} \begin{array}{c} \end{array} \\ \begin{pmatrix} \Sigma_B \\ O_B^{(2)} \end{pmatrix} \end{array}, \tag{4.3}$$

$$VCQ = \begin{array}{c} \begin{array}{cc} n-t_1 & t_1 \end{array} \\ (\Sigma_C, \quad O_C^{(2)}) \end{array}, \tag{4.4}$$

$$\Sigma_A = \begin{array}{c} \begin{array}{ccccc} j & k & l & r & s_1 \end{array} \\ \begin{array}{c} j \\ k \\ l \\ r \\ s_2 \end{array} \begin{pmatrix} I_j & & & & \\ & I_k & & & \\ & & I_l & & \\ & & & S_A & \\ & & & & O_A^{(1)} \end{pmatrix} \end{array}, \tag{4.5}$$

$$\Sigma_B = \begin{array}{c} \begin{array}{ccccc} j & p-j-r-s_2 & r & s_2 \end{array} \\ \begin{array}{c} j \\ k+l \\ r \\ s_2 \end{array} \begin{pmatrix} I_j & & & \\ & O_B^{(1)} & & \\ & & S_B & \\ & & & I_{s_2} \end{pmatrix} \end{array}, \tag{4.6}$$

$$\Sigma_C = \begin{array}{c} \begin{array}{cccc} j+k & l & r & s_1 \end{array} \\ \begin{array}{c} q-l-r-s_1 \\ l \\ r \\ s_1 \end{array} \begin{pmatrix} O_C^{(1)} & & & \\ & I_l & & \\ & & S_C & \\ & & & I_{s_1} \end{pmatrix} \end{array}, \tag{4.7}$$

where $S_A = \operatorname{diag}(\alpha_i)$, $S_B = \operatorname{diag}(\beta_i)$, $S_C = \operatorname{diag}(\gamma_i)$, *and*

$$\alpha_i^2 + \beta_i^2 + \gamma_i^2 = 1, \qquad i = s+1, \cdots, s+r \tag{4.8}$$

where we denote $s = j + k + l$; *furthermore*

$$1 > \alpha_i \geq \alpha_{i+1} > 0\,, \quad 0 < \beta_i \leq \beta_{i+1} < 1\,, \quad 1 > \gamma_i \geq \gamma_{i+1} > 0 \tag{4.9}$$

and

$$\frac{\alpha_i}{\beta_i \gamma_i} \geq \frac{\alpha_{i+1}}{\beta_{i+1}\gamma_{i+1}}, \qquad i = s+1, \cdots, s+r-1. \tag{4.10}$$

Proof. The proof is constructive and consists of four steps. The transformations of each step, according to Lemma 4.1, are of the following form:

$$A^{(k+1)} = P^{(k)} A^{(k)} Q^{(k)}, \tag{4.11}$$

$$B^{(k+1)} = P^{(k)} B^{(k)} U^{(k)}, \tag{4.12}$$

$$C^{(k+1)} = V^{(k)} C^{(k)} Q^{(k)} \tag{4.13}$$

where $P^{(k)}$ and $Q^{(k)}$ are nonsingular matrices, $U^{(k)}$ and $V^{(k)}$ are unitary matrices. In each step we only specify the $P^{(k)}$, $Q^{(k)}$, $U^{(k)}$, and $V^{(k)}$ and the resulted $A^{(k+1)}$, $B^{(k+1)}$, and $C^{(k+1)}$. Set

$$A^{(1)} = A, \quad B^{(1)} = B, \quad C^{(1)} = C.$$

Step 1. Using Theorem 3.4, let the GSVD of $(A^{(1)}, C^{(1)})$ be

$$U_1 A^{(1)} Q_1 = \begin{array}{c} \\ j+k \\ l+r \\ m-j-k-l-r \end{array} \begin{array}{c} \begin{array}{cccc} j+k & l+r & s_1 & t_1 \end{array} \\ \begin{pmatrix} I_{j+k} & O & O & O \\ O & S_A^{(1)} & O & O \\ O & O & O & O \end{pmatrix} \end{array},$$

$$V_1 C^{(1)} Q_1 = \begin{array}{c} \\ j+k \\ l+r \\ s_1 \end{array} \begin{array}{c} \begin{array}{cccc} j+k & l+r & s_1 & t_1 \end{array} \\ \begin{pmatrix} O & O & O & O \\ O & S_C^{(1)} & O & O \\ O & O & I_{s_1} & O \end{pmatrix} \end{array}.$$

Set

$$\begin{aligned} P^{(1)} &= U_1, & Q^{(1)} &= Q_1 \operatorname{diag}(I, (S_C^{(1)})^{-1}, I), \\ U^{(1)} &= I, & V^{(1)} &= V_1, \end{aligned}$$

then

$$A^{(2)} = \begin{array}{c} \\ j+k \\ l+r \\ m-j-k-l-r \end{array} \begin{array}{c} \begin{array}{cccc} j+k & l+r & s_1 & t_1 \end{array} \\ \begin{pmatrix} I_{j+k} & O & O & O \\ O & S_A^{(1)}(S_C^{(1)})^{-1} & O & O \\ O & O & O & O \end{pmatrix} \end{array},$$

$$B^{(2)} = \begin{array}{c} \\ j+k \end{array} \begin{array}{c} \\ \begin{pmatrix} B_1^{(2)} \\ B_2^{(2)} \end{pmatrix} \end{array},$$

$$C^{(2)} = \begin{array}{c} \\ j+k \\ l+r \\ s_1 \end{array} \begin{array}{c} \begin{array}{cccc} j+k & l+r & s_1 & t_1 \end{array} \\ \begin{pmatrix} O & O & O & O \\ O & I_{l+r} & O & O \\ O & O & I_{s_1} & O \end{pmatrix} \end{array}.$$

Step 2. Using Theorem 3.4 let the GSVD of the matrix pair

$$\left(\begin{pmatrix} S_A^{(1)}(S_C^{(1)})^{-1} \\ O \end{pmatrix}, B_2^{(2)} \right)$$

be of the following form:

$$P_2 \begin{pmatrix} S_A^{(1)}(S_C^{(1)})^{-1} \\ O \end{pmatrix} V_2 = \begin{array}{c} \\ l \\ r \\ s_2+t_2 \end{array} \begin{array}{c} \begin{array}{cc} l & r \end{array} \\ \begin{pmatrix} I_l & O \\ O & S_4^{(2)} \\ O & O \end{pmatrix} \end{array},$$

$$P_2 B_2^{(2)} U_2 = \begin{array}{c} \\ l \\ r \\ s_2 \\ t_2 \end{array} \begin{array}{c} \begin{array}{ccc} p-r-s_2 & r & s_2 \end{array} \\ \begin{pmatrix} O & O & O \\ O & S_B^{(2)} & O \\ O & O & I_{s_2} \\ O & O & O \end{pmatrix} \end{array}$$

where $S_A^{(2)} = \operatorname{diag}(s_1, \cdots, s_r)$, $S_B^{(2)} = \operatorname{diag}(t_1, \cdots, t_r)$ and $s_i^2 + t_i^2 = 1$, $1 > s_1 \geq \cdots \geq s_r > 0$ and $0 < t_1 \leq \cdots \leq t_r < 1$. Set

$$P^{(2)} = \operatorname{diag}(I, P_2), \qquad Q^{(2)} = \operatorname{diag}(I, V_2, I),$$
$$U^{(2)} = U_2, \qquad V^{(2)} = \operatorname{diag}(I, V_2^H, I),$$

then

$$A^{(3)} = \begin{array}{c} \\ j+k \\ l \\ r \\ m-j-k-l-r \end{array} \begin{array}{c} \begin{array}{ccccc} j+k & l & r & s_1 & t_1 \end{array} \\ \begin{pmatrix} I_{j+k} & O & O & O & O \\ O & I_l & O & O & O \\ O & O & S_A^{(2)} & O & O \\ O & O & O & O & O \end{pmatrix} \end{array},$$

$$B^{(3)} = \begin{array}{c} \\ j+k \\ l \\ r \\ s_2 \\ t_2 \end{array} \begin{array}{c} \begin{array}{ccc} p-r-s_2 & r & s_2 \end{array} \\ \begin{pmatrix} B_1^{(3)} & B_2^{(3)} & B_3^{(3)} \\ O & O & O \\ O & S_B^{(2)} & O \\ O & O & I_{s_2} \\ O & O & O \end{pmatrix} \end{array},$$

and

$$C^{(3)} = C^{(2)}.$$

Step 3. Set

$$P^{(3)} = \begin{pmatrix} I & O & -B_2^{(3)}(S_B^{(2)})^{-1} & -B_3^{(3)} & O \\ O & I & O & O & O \\ O & O & I & O & O \\ O & O & O & I & O \\ O & O & O & O & I \end{pmatrix},$$

$$Q^{(3)} = \begin{pmatrix} I & O & -B_2^{(3)}(S_B^{(2)})^{-1}S_A^{(2)} & O \\ O & I & O & O \\ O & O & I & O \\ O & O & O & I \end{pmatrix},$$

and

$$U^{(3)} = I, \qquad V^{(3)} = I,$$

then

$$\begin{aligned} A^{(4)} &= A^{(3)}, \\ B^{(4)} &= \begin{pmatrix} B_1^{(3)} & O & O \\ O & O & O \\ O & S_B^{(2)} & O \\ O & O & I_{s_2} \\ O & O & O \end{pmatrix}, \\ C^{(4)} &= C^{(3)}. \end{aligned}$$

Step 4. Let the SVD of $B_1^{(3)}$ be

$$U_3 B_1^{(3)} V_3 = \begin{matrix} j \\ k \end{matrix} \begin{pmatrix} \Sigma_B^{(2)} & O \\ O & O \end{pmatrix}$$

where $\Sigma_B^{(2)}$ is nonsingular. Let $s = j + k + l$ and

$$\begin{aligned} \alpha_{s+i} &= \frac{s_i^2}{(1+s_i^2)^{1/2}}, \\ \beta_{s+i} &= t_i, \qquad\qquad (i = 1, \cdots, r). \\ \gamma_{s+i} &= \frac{s_i}{(1+s_i^2)^{1/2}} \end{aligned}$$

It is easy to verify that $\{\alpha_{s+i}\}$, $\{\beta_{s+i}\}$, and $\{\gamma_{s+i}\}$ satisfy (4.8)–(4.11). Let $S_C = \text{diag}\,(\gamma_{s+i})$, $S_A = S_A^{(2)} S_C$, and $S_B = S_B^{(2)}$, in addition, set

$$\begin{aligned} P^{(4)} &= \text{diag}\,((\Sigma_B^{(2)})^{-1}, I)\ \text{diag}\,(U_3, I), \\ Q^{(4)} &= \text{diag}\,(I, S_C, I)\ \text{diag}\,((\Sigma_B^{(2)})^{-1}, I)\ \text{diag}\,(U_3^H, I), \\ U^{(4)} &= \text{diag}\,(V_3, I), \\ V^{(4)} &= I. \end{aligned}$$

After some manipulation, we obtain the results as stated in (4.2)–(4.4). The proof is completed. □

Remark 4.1. We can also use D_1 and D_2 positive-definite diagonal matrices to scale (S_A, S_B, S_c) to $(D_1 S_A D_2, D_s S_B, S_C D_2)$. For example, we can choose D_1 and D_2 such that $D_1 S_B$ and $S_C D_2$ are identity matrices.

Similar to (3.8) we define

$$\begin{array}{llll} \alpha_i = 1, & \beta_i = 1, & \gamma_i = 0, & i = 1, \cdots, j, \\ \alpha_i = 1, & \beta_i = 0, & \gamma_i = 0, & i = j+1, \cdots, j+k, \\ \alpha_i = 1, & \beta_i = 0, & \gamma_i = 1, & i = j+k+1, \cdots, s, \\ \alpha_i, \beta_i, \gamma_i & \text{as in } S_A, & S_B, \text{ and} S_C, & i = s+1, \cdots, s+r \\ \alpha_i = 0, & \beta_i = 1, & \gamma_i = 1, & i = s+r+1, \cdots, s+r+\min(s_1, s_2) \end{array}$$

to be the nontrivial RSV triplets of (A, B, C).

The following theorem relates Theorem 4.1 with the concept of RSVs and justifies the above definition and calling Theorem 4.1 the RSVD theorem.

THEOREM 4.2. *With the notation as in Theorem* 4.1 *and the above definition the following statements are true*:

(1)

$$\sigma_i(A,B,C) = \frac{\alpha_i}{\beta_i \gamma_i}, \quad i = 1, \cdots, s + r + \min(s_1, s_2), \tag{4.14}$$

$$\sigma_i(A,B,C) = 0, \quad i = n - (s + r + \min(s_1, s_2)) + 1, \cdots, n. \tag{4.15}$$

(2) *Let*

$$l = \text{ rank } (A,B) + \text{ rank } \begin{pmatrix} A \\ B \end{pmatrix} - \text{ rank } \begin{pmatrix} A & B \\ C & O \end{pmatrix},$$

$$u = \min \left(\text{ rank } (A,B), \text{ rank } \begin{pmatrix} A \\ C \end{pmatrix} \right);$$

then for all $D \in \mathcal{C}^{p \times q}$

$$l \leq \text{ rank } (A + BDC) \leq u \tag{4.16}$$

and for all integers k *satisfying* $l \leq k \leq n$, *there exists matrix* $D_k \in \mathcal{C}^{p \times q}$ *such that*

$$\text{rank } (A + BD_kC) = k \, .$$

Proof. (1) Let $U^H D V^H = (D_{ij})_{i,j=1}^4$ be a block matrix partitioned conformally with the partitionings of Σ_B and Σ_C.

$$\begin{aligned}
&\text{rank } (A + BDC) \\
&= \text{ rank } (PAQ + PBUU^H DV^H VCQ) \\
&= \text{ rank } \begin{pmatrix} I_j & O & D_{12} & D_{13}S_C & D_{14} & O \\ O & I_k & O & O & O & O \\ O & O & I_l & O & O & O \\ O & O & S_B D_{32} & S_A + S_B D_{33} S_C & S_B D_{34} & O \\ O & O & D_{42} & D_{44} S_C & D_{44} & O \\ O & O & O & O & O & O \end{pmatrix} \\
&= j + k + l + \text{ rank } \left(\begin{pmatrix} S_B^{-1} S_A S_C^{-1} & O \\ O & O \end{pmatrix} + \begin{pmatrix} D_{33} & D_{34} \\ D_{43} & D_{44} \end{pmatrix} \right);
\end{aligned}$$

using Theorem 3.1, the proof of this part is completed.

(2) For the upper bound, note that

$$A + BDC = (A,B) \begin{pmatrix} I \\ DC \end{pmatrix} = (I, BD) \begin{pmatrix} A \\ C \end{pmatrix}$$

hence

$$\text{rank } (A + BDC) \leq \text{ rank } (A,B)$$

and

$$\text{rank}\,(A+BDC) \leq \text{rank}\begin{pmatrix} A \\ C \end{pmatrix}.$$

For the lower bound, we can verify that

$$\begin{aligned} \text{rank}\,(A,B) &= s+r+s_2, \\ \text{rank}\begin{pmatrix} A \\ C \end{pmatrix} &= s+r+s_1, \\ \text{rank}\begin{pmatrix} A & B \\ C & O \end{pmatrix} &= s+2r+s_1+s_2, \end{aligned}$$

hence

$$s = \text{rank}\,(A,B) + \text{rank}\begin{pmatrix} A \\ C \end{pmatrix} - \text{rank}\begin{pmatrix} A & B \\ C & O \end{pmatrix}.$$

The proof of the theorem is completed. □

Remark 4.2. From the following linear system

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 2 & 1 & 1 \end{pmatrix} \begin{pmatrix} j \\ k \\ l \\ r \\ s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} \text{rank}\,(A) \\ \text{rank}\,(B) \\ \text{rank}\,(C) \\ \text{rank}\,(A,B) \\ \text{rank}\begin{pmatrix} A \\ C \end{pmatrix} \\ \text{rank}\begin{pmatrix} A & B \\ C & O \end{pmatrix} \end{pmatrix}$$

we obtain the following expressions for the integer indices in Theorem 4.1:

$$\begin{aligned} j &= \text{rank}\begin{pmatrix} A \\ C \end{pmatrix} + \text{rank}\,(B) - \text{rank}\begin{pmatrix} A & B \\ C & O \end{pmatrix}, \\ k &= \text{rank}\begin{pmatrix} A & B \\ C & O \end{pmatrix} - \text{rank}\,(B) - \text{rank}\,(C), \\ l &= \text{rank}\,(A,B) + \text{rank}\,(C) - \text{rank}\begin{pmatrix} A & B \\ C & O \end{pmatrix}, \\ r &= \text{rank}\begin{pmatrix} A & B \\ C & O \end{pmatrix} + \text{rank}\,(A) - J\,\text{rank}\,(A,B) - \text{rank}\begin{pmatrix} A \\ C \end{pmatrix}, \\ s_1 &= \text{rank}\begin{pmatrix} A \\ C \end{pmatrix} - \text{rank}\,(A), \\ s_2 &= \text{rank}\,(A,B) - \text{rank}\,(A), \end{aligned}$$

in addition, it is easy to see that

$$\begin{aligned} t_1 &= n - \text{rank}\begin{pmatrix} A \\ C \end{pmatrix}, \\ t_2 &= m - \text{rank}\,(A,B). \end{aligned}$$

If we use $R_r(A)(R_C(A))$ and $N_r(A)(N_C(A))$ to denote the subspace spanned by the rows (or columns) of A and the row (column) null space of A, respectively, and

furthermore, $S\backslash T$ denotes the complement subspace of T in S, such that $S\backslash T \oplus T = S$ and $\dim(S)$ is the dimension of the subspace S, then we can express the above integer index using the following geometric terms:

$$\begin{aligned}
j &= \dim\left(R_C\begin{pmatrix}A\\C\end{pmatrix} \cap R_C\begin{pmatrix}B\\O\end{pmatrix}\right),\\
k &= \dim\left(N_C(C)\backslash N_C\begin{pmatrix}A\\C\end{pmatrix}\right) - \dim\left(R_C\begin{pmatrix}A\\C\end{pmatrix} \cap R_C\begin{pmatrix}B\\O\end{pmatrix}\right)\\
&= \dim(N_r(B)\backslash N_r(A,B)) - \dim(R_r(A,B) \cap R_r(C,O)),\\
l &= \dim(R_r(A,B) \cap R_r(C,O)),\\
r &= \dim(R_r(A) \cap R_r(C)) - \dim(R_r(A,B) \cap R_r(C,O))\\
&= \dim(R_C(A) \cap R_C(B)) - \dim\left(R_C\begin{pmatrix}A\\C\end{pmatrix} \cap R_C\begin{pmatrix}B\\O\end{pmatrix}\right),\\
s_1 &= \dim\left(N_C(A)\backslash N_C\begin{pmatrix}A\\C\end{pmatrix}\right),\\
s_2 &= \dim(N_r(A)\backslash N_r(A,B)),\\
t_1 &= \dim\left(N_C\begin{pmatrix}A\\C\end{pmatrix}\right),\\
t_2 &= \dim(N_r(A,B)).
\end{aligned}$$

The above expressions can serve as the basis of a geometrical derivation of RSVD.

Before we discuss another two special cases of RSVD, we consider the uniqueness problem of the RSVD in Theorem 4.1.

THEOREM 4.3. *Let the following:*

$$P_i A Q_i = \begin{array}{c} \\ m-t_1 \\ t_2 \end{array}\begin{array}{c} \begin{array}{cc} n-t_1 & t_1 \end{array} \\ \begin{pmatrix} \Sigma_A & \\ & O_A^{(2)} \end{pmatrix} \end{array},$$

$$P_i B U_i = \begin{array}{c} \\ t_2 \end{array}\begin{array}{c} \\ \begin{pmatrix} \Sigma_B \\ O_B^{(2)} \end{pmatrix} \end{array},$$

$$V C Q = \begin{array}{c} \begin{array}{cc} n-t_1 & t_1 \end{array} \\ \begin{pmatrix} \Sigma_C, & O_C^{(2)} \end{pmatrix} \end{array}$$

be two RSVD*s of* (A,B,C) *in the form of Theorem* 4.1. *Furthermore, let*

$$S_B^{-1} S_A S_C^{-1} = \operatorname{diag}(\sigma_{i_1} I_{r_1}, \cdots, \sigma_{i_w} I_{r_w}),$$
$$\sigma_{i_1} >, \cdots, > \sigma_{i_w}, \quad \sum_{j=1}^{w} r_j = r;$$

then

$$P_1^{-1} P_2 = \begin{array}{c} \\ j \\ k \\ l \\ r \\ s_2 \\ t_2 \end{array}\begin{array}{c} \begin{array}{cccccc} j & k & l & r & s_2 & t_2 \end{array} \\ \begin{pmatrix} U_{11} & P_{12} & P_{13} & O & O & P_{16} \\ O & P_{22} & P_{23} & O & O & P_{26} \\ O & O & V_{22} & O & O & P_{36} \\ O & O & O & U_{33} & O & P_{46} \\ O & O & O & O & U_{44} & P_{56} \\ O & O & O & O & O & P_{66} \end{pmatrix} \end{array}, \tag{4.17}$$

$$Q_2^{-1}Q_1 = \begin{array}{c} \\ j \\ k \\ l \\ r \\ s_1 \\ t_1 \end{array} \begin{array}{c} \begin{array}{cccccc} j & k & l & r & s_1 & t_1 \end{array} \\ \begin{pmatrix} U_{11} & P_{12} & P_{13} & O & O & O \\ O & P_{22} & P_{23} & O & O & O \\ O & O & V_{22} & O & O & O \\ O & O & O & U_{33} & O & O \\ O & O & O & O & V_{44} & O \\ Q_{61} & Q_{62} & Q_{63} & Q_{64} & Q_{65} & Q_{66} \end{pmatrix} \end{array}, \tag{4.18}$$

$$U_2^H U_1 = \operatorname{diag}(U_{11}, U_{22}, U_{33}, U_{44}), \tag{4.19}$$

$$V_2 V_1^T = \operatorname{diag}(V_{11}, V_{22}, U_{33}, V_{44}) \tag{4.20}$$

where U_{ii} $(i = 1,2,3,4)$ *and* V_{ii} $(i = 1,2,4)$ *are unitary;* $U_{33} = \operatorname{diag}(\tilde{U}_1, \cdots, \tilde{U}_w)$, *and* $\tilde{U}_i \in \mathcal{C}^{r_i \times r_i}$, $(i = 1, \cdots, w)$; P_{22}, P_{66}, *and* Q_{66} *are nonsingular.*

Proof. We have

$$\begin{aligned} (P_2 P_1^{-1}) \begin{pmatrix} \Sigma_A & O \\ O & O \end{pmatrix} &= \begin{pmatrix} \Sigma_A & O \\ O & O \end{pmatrix} (Q_2^{-1}Q_1), \\ (P_2 P_1^{-1}) \begin{pmatrix} \Sigma_B \\ O \end{pmatrix} &= \begin{pmatrix} \Sigma_B \\ O \end{pmatrix} (U_2^H U_1), \\ (V_2 V_1^H)(\Sigma_C, O) &= (\Sigma_C, O)(Q_2^{-1}Q_1). \end{aligned}$$

Let

$$\begin{aligned} P &:= P_2 P_1^{-1} = (P_{ij})_{1,j=1}^6, \\ Q &:= Q_2^{-1}Q_1 = (Q_{ij})_{1,j=1}^6, \\ U &:= U_2^H P_1 = (U_{ij})_{1,j=1}^4, \\ V &:= V_2 V_1^H = (V_{ij})_{1,j=1}^4 \end{aligned}$$

be block matrices partitioned as in (4.17)–(4.20). The equation

$$\begin{pmatrix} P_{11} & P_{12} & P_{13} & P_{14}S_A & O & O \\ P_{21} & P_{22} & P_{23} & P_{24}S_A & O & O \\ P_{31} & P_{32} & P_{33} & P_{34}S_A & O & O \\ P_{41} & P_{42} & P_{43} & P_{44}S_A & O & O \\ P_{51} & P_{52} & P_{53} & P_{54}S_A & O & O \\ P_{11} & P_{12} & P_{13} & P_{14}S_A & O & O \end{pmatrix}$$

$$= \begin{pmatrix} Q_{11} & Q_{12} & Q_{13} & Q_{14} & Q_{15} & Q_{16} \\ Q_{21} & Q_{22} & Q_{23} & Q_{24} & Q_{25} & Q_{26} \\ Q_{31} & Q_{32} & Q_{33} & Q_{34} & Q_{35} & Q_{36} \\ S_A Q_{41} & S_A Q_{42} & S_A Q_{43} & S_A Q_{44} & S_A Q_{45} & S_A Q_{46} \\ O & O & O & O & O & O \\ O & O & O & O & O & O \end{pmatrix}$$

yields

$$\begin{aligned} P_{ij} &= Q_{ij}, && (i = 1,2,3), && (j = 1,2,3), \\ P_{ij} &= O, && (i = 5,6), && (j = 1,2,3,4), \\ Q_{ij} &= O, && (i = 1,2,3,4), && (j = 5,6), \\ P_{i4}S_A &= Q_{i4}, && (i = 1,2,3), \\ P_{4j} &= S_A Q_{4j}, && (j = 1,2,3), \\ P_{44}S_A &= S_A Q_{44}. \end{aligned}$$

Similarly, the equation

$$\begin{pmatrix} P_{11} & O & P_{14}S_B & P_{15} \\ P_{21} & O & P_{24}S_B & P_{25} \\ P_{31} & O & P_{34}S_B & P_{35} \\ P_{41} & O & P_{44}S_B & P_{45} \\ P_{51} & O & P_{54}S_B & P_{55} \\ P_{61} & O & P_{64}S_B & P_{65} \end{pmatrix} = \begin{pmatrix} U_{11} & U_{12} & U_{13} & U_{14} \\ O & O & O & O \\ O & O & O & O \\ S_BU_{31} & S_BU_{32} & S_BU_{33} & S_BU_{34} \\ U_{41} & U_{42} & U_{43} & U_{44} \\ O & O & O & O \end{pmatrix}$$

yields

$$\begin{array}{lll} P_{ij} = O, & (i = 2, 3, 6), & (j = 1, 4, 5), \\ P_{11} = U_{11} & P_{14}S_B = U_{13}, & P_{15} = U_{14}, \\ P_{41} = S_BU_{31}, & P_{44}S_B = S_BU_{33}, & P_{45} = S_BU_{34}, \\ P_{51} = U_{41}, & P_{54}S_B = U_{43}, & P_{55} = U_{44}, \end{array}$$

and $U_{12} = O, U_{32} = O, U_{42} = O$. Because U is unitary, we must have $U_{21} = O, U_{23} = O$, and $U_{24} = O$. Likewise the equation

$$\begin{pmatrix} O & O & V_{12} & V_{13}S_C & V_{14} & O \\ O & O & V_{22} & V_{23}S_C & V_{24} & O \\ O & O & V_{32} & V_{33}S_C & V_{34} & O \\ O & O & V_{42} & V_{43}S_C & V_{44} & O \end{pmatrix}$$

$$= \begin{pmatrix} O & O & O & O & O & O \\ Q_{31} & Q_{32} & Q_{33} & Q_{34} & Q_{35} & Q_{36} \\ S_CQ_{41} & S_CQ_{42} & S_CQ_{43} & S_CQ_{44} & S_CQ_{45} & S_CQ_{46} \\ Q_{51} & Q_{52} & Q_{53} & Q_{54} & Q_{55} & Q_{56} \end{pmatrix}$$

yields

$$\begin{array}{lll} Q_{ij} = O, & (i = 3, 4, 5), & (j = 1, 2, 6), \\ V_{22} = Q_{33}, & V_{23}S_C = Q_{34}, & V_{24} = Q_{35}, \\ V_{42} = Q_{53}, & V_{43}S_C = Q_{54}, & V_{44} = Q_{55}, \\ V_{32} = S_CQ_{43}, & V_{33}S_C = S_CQ_{44}, & V_{34} = S_CQ_{45}, \end{array}$$

and $V_{12} = O, V_{13} = O, V_{14} = O$. Because V is unitary, it follows that $V_{21} = O, V_{31} = O$, and $V_{41} = O$.

Furthermore, because $U_{41} = P_{51} = O, U_{42} = O$, and $U_{43} = P_{54}S_B = O$, we conclude $U_{14} = O$ and $U_{34} = O$; hence $P_{15} = O$ and $P_{45} = O$. Similarly, because $V_{14} = O, V_{24} = Q_{35} = O$, and $V_{34} = S_CQ_{45} = O$, we obtain $V_{42} = O$ and $V_{43} = O$; hence $Q_{53} = O$ and $Q_{54} = O$.

Moreover, $P_{41} = S_AQ_{41} = O, P_{42} = S_AQ_{42} = O$, hence $U_{31} = P_{41}S_B^{-1} = O$; and $U_{13} = O$. Therefore $P_{14} = S_B^{-1}U_{13} = O$. Similarly, $Q_{14} = P_{14}S_A = O, Q_{24} = P_{24}S_A = O$ and $Q_{34} = P_{34}S_A = O$. Because $V_{23} = Q_{34}S_C^{-1} = O$, we must have $V_{32} = O$; hence $Q_{43} = S_C^{-1}V_{32} = O$, and $P_{43} = S_AQ_{43} = O$. Additionally, $P_{32} = Q_{32} = O$.

Finally, we have $P_{44} = S_BU_{33}S_B^{-1}, Q_{44} = S_C^{-1}V_{33}S_C$ and $P_{44} = S_AQ_{44}S_A^{-1}$; hence

$$U_{33}(S_B^{-1}S_AS_C^{-1}) = (S_B^{-1}S_AS_C^{-1})V_{33},$$

which implies

$$U_{33} = V_{33} = \text{ diag } (\tilde{U}_1, \cdots, \tilde{U}_w).$$

As in the proof of Theorem 4.2, we can choose $\{s_i\}$ and $\{t_i\}$ such that

$$\begin{array}{c} 1 > s_1 \geq, \cdots, \geq s_r > 0, \\ 0 < s_1 \leq, \cdots, \leq t_r < 1, \\ s_i^2 + t_i^2 = 1, \qquad (i = 1, \cdots, r), \end{array}$$

and

$$\begin{aligned} \alpha_{s+i} &= s_i^2/(1+s_i^2)^{1/2} \\ \beta_{s+i} &= t_i \qquad\qquad (i=1,\cdots,r) \\ \gamma_{s+i} &= s_i/(1+s_i^2)^{1/2}. \end{aligned}$$

It is then easy to check that

$$\begin{aligned} S_A &= \operatorname{diag}(\alpha_{s+i_1} I_{r_1},\cdots,\alpha_{s+i_w} I_{r_w}), \\ S_B &= \operatorname{diag}(\beta_{s+i_1} I_{r_1},\cdots,\beta_{s+i_w} I_{r_w}), \\ S_C &= \operatorname{diag}(\gamma_{s+i_1} I_{r_1},\cdots,\gamma_{s+i_w} I_{r_w}), \end{aligned}$$

hence $P_{44} = Q_{44} = U_{33}$. The whole proof is completed. □

COROLLARY 4.1. *Let A be nonsingular and the nonzero* SV*s of $CA^{-1}B$ be*

$$\sigma_1 \geq \cdots \geq \sigma_r > 0,$$

then (A,B,C) has $(n-r)$ infinite RSV*s and the r finite* RSV*s are*

$$\frac{1}{\sigma_r} \geq \cdots \geq \frac{1}{\sigma_1} > 0\,.$$

Proof. Using the decomposition of Theorem 4.2, we can show that

$$V(CA^{-1}B)U = \begin{pmatrix} O & & & \\ & O & & \\ & & O & \\ & & & S_C S_A^{-1} S_B \end{pmatrix}.$$

The diagonal elements of $S_C S_A^{-1} S_B$ are nonzero SVs of $CA^{-1}B$. □

COROLLARY 4.2 (PSVD [3]). *Let $B \in \mathcal{C}^{m\times p}$ and $C \in \mathcal{C}^{p\times n}$; then there exist unitary matrices U and V and nonsingular matrix T such that*

$$UBT = \begin{pmatrix} I_j & & \\ & O_B & \\ & & \Sigma_B \end{pmatrix}, \tag{4.21}$$

$$T^{-1}CV = \begin{pmatrix} O_C & & \\ & I_l & \\ & & \Sigma_C \end{pmatrix} \tag{4.22}$$

where

$$\begin{aligned} \Sigma_B &= \operatorname{diag}(s_1,\cdots,s_r)\,, & \Sigma_C &= \operatorname{diag}(t_1,\cdots,t_r), \\ 1 > s_1 \geq \cdots \geq s_r &> O\,, & 1 > t_i \geq \cdots \geq t_r &> O, \\ s_i^2 + t_i^{-2} &= 1\,, & i &= 1,\cdots,r\,. \end{aligned}$$

Proof. Using Theorem 4.1, let the RSVD of (I_p, B^H, C^H) be

$$\begin{aligned} PI_pQ &= \operatorname{diag}(I_j, I_k, I_l, S_A), \\ PB^H\tilde{U} &= \begin{pmatrix} I_j & & \\ & O_B^{(1)} & \\ & & S_B \end{pmatrix}, \\ \tilde{V}C^HQ &= \begin{pmatrix} O_B^{(1)} & & \\ & I_l & \\ & & S_C \end{pmatrix}. \end{aligned}$$

Set $\tilde{Q} = Q \operatorname{diag}(I, S_A^{-1})$; then

$$\begin{aligned} P\tilde{Q} &= I_p, \\ PB^H\tilde{U} &= \begin{pmatrix} I_j & & \\ & O_B^{(1)} & \\ & & S_B \end{pmatrix}, \\ \tilde{V}C^H\tilde{Q} &= \begin{pmatrix} O_C^{(1)} & & \\ & I_l & \\ & & S_C S_A^{-1} \end{pmatrix}. \end{aligned}$$

The proof is finished if we set $U = \tilde{U}^H$, $V = \tilde{V}^H$, $T = P^H$, $O_B = (O_B^{(1)})^H$, $O_C = (O_C^{(1)})^H$, $\Sigma_B = S_B$, and $\Sigma_C = S_C S_A^{-1}$. □

Remark 4.3. Corollary 4.2 is a simplified version of the product-induced SVD (PSVD) in [3]. We can also use the techniques established in proving Theorem 3.2 and Theorem 4.1 to give a direct proof of it.

In the following we give the relation between the RSVD of (A,B,C) and the eigenstructure problem of

$$\left(\begin{pmatrix} O & A \\ A^H & O \end{pmatrix}, \begin{pmatrix} BB^H & O \\ O & C^HC \end{pmatrix} \right).$$

From Theorem 4.1, after suitable permutation Π we obtain

$$\begin{aligned} &\Pi \begin{pmatrix} P & O \\ O & Q^H \end{pmatrix} \left(\begin{pmatrix} O & A \\ A^H & O \end{pmatrix} - \lambda \begin{pmatrix} BB^H & O \\ O & C^HC \end{pmatrix} \right) \begin{pmatrix} P^H & O \\ O & Q \end{pmatrix} \Pi^T \\ &\quad = \operatorname{diag} \left\{ \begin{pmatrix} -\lambda I_j & I_j \\ I_j & O \end{pmatrix}, \begin{pmatrix} O & I_k \\ I_k & O \end{pmatrix}, \begin{pmatrix} O & I_l \\ I_l & -\lambda I_l \end{pmatrix}, \right. \\ &\qquad \left. \begin{pmatrix} -\lambda S_B^2 & S_A \\ S_A & -\lambda S_C^2 \end{pmatrix}, \begin{pmatrix} -\lambda I_{s_1} & O \\ O & -\lambda I_{s_2} \end{pmatrix}, O \right\}, \end{aligned}$$

therefore the eigenstructure of the symmetric matrix pencil is the following:

(i) $2(j + l)$ infinite eigenvalues corresponding to Jordan block of order 2 ($(j + l)$ 2×2 Jordan blocks).
(ii) $2k$ infinite eigenvalues corresponding to Jordan block of order 1.
(iii) $2r$ nonzero finite eigenvalues $\pm\alpha_i/\beta_i\gamma_i, i = s + 1, \cdots, s + r$.
(iv) $s_1 + s_2$ zero eigenvalues.
(v) $(m + n) - 2(j + l + k + s) - s_1 - s_2$ Kronecker blocks of order zero.

5. Concluding remarks. In this paper we introduce the concept of restricted singular values of matrix triplets. A main theorem called restricted singular value decomposition (RSVD) is proved for general matrix triplets. Three special cases of restricted singular values, i.e., the well-known singular values, the generalized singular values and the recently proposed product induced singular values are also discussed. Numerical algorithms for computing the RSVD of a general matrix triplet and applications of RSVD to the total least squares problem and the regularization problem of general Gauss–Markov linear model will appear in separate papers. Perturbation analysis and further applications of RSVD will be the topics of future research. We hope that RSVD will be important not only as a useful theoretical tool for analysing problems in numerical linear algebra, statistics, and control and system theory, but that its algorithmic aspects will also find applications in computer-based methods to solve realworld problems.

Acknowledgments. The author wishes to thank Professor Dr. P. Deuflhard and Professor J. Vandewalle for their encouragement and support, Professor G. Golub for his helpful advice, Professor J. Demmel for pointing out [2] and an alternative approach for solving the rank minimization problem, and Mrs. S. Wacker for her careful and excellent typing of the first version of the manuscript.

REFERENCES

[1] H. BEN-ISRAEL AND T. N. E. GREVILLE, *Generalized Inverses: Theory and Applications,* Wiley-Interscience, New York, 1974.

[2] J. DEMMEL, *The smallest perturbation of a submatrix which lowers the rank and constrained total least squares problems,* SIAM J. Numer. Anal., 24 (1987), pp. 199–206.

[3] K. V. FERNANDO AND S. HAMMARLING, *A product induced singular value decomposition* (PSVD) *for two matrices and balanced realization,* in Proc. Conference on Linear Algebra in Signals, Systems and Control, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1989, pp. 128–140.

[4] G. GOLUB AND C. VAN LOAN, *Matrix Computations,* The Johns Hopkins University Press, Baltimore, MD, 1983.

[5] L. MIRSKY, *Symmetric gauge functions and unitary invariant norms,* Quart. J. Math., 11 (1960), pp. 50–59.

[6] C. C. PAIGE AND M. A. SAUNDERS, *Towards a generalized singular value decomposition,* SIAM J. Numer. Anal., 18 (1981), pp. 398–405.

[7] G. W. STEWART *Rank degeneracy,* SIAM J. Sci. Statist. Comput., 5 (1984), pp. 403–413.

[8] C. VAN LOAN, *Generalizing the singular value decomposition,* SIAM J. Numer. Anal., 13 (1976), pp. 76–83.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 2, pp. 195–204, April 1991

LINEAR OPERATORS PRESERVING CERTAIN EQUIVALENCE RELATIONS ON MATRICES *

ROGER A. HORN †, CHI-KWONG LI ‡, AND NAM-KIU TSING§

Abstract. Using a uniform approach, characterizations are obtained of linear operators on matrix spaces that preserve certain equivalence relations such as consimilarity, $*$-congruence, nonsingular equivalence, and unitary equivalence.

Key words. linear operator, equivalence relation, consimilarity, congruence

AMS(MOS) subject classification. 15A04

1. Introduction.

Let $\sim$ be an equivalence relation on a matrix space $\mathcal{M}$. We are interested in studying the structure of a linear operator $T : \mathcal{M} \to \mathcal{M}$ that preserves $\sim$, that is,

$$T(A) \sim T(B) \text{ whenever } A \sim B.$$

Such an operator T is often called a *linear preserver.* Hiai in [9] studied this problem and obtained complete characterizations of T in two important cases:

(i) $\mathcal{M}$ is the set of all $n \times n$ complex matrices and $\sim$ is similarity;

(ii) $\mathcal{M}$ is the set of all $n \times n$ Hermitian matrices and $\sim$ is unitary similarity.

In this paper we extend Hiai's techniques to treat three additional cases:

(iii) $\mathcal{M}$ is the set of all $n \times n$ complex matrices and $\sim$ is consimilarity;

(iv) $\mathcal{M}$ is either the set of all $n \times n$ complex matrices or $n \times n$ Hermitian matrices and $\sim$ is $*$-congruence; and

(v) $\mathcal{M}$ is the set of all $m \times n$ complex or real matrices and $\sim$ is equivalence or unitary equivalence.

For each case, our general strategy is:

(a) Characterize the kernel of T;

(b) Modify T to obtain a new operator T' that is nonsingular and preserves a certain subset $\mathcal{S}$ of $\mathcal{M}$;

(c) Characterize the linear operators T' on $\mathcal{M}$ that satisfy $T'(\mathcal{S}) = \mathcal{S}$;

(d) Use (b) and (c) to characterize T.

Our approach to (a) and (b) is to analyse the *orbits* under the equivalence relation $\sim$,

$$\mathcal{O}(A; \sim) = \{X \in \mathcal{M} : X \sim A\},$$

and the corresponding tangent space $\mathcal{T}_A$ at A. When there is no ambiguity about $\sim$, we shall write $\mathcal{O}(A)$ instead of $\mathcal{O}(A; \sim)$. It is known (e.g., see [2]) that $\mathcal{O}(A)$ is a

*Received by the editors July 9, 1989; accepted for publication (in revised form) March 12, 1990.

†Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, Maryland 21218.

‡Department of Mathematics, The College of William and Mary, Williamsburg, Virginia 23185. This research was partially supported by National Science Foundation grant DMS 89-00922.

§Systems Research Center and Electrical Engineering Department, University of Maryland, College Park, Maryland 20742. This research was supported in part by National Science Foundation's Engineering Research Centers Program grant NSF CDR 88-03012 and by National Science Foundation grant DMC 84-51515.

homogeneous differentiable manifold if $\sim$ is any of the equivalence relations described in the preceding cases (i)–(v).

In §§2, 3, and 4 we discuss the cases in which the equivalence relation $\sim$ is, respectively, consimilarity, $*$-congruence, and equivalence. In the final section we describe some related results and problems.

We shall use the following notation throughout the paper:

$\mathbb{F}_{m\times n}$: the linear space of all $m \times n$ matrices over $\mathbb{F}$, where $\mathbb{F}$ is the complex field $\mathbb{C}$ or the real field $\mathbb{R}$.

$\mathcal{H}_n$: the real linear space of all $n \times n$ Hermitian matrices.

$U_n(\mathbb{F})$: the group of all $n \times n$ unitary or real orthogonal matrices according as $\mathbb{F} = \mathbb{C}$ or $\mathbb{F} = \mathbb{R}$.

$\{E_{11}, \cdots, E_{mn}\}$: the standard basis of $\mathbb{F}_{m\times n}$, i.e., E_{ij} has a one in the (i,j) position and zeros elsewhere.

X^t : the transpose of $X \in \mathbb{F}_{m\times n}$.

$\overline{X}$: the complex conjugate of $X \in \mathbb{C}_{m\times n}$.

tr X : the trace of $X \in \mathbb{F}_{n\times n}$.

Im(L) : the range (image) of the linear transformation L .

The following simple principle will be used to show that most of the linear preservers we study are nonsingular.

LEMMA 1.1. *Suppose the equivalence relation $\sim$ on the matrix space $\mathcal{M}$ has the property that $A \sim 0$ if and only if $A = 0$, and suppose a given linear operator $T : \mathcal{M} \to \mathcal{M}$ preserves $\sim$. Then*

(a) span $\mathcal{O}(A) \subset \ker(T)$ *for every* $A \in \ker(T)$;

(b) *if there is some* $A \in \ker(T)$ *such that* span $\mathcal{O}(A) = \mathcal{M}$, *then* $T = 0$;

(c) *if $\sim$ is such that* span $\mathcal{O}(A) = \mathcal{M}$ *for every nonzero* $A \in \mathcal{M}$ *and if T is nonzero, then T is nonsingular.*

The hypothesis that $A \sim 0$ if and only if $A = 0$ is clearly met for similarity, consimilarity, $*$-congruence, and equivalence, which are the equivalence relations we consider in this paper.

2. Consimilarity. In this section the matrix space $\mathcal{M}$ is $\mathbb{C}_{n\times n}$ and $\sim$ is *consimilarity*, i.e., $A \sim B$ if there exists a nonsingular $S \in \mathbb{C}_{n\times n}$ such that $A = \overline{S}BS^{-1}$. The main result is the following theorem.

THEOREM 2.1. *A linear operator $T : \mathbb{C}_{n\times n} \to \mathbb{C}_{n\times n}$ satisfies*

$$T(A) \text{ is consimilar to } T(B) \quad \text{whenever} \quad A \text{ is consimilar to } B$$

if and only if there exist a nonsingular $S \in \mathbb{C}_{n\times n}$ and a real number $\alpha \geq 0$ such that either

$$T(X) = \alpha \overline{S} X S^{-1} \quad \text{for all } X \in \mathbb{C}_{n\times n},$$

or

$$T(X) = \alpha \overline{S} X^t S^{-1} \quad \text{for all } X \in \mathbb{C}_{n\times n}.$$

We divide the proof of Theorem 2.1 into several lemmata.

LEMMA 2.2. (a) $\mathcal{O}(E_{11}) = \{xy^t : x, y \in \mathbb{C}^n,\ y^*x = 1\}$.

(b) $\mathcal{O}(E_{12}) = \{xy^t : x, y \in \mathbb{C}^n,\ y^*x = 0\}$.

Proof. (a) Observe that if e_i is the ith column of the $n \times n$ identity matrix, then

$$\begin{aligned}\mathcal{O}(E_{11}) &= \{\overline{S}E_{11}S^{-1} : S \text{ is invertible}\} \\ &= \{xy^t : S \text{ is invertible},\ x = \overline{S}e_1, \text{ and } y^t = e_1^t S^{-1}\} \\ &= \{xy^t : x, y \in \mathbb{C}^n,\ y^*x = 1\}.\end{aligned}$$

(b) The proof of (b) is similar to that of (a). □

LEMMA 2.3. Span $\mathcal{O}(A) = \mathbb{C}_{n\times n}$ *for every nonzero* $A \in \mathbb{C}_{n\times n}$.

Proof. Let $A \in \mathbb{C}_{n\times n}$ be a given nonzero matrix. Since A is consimilar to a real matrix [10, Thm. 4.9], to prove the lemma, we may assume that A is real. As A is nonzero, there exists a nonsingular $R \in \mathbb{R}_{n\times n}$ such that the (1, 1) entry of $RAR^{-1} \in \mathcal{O}(A)$ is nonzero. Let $P_1, \cdots, P_k$ with $k = 2^n$ be all the distinct diagonal matrices with diagonal entries equal to 1 or -1. Then $B = \sum_{i=1}^{k} P_i A P_i^{-1} \in$ span $\mathcal{O}(A)$ is a diagonal matrix with nonzero (1, 1) entry. It follows that for $S =$ diag$(1, i, \cdots, i)$, $B + \overline{S}BS^{-1} = \lambda E_{11} \in$ span $\mathcal{O}(A)$ for some nonzero λ. Thus, $E_{11} \in$ span $\mathcal{O}(A)$ and hence $\mathcal{O}(E_{11}) \subset$ span $\mathcal{O}(A)$. By Lemma 2.2(a), $E_{ii}, E_{ii} + E_{ij} \in$ span $\mathcal{O}(A)$ for all i, j. As a result, span $\mathcal{O}(A) = \mathbb{C}_{n\times n}$. □

The following result now follows from Lemma 1.1(c).

LEMMA 2.4. *Let* $T : \mathbb{C}_{n\times n} \to \mathbb{C}_{n\times n}$ *be a nonzero linear operator that preserves consimilarity. Then* T *is nonsingular.*

LEMMA 2.5. *Let* $A \in \mathbb{R}_{n\times n}$ *be a nonzero matrix. Then the tangent space to* $\mathcal{O}(A)$ *at* A *is*

$$\mathcal{T}_A = \{\overline{X}A - AX : X \in \mathbb{C}_{n\times n}\}.$$

Moreover, the real dimension of $\mathcal{T}_A$ *equals:*

(a) n^2 *if* $A = \lambda I$,

(b) $4n - 3$ *if* $A = \lambda E_{11}$,

(c) $4n - 4$ *if* $A = \lambda E_{12}$,

(d) $p \geq n^2$ *if* rank$(A - \alpha I) = 1$ *for some nonzero* $\alpha \in \mathbb{C}$, *and equality holds if and only if* A *is consimilar to* αI *or* A *is a* 2×2 *matrix that is consimilar to* $|\alpha|(E_{12} - E_{21})$,

(e) $q \geq 4n$ *if* rank$(A - \alpha I) > 1$ *for all* $\alpha \in \mathbb{C}$.

Proof. The asserted form of $\mathcal{T}_A$ follows immediately from the power series expansion of $\overline{S}AS^{-1}$ with $S = e^{\varepsilon X} = \sum_{p=0}^{\infty} (\varepsilon X)^p / p!$, $X \in \mathbb{C}_{n\times n}$.

Note that every rank one matrix is consimilar either to λE_{11} or to λE_{12} by Lemma 2.2, so the five cases listed in the lemma are exhaustive. Let $X = X_1 + iX_2$ with $X_1, X_2 \in \mathbb{R}_{n\times n}$. Then

$$\overline{X}A - AX = (X_1 A - AX_1) - i(AX_2 + X_2 A)$$

and hence the real dimension of $\mathcal{T}_A$ equals rank(L_1)+ rank(L_2), where L_1 and L_2 are the linear operators on $\mathbb{R}_{n\times n}$ defined by

$$L_1(Y) = YA - AY \quad \text{and} \quad L_2(Y) = AY + YA.$$

Since A is real, the rank of L_1 over $\mathbb{R}$ is the same as its rank over $\mathbb{C}$.

If $A = \lambda I$, then it is clear that rank$(L_1) = 0$ and rank$(L_2) = n^2$, so rank(L_1) + rank$(L_2) = n^2$.

Now suppose A is not a multiple of the identity matrix. By [9, Lem. 1.3], we have rank$(L_1) = 2n - 2$ if rank$(A - \alpha I) = 1$ for some $\alpha \in \mathbb{C}$, and rank$(L_1) \geq 2n$ otherwise. Next consider rank(L_2). If $A = \lambda E_{11}$, then Im$(L_2) =$ span$\{E_{ij} : i = 1$ or $j = 1\}$, and hence rank$(L_2) = 2n - 1$. If $A = \lambda E_{12}$, then Im$(L_2) =$ span$(\{E_{11} + E_{22}\} \cup \{E_{i2} : i \geq 3\} \cup \{E_{1j} : j \geq 2\})$, and hence rank$(L_2) = 2n - 2$. If rank$(A - \alpha I) = 1$ for some nonzero $\alpha \in \mathbb{C}$, then $A = S(\nu E_{11} + \mu E_{12} + \alpha \sum_{i=2}^{n} E_{ii})S^{-1}$ for some nonsingular $S \in \mathbb{C}_{n\times n}$. Notice that if A has eigenvalues $\lambda_1, \cdots, \lambda_n$, then L_2 has eigenvalues $\lambda_i + \lambda_j$ $(1 \leq i, j \leq n)$. It follows that if $\nu \neq -\alpha$ and $\nu \neq 0$, then L_2 has n^2 nonzero eigenvalues, and hence the rank of L_2 is n^2. If $\nu = 0$, then L_2 has $n^2 - 1$ nonzero eigenvalues, so the rank of L_2 is at least $n^2 - 1$. In both cases, we have rank(L_1) + rank$(L_2) > n^2$.

Now suppose $\nu = -\alpha$. Since the complex eigenvalues of A occur in conjugate pairs, we have either (i) $\nu = -\alpha \in \mathbb{R}$; or (ii) $n = 2$ and $\nu = \overline{\alpha}$ are pure imaginary. If (i) holds, then A is similar to $\alpha(I - 2E_{11})$ (via a real matrix), which in turn is consimilar to αI (via $S = (i-1)E_{11} - I$), and hence A is consimilar to αI. If (ii) holds, then A is similar to $|\alpha|(E_{12} - E_{21})$ via a real matrix. In both cases, one can check that $\text{rank}(L_1) + \text{rank}(L_2) = n^2$. Finally, if $\text{rank}(A - \alpha I) > 1$ for all $\alpha \in \mathbb{C}$, the same arguments used in the proof of [9, Lem. 1.3] show that $\text{rank}(L_2) \geq 2n$ over $\mathbb{C}$, and hence $\text{rank}(L_2) \geq 2n$ over $\mathbb{R}$ as well. Thus, $\text{rank}(L_1) + \text{rank}(L_2) \geq 4n$. □

LEMMA 2.6. *Suppose the linear operator $T : \mathbb{C}_{n\times n} \to \mathbb{C}_{n\times n}$ is nonsingular and satisfies $T(\mathcal{O}(E_{11})) \subset \mathcal{O}(E_{11})$. Then there exists a nonsingular $S \in \mathbb{C}_{n\times n}$ such that either*

$$T(X) = \overline{S}XS^{-1} \qquad \textit{for all} \quad X \in \mathbb{C}_{n\times n},$$

or

$$T(X) = \overline{S}X^tS^{-1} \qquad \textit{for all} \quad X \in \mathbb{C}_{n\times n}.$$

Proof. Let e_i be the ith column of the identity matrix. Set $x_1 = e_1, x_i = e_1 + e_i$ *for* $i = 2, \cdots, n$. We may assume $T(x_1x_1^t) = x_1x_1^t$; otherwise consider T' defined by $T'(X) = \overline{S}T(X)S^{-1}$ for some suitable invertible S. For any $a = (a_2, \cdots, a_n)^t \in \mathbb{C}^{n-1}$, let $y_a = (1, a_2, \cdots, a_n)^t$. Then $x_1(\mu x_1 + (1-\mu)y_a)^t \in \mathcal{O}(E_{11})$ by Lemma 2.2. It follows that $\mu x_1x_1^t + (1-\mu)T(x_1y_a^t) \in \mathcal{O}(E_{11})$, and hence either

$$T(x_1y_a^t) \in \left\{E_{11} + \sum_{i=2}^n b_iE_{1i} : b_i \in \mathbb{C}\right\} \quad \text{or} \quad T(x_1y_a^t) \in \left\{E_{11} + \sum_{i=2}^n b_iE_{i1} : b_i \in \mathbb{C}\right\}.$$

Choose any fixed nonzero $c \in \mathbb{C}^{n-1}$. We may assume $T(x_1y_c^t) = E_{11} + \sum_{i=2}^n d_iE_{1i}$ for some $d_i \in \mathbb{C}$; otherwise consider T' defined by $T'(X) = T(X)^t$. Since T is invertible and c is nonzero, some of the d_i's must be nonzero. Now take any $a \in \mathbb{C}^{n-1}$ and consider $e = (a+c)/2$. Since $x_1y_e^t \in \mathcal{O}(E_{11})$, it follows that $T(x_1y_a^t)/2 + T(x_1y_c^t)/2 = T(x_1y_e^t)$ is also in $\mathcal{O}(E_{11})$ and hence has rank one. It then follows that $T(x_1y_a^t) \in \{E_{11} + \sum_{i=2}^n b_iE_{1i} : b_i \in \mathbb{C}\}$, also. Since T is invertible, the mapping $T_0 : \mathbb{C}^{n-1} \to \mathbb{C}^{n-1}$ defined by

$$T_0(a) = b \quad \text{if} \quad T(x_1y_a)^t = E_{11} + \sum_{i=2}^n b_iE_{1i}$$

is linear and invertible. Let $R \in \mathbb{C}_{(n-1)\times(n-1)}$ be such that $T_0(a) = R^ta$ for all $a \in \mathbb{C}^{n-1}$. Consider T' defined by $T'(X) = \overline{S}T(X)S^{-1}$ with $S = (1) \oplus R$. Then $T'(x_1x_i^t) = x_1x_i^t$ for $i = 1, \cdots, n$. For notational convenience, we write T instead of T'; we shall prove that $T(X) \equiv X$. Now for any $\mu, \nu \geq 0$ and any $i, j \neq 1$,

$$x_1x_1^t + \nu(x_1x_j^t) + \mu T(x_ix_1^t) + \mu\nu T(x_ix_j^t) = T((x_1 + \mu x_i)(x_1 + \nu x_j)^t) \in k\mathcal{O}(E_{11}),$$

where $k = (1+\mu)(1+\nu) + \mu\nu\delta_{ij}$. Taking $\nu = 0$, we see that $T(x_ix_1^t) = z_ix_1^t \in \mathcal{O}(E_{11})$ for some $z_i \in \mathbb{C}^n$ with $x_1^*z_i = 1$ by Lemma 2.5. (Since T is one-to-one, it follows that $T(x_ix_1^t)$ cannot be in $\{E_{11} + \sum_{i=2}^n b_iE_{1i} : b_i \in C\}$.) Thus

$$T((x_1 + \mu x_i)(x_1 + \nu x_j)^t) = (x_1 + \mu z_i)(x_1 + \nu x_j)^t + \mu\nu(T(x_ix_j^t) - z_ix_j^t)$$

always has rank one. It follows that $T(x_ix_j^t) = z_ix_j^t$ for all $i, j \geq 2$. As a result, if $i, j \geq 2, i \neq j$, then $x_j^*x_i = 1 = x_j^*z_i$; and if $i = j \geq 2$, then $x_j^*x_i = 2 = x_j^*z_i$.

Consequently, we must have $z_i = x_i$ and hence $T(x_i x_j^t) = x_i x_j^t$ for all $i, j \geq 1$. Therefore, $T(X) = X$ for all $X \in \mathbb{C}_{n\times n}$, as required. □

Proof of Theorem 2.1. ($\Leftarrow$) The sufficiency part of the theorem can be verified readily.

($\Rightarrow$) If $T = 0$, then the conclusion holds with $\alpha = 0$. If $T \neq 0$, T is nonsingular by Lemma 2.4. If $n > 3$, let $A \equiv T^{-1}(E_{12})$. Since $T(\mathcal{T}_A) \subset \mathcal{T}_{E_{12}}$ and T is nonsingular, $\dim \mathcal{T}_A \leq \dim \mathcal{T}_{E_{12}}$. By Lemma 2.5 and the fact that $\mathcal{O}(\lambda E_{12}) = \mathcal{O}(E_{12})$ if $\lambda \neq 0$, we have $A \in \mathcal{O}(E_{12})$, and hence $T(\mathcal{O}(A)) = T(\mathcal{O}(E_{12})) \subset \mathcal{O}(E_{12})$. Next consider $B \equiv T^{-1}(E_{11})$. Since $T(\mathcal{T}_B) \subset \mathcal{T}_{E_{11}}$, it follows that $\dim \mathcal{T}_B \leq \dim \mathcal{T}_{E_{11}}$. By Lemma 2.5, $B \in \mathcal{O}(E_{12})$ or $B \in \mathcal{O}(\mu E_{11})$ for some $\mu > 0$. Since $T(\mathcal{O}(E_{12})) \subset \mathcal{O}(E_{12})$, we must have $B \in \mathcal{O}(\mu E_{11})$ for some $\mu > 0$. Thus $\mu T(\mathcal{O}(E_{11})) \subset \mathcal{O}(E_{11})$, and the result now follows from Lemma 2.6.

Suppose $n = 3$. Using arguments similar to those preceding, we have $T(\mathcal{O}(E_{12})) \subset \mathcal{O}(E_{12})$. Now set $B = T^{-1}(E_{11})$. Since $\dim \mathcal{T}_B \leq \dim \mathcal{T}_{E_{11}}$, by Lemma 2.5 we have $B \in \mathcal{O}(\mu E_{11})$ or $B \in \mathcal{O}(\mu I)$, for some $\mu > 0$. If $B \in \mathcal{O}(\mu E_{11})$, then the result follows as before. If $B \in \mathcal{O}(\mu I)$, then $\cup_{\mu>0} T(\mathcal{O}(\mu I)) \subset \cup_{\mu>0} \mathcal{O}(\mu E_{11})$. Thus, $T^{-1}(I)$ must be of the form μE_{11}, and hence $\cup_{\mu>0} T(\mathcal{O}(\mu E_{11})) \subset \cup_{\mu>0} \mathcal{O}(\mu I)$. But then the matrices $C_1 = T(I)$ and $C_2 = T(E_{11} - E_{22} - E_{33})$ are both in $\cup_{\mu>0} \mathcal{O}(\mu E_{11})$ and have rank one. Thus, $2T(E_{11}) = C_1 + C_2$ must be singular, which contradicts the fact that $2T(E_{11}) \in \cup_{\mu>0} \mathcal{O}(\mu I)$.

Suppose $n = 2$, and let $F_{12} \equiv E_{12} - E_{21}$. By comparing the dimensions of the tangent spaces and using the fact that T is nonsingular, we can conclude that there exist $\mu, \nu > 0$ such that $T^{-1}(E_{12})$, $T^{-1}(I)$, $T^{-1}(F_{12})$ are lying in the different orbits $\mathcal{O}(\mu I)$, $\mathcal{O}(\nu F_{12})$, $\mathcal{O}(E_{12})$. It follows that

$$T(\cup_{\mu>0}\mathcal{O}(\mu I) \cup [\cup_{\nu>0}\mathcal{O}(\nu F_{12})] \cup \mathcal{O}(E_{12})) \subset \cup_{\mu>0}\mathcal{O}(\mu I) \cup [\cup_{\nu>0}\mathcal{O}(\nu F_{12})] \cup \mathcal{O}(E_{12}).$$

Consequently, $T^{-1}(E_{11}) \in \eta\mathcal{O}(E_{11})$ for some $\eta > 0$, and the result follows. □

In $\mathbb{R}_{n\times n}$, consimilarity is the same as ordinary *similarity*. Although Hiai characterized only the linear operators that preserve similarity on $\mathbb{C}_{n\times n}$, the proof in [9] can be modified to yield the same result in the real case as in the complex case. We summarize the results in the following theorem.

THEOREM 2.7. *Let* $\mathbb{F} = \mathbb{R}$ *or* $\mathbb{C}$. *A linear operator* $T : \mathbb{F}_{n\times n} \to \mathbb{F}_{n\times n}$ *satisfies*

$$T(A) \text{ is similar to } T(B) \quad \text{whenever} \quad A \text{ is similar to } B$$

if and only if one of the following happens :

(a) *there exists* $A_0 \in \mathbb{F}_{n\times n}$ *such that*

$$T(X) = (tr\ X)A_0 \qquad \text{for all} \quad X \in \mathbb{F}_{n\times n};$$

(b) *there exist a nonsingular* $S \in \mathbb{F}_{n\times n}$ *and* $\alpha, \beta \in \mathbb{F}$ *such that either*

$$T(X) = \alpha SXS^{-1} + \beta(tr\ X)I \qquad \text{for all} \quad X \in \mathbb{F}_{n\times n},$$

or

$$T(X) = \alpha SX^tS^{-1} + \beta(tr\ X)I \qquad \text{for all} \quad X \in \mathbb{F}_{n\times n}.$$

3. Congruence. In this section we take $\mathcal{M} = \mathbb{C}_{n\times n}$ or $\mathcal{H}_n$ and let $\sim$ be $*$-*congruence*, i.e., $A \sim B$ if there exists a nonsingular S in $\mathbb{C}_{n\times n}$ such that $A = SBS^*$. Our main result is the following theorem.

THEOREM 3.1. *Let $\mathcal{M}$ be $\mathbb{C}_{n\times n}$ or $\mathcal{H}_n$. A linear operator $T : \mathcal{M} \to \mathcal{M}$ satisfies*

$$T(A) \text{ is } *\text{-congruent to } T(B) \text{ whenever } A \text{ is } *\text{-congruent to } B$$

if and only if there exist a nonsingular $S \in \mathbb{C}_{n\times n}$ and $\alpha \in \mathbb{F}$ (where $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$ according as $\mathcal{M} = \mathbb{C}_{n\times n}$ or $\mathcal{H}_n$), with $\alpha = 0$ or $|\alpha| = 1$ such that either

$$T(X) = \alpha SXS^* \qquad \text{for all } X \in \mathcal{M},$$

or

$$T(X) = \alpha SX^tS^* \qquad \text{for all } X \in \mathcal{M}.$$

Again, we divide the proof into several lemmata.

LEMMA 3.2. *Let $\mathcal{M}$ be $\mathbb{C}_{n\times n}$ or $\mathcal{H}_n$. If $A \in \mathcal{M}$ is nonzero, then* span $\mathcal{O}(A) = \mathcal{M}$.

Proof. Let $A \in \mathcal{M}$ be a nonzero matrix. If $A \neq \lambda I$ and tr $A \neq 0$, then (e.g., see [25] and the proof of Theorem 2.1 in [9]) span$\{UAU^* : U \in U_n(\mathbb{C})\} = \mathcal{M}$. If $A = \lambda I$ or tr $A = 0$, there is a nonsingular $S \in \mathbb{C}_{n\times n}$ such that $SAS^* \neq \lambda I$ and $\operatorname{tr}(SAS^*) \neq 0$. Then span $\mathcal{O}(A) =$ span $\mathcal{O}(SAS^*) = \mathcal{M}$. □

One may also prove Lemma 3.2 by arguments similar to those in the proof of Lemma 2.3.

The desired nonsingularity of T now follows from Lemma 1.1 (c).

LEMMA 3.3. *Let $\mathcal{M}$ be $\mathbb{C}_{n\times n}$ or $\mathcal{H}_n$. Suppose $T : \mathcal{M} \to \mathcal{M}$ is a nonzero linear operator that preserves $*$-congruence. Then T is nonsingular.*

LEMMA 3.4. *Let $\mathcal{M}$ be $\mathbb{C}_{n\times n}$ or $\mathcal{H}_n$. Suppose $A \in \mathcal{M}$ is nonzero. Then the tangent space to $\mathcal{O}(A)$ at A is*

$$\mathcal{T}_A = \{XA + AX^* : X \in \mathbb{C}_{n\times n}\}.$$

Moreover, the real dimension of $\mathcal{T}_A$ is at least $2n-1$, and it equals $2n-1$ if and only if $A = \alpha SE_{11}S^*$ for some $\alpha \in \mathbb{F}$ and some nonsingular $S \in \mathbb{C}_{n\times n}$.

Proof. The asserted form for $\mathcal{T}_A$ follows immediately from the power series expansion of SAS^* with $S = e^{\varepsilon X} = \sum_{p=0}^{\infty}(\varepsilon X)^p/p!, X \in \mathbb{C}_{n\times n}$.

To prove the second part of the lemma, we first consider the case of $\mathcal{M} = \mathcal{H}_n$. If A has r positive eigenvalues, s negative eigenvalues, and t zero eigenvalues, then there exists a nonsingular S such that $SAS^* = I_r \oplus -I_s \oplus 0_t$. Since $\mathcal{O}(A)$ is a homogeneous manifold, we may assume $A = I_r \oplus -I_s \oplus 0_t$ in order to compute the dimension of $\mathcal{T}_A$. In this case, $\mathcal{T}_A$ is just the collection of Hermitian matrices whose (i,j) entries equal 0 if $i \geq r+s$ and $j \geq r+s$. The real dimension of $\mathcal{T}_A$ is evidently $n^2 - t^2$, and the minimum occurs when $t = n-1$. The result follows.

Now suppose $\mathcal{M} = \mathbb{C}_{n\times n}$. Let $A = A_1 + iA_2$ with $A_1, A_2 \in \mathcal{H}_n$. We may assume $A_1 \neq 0$; otherwise consider μA for some nonzero $\mu \in \mathbb{C}$. Then

$$\mathcal{T}_A = \{(XA_1 + A_1X^*) + i(XA_2 + A_2X^*) : X \in \mathbb{C}_{n\times n}\},$$

and hence

$$\dim \mathcal{T}_A \geq \dim \mathcal{T}_{A_1} \geq 2n-1.$$

Moreover, if A is not a scalar multiple of a matrix of the form $SE_{11}S^*$ with $S \in \mathbb{C}_{n\times n}$, we can find a nonsingular $R \in \mathbb{C}_{n\times n}$ and a nonzero $\nu \in \mathbb{C}$ such that the matrix $B = \nu RAR^* + \bar{\nu}RA^*R^*$ has rank at least 2. It follows that

$$\dim \mathcal{T}_A \geq \dim \mathcal{T}_B > 2n-1. \qquad \square$$

The following result is in [12].

LEMMA 3.5. *Suppose the linear operator* $T : \mathcal{H}_n \to \mathcal{H}_n$ *is nonsingular and satisfies* $T(\mathcal{O}(E_{11})) \subset \mathcal{O}(E_{11})$. *Then there exists a nonsingular* $S \in \mathbb{C}_{n\times n}$ *such that either*

$$T(X) = SXS^* \qquad \text{for all } X \in \mathcal{H}_n,$$

or

$$T(X) = SX^tS^* \qquad \text{for all } X \in \mathcal{H}_n.$$

Proof of Theorem 3.1. ($\Leftarrow$) The sufficiency part of the theorem can be verified readily.

($\Rightarrow$) If $T = 0$, then the conclusion holds with $\alpha = 0$. If $T \neq 0$, then T is nonsingular by Lemma 3.3.

Suppose $\mathcal{M} = \mathcal{H}_n$ and $A = T^{-1}(E_{11})$. Since $T(\mathcal{T}_A) \subset \mathcal{T}_{E_{11}}$ and T is nonsingular, $\dim \mathcal{T}_A \leq \dim \mathcal{T}_{E_{11}}$. By Lemma 3.4, $A \in \mathcal{O}(\alpha E_{11})$, so $T(\mathcal{O}(\alpha E_{11})) \subset \mathcal{O}(E_{11})$. Notice that $\mathcal{O}(\alpha E_{11}) = \mathcal{O}(E_{11})$ if $\alpha > 0$, and $\mathcal{O}(\alpha E_{11}) = \mathcal{O}(-E_{11})$ if $\alpha < 0$. The result now follows from Lemma 3.5.

Now suppose $\mathcal{M} = \mathbb{C}_{n\times n}$ and $A = T^{-1}(E_{11})$. By the same arguments we have already made, $A \in \mathcal{O}(\alpha E_{11})$ for some nonzero $\alpha \in \mathbb{C}$. Thus, $T'(\mathcal{O}(E_{11})) \subset \mathcal{O}(E_{11})$, where $T' = \alpha T$. Since $\mathcal{H}_n$ is the real span of $\mathcal{O}(E_{11})$, it follows that T' maps $\mathcal{H}_n$ into $\mathcal{H}_n$. Regarding T' as a real linear operator on $\mathcal{H}_n$, T' satisfies the conclusion of the theorem. Since $\mathbb{C}_{n\times n}$ is the complex span of $\mathcal{H}_n$, the complex linear operator T', and hence T, is of the required form. □

If S is unitary, $*$-congruence via S is *unitary similarity*. Hiai [9] characterized the real linear operators $T : \mathcal{H}_n \to \mathcal{H}_n$ that preserve unitary similarity. Using our method in the proof of Theorem 3.1, one can extend the result of Hiai to $\mathbb{C}_{n\times n}$. We summarize the result in the following theorem.

THEOREM 3.6. *Let* $\mathcal{M}$ *be* $\mathbb{C}_{n\times n}$ *or* $\mathcal{H}_n$. *A linear operator* $T : \mathcal{M} \to \mathcal{M}$ *satisfies*

$$T(A) \text{ is unitarily similar to } T(B) \quad \text{whenever} \quad A \text{ is unitarily similar to } B$$

if and only if one of the following happens:

(a) *there exists* $A_0 \in \mathcal{M}$ *such that*

$$T(X) = (\operatorname{tr} X)A_0 \qquad \text{for all } X \in \mathcal{M};$$

(b) *there exist* $U \in U_n(\mathbb{C})$ *and* $\alpha, \beta \in \mathbb{F}$, *where* $\mathbb{F} = \mathbb{C}$ *or* $\mathbb{R}$ *according as* $\mathcal{M} = \mathbb{C}_{n\times n}$ *or* $\mathcal{H}_n$, *such that either*

$$T(X) = \alpha UXU^* + \beta(\operatorname{tr} X)I \qquad \text{for all } X \in \mathcal{M},$$

or

$$T(X) = \alpha UX^tU^* + \beta(\operatorname{tr} X)I \qquad \text{for all } X \in \mathcal{M}.$$

4. Equivalence. In this section we take $\mathcal{M} = \mathbb{F}_{m\times n}$ with $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, and let $\sim$ be (*nonsingular*) *equivalence*, i.e., $A \sim B$ if there exist nonsingular $M \in \mathbb{F}_{m\times m}$ and $N \in \mathbb{F}_{n\times n}$ such that $A = MBN$. Our principal result is the following theorem.

THEOREM 4.1. *A nonzero linear operator* $T : \mathbb{F}_{m\times n} \to \mathbb{F}_{m\times n}$ *satisfies*

$$T(A) \text{ is equivalent to } T(B) \quad \text{whenever} \quad A \text{ is equivalent to } B$$

if and only if there exist nonsingular $M \in \mathbb{F}_{m\times m}$ *and* $N \in \mathbb{F}_{n\times n}$ *such that either*

$$T(X) = MXN \qquad \text{for all } X \in \mathbb{F}_{m\times n},$$

or $m = n$ *and*

$$T(X) = MX^tN \qquad \text{for all } X \in \mathbb{F}_{m\times n}.$$

If M and N are unitary, then equivalence via M and N is *unitary equivalence.* We have the following theorem in this important special case.

THEOREM 4.2. *A nonzero linear operator* $T : \mathbb{F}_{m\times n} \to \mathbb{F}_{m\times n}$ *satisfies*

$T(A)$ *is unitarily equivalent to* $T(B)$ *whenever* A *is unitarily equivalent to* B

if and only if there exist $\alpha > 0$, $U \in U_m(\mathbb{F})$, *and* $V \in U_n(\mathbb{F})$ *such that either*

$$T(X) = \alpha UXV \qquad \text{for all } X \in \mathbb{F}_{m\times n},$$

or $m = n$ *and*

$$T(X) = \alpha UX^tV \qquad \text{for all } X \in \mathbb{F}_{m\times n}.$$

Since the proof of Theorem 4.1 is similar to those of Theorems 2.1 and 3.1, we just list the lemmata required and omit their proofs.

LEMMA 4.3. Span $\mathcal{O}(A) = \mathbb{F}_{m\times n}$ *for every nonzero* $A \in \mathbb{F}_{m\times n}$.

LEMMA 4.4. *Let* $T : \mathbb{F}_{m\times n} \to \mathbb{F}_{m\times n}$ *be a nonzero linear operator that preserves equivalence. Then* T *is nonsingular.*

LEMMA 4.5. *Let* $A \in \mathbb{F}_{m\times n}$ *be a nonzero matrix. Then the tangent space to* $\mathcal{O}(A)$ *at* A *is*

$$\mathcal{T}_A = \{XA + AY : X \in \mathbb{F}_{m\times m} \text{ and } Y \in \mathbb{F}_{n\times n}\}.$$

Moreover, the dimension of $\mathcal{T}_A$ *(over* $\mathbb{F}$*) is at least* $m+n-1$, *and it equals* $m+n-1$ *if and only if* A *is a rank one matrix.*

In the present case in which $\sim$ is equivalence, $\mathcal{O}(E_{11})$ is simply the set of all rank one matrices. The following result is in [1] and [4].

LEMMA 4.6. *If a linear operator* $T : \mathbb{F}_{m\times n} \to \mathbb{F}_{m\times n}$ *is nonsingular and preserves rank one matrices, then* T *is of the form described in Theorem* 4.1.

Using these lemmata, we can prove Theorem 4.1. In the special case of unitary equivalence, there are results similar to Lemmata 4.3 and 4.4. The analog of Lemma 4.5 is the following.

LEMMA 4.7. *Let* $A \in \mathbb{F}_{m\times n}$ *be a nonzero matrix, and let* $\sim$ *be unitary equivalence.*

(a) *If* $\mathbb{F} = \mathbb{C}$, *then the tangent space to* $\mathcal{O}(A)$ *at* A *is*

$$\mathcal{T}_A = \{i(XA + AY) : X \in \mathcal{H}_m \text{ and } Y \in \mathcal{H}_n\}.$$

Moreover, the real dimension of $\mathcal{T}_A$ *is at least* $2(m+n)-3$, *and it equals* $2(m+n)-3$ *if and only if* A *is a rank one matrix.*

(b) *If* $\mathbb{F} = \mathbb{R}$, *then the tangent space to* $\mathcal{O}(A)$ *is*

$$\mathcal{T}_A = \{XA + AY : X = -X^t \in \mathbb{R}_{m\times m} \text{ and } Y = -Y^t \in \mathbb{R}_{n\times n}\}.$$

Moreover, the real dimension of $\mathcal{T}_A$ *is at least* $m+n-2$, *and it equals* $m+n-2$ *if and only if* A *is a rank one matrix.*

Now we are ready to prove Theorem 4.2.

Proof of Theorem 4.2. ($\Leftarrow$) The sufficiency part of the theorem can be verified readily.

($\Rightarrow$) Suppose T is nonzero. Then T is nonsingular and $T^{-1}(E_{11})$ must have rank one by Lemma 4.7. Thus $\cup_{\mu>0}T(\mathcal{O}(\mu E_{11})) \subset \cup_{\mu>0}\mathcal{O}(\mu E_{11})$, and hence T preserves rank one matrices. By Lemma 4.6, T has the form described in Theorem 4.1. Let the matrices M and N have singular value decompositions $X_1D_1X_2$ and $Y_1D_2Y_2$, respectively, where $X_1, X_2 \in U_m(\mathbb{F})$, $Y_1, Y_2 \in U_n(\mathbb{F})$, and D_1, D_2 are positive diagonal matrices with diagonal entries arranged in nonincreasing order. If $T(X) = MXN$, let $A_1 = X_2^*E_{11}Y_1^*$ and $A_2 = X_2^*E_{mn}Y_1^*$. Since A_1 and A_2 are unitarily equivalent, so are $T(A_1)$ and $T(A_2)$. It follows that all of the singular values of M (respectively, N) are the same, so M and N are both multiples of unitary matrices and hence $T(X) = \alpha UXV$, as asserted by Theorem 4.2. If $T(X) = MX^tN$, the same argument shows that $T(X) = \alpha UX^tV$, as asserted by Theorem 4.2. □

5. Related results and questions. In this paper we have characterized those linear operators T satisfying

$$T(\mathcal{O}(A)) \subset \mathcal{O}(T(A)) \quad \text{for every} \quad A \in \mathcal{M}$$

for several choices of $\sim$ and $\mathcal{M}$. In [11], characterizations are given for the linear operators on various matrix spaces that preserve *t-congruence*, i.e., $A \sim B$ if $A = SBS^t$ for some nonsingular matrix S. One may also consider the problem of characterizing linear operators T such that

$$T(\mathcal{O}(A)) = \mathcal{O}(A) \quad \text{or} \quad T(\mathcal{O}(A)) \subset \mathcal{O}(A),$$

where A is a given fixed matrix. Even more generally, one might try to determine the conditions on a pair of matrices A and B so that there is a linear operator with

$$T(\mathcal{O}(A)) = \mathcal{O}(B) \quad \text{or} \quad T(\mathcal{O}(A)) \subset \mathcal{O}(B),$$

and then characterize T. In fact, many authors have studied these linear preserver problems under different settings, and many results have been obtained. We list only a few references below indicating the source of some results on each of the indicated problems (see also [6] and [20]).

For linear operators preserving an orbit of the consimilarity relation on $\mathbb{C}_{n\times n}$, see our Lemma 2.6.

For linear operators preserving an orbit of the similarity relation on $\mathbb{F}_{n\times n}$, see [22] .

For linear operators preserving an orbit of the $*$-congruence relation on $\mathcal{H}_n$ (the orbit is then an inertia class), see [8], [12], [18] .

For linear operators preserving an orbit of the unitary similarity relation on $\mathcal{H}_n$ or $\mathbb{C}_{n\times n}$, see [13] .

For linear operators preserving an orbit of the (nonsingular) equivalence relation on $\mathbb{F}_{m\times n}$ (the orbit is then a set of matrices with a fixed rank), see [1], [3], [4], [5], [16], [17], [21], [24].

For linear operators preserving an orbit of the unitary equivalence relation on $\mathbb{F}_{m\times n}$ (the orbit is then a set of matrices with prescribed singular values), see [7], [14], [19], [23].

For linear operators preserving an orbit of the t-congruence relation on various matrix spaces, see [15].

REFERENCES

[1] L. R. BEASLEY, *Linear operators on matrices: The invariance of rank-k matrices*, Linear Algebra Appl., 107 (1988), pp. 161–167.

[2] W. M. BOOTHBY, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.

[3] E. P. BOTTA, *Linear maps preserving rank less than or equal to one*, Linear and Multilinear Algebra, 20 (1987), pp. 197–201.

[4] G. H. CHAN AND M. H. LIM, *Linear transformations on tensor spaces*, Linear and Multilinear Algebra, 14 (1983), pp. 3–9.

[5] D. Z. DJOKOVIC, *Linear transformations of tensor products preserving a fixed rank*, Pacific J. Math., 30 (1969), pp. 411–414.

[6] R. GRONE, *Isometries of Matrix Algebras*, Ph.D. thesis, Department of Mathematics, University of California, Santa Barbara, CA, 1976.

[7] ———, *The invariance of partial isometries*, Indiana Univ. Math. J., 28 (1979), pp. 445–449.

[8] J. W. HELTON AND L. RODMAN, *Signature preserving linear maps of Hermitian matrices*, Linear and Multilinear Algebra, 17 (1985), pp. 29–37.

[9] F. HIAI, *Similarity preserving linear maps on matrices*, Linear Algebra Appl., 97 (1987), pp. 127–139.

[10] Y. P. HONG AND R. A. HORN, *A canonical form for matrices under consimilarity*, Linear Algebra Appl., 102 (1988), pp. 143–168.

[11] Y. P. HONG, R. A. HORN, AND C. K. LI, *Linear operators preserving t-congruence on matrices*, in preparation.

[12] C. R. JOHNSON AND S. PIERCE, *Linear maps on Hermitian matrices: The stabilizer of an inertia class* II, Linear and Multilinear Algebra, 19 (1986), pp. 21–31.

[13] C. K. LI AND N. K. TSING, *Duality between some linear preserver problems: The invariance of the C-numerical range, the C-numerical radius and certain matrix sets*, Linear and Multilinear Algebra, 23 (1988), 353–362.

[14] ———, *Duality between some linear preserver problems* II*: Isometries with respect to c-spectral norms and matrices with fixed singular values*, Linear Algebra Appl., 110 (1988), pp. 181–212.

[15] ———, *Duality between some linear preserver problems* III*: c-spectral norms and symmetric or skew-symmetric matrices with fixed singular values*, Linear Algebra Appl., to appear.

[16] M. H. LIM, *Linear transformations on symmetric matrices*, Linear and Multilinear Algebra, 7 (1979), pp. 45–57.

[17] ———, *Rank preservers of skew symmetric matrices*, Pacific. J. Math., 35 (1970), pp. 169–174.

[18] R. LOEWY, *Linear maps which preserve an inertia class*, SIAM J. Matrix Anal. Appl., to appear.

[19] M. MARCUS, *All linear operators leaving the unitary group invariant*, Duke Math. J., 26 (1959), pp. 155–163.

[20] ———, *Linear transformations on matrices*, J. Research Nat. Bur. Standards, 75B (1971), pp. 107–113.

[21] M. MARCUS AND B. MOYLS, *Transformations on tensor product spaces*, Pacific J. Math., 9 (1959), pp. 1215–1221.

[22] W. WATKINS, *Linear transformations that preserve a similarity class of matrices*, Linear and Multilinear Algebra, 11 (1982), pp. 19–22.

[23] A. WEI, *Linear transformations preserving the real orthogonal group*, Canad. J. Math., 27 (1975), pp. 561–572.

[24] R. WESTWICK, *Transformations on tensor spaces*, Pacific J. Math., 23 (1967), pp. 613–620.

[25] B. S. TAM, *A simple proof of the Goldberg–Straus theorem on numerical radii*, Glasgow Math. J., 28 (1986), pp. 139–141.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 2, pp. 205–208, April 1991

002

TWO SIMPLE RESIDUAL BOUNDS FOR THE EIGENVALUES OF A HERMITIAN MATRIX*

G. W. STEWART†

Abstract. Let A be Hermitian and let the orthonormal columns of X span an approximate invariant subspace of X. Then the residual $R = AX - XM$ ($M = X^{\mathrm{H}}AX$) will be small. The theorems of this paper bound the distance of the spectrum of M from the spectrum of A in terms of appropriate norms of R.

Key words. eigenvalue, invariant subspace, perturbation theory, residual bounds

AMS(MOS) subject classifications. 15A18, 15A42

Let A be a Hermitian matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$. If X is a matrix with orthonormal columns that spans an invariant subspace of A and

$$M = X^{\mathrm{H}}AX, \tag{1}$$

then $AX - XM = 0$.

Now suppose that the columns of X span an *approximate* invariant subspace of A. Then the matrix

$$R = AX - XM$$

will be small, say in the spectral norm $\|\cdot\|$ defined by $\|R\| = \max_{\|x\|=1} \|Rx\|$, where $\|x\|$ is the Euclidean norm of x.[1] If the eigenvalues of M are $\mu_1 \geq \cdots \geq \mu_k$, then we should expect the μ_i to be near k of the λ_i. The problem treated in this note is to derive a bound in terms of the matrix R.

An important result, due to Kahan [3] (see also [6, p. 219]), states that there are eigenvalues $\lambda_{j_1}, \cdots, \lambda_{j_k}$ of A such that

$$|\mu_i - \lambda_{j_i}| \leq \|R\|, \qquad i = 1, \cdots, k. \tag{2}$$

If nothing further is known about the spectrum of A, this bound is generally satisfactory, although it can be improved somewhat [5]. However, it frequently happens (e.g., in the Lanczos algorithm or simultaneous iteration [6, Chaps. 13,14]) that we know that $n-k$ of the eigenvalues of A are well separated from the eigenvalues of M: specifically, if we know that

(3) there is a number $\delta > 0$ such that exactly $n-k$ of the eigenvalues of A lie outside the interval $[\mu_k - \delta, \mu_1 + \delta]$,

then the bound in (2) can be replaced by a bound of order $\|R\|^2$. Bounds of this kind have been given by Temple, Kato, and Lehman (see [6, Chap. 10] and [1, §6.5]). Early bounds of this kind dealt only with a single eigenvalue and eigenvector. Lehman's bounds are in some sense optimal, but they are quite complicated.

* Received by the editors January 25, 1990; accepted for publication (in revised form) June 13, 1990.

† Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, Maryland 20742. This work was supported in part by Air Force Office of Scientific Research contract AFOSR-87-0188.

[1] In fact, the choice (1) of M minimizes $\|R\|$, although we will not make use of this fact here.

The purpose of this note is to give two other bounds derived from bounds on the accuracy of the column space of X as an invariant subspace of A. They are very simple to state and yet are asymptotically sharp. In addition, they can be established by appealing to results readily available in the literature.

THEOREM 1. *With the above definitions, assume that A and M satisfy* (3). *If*

$$\rho \equiv \frac{\|R\|}{\delta} < 1,$$

then there is an index j such that $\lambda_j, \cdots, \lambda_{j+k-1} \in (\mu_k - \delta, \mu_1 + \delta)$ and

$$|\mu_i - \lambda_{j+i-1}| \leq \frac{1}{1-\rho^2}\frac{\|R\|^2}{\delta}, \qquad i = 1, \cdots, k.$$

Proof. Let $(X\ Y)$ be unitary. Then

$$\begin{pmatrix} X^{\mathrm{H}} \\ Y^{\mathrm{H}} \end{pmatrix} A(X\ Y) = \begin{pmatrix} M & S^{\mathrm{H}} \\ S & N \end{pmatrix}$$

where $\|S\| = \|R\|$. By the "$\sin\Theta$" theorem of Davis and Kahan [2], there is a matrix P satisfying

$$\|P(I + P^{\mathrm{H}}P)^{1/2}\| \leq \rho \tag{4}$$

such that the columns of

$$\hat{X} = (X + YP)(I + P^{\mathrm{H}}P)^{-1/2}$$

(which are orthonormal) span an invariant subspace of A. From (4) it follows that

$$\frac{\|P\|}{\sqrt{1 + \|P\|^2}} \leq \rho,$$

and since $\rho < 1$

$$\|P\| \leq \frac{\rho}{\sqrt{1-\rho^2}}. \tag{5}$$

Let $\hat{Y} = (Y - XP^{\mathrm{H}})(I + PP^{\mathrm{H}})^{-1/2}$. Then $(\hat{X}\ \hat{Y})$ is unitary. Since the columns of $\hat{X}$ span an invariant subspace of A, we have $\hat{Y}^{\mathrm{H}}A\hat{X} = 0$. Hence

$$\begin{pmatrix} \hat{X}^{\mathrm{H}} \\ \hat{Y}^{\mathrm{H}} \end{pmatrix} A(\hat{X}\ \hat{Y}) = \begin{pmatrix} \hat{M} & 0 \\ 0 & \hat{N} \end{pmatrix}.$$

In [7] it is shown that

$$\hat{M} = (I + P^{\mathrm{H}}P)^{1/2}(M + S^{\mathrm{H}}P)(I + P^{\mathrm{H}}P)^{-1/2}.$$

The eigenvalues of $\hat{M}$ are eigenvalues of A. Since $\rho < 1$, it follows from (2) that they lie in the interval $(\mu_k - \delta, \mu_1 + \delta)$, and hence are $\lambda_j, \cdots, \lambda_{j+k-1}$ for some index j. By a result of Kahan [4] on non-Hermitian perturbations of Hermitian matrices,

$$|\mu_i - \lambda_{j+i-1}| \leq \|(I + P^{\mathrm{H}}P)^{1/2}\|\|(I + P^{\mathrm{H}}P)^{-1/2}\|\|S\|\|P\|, \qquad i = 1, \cdots, k.$$

The theorem now follows on noting that $\|(I+P^{\mathrm{H}}P)^{-1/2}\| \leq 1$ and inserting the bound (5) for $\|P\|$. □

There are two remarks to be made about this theorem. First, it extends to operators in Hilbert space, provided X (now itself an operator) has a finite-dimensional domain. Second, the bound is asymptotically sharp, as may be seen by letting $X = (1\ 0)^{\mathrm{T}}$ and

$$A = \begin{pmatrix} 0 & \epsilon \\ \epsilon & 1 \end{pmatrix}$$

(the eigenvalues of A are asymptotic to ϵ^2 and $1 - \epsilon^2$).

The requirement (3) unfortunately does not allow the eigenvalues of M to be scattered through the spectrum of A. If we pass to the Frobenius norm defined by $\|X\|_{\mathrm{F}}^2 = \mathrm{trace}(X^{\mathrm{H}}X)$, then we can obtain a Hoffman–Wielandt type residual bound. Specifically, if

$$\delta = \min\{|\lambda_i - \mu_j| : \lambda_i \in \lambda(A), \mu_j \in \lambda(M)\} > 0, \tag{6}$$

then a variant of the $\sin\Theta$ theorem shows that there is a matrix P satisfying

$$\|P(I + P^{\mathrm{H}}P)^{1/2}\| \leq \|P(I + P^{\mathrm{H}}P)^{1/2}\|_{\mathrm{F}} \leq \frac{\|R\|_{\mathrm{F}}}{\delta}$$

such that the columns of

$$\hat{X} = (X + YP)(I + P^{\mathrm{H}}P)^{-1/2}$$

span an invariant subspace of A. By a variant of Kahan's theorem due to Sun [9], [8], the eigenvalues $\lambda_{j_1}, \cdots, \lambda_{j_k}$ of $\hat{M}$ may be ordered so that

$$\sqrt{\sum_{i=1}^{k} (\mu_i - \lambda_{j_i})^2} \leq \|(I + P^{\mathrm{H}}P)^{1/2}\| \|(I + P^{\mathrm{H}}P)^{-1/2}\| \|S\|_{\mathrm{F}} \|P\|.$$

Hence we have the following theorem.

THEOREM 2. *With the above definitions, assume that A and M satisfy* (6). *If*

$$\rho_{\mathrm{F}} \equiv \frac{\|R\|_{\mathrm{F}}}{\delta} < 1,$$

then there are eigenvalues $\lambda_{j_1}, \cdots, \lambda_{j_k}$ of A such that

$$\sqrt{\sum_{i=1}^{k} (\mu_i - \lambda_{j_i})^2} \leq \frac{1}{1 - \rho_{\mathrm{F}}^2} \frac{\|R\|_{\mathrm{F}}^2}{\delta}.$$

REFERENCES

[1] F. CHATELIN, *Spectral Approximation of Linear Operators*, Academic Press, New York, 1983.

[2] C. DAVIS AND W. M. KAHAN, *The rotation of eigenvectors by a perturbation.* III, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.

[3] W. KAHAN, *Inclusion theorems for clusters of eigenvalues of Hermitian matrices*, Tech. Report, Computer Science Department, University of Toronto, Toronto, Ontario, Canada, 1967.

[4] W. KAHAN, *Spectra of nearly Hermitian matrices*, Proc. Amer. Math. Soc., 48 (1975), pp. 11–17.
[5] N. J. LEHMANN, *Optimale Eigenwerteinschiessungen*, Numer. Math., 5 (1963), pp. 246–272.
[6] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[7] G. W. STEWART, *Error bounds for approximate invariant subspaces of closed linear operators*, SIAM J. Numer. Anal., 8 (1971), pp. 796–808.
[8] G. W. STEWART AND G.-J. SUN, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.
[9] J.-G. SUN, *On the perturbation of the eigenvalues of a normal matrix*, Math. Numer. Sinica, 6 (1984), pp. 334–336.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 2, pp. 209–219, April 1991

003

INERTIA-PRESERVING MATRICES*

ABRAHAM BERMAN† AND DAFNA SHASHA†

Abstract. A real matrix A is *inertia preserving* if in AD = in D, for every invertible diagonal matrix D. This class of matrices is a subset of the D-stable matrices and contains the diagonally stable matrices.

In order to study inertia-preserving matrices, matrices that have no imaginary eigenvalues are characterized. This is used to characterize D-stability of stable matrices. It is also shown that irreducible, acyclic D-stable matrices are inertia preserving.

Key words. inertia-preserving matrices, diagonally stable and semistable matrices, D-stable matrices, irreducible acyclic matrices

AMS(MOS) subject classifications. 15A18, 15A99, 93D05

1. Introduction. The inertia, in A, of a square matrix A is a triple $(i_+(A), i_0(A), i_-(A))$, where $i_+(A)$ is the number of eigenvalues of A in the right open halfplane, $i_0(A)$ is the number of pure imaginary eigenvalues of A, and $i_-(A)$ is the number of eigenvalues in the left open halfplane.

A matrix $A \in R^{n\times n}$ is (*positive*) *stable* if $i_+(A) = n$. A is *D-stable* if AD is stable for every positive diagonal matrix (a diagonal matrix whose diagonal entries are positive) D. A is (*Lyapunov*) *diagonally* (*semi*)*stable* if there exists a positive diagonal matrix D such that $AD + DA^T$ is positive (semi)definite. It is known, e.g., [2], that diagonally stable matrices are D-stable. Stable, D-stable, and diagonally stable matrices arise in problems in ecology, chemistry, and economics, e.g., [2], [8], [7]. A real matrix A is *inertia preserving* if for every invertible diagonal matrix D, in AD = in D. In this paper we study these matrices. This is of interest because, clearly, inertia-preserving matrices are D-stable. In § 3, which follows a section of notation and preliminaries, we compare inertia-preserving matrices with special D-stable matrices and, in particular, show that diagonally stable matrices are inertia preserving.

If A is inertia preserving then necessarily $i_0(A) = 0$. In § 4 we characterize this important condition. In § 5 we restrict our discussion to diagonally semistable matrices and finally, in § 6, we prove that acyclic irreducible D-stable matrices are inertia preserving.

2. Notation and preliminaries. In this section we collect definitions and results needed in the paper. Some notation and results are given only when needed, particularly, in § 6.

The definitions and preliminaries are divided into four groups: general notation, stability and inertia, cones and consistency, and graph theoretical notation.

2.1. General notation. For positive integers n, m, we denote by:
$R[C]$ the set of all real [complex] numbers,
$R^n[C^n]$ the set of all real [complex] n-dimensional (column) vectors,
$R^{n\times m}[C^{n\times m}]$ the set of all real [complex] $n \times m$ matrices,
$\langle\ ,\ \rangle$ an inner product.

* Received by the editors August 7, 1989; accepted for publication (in revised form) February 7, 1990.

† Department of Mathematics, Technion–Israel Institute of Technology, Haifa 32000, Israel (MAR64AA@TECHNION and MAR31AA@TECHNION). The research of the first author was supported by Technion V.P.R.-Fund, Coleman Cohen Research Fund.

In general, almost all the matrices in this paper are real, with the exception of complex eigenvectors x of real matrices and the corresponding square matrices xx^*, where

A^T is the transpose of a matrix A, and,

A^* is $\overline{A^T}$, the complex conjugate of A^T.

Let A be an $n \times n$ matrix, and let α be a nonempty subset of $\{1, \cdots, n\}$. We denote by:

$A[\alpha]$ the principal submatrix of A whose rows and columns are indexed by α in their natural order,

A^i is the ith column of A,

A_i is the ith row of A,

tr A is the trace of A.

The notation $A > 0$ [$\geqq 0$] means that A is positive definite [positive semidefinite].

Denote by $D = \operatorname{diag}\{d_1, \cdots, d_n\}$ the diagonal matrix D whose diagonal entries are $(D)_{ii} = d_i$.

A real diagonal matrix $E = \operatorname{diag}\{e_1, \cdots, e_n\}$ is called a *signature* matrix if $|e_i| = 1$, $i = 1, \cdots, n$.

2.2. Stability and inertia. A *scaling factor* of a diagonally semistable matrix A is a positive diagonal matrix D, such that the matrix $AD + DA^T$ is positive semidefinite.

A property of a matrix A is an *inherited property* if every principal submatrix of A shares it. Diagonal stability and semistability, for example, are inherited properties but D-stability is not.

We denote by P [P_0] the class of $n \times n$ real matrices all of whose principal minors are positive [nonnegative], and by P_0^+ the subclass of P_0 of the matrices with at least one positive principal minor of every order.

It is well known that a diagonally stable [diagonally semistable] matrix must be in P [P_0], and that a D-stable matrix must be in P_0^+.

A key tool in our study is the main inertia theorem due to Tausky [16] and (independently) Ostrowski and Schneider [15].

MAIN INERTIA THEOREM 2.1 [15], [16]. *For a given matrix A, there exists a Hermitian matrix H such that*

$$AH + HA^* > 0$$

if and only if $i_0(A) = 0$. *If* $AH + HA^* > 0$, *then* in A = in H.

We shall also need the following lemma from [6].

LEMMA 2.2 [6]. *Suppose* $A \in C^{n\times n}$, $i_0(A) = 0$ *and H is a nonsingular Hermitian matrix such that* $AH + HA^* \geqq 0$. *Then*, in A = in H.

2.3. Positive-semidefinite matrices. We shall denote by PSD the cone of real positive-semidefinite matrices. The interior of PSD consists of positive-definite real matrices and will be denoted by PD.

Two subcones of PSD, which are of interest, are

$$B_{-0}(A) = \{B \in \mathrm{PSD} \mid (BA)_{ii} \leqq 0, i = 1, \cdots, n\}, \quad \text{and}$$

$$B_0(A) = \{B \in \mathrm{PSD} \mid (BA)_{ii} = 0, i = 1, \cdots, n\}.$$

Obviously, $B_0(A) \subseteq B_{-0}(A)$.

2.4. Graph-theoretical notation. With an $n \times n$ matrix A we associate a directed graph $D(A)$ and a nondirected graph $G(A)$:

$$V(G(A)) = V(D(A)) = \{1, \cdots, n\}.$$

The edges of $D(A)$ are

$$E(D(A)) = \{(i,j); i \neq j; a_{ij} \neq 0\},$$

and the edges of $G(A)$ are

$$E(G(A)) = \{(i,j); i \neq j; a_{ij} \neq 0 \text{ or } a_{ji} \neq 0\}.$$

A matrix is *acyclic* if its nondirected graph $G(A)$ contains no cycles.

It is well known that a matrix is irreducible if and only if $D(A)$ is connected. Let α be a maximal connected subset of $V(D(A))$. Then $A[\alpha]$ is called an irreducible component of A.

3. Classes of *D*-stable matrices. As was pointed out in the introduction, inertia-preserving matrices are D-stable. The converse is not true, as shown by the following example.

Example 3.1.

$$A = \begin{bmatrix} 1 & 0 & -50 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

is D-stable [11]. Using Routh's scheme [9, II, p. 180], we see that for $D = \operatorname{diag}\{-1, 3, -1\}$ the matrix

$$AD = \begin{bmatrix} -1 & 0 & 50 \\ -1 & 3 & 0 \\ -1 & 3 & -1 \end{bmatrix}$$

is stable. Thus A is not inertia preserving.

A subclass of D-stable matrices is the class of *Arrow–McManus D-stable matrices* [1]: matrices A such that AD is stable, where D is a diagonal matrix, if and only if D is positive. Again, it is clear that inertia-preserving matrices are Arrow–McManus D-stable.

Example 3.1 is also an example of a D-stable matrix which is not Arrow–McManus D-stable. Observe that it is not diagonally semistable [14]. In fact, we shall see in § 5 that for diagonally semistable matrices, D-stability and Arrow–McManus D-stability coincide.

We shall wait until § 5 to prove the following.

Example 3.2.

$$A = \begin{bmatrix} 2 & 1 & -2 \\ 3 & 2 & 0 \\ 6 & 4 & 2 \end{bmatrix}$$

is Arrow–McManus D-stable but not inertia preserving.

A real matrix A is *strongly inertia preserving* if for every real diagonal (not necessarily invertible) matrix D, in AD = in D.

Observe that A is strongly inertia preserving if and only if all its principal submatrices are inertia preserving.

Example 3.3. The matrix

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}$$

is inertia preserving but not strongly inertia preserving.

An important class of inertia preserving (and even strongly inertia preserving) matrices is the class of diagonally stable matrices.

THEOREM 3.4. *A diagonally stable matrix is strongly inertia preserving.*

Proof. Diagonal stability is an inherited property. Thus it is enough to prove that A is inertia preserving. Let F be a nonsingular diagonal matrix. Since A is diagonally stable, it has a scaling factor D such that $AD + DA^T = (AF)F^{-1}D + DF^{-1}(FA^T)$ is positive semidefinite. Since $F^{-1}D$ is invertible and Hermitian, it follows by the main inertia theorem (Theorem 2.1), that in AF = in $F^{-1}D$ = in F. □

QUESTION 3.5. Is the converse true? Is every strongly inertia-preserving matrix diagonally stable?

A final remark. Diagonal semistability and being a P matrix are also inherited properties. In § 5 we shall strengthen Example 3.3 by giving an example (Example 5.5) of an inertia-preserving matrix, which is a P-matrix and is diagonally semistable but not strongly inertia preserving.

4. Matrices that have no imaginary eigenvalues. In the beginning of this section we prove some useful properties of eigenvectors of a pure imaginary eigenvalue of a matrix $A \in R^{n\times n}$.

THEOREM 4.1. (a) *Let $A \in R^{n\times n}$ and suppose that $x = c + id$ is an eigenvector of A^T:*

$$A^T X = \alpha i x = \alpha i(c+id), \qquad \alpha \in R, \qquad c,d \in R^n.$$

Denote $B = cc^T + dd^T$. Then

$$BA + A^T B = 0.$$

(b) *If in addition, $\alpha \neq 0$, then c and d are linearly independent.*

Proof. (a)

$$A^T x = \alpha i x, \qquad \alpha \in R, \qquad X \neq 0.$$

Multiplying by $x^\star$ we get

$$A^T x x^\star = \alpha i x x^\star, \tag{4.2}$$

$$x x^\star A = -\alpha i x x^\star. \tag{4.3}$$

Adding (4.2) and (4.3) yields

$$x x^\star A + A^T x x^\star = 0. \tag{4.4}$$

Observe that

$$x x^\star = (c+id)(c^T - id^T) = cc^T + dd^T + i(cd^T - dc^T).$$

Denoting

$$B = cc^T + dd^T,$$

we observe that

$$1 \leqq \text{rank } B \leqq 2 \quad (\text{since } x \neq 0) \tag{4.5}$$

and that

$$BA + A^T B = 0$$

(since $BA + A^T B$ is the real part of the expression in (4.4)).

(b) Suppose $d = kc$ for some scalar k. Then

$$x = (1+ki)c$$

is nonzero and $c = (1 + ki)^{-1}x$ is a real eigenvector of A^T corresponding to the same eigenvalue αi. So,

$$A^T c = \alpha i c,$$

which is impossible since $A^T c$ is a real vector. □

Clearly, if A is inertia preserving, then $i_0(A) = 0$. The following theorem characterizes the real square matrices A such that $i_0(A) = 0$. The theorem is followed by corollaries which are of interest on their own.

THEOREM 4.6. *The following properties of* $A \in R^{n\times n}$ *are equivalent*:

(a) $i_0(A) = 0$.

(b) $B \in \text{PSD}$, $BA + A^T B = 0 \Leftrightarrow B = 0$.

(c) $B \in \text{PSD}$, $BA + A^T B = 0$, rank $B \leqq 2 \Leftrightarrow B = 0$.

(d) $B \in B_0(A)$, $BA + A^T B = 0 \Leftrightarrow B = 0$.

(e) $B \in B_0(A)$, $BA + A^T B = 0$, rank $B \leqq 2 \Leftrightarrow B = 0$.

Proof. a $\Rightarrow$ b. Let U be a unitary matrix such that $A' = UAU^\star$ is upper triangular, and let $B \in \text{PSD}$ be such that $BA + A^T B = 0$. Let B' be the positive-semidefinite matrix $UBU^\star$. Then $B'A' + A'^\star B = 0$, and thus $\alpha = (B'A')_{11}$ has to be imaginary. Since A' is upper triangular, it follows that $\alpha = B'_{11}A'_{11}$, and since $B' \in \text{PSD}$ and $i_0(A) = 0$ it follows that $B'_{11} = 0$. Therefore the first row and column of B' are zero. Applying the same argument now to the second, third, and so on, rows and columns of B', we obtain $B' = 0$, and thus, $B = 0$.

(c) $\Rightarrow$ (a) follows directly from Theorem 4.1.

The implications (b) $\Leftrightarrow$ (d), (b) $\Rightarrow$ (c), and (c) $\Leftrightarrow$ (e) are obvious. □

For matrices which have no pure imaginary eigenvalues we have Theorem 4.7.

THEOREM 4.7. *Suppose* $i_0(A) = 0$. *Then*

(a) in A = in AD *for every positive diagonal matrix* D *if and only if*

(b) $i_0(AD) = 0$ *for every positive diagonal matrix* D.

Proof. The proof of (a) $\Rightarrow$ (b) is trivial.

(b) $\Rightarrow$ (a) Suppose there exists a positive diagonal matrix D such that in $A \neq$ in AD. Let

$$D_t = (1-t)I + tD, \qquad A_t = AD_t, \qquad 0 \leqq t \leqq 1.$$

Then, $D_0 = I$; $D_1 = D$, $A_0 = A$, and $A_1 = AD$. By continuity considerations $i_0(A_t) \neq 0$ for some t, which contradicts (b). □

In Theorem 5.3 of [14] it was proved that a real square matrix A is D-stable if and only if for every $B \in B_{-0}(A)$, and every positive diagonal matrix D,

$$-(BAD + DA^T B) \in \text{PSD} \Leftrightarrow B = 0.$$

Given that a matrix A is stable, we obtain here a simpler characterization of D-stability as a corollary of Theorems 4.6 and 4.7.

COROLLARY 4.8. *Let* A *be a real stable matrix. Then* A *is* D*-stable if and only if for every* $B \in B_0(A)$, *of rank less than or equal to two and for every positive diagonal matrix* D,

$$BAD + DA^T B = 0 \Leftrightarrow B = 0.$$

Example 4.9. The matrix

$$A = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & 0 \\ 0 & -1 & 4 \end{bmatrix}$$

is stable. To show that it is D-stable we observe that $B \in B_0(A)$ if and only if it is of the form

$$B = \begin{bmatrix} a & 0 & -4c \\ 0 & b & 0 \\ -4c & 0 & c \end{bmatrix}, \qquad a, b, c \geqq 0, \qquad a \geqq 16c.$$

In this case

$$BA = \begin{bmatrix} 0 & (4c-a) & (a-16c) \\ b & 0 & 0 \\ 0 & 3c & 0 \end{bmatrix}$$

and if BAD is skew symmetric for an invertible matrix D, then necessarily, c, a, and b are equal to zero, and so $BA = 0$, which implies that $B = 0$.

A possible extension of Corollary 4.8 could have been: If A is stable and $i_0(AD) = 0$ for every invertible diagonal matrix D, then, A is inertia preserving. This is unfortunately not true. Let $D = \text{diag}\{-1, -1, 1\}$ and let A be the matrix of Example 4.9. Then AD is stable so A is not Arrow–McManus D-stable and therefore not inertia preserving.

5. Diagonally semistable matrices. Since inertia-preserving matrices are D-stable and include the diagonally stable matrices, it is natural to ask whether D-stable diagonally semistable matrices are inertia preserving.

Example 5.1 answers this question in the negative. However, we shall prove in the next section that the two classes coincide for irreducible acyclic matrices.

Example 5.1. The matrix

$$A = \begin{bmatrix} 2 & 1 & -2 \\ 3 & 2 & 0 \\ 6 & 4 & 2 \end{bmatrix}$$

of Example 3.2 is stable and diagonally semistable since $A + A^T$ is positive semidefinite. The cone $B_0(A)$ consists of the matrices B of the form

$$B = \alpha \begin{bmatrix} 3 & -4 & 1 \\ -4 & 6 & -2 \\ 1 & -2 & 1 \end{bmatrix}, \qquad \alpha \geqq 0;$$

so,

$$BA = \alpha \begin{bmatrix} 0 & -1 & -4 \\ -2 & 0 & 4 \\ 2 & 1 & 0 \end{bmatrix}, \qquad \alpha \geqq 0.$$

By Corollary 4.8 A is D-stable since for $B \neq 0$ and for every positive diagonal matrix D, $BAD + DA^TB \neq 0$. A is not inertia preserving since, for $F = \text{diag}\{-2, 4, -1\}$, $BAF + FA^TB = 0$, so by Theorem 4.6 $i_0(AF) \neq 0$.

The following property of diagonally semistable matrices is of great importance.

THEOREM 5.2. *Let $A \in R^{n\times n}$ be a diagonally semistable matrix and let F be an invertible diagonal matrix. The following are equivalent*:

(a) $\text{In } AF = \text{in } F$.

(b) $i_0(AF) = 0$.

(c) $BAF + FA^TB \neq 0$ *for every nonzero* $B \in B_0(A)$ *subject to* $\text{rank } B \leqq 2$.

(d) $BAF + FA^TB \neq 0$ *for every nonzero* $B \in B_0(A)$.

Proof. Conditions (b)–(d) are equivalent by Theorem 4.6.

(a) $\Rightarrow$ (b) since F is invertible.

(b) $\Rightarrow$ (a). Let D be a scaling factor of A. Then,

$$AD + DA^T = (AF)F^{-1}D + DF^{-1}(FA^T) \in \text{PSD}.$$

By Lemma 2.2, in AF = in $F^{-1}D$, and since D is a positive-definite diagonal matrix, (a) follows. □

Observe that the implication (b), (c), (d) $\Rightarrow$ (a), in Theorem 5.2 does not hold in general as shown by Example 3.1:

The matrix A of Example 3.1 is a D-stable matrix which is not Lyapunov diagonally semistable (see the remark following Corollary 5.11 of [14]). In that example, $i_0(AD) = 0$ since AD is stable, but in $AD \neq$ in D, as $D - \text{diag}\{-1, 3, -1\}$ is not stable.

Example 5.3. For the matrix A of Example 5.1, in AS = in S for every signature matrix S. This follows from the fact that for every nonzero $B \in B_0(A)$, BAS is not skew symmetric.

COROLLARY 5.4. *Let A be a diagonally semistable matrix. The following are equivalent*:

(a) *A is inertia preserving.*

(b) *$i_0(AF) = 0$ for every real invertible diagonal matrix F.*

(c) *$BAF + FA^TB \neq 0$ for every nonzero $B \in B_0(A)$ such that* rank $B \leqq 2$, *and for every real diagonal invertible matrix F.*

(d) *$BAF + FA^TB \neq 0$ for every nonzero $B \in B_0(A)$, and for every real diagonal invertible matrix F.*

Example 5.5. Let

$$A_d = \begin{bmatrix} 1 & 2 & 0 & 2 \\ 0 & 1 & 2 & 0 \\ 2 & 0 & 1 & 2 \\ 0 & 2 & 0 & d \end{bmatrix}, \qquad d > 1.$$

The cone $B_0(A)$ consists of matrices B of the form

$$B = \alpha \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \alpha \geqq 0.$$

A simple computation shows that A is inertia preserving. A is not strongly inertia preserving, since its principal submatrix

$$\begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 2 \\ 2 & 0 & 1 \end{bmatrix}$$

is not stable.

Another corollary of Theorem 5.2 is that in the case of diagonally semistable matrices there is no difference between D-stability and Arrow–McManus D-stability.

COROLLARY 5.6. *Let A be a diagonally semistable matrix. Then A is Arrow–McManus D-stable if and only if A is D-stable.*

Proof. Obviously, if A is Arrow–McManus D-stable, it is D-stable. Suppose that AF is stable, where F is a real invertible diagonal matrix. We want to show that F is positive. Since AF is stable, $i_0(AF) = 0$, so by Theorem 5.2 in AF = in F and F is positive. □

Observe that Corollary 5.6 justifies the claim we made in Example 3.2 (Example 5.1).

The following question is analogous to Question 3.5.

QUESTION 5.7. Is every irreducible inertia-preserving matrix diagonally semistable?

We conclude the section by remarking that a matrix is D-stable if and only if its irreducible components are D-stable and is inertia preserving if and only if its irreducible components are inertia preserving. Thus, the results which were proved for diagonally semistable matrices hold for matrices whose irreducible components are diagonally semistable.

6. Acyclic matrices. Example 5.1 is a diagonally semistable matrix which is D-stable but not inertia preserving. In this section it will be proved that irreducible, acyclic D-stable matrices are inertia preserving. This is done by using results and methods of [4].

Let A be an $n \times n$ matrix. Then A is *combinatorially symmetric* if

$$a_{ij} \neq 0 \Leftrightarrow a_{ji} \neq 0.$$

Observe that if A is combinatorially symmetric and F is an invertible diagonal matrix, then AF is also combinatorially symmetric. A is *symmetric in modulus* if $|a_{ij}| = |a_{ji}|$ for every i and j. A is *symmetric in signs* if $a_{ij}a_{ji} \geqq 0$ for every i and j.

The following proposition is well known.

PROPOSITION 6.1 [3], [12], [14]. *Let A be an acyclic irreducible matrix. Then A is diagonally semistable if and only if $A \in P_0$. In this case, there exists a positive diagonal matrix D such that AD is symmetric in modulus.*

The characterization of irreducible diagonally semistable acyclic matrices can be found in Theorem 3 of [12]. The existence of the matrix D is asserted in [3] and stated explicitly as Lemma 3.1 in [14].

For the sake of convenience, we now state some of the definitions and notation of [4].

For a nondirected graph G, $\Omega(G)$ denotes the collection of the following sets:

$$\Omega(G) = \{ w \subseteq V(G); \forall i \in V(G) | \{ i \} UN(i) \setminus w | \neq 1 \}.$$

The above sets w are called Ω-*sets of* G.

For a matrix A, the set of edges $E(G(A))$ can be partitioned into two sets:

$$H(A) = \{ (i,j) \in E(G(A)) : a_{ij}a_{ji} \geqq 0 \}.$$

$$S(A) = \{ (i,j) \in E(G(A)) : a_{ij}a_{ji} < 0 \}.$$

Vertices k and l are H-connected [S-connected] if there is a path of edges in $H(A)$ [$S(A)$] leading from k to l. Denote by $(i)_H$[$(i)_S$] the set of vertices which includes i and all vertices which are H-connected [S-connected] to i. Also, let $G_H(A)$ [$G_S(A)$] be the graph obtained from $G(A)$ by deleting the edges in $S(A)$ [$H(A)$].

An Ω-*set of* A is an Ω-set $w \in \Omega(G_S(A))$ such that

$$i \in w, \qquad (i,j) \in H(A) \Rightarrow j \in w.$$

The class of all Ω-sets of A is denoted by $\Omega(A)$.

Observe that if i belongs to an Ω-set of A, then $(i)_H$ is contained in that set. Let α be a set of vertices of $G(A)$. The *closure of* α, cl α, is defined as the smallest Ω-set of A which contains α.

In the proof of the main theorem of this section, we use the following results from [4], which we state here.

PROPOSITION 6.2 [4]. *Let $H \in P_0$ be an $n \times n$ irreducible, acyclic, symmetric-in-signs matrix. Then all its principal minors of order less than n are positive.*

THEOREM 6.3 [4]. *Let $A \in P_0^+$ be an $n \times n$ irreducible acyclic matrix. Denote*

$$P(A) = \{ i: \det A[(i)_H] > 0 \}.$$

Then A is D-stable if and only if cl $P(A) = \{1, \cdots, n\}$.

We also need the following definition and propositions.

For an $n \times n$ matrix C denote $W(C) = \{ i; C_i = 0 \text{ and } C^i = 0 \}$.

PROPOSITION 6.4. *Let S and B be $n \times n$ combinatorially symmetric matrices such that $s_{ii} = 0$ for every i, and such that*

(6.5) *BS is combinatorially symmetric, and $W(B) = W(BS)$.*

Then

$$W(B) \in \Omega(G(S)).$$

Proof. Suppose that

$$\{i\} \cup N(i) \backslash W(B) = \{j\}$$

for some $i, j \in G(S)$. Then,

$$(BS)_{ki} = \sum_{r=1}^{n} b_{kr} s_{ri} = b_{kj} s_{ji}, \qquad k = 1, 2, \cdots, n. \tag{6.6}$$

If $i \neq j$, then $i \in W(B)$ so

$$(BS)_{ki} = 0, \qquad k = 1, 2, \cdots, n. \tag{6.7}$$

But, as $j \notin W(B)$, $j \in N(i)$, and since B and S are combinatorially symmetric, there exists k, such that $b_{kj} s_{ji} \neq 0$, which contradicts (6.6) and (6.7).

Suppose $i = j$, then $i \notin W(B)$, and by (6.5), $i \notin W(BS)$. Since BS is combinatorially symmetric, there exists k such that, $(BS)_{ki} \neq 0$, but, observing that $(BS)_{ki} = b_{ki} s_{ii}$ and that $s_{ii} = 0$, we again have a contradiction.

We proved that for every i, $|\{i\} \cup N(i) \backslash W(B)| \neq 1$, so $W(B) \in \Omega(G(S))$. □

PROPOSITION 6.8. *Let $A, F \in R^{n \times n}$, and suppose that F is an invertible diagonal matrix such that $i_0(AF)$ is nonzero. Let $x = c + id$ be an eigenvector of FA^T:*

$$FA^T x = \alpha i x, \tag{6.9}$$

where $\alpha \in R$, $c, d \in R^n$. Denote

$$B = cc^T + dd^T.$$

Then,

(a) *BA is combinatorially symmetric.*

If α is nonzero, then,

(b) *$r \in W(B)$ if and only if*

$$x_r = (c + id)_r = 0.$$

(c) $W(B) = W(BA)$.

Proof. (a) By Theorem 4.1, $BAF + FA^T B = 0$, so, BAF is skew symmetric and therefore BA is combinatorially symmetric.

(b) Observe that B is a nonzero positive-semidefinite matrix, so, $r \in W(B)$ if and only if $b_{rr} = 0$.

$b_{rr} = (c_r)^2 + (d_r)^2 \Rightarrow$

$$b_{rr} = 0 \Leftrightarrow x_r = (c + id)_r = 0 \Leftrightarrow c_r = d_r = 0. \tag{6.10}$$

(c) $W(B)CW(BA)$, since BA is combinatorially symmetric. Suppose there exists $r \in [W(BA) \backslash W(B)]$. In that case

$$(BA)^r = 0 \quad \text{and} \quad (BA)_r = 0.$$

$(BA)^r = 0 \Rightarrow$

$$(BA)_{kr} = [(cc^T + dd^T)A]_{kr} = 0; k = 1, \cdots, n,$$

or in other words

$$c_k(A^T c)_r + d_k(A^T d)_r = 0, \qquad k = 1, \cdots, n.$$

Denote

$$\gamma = (A^T c)_r, \qquad \delta = (A^T d)_r.$$

Then

$$\gamma c + \delta d = 0.$$

By Theorem 4.1, $\gamma = \delta = 0$, since c and d are linearly independent.

Observe that by (6.9),

$$\gamma + i\delta = (A^T c + iA^T d)_r = (F^{-1}\alpha i x)_r,$$

which, by (6.10), is not zero since $r \notin W(B)$. So, γ and δ are not simultaneously equal to zero. □

THEOREM 6.11. *Let A be an acyclic irreducible matrix. Then A is D-stable if and only if A is inertia preserving.*

Proof. We have to show that if A is D-stable, then A is inertia preserving. Let A be a D-stable irreducible acyclic matrix. By Proposition 6.1, A is Lyapunov diagonally semistable, and there exists a positive diagonal matrix D, such that AD is symmetric in modulus, and

$$H = (1/2)(AD + DA^T) \geqq 0.$$

Denote

$$S = (1/2)(AD - DA^T).$$

Observe that, since AD is symmetric in modulus, and D is a positive diagonal matrix,

$$\det A[(i)_H] > 0 \Leftrightarrow \det AD[(i)_H] > 0 \Leftrightarrow \det H[(i)_H] > 0,$$

and so,

$$P(A) = P(AD) = P(H). \tag{6.12}$$

Moreover, $G(S) = G_S(A)$.

We have to show that $i_0(AF) = 0$ for every invertible diagonal matrix F. Assume that there exists a real number k such that ki is an eigenvalue of AF. Let $x = c + id$ be an eigenvector of FA^T

$$FA^T x = kix,$$

where $c, d \in R^n$. Denote

$$B = cc^T + dd^T.$$

Clearly, $B \neq 0$. Since A is D-stable A is nonsingular, and hence $k \neq 0$. By Proposition 6.8, $W(B) = W(BA)$; BA is combinatorially symmetric. By Theorem 4.1, $B \in B_0(A)$, and by Theorem 3.9 of [17],

$$BH = 0. \tag{6.13}$$

Observe that by (6.13), $BA = B(H + S)D^{-1} = BSD^{-1}$. So, $W(B) = W(BA) = W(BSD^{-1})$, or $W(B) = W(BS)$, where B, S, BS are combinatorially symmetric, and $s_{ii} = 0$. By Proposition 6.4 $W(B) \in \Omega(G(S)) = \Omega(G_S(A))$.

By (6.12) and (6.13),

$$P(A) \subseteq W(B). \tag{6.14}$$

Let $i \in W(B)$, and let $(i, j) \in H(A)$. The submatrix $H[(i)_H]$ is a principal submatrix of AD and as such belongs to P_0. By Proposition 6.2, $\det H[(i)_H - \{i\}] > 0$, and since $i \in W(B)$, it follows from (6.13) that $(i)_H \subseteq W(B)$. Thus, $W(B) \in \Omega(A)$. Since A is D-stable it follows that $\operatorname{cl} P(A) = \{1, \cdots, n\}$, and it follows from (6.14) that $W(B) = \{1, \cdots, n\}$, so $B = 0$, which is a contradiction. Therefore, $i_0(AF) = 0$. □

COROLLARY 6.15. *If the irreducible components of A are acyclic, then A is inertia preserving if and only if it is D-stable.*

Acknowledgment. We thank Professor Daniel Hershkowitz for many suggestions which improved the paper, and in particular for Example 4.9 and the proof of Theorem 4.6, which is simpler than our original proof.

REFERENCES

[1] K. J. ARROW AND M. MCMANUS, *A note on dynamic stability*, Econometrica, 26 (1958), pp. 448–454.
[2] G. P. BARKER, A. BERMAN, AND R. J. PLEMMONS, *Positive diagonal solutions to the Lyapunov equations*, Linear Multilinear Algebra, 5 (1978), pp. 249–256.
[3] A. BERMAN AND D. HERSHKOWITZ, *Matrix diagonal stability and its implications*, SIAM J. Algebraic Discrete Methods, 4 (1983), pp. 377–382.
[4] ———, *Characterization of acyclic D-stable matrices*, Linear Algebra Appl., 58 (1984), pp. 17–31.
[5] D. CARLSON, B. N. DATTA, AND C. R. JOHNSON, *A semi-definite Lyapunov theorem and the characterization of tridiagonal D-stable matrices*, SIAM J. Algebraic Discrete Methods, 3 (1982), pp. 293–304.
[6] D. CARLSON AND H. SCHNEIDER, *Inertia theorems for matrices: The semidefinite case*, J. Math Anal. Appl., 6 (1963), pp. 430–446.
[7] B. L. CLARKE, *D-stability and chemical reaction networks*, a talk at the Combinatorial Matrix Analysis Conference, Victoria, British Columbia, Canada, 1987.
[8] G. W. CROSS, *Three types of matrix stability*, Linear Algebra Appl., 20 (1978), pp. 253–263.
[9] F. R. GANTMACHER, *The Theory of Matrices* I, II, Chelsea, New York, 1959.
[10] M. HALL, *Combinatorial Theory*, Blaisdell, Waltham, MA, 1967.
[11] D. J. HARTFIEL, *Concerning the interior of the D-stable matrices*, Linear Algebra Appl., 30 (1980), pp. 201–207.
[12] D. HERSHKOWITZ, *Stability of acyclic matrices*, Linear Algebra Appl., 73 (1986), pp. 157–169.
[13] ———, *Lyapunov diagonal semistability of acyclic matrices*, Linear Multilinear Algebra, 22 (1988), pp. 267–283.
[14] D. HERSHKOWITZ AND D. SHASHA, *Cones of real positive semidefinite matrices associated with matrix stability*, Linear Multilinear Algebra, 23 (1988), pp. 165–181.
[15] A. OSTROWSKI AND H. SCHNEIDER, *Some theorems on the inertia of general matrices*, J. Math. Anal. Appl., 4 (1962), pp. 72–84.
[16] O. TAUSKY, *A generalization of a theorem by Lyapunov*, J. Soc. Indust. Appl. Math., 9 (1961), pp. 640–643.
[17] D. SHASHA AND A. BERMAN, *On the uniqueness of the Lyapunov scaling factors*, Linear Algebra Appl., 91 (1987), pp. 53–63.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 2, pp. 220–238, April 1991

004

TRIDIAGONAL APPROACH TO THE ALGEBRAIC ENVIRONMENT OF TOEPLITZ MATRICES, PART I: BASIC RESULTS*

P. DELSARTE† AND Y. GENIN†

Abstract. This paper contains a thorough investigation of a family of symmetric "predictor polynomials" associated with a nonnegative-definite Toeplitz matrix. These polynomials are constructed from the classical predictors and from the values assumed by some dual predictors in a fixed point of unit modulus; the appropriate duality is induced by changing the sequence of reflection coefficients into its conjugate mirror image, within a unit modulus factor. The central theme of the paper is a well-defined three-term recurrence relation satisfied by these symmetric polynomials; it motivates the "tridiagonal" terminology. The properties of the recurrence are studied in detail; special attention is paid to the important issue of computing the recurrence coefficients from the reflection coefficients. It is shown how this three-term recurrence formula produces an efficient solution method, called the split Levinson algorithm, for the linear prediction problem.

Key words. nonnegative-definite Toeplitz matrices, three-term recurrence, symmetric predictor polynomials, split Levinson algorithm

AMS(MOS) subject classifications. 65F05, 42C05, 60G10

1. Introduction.

Nonnegative-definite Toeplitz matrices play a prominent role in various areas of applied mathematics. From a theoretical viewpoint, they can be defined as covariance matrices of stationary stochastic processes [17]. In digital signal processing applications and the like, they are generally obtained, within a constant diagonal shift, as autocorrelation matrices of sampled signal records.

The simplest and most central problem relative to a positive-definite Toeplitz matrix is the *linear prediction problem* [20]; it amounts to computing the first column of the inverse of that matrix. (There is a natural generalization to the nonnegative definite case.) This problem is classically solved by means of the *Levinson algorithm* [16], which is based on a recurrence relation first discovered in the framework of Szegö's theory of *polynomials orthogonal on the unit circle* [17], [23]. The same recurrence relation, used in reverse order, underlies the *Schur–Cohn polynomial stability test* [21]. It is worth mentioning that the algebraic results alluded to above have far-reaching applications in the field of positive (Carathéodory) and of bounded (Schur) functions [1], which themselves are relevant to some important modelling problems in digital signal processing (among other areas). In particular, the *Schur algorithm*, which is a possible substitute for the Levinson algorithm to compute reflection coefficients [19], can be viewed as an implementation of Schur's recursion for bounded functions [1].

Within the last few years, a novel approach to the whole mathematical environment of nonnegative-definite Toeplitz matrices has been introduced and examined in some detail; it is called either the *split approach* [8], [10], [11] or, equivalently, the *tridiagonal approach* [7]. Roughly speaking, the first terminology refers to a "splitting" of the classical predictor into symmetric and antisymmetric parts, while the second one refers to the tridiagonal matrix representation of the recursive structure of the new theory. In fact, the basic idea of such an approach is to replace the one-step recurrence relation underlying the Levinson algorithm, which involves predictor polynomials and their reciprocals, by an appropriate two-step recurrence relation, of the Frobenius type [15], which involves "symmetric predictor polynomials." This approach provides new efficient numerical

* Received by the editors March 24, 1989; accepted for publication (in revised form) December 13, 1989.
† Philips Research Laboratory, Av. Albert Einstein 4, B-1348 Louvain-la-Neuve, Belgium (phd@prlb.philips.be and yg@prlb.philips.be).

methods to solve the standard linear prediction problem and various related problems. In addition, it is of noticeable interest from a theoretical viewpoint due to its connections with the theory of orthogonal polynomials on the real line and with the theory of positive and bounded functions. Among the contributions in the field, papers by Bube and Burridge [4], Bistritz [2], and Delsarte and Genin [5] deserve a special mention. A comprehensive survey of recent results, with an extensive bibliography, is given in [8].

The present paper and its companions [7], [12] aim at providing a thorough investigation of a natural and general setting of the tridiagonal approach to the Toeplitz environment. While [7] is mainly devoted to function theoretic aspects, this paper and [12] are almost exclusively concerned with the algebraic components of the subject. More precisely, they deal with nonnegative-definite Hermitian Toeplitz matrices of nullity one or, equivalently, with sequences of reflection coefficients having modulus smaller than unity, except the last of them, which has modulus equal to unity. As for the distinction between both "algebraic parts," it can be explained roughly as follows. This paper is concerned with rational problems only, while its companion (and successor) [12] is concerned with algebraic problems involving zeros (of symmetric predictor polynomials) and eigenvalues (of unitary Hessenberg matrices).

Most results given in this paper are stated in the form of *polynomial identities*. However, it should be emphasized that our study is essentially concerned with *matrix theory problems*. In that respect, let us recall that the classical inversion methods for Toeplitz matrices, such as the celebrated Trench formula, can be derived and interpreted in a transparent manner in a polynomial framework (see especially [18]). A similar observation can be made regarding the results of the present contribution, in the sense that they provide new efficient algorithms to compute the ingredients involved in the classical Toeplitz inversion formulas (for example). It is worth mentioning here that the symmetric predictor polynomials provide a reduction of a Toeplitz matrix to a *tridiagonal form*, in contrast to the usual predictors, which yield a *diagonal form* reduction. Furthermore, it should be stressed that the material of this paper serves, for a good deal, as a preparation for the companion paper [12], which pertains to matrix theory in a more obvious manner.

In § 2 we introduce the ρ_n-*symmetric predictor polynomials* $b_k(z)$, for $0 \leqq k \leqq n$, relative to a sequence $(\rho_j)_{j=1}^{n}$ of complex reflection coefficients ρ_j, with $|\rho_n| = 1$ and $|\rho_j| < 1$ for $j = 1, \cdots, n-1$, and relative to a given complex number ζ_0, referred to as the circle parameter, with $|\zeta_0| = 1$. In particular, $b_n(z)$ is equal to the last predictor polynomial generated by the reflection coefficients ρ_j (in the classical sense).

In § 3 we derive a simple one-parameter three-term recurrence formula for normalized versions, $p_k(z) = g_k b_k(z)$, of the ρ_n-symmetric polynomials $b_k(z)$. These polynomials $p_k(z)$ are *symmetric*; they are the main objects considered in the sequel. We explain how the coefficients of the recurrence can be computed either from the reflection coefficients, or from the entries of the associated Toeplitz matrix; the latter method is a general version of the *split Levinson algorithm*.

In § 4 we investigate the *singular case* of the theory, characterized by the fact that the circle parameter ζ_0 is a zero of $p_n(z)$. In this case, $p_k(\zeta_0)$ vanishes for all k, and we show that suitably normalized versions of the polynomials $p_k(z)/(\zeta_0 - z)$ belong exactly to our theory of "symmetric predictors" (with respect to a certain Toeplitz matrix which can be constructed explicitly from the original one).

In § 5 we examine the *duality* induced by changing the reflection coefficient ρ_j into $\rho_n \bar{\rho}_{n-j}$, for $j = 1, \cdots, n$. This preserves the last classical predictor polynomial (of degree n). In this context we explain how the regular case (i.e., $p_n(\zeta_0) \neq 0$) can be extended to the singular case, in some sense, and we show how the split Levinson algorithm can be extended so as to produce the ρ_n-symmetric predictor $b_n(z)$.

2. Symmetrization of predictor polynomials. For a positive integer n, let there be given a sequence of n complex numbers $\rho_1, \cdots, \rho_n$ subject to the constraints

$$|\rho_k| < 1 \quad \text{for } k = 1, \cdots, n-1, \qquad |\rho_n| = 1. \tag{2.1}$$

From this sequence we construct the family of polynomials $a_k(z)$ of formal degree k, for $k = 0, \cdots, n$, via the *Szegö–Levinson recurrence formula*

$$a_k(z) = a_{k-1}(z) + \rho_k z \hat{a}_{k-1}(z), \tag{2.2}$$

with the initialization $a_0(z) = 1$. Here and in the sequel, the notation $\hat{v}_k(z)$ stands for the *reciprocal* (conjugate mirror image) of a complex polynomial $v_k(z)$ of formal degree k; it is defined by $\hat{v}_k(z) = z^k \bar{v}_k(1/\bar{z})$. Note that the polynomial $a_k(z)$ is *comonic*, in the sense that it satisfies $a_k(0) = 1$. Equivalently, its reciprocal $\hat{a}_k(z)$ is monic (the leading coefficient equals unity).

In view of the property $|\rho_n| = 1$, it follows from (2.2) that $a_n(z)$ is ρ_n-*symmetric*, in the sense that it satisfies

$$\hat{a}_n(z) = \bar{\rho}_n a_n(z). \tag{2.3}$$

The polynomials $a_k(z)$ with $0 < k < n$ are quite different from $a_n(z)$ in that respect. Note that the coefficient of z^k in $a_k(z)$ is equal to ρ_k. The monic polynomials $\hat{a}_k(z)$ are often called *Szegö polynomials*; this refers to the fact that they are pairwise orthogonal on the unit circle with respect to a certain positive measure (details are given in [12]).

Given any positive real number c_0, let us denote by C_k the *Hermitian Toeplitz matrix* of order $k+1$ (with $0 \leqq k \leqq n$) having c_0 as its diagonal element and admitting $(\rho_1, \cdots, \rho_k)$ as the sequence of its *Schur–Szegö parameters* (or *reflection coefficients*). As explained below, this well-known relationship between Toeplitz matrices and reflection coefficient sequences can be made explicit by use of some "second-kind predictors." By definition, C_k has the form

$$C_k = \begin{bmatrix} c_0 & c_{-1} & \cdots & c_{-k} \\ c_1 & c_0 & \cdots & c_{1-k} \\ \vdots & \vdots & & \vdots \\ c_k & c_{k-1} & \cdots & c_0 \end{bmatrix}, \tag{2.4}$$

i.e., $C_k = [c_{i-j} : 0 \leqq i, j \leqq k]$, with the Hermitian property $\bar{c}_i = c_{-i}$ for all i. As a consequence of the assumption (2.1), the Toeplitz matrix C_k is *positive definite* for $0 \leqq k \leqq n-1$, whereas C_n is *nonnegative definite and singular* (of nullity one). Let us stress that C_{k-1} is obtained from C_k by deleting its first (or last) row and column.

In the context of linear prediction for stationary stochastic processes, $a_k(z)$ is usually referred to as the (first-kind) *predictor polynomial* relative to C_k. This means that the coefficient vector $\mathbf{a}_k = [a_{k,0}, \cdots, a_{k,k}]^T$ of $a_k(z) = \sum_{i=0}^{k} a_{k,i} z^i$ satisfies the system of linear equations

$$C_k \mathbf{a}_k = [\sigma_k, 0, \cdots, 0]^T, \tag{2.5}$$

where σ_k is a nonnegative real number, called prediction error squared norm, which is uniquely determined from the data $c_0, \rho_1, \cdots, \rho_k$. In fact, we have

$$\sigma_0 = c_0 \quad \text{and} \quad \sigma_k = \sigma_{k-1}(1 - |\rho_k|^2), \tag{2.6}$$

for $k = 1, \cdots, n$. Thus, σ_k is positive for $0 \leqq k \leqq n-1$ while σ_n equals zero. Note the property $\sigma_k = \det C_k / \det C_{k-1}$. In particular, it follows from (2.5) that $a_n(z)$ is the unique comonic polynomial satisfying the singular homogeneous system

$$C_n \mathbf{a}_n = \mathbf{0}. \tag{2.7}$$

Let us now explain how the entries $c_1, \cdots, c_n$ of C_n can actually be computed from c_0 and $\rho_1, \cdots, \rho_n$. To that end, we introduce the *second-kind predictor polynomials* $r_k(z)$, for $0 \leqq k \leqq n$, by means of the recurrence formula

$$r_k(z) = r_{k-1}(z) - \rho_k z \hat{r}_{k-1}(z), \tag{2.8}$$

which is the same as (2.2) except that the reflection coefficient ρ_k is replaced by $-\rho_k$. The initial value is $r_0(z) = c_0$; this implies $r_k(0) = c_0$ for all k. Define the rational function

$$f(z) = r_n(z)/a_n(z). \tag{2.9}$$

It is a Carathéodory function [1], which means that it is analytic and that its real part is nonnegative in the unit disc $|z| < 1$. More precisely, $f(z)$ is a *lossless* function, i.e., a Carathéodory function having imaginary values almost everywhere on the unit circle (see details in [12]). Note that $f(z)$ has degree n exactly.

Now consider the Maclaurin expansion

$$f(z) = c_0 + 2 \sum_{k=1}^{\infty} c_k z^k. \tag{2.10}$$

It turns out that the numbers $c_0, c_1, \cdots, c_n$ thus defined are precisely the entries of the Toeplitz matrix C_n mentioned above. (Furthermore, the coefficients c_k with $n + 1 \leqq k \leqq m$ in (2.10) yield the unique nonnegative-definite Toeplitz extension C_m of C_n, for every $m > n$.) Recall the property

$$r_k(z) = f(z) a_k(z) + O(z^{k+1}), \tag{2.11}$$

for $0 \leqq k \leqq n$. This can be used as an explicit definition of the second-kind predictor $r_k(z)$ in terms of the first-kind predictor $a_k(z)$ and the Toeplitz matrix C_k.

After this standard material, let us introduce some ingredients that are somewhat less classical (see [10], [13]). The basic idea is to consider the *dual Schur–Szegö sequence* $(\rho_k^{\#})_{k=1}^n$; it is defined from the original sequence $(\rho_k)_{k=1}^n$ by

$$\rho_k^{\#} = \rho_n \bar{\rho}_{n-k} \quad \text{for } k = 1, \cdots, n, \tag{2.12}$$

with the natural convention $\rho_0 = 1$, yielding $\rho_n^{\#} = \rho_n$. Then we define the corresponding family of comonic polynomials $s_k(z) = a_k^{\#}(z)$, *dual* of the predictors $a_k(z)$, by means of the recurrence formula

$$s_k(z) = s_{k-1}(z) + \rho_k^{\#} z \hat{s}_{k-1}(z). \tag{2.13}$$

Combining the Szegö–Levinson recurrences (2.2) and (2.13) we obtain a result that plays an important role in the theory.

PROPOSITION 1. *The comonic polynomials $a_i(z)$ and $s_j(\zeta)$, in the independent variables z and ζ, satisfy the duality relation*

$$s_{n-k}(\zeta) a_{k-1}(z) + \rho_n \hat{s}_{n-k}(\zeta) z \hat{a}_{k-1}(z) = s_{n-k-1}(\zeta) a_k(z) + \rho_n \zeta \hat{s}_{n-k-1}(\zeta) \hat{a}_k(z). \tag{2.14}$$

If $\zeta = z$, then both sides of (2.14) are necessarily independent of k. Hence, setting $k = n - 1$ in the right-hand side and using (2.2), we obtain the identity

$$s_{n-k}(z) a_{k-1}(z) + \rho_n z \hat{s}_{n-k}(z) \hat{a}_{k-1}(z) = a_n(z), \tag{2.15}$$

for $k = 1, \cdots, n$. In the case $k = 1$, this yields $s_n(z) = a_n(z)$. It is also interesting to mention the two-variable expansion

$$s_{n-k}(\zeta) a_{k-1}(z) + \rho_n \hat{s}_{n-k}(\zeta) z \hat{a}_{k-1}(z) = a_n(\zeta) + \rho_n (z - \zeta) \sum_{j=0}^{k-1} \hat{s}_{n-j-1}(\zeta) \hat{a}_j(z). \tag{2.16}$$

This follows from (2.14) by summation.

It is easily checked that formulas (2.14)–(2.16) remain valid when the second-kind predictors $r_k(z)$ are substituted for the first-kind predictors $a_k(z)$, provided ρ_n is replaced by $-\rho_n$. Thus it is seen from (2.15) that the lossless function (2.9) can be represented in the form

$$f(z)=[r_{k-1}(z)-z\psi_k(z)\hat{r}_{k-1}(z)]/[a_{k-1}(z)+z\psi_k(z)\hat{a}_{k-1}(z)], \tag{2.17}$$

for $k = 1, \cdots, n$, where $\psi_k(z)$ is a rational *Schur function*, of the *inner* type, of degree $n - k$, given by

$$\psi_k(z)=\rho_n\hat{s}_{n-k}(z)/s_{n-k}(z). \tag{2.18}$$

Without going into detail, let us point out that representations such as (2.17) occur classically in the framework of the Carathéodory–Fejér interpolation problem [14], [24]. From (2.13) we deduce the *Schur-type recurrence relation* [1]

$$\psi_k(z)=[\rho_k+z\psi_{k+1}(z)]/[1+\bar{\rho}_k z\psi_{k+1}(z)]. \tag{2.19}$$

We now introduce the concept of "symmetric Szegö polynomials," which is the main theme of this paper. Let ζ_0 be *any complex number of unit modulus*; it will be referred to in the sequel as the *circle parameter*. For $k = 1, \cdots, n$, define the polynomial

$$b_k(z)=\zeta_0^{(k-n)/2}[s_{n-k}(\zeta_0)a_{k-1}(z)+\rho_n\hat{s}_{n-k}(\zeta_0)z\hat{a}_{k-1}(z)], \tag{2.20}$$

in terms of the predictors and their duals. Throughout this paper, $\zeta_0^{1/2}$ denotes either of the square roots of ζ_0, and $\zeta_0^{m/2}$ stands for the mth power of $\zeta_0^{1/2}$. (Note that replacing $\zeta_0^{1/2}$ by $-\zeta_0^{1/2}$ amounts to multiplying $b_k(z)$ by $(-1)^{n-k}$.) It is clear that $b_k(z)$ has degree k exactly. (Indeed, we have $s_{n-k}(\zeta_0)\neq 0$, for $k \geqq 1$, in view of the Schur–Cohn criterion.) Using (2.14) we obtain an alternative expression for $b_k(z)$, involving $a_k(z)$ instead of $a_{k-1}(z)$. This leads us to define the constant $b_0(z)$ to be

$$b_0(z)=\zeta_0^{-n/2}s_n(\zeta_0)=\zeta_0^{-n/2}a_n(\zeta_0). \tag{2.21}$$

It is seen from (2.20) and (2.21) that the polynomials $b_k(z)$ are ρ_n-symmetric: they enjoy the same property $\hat{b}_k(z) = \bar{\rho}_n b_k(z)$, for every degree k, as the classical predictor $a_n(z)$ of degree n. By lack of a better terminology we shall say that the polynomials $b_k(z)$ constitute a family of (nonnormalized) *ρ_n-symmetric predictor polynomials*, for $k = 0, 1, \cdots, n$. Recall that this family is defined from the Schur–Szegö parameters $\rho_1, \cdots, \rho_n$ and the (square root of) the circle parameter ζ_0. It can be viewed as an *embedding* of the predictor $a_n(z)$; indeed, as a direct consequence of (2.20) and (2.2), we have

$$b_n(z)=a_n(z)=s_n(z). \tag{2.22}$$

Let us write down the value of $b_k(z)$ for the two distinguished points, $z = 0$ and $z = \zeta_0$; in view of (2.20) and (2.15) we have

$$b_k(0)=\zeta_0^{(k-n)/2}s_{n-k}(\zeta_0), \tag{2.23}$$

$$b_k(\zeta_0)=\zeta_0^{(k-n)/2}a_n(\zeta_0). \tag{2.24}$$

In certain parts of the theory it is interesting (or necessary) to separate the *singular case* $a_n(\zeta_0) = 0$ from the *regular case* $a_n(\zeta_0)\neq 0$ (see [10]). It follows from (2.15) that $b_k(\zeta_0)$ equals zero for all k (including $k = 0$) in the singular case, and differs from zero for all k in the regular case. A detailed discussion of the singular case is given in § 4.

The classical predictor $a_k(z)$ can be recovered from the ρ_n-symmetric predictors $b_k(z)$ and $b_{k+1}(z)$ in an easy manner (for $0 \leqq k \leqq n-1$). In fact, simple algebraic manipulations using (2.14), (2.20), and (2.23) yield the formula

$$b_{k+1}(0)(1-\zeta_0^{-1}z)a_k(z)=b_{k+1}(z)-\zeta_0^{-1/2}zb_k(z). \tag{2.25}$$

As a consequence, the Schur–Szegö parameters $\rho_k = a_{k,k}$ with $1 \leqq k \leqq n-1$ can be computed from the numbers $b_k(0)$, $b_{k+1}(0)$, and $\rho_n\,(=b_{n,n})$. Indeed, equating the leading coefficients in both sides of (2.25) and using ρ_n-symmetry, we obtain

$$b_{k+1}(0)\rho_k = \zeta_0^{1/2}\rho_n[\bar{b}_k(0) - \zeta_0^{1/2}\bar{b}_{k+1}(0)]. \tag{2.26}$$

Of course it is tacitly assumed here (as everywhere in the paper) that the circle parameter ζ_0 is *given*.

Note that both families of polynomials $a_k(z)$ and $b_k(z)$ are determined uniquely from the n complex numbers $s_1(\zeta_0), \cdots, s_n(\zeta_0)$. In view of the results above, to prove this property, we have only to check that ρ_n can be computed from these data. Appropriate expressions, resulting from (2.13), are $\rho_n = s_n(\zeta_0)/\hat{s}_n(\zeta_0)$ in the regular case ($s_n(\zeta_0) \neq 0$) and $\rho_n = -s_{n-1}(\zeta_0)/\zeta_0\hat{s}_{n-1}(\zeta_0)$ in the singular case.

For future use let us introduce the *pseudoreflection coefficient* ω_k involved in formula (2.20); it is defined by

$$\omega_k = \psi_k(\zeta_0) = \rho_n\hat{s}_{n-k}(\zeta_0)/s_{n-k}(\zeta_0), \tag{2.27}$$

for $k = 1, \cdots, n$, where $\psi_k(z)$ is the rational inner function (2.18). Note that ω_k has unit modulus. In particular, we have $\omega_n = \rho_n$. From (2.23) it follows that ω_k equals $\rho_n\bar{b}_k(0)/b_k(0)$. As a consequence of this identity, together with relations (2.19), (2.20), and (2.27), we obtain the following result.

PROPOSITION 2. *The ρ_n-symmetric predictor $b_k(z)$ can be written in the form*

$$b_k(z) = b_k(0)[a_{k-1}(z) + \omega_k z\hat{a}_{k-1}(z)]. \tag{2.28}$$

The sequence of pseudoreflection coefficients $(\omega_n, \cdots, \omega_1)$ can be computed from the sequence $(\rho_n, \cdots, \rho_1)$ by means of the formula

$$\omega_k = (\rho_k + \zeta_0\omega_{k+1})/(1 + \zeta_0\omega_{k+1}\bar{\rho}_k), \tag{2.29}$$

for $k = n-1, \cdots, 1$, with the initial value $\omega_n = \rho_n$.

The name "pseudoreflection coefficient" is suggested by a comparison between the roles of ω_k and ρ_k in the relations (2.28) and (2.2), respectively.

It is interesting to associate a family of *second-kind polynomials*, denoted here by $t_k(z)$, with the "first-kind polynomials" $b_k(z)$. The definition is formally the same as above (within a factor c_0) except that the Schur–Szegö parameters ρ_k are replaced by $-\rho_k$. More precisely, $t_k(z)$ can be defined by

$$t_k(z) = b_k(0)[r_{k-1}(z) - \omega_k z\hat{r}_{k-1}(z)], \tag{2.30}$$

for $k = 1, \cdots, n$, where $r_k(z)$ is the second-kind predictor associated with $a_k(z)$. For $k = 0$, the natural definition is $t_0(z) = \zeta_0^{-n/2}r_n(\zeta_0)$. The second-kind predictor $t_k(z)$ can be determined from its first-kind associate $b_k(z)$, for $1 \leqq k \leqq n$, by means of the relation

$$t_k(z) = f(z)b_k(z) + O(z^k), \tag{2.31}$$

supplemented with the identity $t_{k,k} = -\rho_n\bar{t}_{k,0} = -c_0 b_{k,k}$. Note that (2.31) is slightly weaker than its classical counterpart (2.11).

It is clear that $t_k(z)$ enjoys the ρ_n-*antisymmetry* property $\hat{t}_k(z) = -\bar{\rho}_n t_k(z)$, for $0 \leqq k \leqq n$. Next we observe that $\zeta_0^{-n/2}t_k(\zeta_0)$ is independent of k, as in (2.24). Furthermore, since ω_n equals ρ_n, we deduce from (2.30) and (2.8) the *embedding* property $t_n(z) = r_n(z)$. Therefore, the ratio $t_n(z)/b_n(z)$ equals the lossless function $f(z)$ in (2.9). More generally, it is easily seen that $t_k(z)/b_k(z)$ is a lossless rational function of degree k, for $0 \leqq k \leqq n$. Note that this function equals the right-hand side of (2.17), where $\psi_k(z)$ is replaced by the constant $\psi_k(\zeta_0)$.

Remark. It is interesting to note that the circle parameter ζ_0 can be normalized to the value $\zeta_0' = 1$ at the cost of the simple data transformation $\rho_k \to \rho_k' = \zeta_0^k \rho_k$ for $k = 1, \cdots, n$. It is easily seen that this transformation produces the polynomials $a_k'(z) = a_k(\zeta_0 z)$ and $s_k'(z) = s_k(\zeta_0 z)$ instead of $a_k(z)$ and $s_k(z)$, whence the ρ_n-symmetric predictor $b_k'(z) = \zeta_0^{(n-k)/2} b_k(\zeta_0 z)$ instead of $b_k(z)$. Similar observations could be made at many places in the paper.

3. Recurrence relations and Toeplitz systems. One of the most significant properties of the ρ_n-symmetric polynomials $b_k(z)$ lies in the fact that they are linked by a three-term recurrence relation (of the Frobenius type [15]). This result can be derived either from the Szegö–Levinson formula (2.2) or from the Toeplitz linear system (2.5). Let us now explain the first method. For $k = 0, 1, \cdots, n-1$, define the complex number

$$\beta_k = b_{k+1}(0)/b_k(0) = \zeta_0^{1/2} s_{n-k-1}(\zeta_0)/s_{n-k}(\zeta_0), \tag{3.1}$$

with the convention $\beta_0 = \infty$ in the singular case $s_n(\zeta_0) = 0$. Let us rewrite (2.2) in terms of the $b_k(z)$ polynomials, with the help of (2.25); owing to the ρ_n-symmetry property, we obtain the identity

$$\begin{aligned} b_{k+1}(z) = &\{\beta_k + b_k^{-1}(0)[\zeta_0^{-1/2}\bar{b}_k(0) - \zeta_0^{-1}\bar{\rho}_n\rho_k b_{k+1}(0)]z\} b_k(z) \\ &- |\beta_k|^2 \bar{b}_{k+1}^{-1}(0)[\bar{\rho}_n\rho_k b_k(0) - \bar{b}_k(0)] z b_{k-1}(z). \end{aligned} \tag{3.2}$$

Using (2.26) to simplify the expressions between square brackets in (3.2), we deduce the *three-term recurrence relation*

$$b_{k+1}(z) = (\beta_k + \bar{\beta}_k z) b_k(z) - (1 - |\rho_k|^2)|\beta_k|^2 z b_{k-1}(z). \tag{3.3}$$

Note that (3.3) remains valid for $k = 0$, in the regular case $a_n(\zeta_0) \neq 0$, with the convention $\rho_0 = 1$.

This allows us to compute the whole family of polynomials $b_k(z)$ from the sequence of recurrence coefficients $\beta_1, \cdots, \beta_{n-1}$ and the initial conditions $b_0(z)$ and $b_1(z)$. Indeed, the cofactor of $z b_{k-1}(z)$ in (3.3) can be expressed in terms of β_k as follows:

$$(1 - |\rho_k|^2)|\beta_k|^2 = \zeta_0^{-1/2}\beta_k + \zeta_0^{1/2}\bar{\beta}_k - 1. \tag{3.4}$$

This follows readily from (3.1) and (2.13). Alternatively, (3.4) can be viewed as a consequence of (2.26), written in the interesting form

$$\rho_k = \zeta_0 \omega_{k+1}(\zeta_0^{-1/2}\bar{\beta}_k^{-1} - 1). \tag{3.5}$$

Next, let us introduce a natural normalization in the theory; the idea is to get rid of the factor (3.4) in the recurrence (3.3). The *normalized symmetric predictor polynomial* of degree k, denoted by $p_k(z)$, is defined by

$$p_k(z) = g_k b_k(z), \tag{3.6}$$

for $k = 0, \cdots, n$, where the *normalizing factor* g_k is a nonzero complex number chosen in such a way that the three-term recurrence relation assumes the simple form

$$p_{k+1}(z) = (\alpha_k + \bar{\alpha}_k z) p_k(z) - z p_{k-1}(z). \tag{3.7}$$

To deal with the normalization problem it is useful to introduce the so-called *Jacobi parameters* $\lambda_1, \cdots, \lambda_n$ (see [7], [10]); they are given by

$$\lambda_k = g_k / g_{k-1}. \tag{3.8}$$

Comparing (3.3) and (3.7) shows that λ_k is *real*; this implies that $\bar{g}_k/g_k$ is independent of k. The *normalized recurrence coefficient* α_k in (3.7) is related to the coefficient β_k in (3.3) by

$$\alpha_k = \beta_k \lambda_{k+1} \quad \text{with } (1 - |\rho_k|^2)|\beta_k|^2 \lambda_k \lambda_{k+1} = 1. \tag{3.9}$$

This implies that all Jacobi parameters λ_k have the *same sign*. As a conclusion, it is seen that the normalization process involves two nonzero constants g_0 and g_1 subject only to the condition that their ratio be real.

It is easy to show that the Jacobi parameters can be expressed in the form

$$\lambda_k = d\sigma_{k-1} |p_k(0)|^2, \tag{3.10}$$

with the real constant $d = \lambda_1 c_0^{-1} |p_1(0)|^{-2}$. Indeed, since $p_{k+1}(0) = \alpha_k p_k(0)$ by (3.7), the identity (3.10) is equivalent to $\lambda_k^{-1}\lambda_{k+1} = (1 - |\rho_k|^2)|\alpha_k|^2$, which itself is a direct consequence of (3.9). Alternatively, the Jacobi parameters can be expressed in terms of the coefficients α_k by means of a simple *continued fraction*, i.e., by means of the recurrence relation

$$\lambda_{k+1} = \zeta_0^{-1/2}\alpha_k + \zeta_0^{1/2}\bar{\alpha}_k - \lambda_k^{-1}, \tag{3.11}$$

which results readily from (3.4) and (3.9). Furthermore, in the regular case, λ_k can be written explicitly as the ratio

$$\lambda_k = \zeta_0^{-1/2} p_k(\zeta_0)/p_{k-1}(\zeta_0). \tag{3.12}$$

Equivalently, g_k is proportional to $\zeta_0^{-k/2} p_k(\zeta_0)$. This follows from (2.24) and (3.6). Note that, in the regular case, we can deduce (3.11) from (3.12) by setting $z = \zeta_0$ in (3.7).

Next, let us explain how the coefficients α_k occurring in the normalized three-term recurrence relation (3.7) can be determined from the reflection coefficients ρ_k occurring in the classical Szegö–Levinson formula (2.2). By use of (3.1) and (3.9) we obtain

$$(1 - |\rho_k|^2)\alpha_{k-1}\alpha_k = \frac{s_{n-k}(\zeta_0)\hat{s}_{n-k}(\zeta_0)}{s_{n-k+1}(\zeta_0)\hat{s}_{n-k-1}(\zeta_0)}. \tag{3.13}$$

Both denominator factors in the right-hand side of (3.13) can be expressed in terms of the numerator factors with the help of the ascending and descending versions of (2.13). This produces the remarkable identity

$$\alpha_{k-1}\alpha_k = \zeta_0(1 + \zeta_0\omega_k\bar{\rho}_{k-1})^{-1}(1 - \bar{\omega}_k\rho_k)^{-1}, \tag{3.14}$$

where ω_k is the pseudoreflection coefficient (2.27). A way of using this result is explained in Proposition 3 below.

Let us now consider the question of the choice of the normalizing factors g_0 and g_1. Interesting simplifications occur in the theory if we set the constraints

$$\lambda_1 = 1, \qquad p_1(0) = \omega_1^{-1/2}, \tag{3.15}$$

with $\omega_1^{1/2}$ either of the square roots of $\omega_1 = \rho_n \hat{s}_{n-1}(\zeta_0)/s_{n-1}(\zeta_0)$. However, it should be stressed that all acceptable choices for g_0 and g_1 are essentially equivalent. (Indeed, the structure of (3.7) is invariant under the transformation $p_k(z) \to \mu_0 p_k(z)$ for even k and $p_k(z) \to \mu_1 p_k(z)$ for odd k, provided μ_1/μ_0 is real.) In view of (2.21), (2.28), and (3.6), the choice (3.15) amounts to

$$g_0 = g_1 = \omega_1^{-1/2}\zeta_0^{(n-1)/2}[s_{n-1}(\zeta_0)]^{-1}, \tag{3.16}$$

$$p_0(z) = \omega_1^{-1/2}\zeta_0^{-1/2} + \omega_1^{1/2}\zeta_0^{1/2}, \qquad p_1(z) = \omega_1^{-1/2} + \omega_1^{1/2} z. \tag{3.17}$$

This implies that $\bar{g}_0 = \rho_n g_0$, whence $\bar{g}_k = \rho_n g_k$ for all k. Therefore, since $b_k(z)$ is ρ_n-symmetric, it follows from (3.6) that $p_k(z)$ enjoys the simple *symmetry property*

$$\hat{p}_k(z) = p_k(z). \tag{3.18}$$

In view of (3.15), the coefficient d in (3.10) equals c_0^{-1}. Therefore, the Jacobi parameters are *positive*. More precisely, the sequence $(\alpha_1, \cdots, \alpha_{n-1})$ that generates the symmetric polynomials $p_k(z)$, via (3.7), corresponds to a sequence $(\rho_1, \cdots, \rho_{n-1})$ satisfying $|\rho_k| < 1$ for $k = 1, \cdots, n-1$ if and only if the Jacobi parameters $\lambda_2, \cdots, \lambda_n$ determined from the recurrence relation (3.11), with $\lambda_1 = 1$, are all positive.

It is clear that the singular case $a_n(\zeta_0) = 0$ is characterized by $\omega_1\zeta_0 = -1$. In the *regular case*, the recurrences (3.7) and (3.11) can be initialized at $k = 0$, with the conventions $p_{-1}(z) = 0$, $\lambda_0 = \infty$, and

$$\alpha_0 = (\zeta_0^{-1/2} + \omega_1\zeta_0^{1/2})^{-1}. \tag{3.19}$$

This allows us to use (3.14) with $k = 1$ by setting $\rho_0 = 1$ as before. Indeed, it produces the correct value $\alpha_1 = \zeta_0^{1/2}(1 - \bar{\omega}_1\rho_1)^{-1}$, as computed with the help of (2.13), (3.1), and (3.9). In the *singular case*, we have to initialize (3.14) at $k = 2$, with the value of α_1 just mentioned. Summarizing the results above, we obtain the following proposition, which plays a major role, especially in the companion paper [12].

PROPOSITION 3. *The numbers α_k involved in* (3.7) *can be determined from the reflection coefficients ρ_j by means of the recurrence relations* (2.29) *and* (3.14). *The initial condition for* (3.14) *is given by* (3.19) *in the regular case and by $\alpha_1 = \zeta_0^{1/2}(1 - \bar{\omega}_1\rho_1)^{-1}$ in the singular case.*

Let us say a few words concerning the normalized versions, denoted by $q_k(z)$, of the second-kind polynomials $t_k(z)$ in (2.30). The definition is

$$q_k(z) = g_k t_k(z), \tag{3.20}$$

with the same normalizing factors g_k as above. It is easily seen that $q_k(z)$ enjoys the *antisymmetry property* $\hat{q}_k(z) = -q_k(z)$, instead of (3.18). In view of this and of (2.31), it is clear that the $q_k(z)$ polynomials are linked by the same three-term recurrence relation as the $p_k(z)$ polynomials, that is,

$$q_{k+1}(z) = (\alpha_k + \bar{\alpha}_k z)q_k(z) - zq_{k-1}(z). \tag{3.21}$$

The difference arises from the initial conditions; instead of (3.17), we have

$$q_0(z) = c_0(\omega_1^{-1/2}\zeta_0^{-1/2} - \omega_1^{1/2}\zeta_0^{1/2}), \qquad q_1(z) = c_0(\omega_1^{-1/2} - \omega_1^{1/2}z). \tag{3.22}$$

It is interesting to note that (3.12) remains valid when p_k and p_{k-1} are replaced by q_k and q_{k-1}, even in the singular case $\omega_1\zeta_0 = -1$. Indeed, in the context of second-kind polynomials, the "critical situation" corresponds to $\omega_1\zeta_0 = 1$ (instead of -1). Thus, (3.11) can be deduced immediately from (3.21) provided $\omega_1\zeta_0$ differs from one.

In certain applications it is useful to introduce some *shifted* second-kind polynomials $q'_k(z)$, having the property that they vanish at the point $z = \zeta_0$ (see [12]). This is actually possible only in the regular case $\omega_1\zeta_0 \neq -1$. The definition is

$$q'_k(z) = q_k(z) + i\gamma p_k(z), \tag{3.23}$$

with the real constant $\gamma = iq_0/p_0$. This implies that $q'_1(\zeta_0) = 0$ by (3.17) and (3.22). It is clear that the shifted polynomials $q'_k(z)$ satisfy the three-term recurrence relation (3.7), (3.21). Therefore, we have $q'_k(\zeta_0) = 0$ for all k, by induction. Note that $q'_k(z)$ is antisymmetric.

In the second part of this section we examine the system of linear equations, with the Toeplitz matrix C_k, having the coefficient vector $\mathbf{p}_k = [p_{k,0}, \cdots, p_{k,k}]^T$ of the symmetric polynomial $p_k(z) = \sum_{i=0}^{k} p_{k,i} z^i$ as its solution. By use of (2.5), (2.14), (2.20), and (3.6), we obtain the system

$$C_k \mathbf{p}_k = [\bar{\tau}_k, 0, \cdots, 0, \tau_k]^T, \tag{3.24}$$

for $k = 1, \cdots, n$, where τ_k is given by

$$\tau_k = \rho_n g_k \sigma_k \zeta_0^{(n-k)/2} \bar{s}_{n-k-1}(\zeta_0). \tag{3.25}$$

Let us now explain how the three-term recurrence formula (3.7) can be interpreted (and could even be established from scratch) on the basis of (3.24). Consider the matrix identity

$$C_{k+1}[\mathbf{p}_{k+1}, \mathbf{p}_k, z\mathbf{p}_k, z\mathbf{p}_{k-1}] = \begin{bmatrix} \bar{\tau}_{k+1} & \bar{\tau}_k & \bar{\nu}_k & \bar{\nu}_{k-1} \\ 0 & 0 & \bar{\tau}_k & \bar{\tau}_{k-1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0 & \tau_k & 0 & \tau_{k-1} \\ \tau_{k+1} & \nu_k & \tau_k & \nu_{k-1} \end{bmatrix}, \tag{3.26}$$

deduced directly from (3.24), for some well-defined numbers ν_j (playing no special role in the sequel). In the left-hand side of (3.26) we make an abuse of notation that amounts to identifying a polynomial with its coefficient vector. Formula (3.7) induces a linear dependence relation between the columns of the right-hand side of (3.26). In particular, the penultimate component yields the important identity

$$\alpha_k \tau_k = \tau_{k-1}. \tag{3.27}$$

Let us stress the fact that the recurrence formula (3.7) can be derived from the linear relations (3.26) by using the argument above in the opposite direction. We shall not go into detail about this question.

Comparing (3.27) with the identity $p_{k+1}(0) = \alpha_k p_k(0)$ obtained by setting $z = 0$ in (3.7), we conclude that *the product* $\tau_{k-1} p_k(0)$ *is independent of* k. With the help of (2.23), (3.6), and (3.25), we obtain an explicit formula for $\tau_{k-1} p_k(0)$, involving the normalizing factors g_{k-1} and g_k. This formula makes sense in the case $k = 1$, where it gives the value $c_0 \zeta_0^{1/2}$ as a consequence of the choice (3.16). Hence we deduce the remarkable relation

$$\tau_{k-1} p_k(0) = c_0 \zeta_0^{1/2}, \tag{3.28}$$

for $k = 1, 2, \cdots, n$, if we set $\tau_0 = c_0 \omega_1^{1/2} \zeta_0^{1/2}$ to agree with (3.17). This value of τ_0 is easily seen to satisfy (3.27) with $k = 1$. Note that (3.26) is not valid as such in the case $k = 1$, since C_2 has only three rows; here the requirement (3.7) yields $c_0 p_0(z) = \tau_0 + \bar{\tau}_0$, with $\tau_0 = \alpha_1 \tau_1$, and this is correct in view of (3.17). Note that (3.25) produces an interesting expression for the pseudoreflection coefficient (2.27), namely,

$$\omega_k = \tau_{k-1} / \zeta_0 \bar{\tau}_{k-1}. \tag{3.29}$$

In view of (3.28), this is equivalent to $\omega_k = \bar{p}_k(0)/p_k(0)$, which itself amounts to the characterization of ω_k expressed by (2.27).

The results above give rise to a recursive method for computing the symmetric polynomials $p_k(z)$, for $k = 0, 1, \cdots, n$, from a given positive-definite Toeplitz matrix C_{n-1} of order n. This method is a generalized version of the *split Levinson algorithm* [5]. Let us select two complex numbers ζ_0 and ω_1 of unit modulus, in an arbitrary manner. (More precisely, we choose the square roots $\zeta_0^{1/2}$ and $\omega_1^{1/2}$.) For a given ζ_0 (the

circle parameter), the choice of ω_1 yields a unique value for the reflection coefficient ρ_n; hence it corresponds to a well-defined singular extension C_n of C_{n-1}. (Details about that question can be found in [10].)

The main ingredients are the relations (3.7) and (3.27), together with the *inner product formula*

$$\tau_k = \sum_{i=0}^{k} c_i \bar{p}_{k,i}, \tag{3.30}$$

which results directly from (3.24). We are now in a position to give a complete description of the *general split Levinson algorithm* [10].

PROPOSITION 4. *The sequence* $(p_k(z))_{k=0}^{n}$ *of symmetric polynomials relative to a positive-definite Toeplitz matrix* C_{n-1} *can be computed by means of the three-term formula* (3.7); *the coefficient* α_k *is obtainable via* (3.27) *from the numbers* τ_{k-1} *and* τ_k *given by* (3.30). *The initial conditions are given by* (3.17), *together with* $\tau_0 = c_0 \omega_1^{1/2} \zeta_0^{1/2}$, *where* $\omega_1^{1/2}$ *and* $\zeta_0^{1/2}$ *denote any two elements of the unit circle.*

We shall comment neither on the computational complexity nor on the numerical stability of this split Levinson algorithm. Let us mention the possibility of defining a split Schur and a split lattice algorithm in close connection with the split Levinson algorithm [6], [7], [8], [11].

Note that the appropriate reflection coefficient ρ_n (corresponding to the choice of ω_1) is available at the end of the procedure, in the form $\rho_n = \bar{p}_n(0)/p_n(0)$. As for the reflection coefficients $\rho_1, \cdots, \rho_{n-1}$ relative to the given Toeplitz matrix C_{n-1}, they can be computed with the help of (3.5), which can be written as

$$\rho_k = \frac{\tau_k}{\bar{\tau}_k}\left(\frac{\lambda_{k+1}}{\zeta_0^{1/2}\bar{\alpha}_k} - 1\right), \tag{3.31}$$

in view of (3.9) and (3.29). The Jacobi parameters λ_k involved in (3.31) are obtainable in a recursive manner by means of (3.11), with $\lambda_1 = 1$. In the regular case $\omega_1 \zeta_0 \neq -1$, they can alternatively be computed from (3.12). The predictor polynomial $a_{n-1}(z)$ relative to C_{n-1} is available from $p_n(z)$ and $p_{n-1}(z)$ in the form

$$a_{n-1}(z) = \frac{p_n(z) - \zeta_0^{-1/2} \lambda_n z p_{n-1}(z)}{p_n(0)(1 - \zeta_0^{-1} z)}. \tag{3.32}$$

This follows from (2.25), by use of (3.6) and (3.8). As for the corresponding prediction error squared norm σ_{n-1}, it can be obtained as $\sigma_{n-1} = c_0 \lambda_n |p_n(0)|^{-2}$, in view of (3.10).

Remark. To help in making comparisons with previous publications devoted to the split Levinson algorithm, let us make the following comments. In the (first-kind) singular case $\omega_1 \zeta_0 = -1$ we have $b_k(\zeta_0) = 0$, for all k, whence

$$\omega_k = -a_{k-1}(\zeta_0)/\zeta_0 \hat{a}_{k-1}(\zeta_0), \tag{3.33}$$

in view of (2.28). (In fact, $s_{n-k}(\zeta_0)$ is proportional to $\sigma_{k-1}^{-1} \bar{a}_{k-1}(\zeta_0)$; see details in § 4.) Similarly, in the "second-kind singular case" $\omega_1 \zeta_0 = 1$, we have $t_k(\zeta_0) = 0$, for all k, whence

$$\omega_k = r_{k-1}(\zeta_0)/\zeta_0 \hat{r}_{k-1}(\zeta_0), \tag{3.34}$$

in view of (2.30). Within normalization, the polynomials $b_k(z)$ that admit the pseudoreflection coefficients (3.33) and (3.34) are natural generalizations of the so-called "singular predictors" processed by the *antisymmetric version* and by the *symmetric version* of the split Levinson algorithm, respectively (see especially [5] and [9]).

It is interesting to see how the symmetric polynomials $p_k(z)$ provide a reduction of the Toeplitz matrix C_{n-1} to a tridiagonal form, by a triangular type congruence transformation. Here we consider only the regular case. From $p_k(z)$ let us construct the Laurent polynomial

$$w_k(z) = c_0^{-1/2}(-1)^k \alpha_k^{(k)} (\zeta_0 z)^{-\lfloor k/2 \rfloor} p_k(z), \tag{3.35}$$

with $\alpha_k^{(k)} = \alpha_k$ or $\bar{\alpha}_k$, depending on whether k is even or odd. For $0 \leqq k \leqq n-1$, define the tridiagonal matrix J_k (playing a major role in [12]) as follows:

$$J_k = \begin{bmatrix} 2\,\mathrm{Re}\,(\bar{\alpha}_0 \zeta_0^{1/2}) & -\zeta_0^{-1/2} & & \\ -\zeta_0^{1/2} & 2\,\mathrm{Re}\,(\bar{\alpha}_1 \zeta_0^{1/2}) & \cdot & \\ & \cdot & \cdot & -\zeta_0^{-1/2} \\ & & -\zeta_0^{1/2} & 2\,\mathrm{Re}\,(\bar{\alpha}_k \zeta_0^{1/2}) \end{bmatrix}. \tag{3.36}$$

Let W_k be the square matrix of order $k+1$ whose successive columns are the coefficient vectors of the Laurent polynomials $w_0(z), w_1(z), \cdots, w_k(z)$. Note that W_k is equivalent to an upper triangular matrix under row permutation. By use of (3.24), (3.27), and (3.28) we can derive the remarkable identity

$$W_k^* C_k W_k = J_k, \tag{3.37}$$

which generalizes a result given in [7]. (Here and in the sequel, the star symbol stands for the conjugate transpose.) This shows that J_k is positive definite together with C_k. Let us emphasize the analogy between (3.37) and the Cholesky factorization of C_k^{-1} provided by the classical predictors $a_l(z)$ (e.g., see [18]) instead of the symmetric polynomials $p_l(z)$.

Finally consider the problem of computing the predictor $a_n(z)$ relative to a given nonnegative-definite Toeplitz matrix C_n (of nullity one). This can be solved by use of the split Levinson algorithm explained above, where an arbitrary value ε (with $|\varepsilon| = 1$) is assigned to the parameter ω_1. If, by accident, ε equals the pseudoreflection coefficient ω_1 that corresponds to the given ρ_n, then $p_n(z)$ is proportional to $a_n(z)$. In the other cases this property does not hold, but the method can be continued one step further so as to produce a symmetric polynomial $p_{n+1}(z)$ that is proportional to $(\zeta_0 - z)a_n(z)$. This result will be proved at the end of § 5. At least from a theoretical viewpoint, the most interesting choice for the ε parameter is the value $\varepsilon = -\zeta_0^{-1}$.

4. Discussion of the singular case. Although it does not give rise to any difficulty in the general theory, the singular case (often mentioned above) deserves special attention for several reasons. We now examine that subject in a detailed manner. Thus, throughout the present section, we assume that the circle parameter ζ_0 obeys the *singularity condition*

$$a_n(\zeta_0) = 0, \quad \text{i.e., equivalently, } \omega_1 \zeta_0 = -1, \tag{4.1}$$

with respect to the given Schur–Szegö parameters $\rho_1, \cdots, \rho_n$.

In view of (2.24) and (3.6), all symmetric (normalized) polynomials $p_k(z)$ are divisible by $\zeta_0 - z$. Thus it is quite natural to consider the family of *reduced polynomials*

$$\tilde{p}_k(z) = \mu_k p_{k+1}(z)/p_1(z), \tag{4.2}$$

for $k = 0, 1, \cdots, n-1$. Here μ_k is a *positive* normalizing factor that is introduced for a technical (or aesthetic) reason of relatively small importance. Note that $\tilde{p}_k(z)$ enjoys the symmetry property (3.18). The crucial requirement is that μ_k *depends only on the parity of* k. In other words, we have $\mu_k = \mu_{k-2}$ for all k. This implies that the polynomials $\tilde{p}_k(z)$

satisfy the normalized three-term recurrence relation (3.7) where the coefficient α_k is replaced by

$$\tilde{\alpha}_k = (\mu_{k+1}/\mu_k)\alpha_{k+1}. \tag{4.3}$$

The parameters μ_0 and μ_1 are now chosen in such a way that the reduced polynomials $\tilde{p}_0(z)$ and $\tilde{p}_1(z)$ have the appropriate form (3.17). First, since $\tilde{p}_1(z) = \mu_1(\alpha_1 + \bar{\alpha}_1 z)$, by (4.2) and (3.7), we must have

$$\tilde{\omega}_1 = \bar{\alpha}_1/\alpha_1, \tag{4.4}$$

with $\alpha_1 = \zeta_0^{1/2}(1 + \zeta_0\rho_1)^{-1}$. Then, choosing the square root $\tilde{\omega}_1^{1/2} = \bar{\alpha}_1/|\alpha_1|$, we readily obtain the following values for the normalizing factors:

$$\mu_k = \begin{cases} |\alpha_1|^{-1}(\zeta_0^{-1/2}\alpha_1 + \zeta_0^{1/2}\bar{\alpha}_1) & \text{even } k, \\ |\alpha_1|^{-1} & \text{odd } k. \end{cases} \tag{4.5}$$

Let us examine the explicit definition (2.20) of $b_k(z)$ under our assumption (4.1). Set $\beta = \sigma_{n-1}/\bar{a}_{n-1}(\zeta_0)$. Using (2.2), we obtain $\beta/\bar{\beta} = -\rho_n\zeta_0^n$. It is easily seen that $s_{n-k-1}(\zeta_0)$ equals $\beta\sigma_k^{-1}\bar{a}_k(\zeta_0)$. This is true by definition for $k = n - 1$ and is proved for all k with the help of the Szegö–Levinson recurrences (2.2) and (2.13). As a result, we can write

$$b_{k+1}(z) = \beta\sigma_k^{-1}\zeta_0^{-(k+n+1)/2}[\zeta_0\hat{a}_k(\zeta_0)a_k(z) - a_k(\zeta_0)z\hat{a}_k(z)]. \tag{4.6}$$

The reader who is familiar with the theory of orthogonal polynomials on the unit circle will recognize the presence of Szegö's "kernel polynomial" in the right-hand side of (4.6). Let us now provide the required information about this classical subject (see [18], [22], [23]).

The *inner product* of two complex polynomials $x(z)$ and $y(z)$, of degree less than or equal to n, relative to the positive measure associated with the Toeplitz matrix C_n, can be expressed in the form

$$\langle x(z), y(z)\rangle = \mathbf{x}^*\mathbf{C}_n\mathbf{y}, \tag{4.7}$$

where $\mathbf{x}$ and $\mathbf{y}$ denote the coefficient vectors of $x(z)$ and $y(z)$. For an integer $k = 0, \cdots, n - 1$, a polynomial $\Phi_k(\zeta, z)$, of degree k in each variable $\bar{\zeta}$ and z, is called a *kernel polynomial* with respect to the inner product (4.7) if it enjoys the *reproducing property*

$$\langle \Phi_k(\zeta, z), x(z)\rangle = x(\zeta) \tag{4.8}$$

for every polynomial $x(z)$ of degree less than or equal to k in z. Applying (4.8) to the monomials $x(z) = z^t$, with $t = 0, \cdots, k$, we see that the kernel polynomial is determined uniquely and that its $\bar{\zeta}^s z^t$ coefficient equals the (t, s) entry of the inverse of the Toeplitz matrix C_k. Thus we have the expression

$$\Phi_k(\zeta, z) = [1, z, \cdots, z^k]C_k^{-1}[1, \zeta, \cdots, \zeta^k]^*. \tag{4.9}$$

Next, consider any *orthonormal basis* $(u_0(z), u_1(z), \cdots, u_k(z))$ of the vector space of complex polynomials of degree less than or equal to k in z. Here, orthonormal means $\langle u_s(z), u_t(z)\rangle = \delta_{s,t}$ for all s and t. A simple computation, based on (4.8), yields the expansion

$$\Phi_k(\zeta, z) = \sum_{t=0}^{k} \bar{u}_t(\zeta)u_t(z). \tag{4.10}$$

Let us apply this to the *degree-graded basis*, consisting of the normalized Szegö polynomials $u_t(z) = \sigma_t^{-1/2}\hat{a}_t(z)$, and to the *delay-graded basis*, consisting of the shifted normalized predictor polynomials $u_t(z) = \sigma_t^{-1/2}z^{k-t}a_t(z)$. In both cases, the orthonormality relations follow directly from the Toeplitz linear systems (2.5). Substituting these bases into (4.10), we deduce both identities

$$\Phi_k(\zeta, z) = \Phi_{k-1}(\zeta, z) + \sigma_k^{-1}\bar{\hat{a}}_k(\zeta)\hat{a}_k(z), \tag{4.11}$$

$$\Phi_k(\zeta, z) = \bar{\zeta}z\Phi_{k-1}(\zeta, z) + \sigma_k^{-1}\bar{a}_k(\zeta)a_k(z). \tag{4.12}$$

By elimination of $\Phi_{k-1}(\zeta, z)$ between (4.11) and (4.12) we obtain the *Christoffel–Darboux formula*, which leads to the *Trench inversion formula* for Toeplitz matrices, via (4.9) [18]. For a point $\zeta = \zeta_0$ on the unit circle ($|\zeta_0| = 1$), the Christoffel–Darboux formula reads as follows:

$$(\zeta_0 - z)\Phi_k(\zeta_0, z) = \sigma_k^{-1}\zeta_0^{-k}[\zeta_0\hat{a}_k(\zeta_0)a_k(z) - a_k(\zeta_0)z\hat{a}_k(z)]. \tag{4.13}$$

Thus we have identified the polynomial (4.6). With $\omega_1^{1/2}\zeta_0^{1/2} = i$, we have $p_1(z) = -i\zeta_0^{-1/2}(\zeta_0 - z)$. Using this expression together with the results above, we can write the reduced polynomial (4.2) in the remarkable form

$$\tilde{p}_k(z) = \nu_k\zeta_0^{k/2}\Phi_k(\zeta_0, z), \tag{4.14}$$

with $\nu_k = c_0\mu_k g_0^{-1}g_{k+1}$ (a positive number). Hence, in view of (4.9), we can interpret the coefficient vector $\tilde{\mathbf{p}}_k$ of $\tilde{p}_k(z)$ as the solution of the Toeplitz linear system

$$C_k\tilde{\mathbf{p}}_k = \nu_k\zeta_0^{k/2}[1, \zeta_0^{-1}, \cdots, \zeta_0^{-k}]^T. \tag{4.15}$$

Systems of that type have been investigated recently by Bruckstein and Kailath [3]. In the special case where the circle parameter ζ_0 equals one and the reflection coefficients ρ_k are real, (4.15) is sometimes referred to as the *discrete Gopinath–Sondhi equation*. An efficient solution algorithm for this equation has been discovered by Bube and Burridge [4]; it is based on a recurrence relation essentially equivalent to the appropriate special case of (3.7); it is closely related to the split Levinson algorithm.

As indicated in the beginning of this section, the family of reduced polynomials $\tilde{p}_k(z)$ fits exactly into the general theory developed in §§ 2 and 3. Recall that we have identified the relevant recurrence coefficients $\tilde{\alpha}_k$ in (4.3), via (4.5). In principle, this allows us to determine all other parameters, within an arbitrary choice for the positive number $\tilde{c}_0$. We shall now identify the nonnegative-definite Toeplitz matrix $\tilde{C}_{n-1}$ (of order n and rank $n-1$) that produces the symmetric polynomials $\tilde{p}_k(z)$ in the sense of our general theory. It is important to note that *the reduced polynomials belong to the regular case* (with respect to the given circle parameter ζ_0). Indeed, $\tilde{p}_{n-1}(\zeta_0)$ does not vanish, since the zeros of $a_n(z)$ are simple; equivalently, we have $\tilde{\omega}_1\zeta_0 \neq -1$, in view of (4.4).

For $1 \leqq k \leqq n-1$, define Z_k to be the $k \times (k+1)$ bidiagonal Toeplitz matrix having the coefficient vector of $p_1(z)$ as its first column, that is,

$$Z_k = -i\zeta_0^{1/2}\begin{bmatrix} 1 & & & \\ -\zeta_0^{-1} & 1 & & \\ & \ddots & \ddots & \\ & & -\zeta_0^{-1} & 1 \\ & & & -\zeta_0^{-1} \end{bmatrix}. \tag{4.16}$$

We shall see that the $k \times k$ Toeplitz section of $\tilde{C}_{n-1}$ can be written in the form $\tilde{C}_{k-1} = Z_k^* C_k Z_k$. In explicit terms, this means that the entries $\tilde{c}_k$ of $\tilde{C}_{n-1}$ are given by

$$\tilde{c}_k = 2c_k - \zeta_0 c_{k+1} - \zeta_0^{-1} c_{k-1}, \tag{4.17}$$

for $1 - n \leqq k \leqq n - 1$. It is clear that (4.2) amounts to the vector identity $Z_k \tilde{\mathbf{p}}_{k-1} = \mu_{k-1} \mathbf{p}_k$ (with k replaced by $k+1$). Hence, setting the Toeplitz matrix $\tilde{C}_{k-1} = Z_k^* C_k Z_k$, and multiplying the linear system (3.24) by Z_k^*, we obtain a system of exactly the same form, with C_k, $\mathbf{p}_k$, and τ_k replaced by $\tilde{C}_{k-1}$, $\tilde{\mathbf{p}}_{k-1}$, and

$$\tilde{\tau}_{k-1} = -i\zeta_0^{1/2} \mu_{k-1} \tau_k. \tag{4.18}$$

Note that the discussion should be slightly different when $k = 1$; however, the conclusion (4.18) remains valid in that case. In view of the uniqueness of the Toeplitz matrix $\tilde{C}_{n-1}$ underlying the family of symmetric polynomials $\tilde{p}_k(z)$, for a given ζ_0, this argument proves the claim (4.17). The consistency of the results can be checked by examining the identities (3.27) and (3.28) for the reduced polynomials. In summary, we have the following result.

PROPOSITION 5. *The reduced polynomial $\tilde{p}_k(z)$ is proportional to the kernel polynomial $\Phi_k(\zeta, z)$ with $\zeta = \zeta_0$ (the circle parameter); its coefficient vector obeys the equation* (4.15). *Furthermore, $\tilde{p}_k(z)$ coincides with the degree k symmetric polynomial that is canonically associated with the Toeplitz matrix $\tilde{C}_{n-1}$ whose entries are given by* (4.17).

The corresponding pseudoreflection coefficients $\tilde{\omega}_k$ can be determined with the help of (3.29) and (4.18); the result is

$$\tilde{\omega}_k = -\zeta_0 \omega_{k+1}. \tag{4.19}$$

The case $k = n - 1$ reads $\tilde{\rho}_{n-1} = -\zeta_0 \rho_n$. Alternatively, this follows from the very definition of the reduced predictor, i.e.,

$$\tilde{a}_{n-1}(z) = a_n(z)/(1 - \zeta_0^{-1} z). \tag{4.20}$$

As for the remaining reflection coefficients $\tilde{\rho}_k$ relative to $\tilde{C}_{n-1}$, they can be determined by means of formula (3.31). In view of (4.2), this involves the Jacobi parameter $\tilde{\lambda}_k = (\mu_k/\mu_{k-1}) l_{k+1}$, where

$$l_{k+1} = \zeta_0^{-1/2} \lim_{z \to \zeta_0} \frac{p_{k+1}(z)}{p_k(z)}, \tag{4.21}$$

by (3.12). From (3.31) we readily obtain

$$\tilde{\rho}_{k-1} = -\zeta_0 \rho_k (l_{k+1} - \zeta_0^{1/2} \bar{\alpha}_k)/(\lambda_{k+1} - \zeta_0^{1/2} \bar{\alpha}_k). \tag{4.22}$$

Note that the sequences $(\lambda_k)_{k=1}^n$ and $(l_k)_{k=1}^n$ satisfy the *same* recurrence relation (3.11), with the initial values $\lambda_1 = 1$ and $l_1 = \infty$. By subtraction, this implies

$$l_{k+1} - \lambda_{k+1} = \lambda_k^{-1} - l_k^{-1}. \tag{4.23}$$

With the family of "first-kind" symmetric polynomials $\tilde{p}_k(z)$, we can associate a family of "second-kind" *antisymmetric polynomials* $\tilde{q}_k(z)$. They satisfy the same three-term recurrence relation as the first-kind polynomials (with the same coefficients $\tilde{\alpha}_k$). The initial conditions $\tilde{q}_0(z)$ and $\tilde{q}_1(z)$ are given by (3.22), where c_0 and $\omega_1^{1/2}$ are replaced by $\tilde{c}_0 = c_0 \mu_0 \mu_1$ and $\tilde{\omega}_1^{1/2} = \mu_1 \bar{\alpha}_1$. We could also consider shifted second-kind polynomials $\tilde{q}_k(z) + i\tilde{\gamma}\tilde{p}_k(z)$, vanishing at the point $z = \zeta_0$, as in (3.23).

5. On duality and its interpretation. Replacing the Schur–Szegö sequence $(\rho_k)_{k=1}^n$ by its dual $(\rho_k^{\#})_{k=1}^n$, with $\rho_k^{\#} = \rho_n \bar{\rho}_{n-k}$ (where $\rho_0 = 1$) as in (2.12), and preserving

the circle parameter ζ_0, we obtain a family of symmetric polynomials $p_k^{\#}(z)$, which we call the *dual* of the original $p_k(z)$ family. This dual family is quite interesting in that it enjoys the *same embedding property*,

$$p_n^{\#}(z) = p_n^{\#}(0) a_n(z), \tag{5.1}$$

as the original family, since we have $a_n^{\#}(z) = s_n(z) = a_n(z)$ in view of (2.15). In other words, $p_n^{\#}(z)$ is proportional to $p_n(z)$, while it is clear that $p_k^{\#}(z)$ is generally not proportional to $p_k(z)$ for $0 < k < n$. Let us now examine some remarkable features of this duality relation.

By definition, $p_k^{\#}(z)$ is the normalized version $p_k^{\#}(z) = g_k^{\#} b_k^{\#}(z)$ of the ρ_n-symmetric predictor polynomial $b_k^{\#}(z)$ given by

$$b_k^{\#}(z) = \zeta_0^{(k-n)/2} [a_{n-k}(\zeta_0) s_{k-1}(z) + \rho_n \hat{a}_{n-k}(\zeta_0) z \hat{s}_{k-1}(z)]. \tag{5.2}$$

Indeed, our duality permutes the roles of $a_k(z)$ and $s_k(z)$; hence (5.2) is the dual of (2.20). The pseudoreflection coefficient $\omega_k^{\#}$ of the polynomial $b_k^{\#}(z)$ is given by $\omega_k^{\#} = \psi_k^{\#}(\zeta_0)$, with the classical inner function $\psi_k^{\#}(z) = \rho_n \hat{a}_{n-k}(z)/a_{n-k}(z)$. Information concerning the connection between the Toeplitz matrices C_k and their duals $C_k^{\#}$ can be found in [13].

In the sequel we consider only the *regular case*, i.e., $a_n(\zeta_0) \neq 0$. Our objective in this section is to explain how the dual family can be interpreted in the original framework. We are mainly interested in the coefficients $\alpha_k^{\#}$ of the dual three-term recurrence (3.7). For $k = 1, \cdots, n, n+1$, define the complex number ω_k^{+} (of unit modulus) by

$$\omega_k^{+} = -a_{k-1}(\zeta_0)/\zeta_0 \hat{a}_{k-1}(\zeta_0). \tag{5.3}$$

In particular, $\omega_1^{+} = -\zeta_0^{-1}$ and $\omega_{n+1}^{+} = -\zeta_0^{-1}\rho_n$. Note that we have $\omega_n^{+} \neq \rho_n$. The dual pseudoreflection coefficients are given by

$$\omega_k^{\#} = -\zeta_0^{-1} \rho_n \bar{\omega}_{n+1-k}^{+}. \tag{5.4}$$

Elementary computation shows that the dual version of formula (3.14) can be written in the form

$$\alpha_{n+1-k}^{\#} \alpha_{n-k}^{\#} = \zeta_0 (1 + \zeta_0 \omega_k^{+} \bar{\rho}_{k-1})^{-1} (1 - \bar{\omega}_k^{+} \rho_k)^{-1}. \tag{5.5}$$

The right-hand side of (5.5) coincides exactly with that of (3.14), except that ω_k is replaced by ω_k^{+}. As a consequence, the numbers α_k^{+} generated by the relation (3.14) thus modified have the form

$$\alpha_k^{+} = \begin{cases} t\alpha_{n-k}^{\#} & \text{even } k, \\ t^{-1}\alpha_{n-k}^{\#} & \text{odd } k, \end{cases} \tag{5.6}$$

for a suitable constant t. Note that (5.6) makes sense for $k = n$, with the value of $\alpha_0^{\#}$ given by (3.19), i.e.,

$$\alpha_0^{\#} = \zeta_0^{1/2} (1 - \bar{\omega}_n^{+} \rho_n)^{-1}. \tag{5.7}$$

The numbers ω_k^{+} defined in (5.3) obey exactly the same Schur-type recurrence relation (2.29) as the "correct" numbers ω_k, except that we have $\omega_1^{+} = -\zeta_0^{-1}$ (instead of $\omega_n = \rho_n$). Hence they correspond to the *distorted sequence of reflection coefficients* ρ_k^{+} given by

$$\rho_k^{+} = \rho_k \quad \text{for } k = 1, 2, \cdots, n-1, \qquad \rho_n^{+} = \omega_n^{+}. \tag{5.8}$$

Let us now consider the family of symmetric polynomials $p_k^{+}(z)$ associated with this distorted sequence. It is generated by the recurrence formula (3.7), with the coefficients

α_k^+ given in (5.6). The initialization is $p_0^+(z) = 0$ and $p_1^+(z) = -i\zeta_0^{1/2}(1 - \zeta_0^{-1}z)$. The first coefficient is found to be $\alpha_1^+ = \zeta_0^{1/2}(1 + \zeta_0\rho_1)^{-1}$. Hence, the dual versions of (3.1) and (3.9) yield $\alpha_{n-1}^{\#} = \lambda_n^{\#}\alpha_1^+$, with $\lambda_n^{\#}$ denoting the dual Jacobi parameter (3.8). This determines the factor t in (5.6) to be the *positive real number* $t = \lambda_n^{\#}$. By construction, the family of polynomials $p_k^+(z)$ belongs to the *singular case*; we have $a_n^+(\zeta_0) = 0$ (as a consequence of $\omega_1^+\zeta_0 = -1$). Thus, as shown in (4.14), the reduced polynomial $\tilde{p}_k^+(z)$ is proportional to the Szegö kernel $\Phi_k(\zeta_0, z)$ relative to the Toeplitz matrix C_k, for $0 \leqq k \leqq n - 1$ (see (5.8)).

The coefficient vector of the polynomial $p_k^+(z)$ obeys a linear system of the form (3.24), with the *same Toeplitz matrix* C_k (even when $k = n$) but with a *different number* τ_k, denoted here by τ_k^+. In particular, let us stress that τ_n^+ is not zero, since $a_n^+(z)$ differs from $a_n(z)$. Our definition of α_n^+, given by (5.6) and (5.7), allows us to go one step further in the recurrence (3.7) and thus to obtain a symmetric polynomial $p_{n+1}^+(z)$ of degree $n + 1$. We shall see that it has the quite remarkable form

$$p_{n+1}^+(z) = p_{n+1}^+(0)(1 - \zeta_0^{-1}z)a_n(z). \tag{5.9}$$

Thus, the extended family of polynomials $p_k^+(z)$ relative to the distorted Schur–Szegö sequence enjoys the same type of embedding property as in the "official theory," except that n is replaced by $n + 1$ and $a_n(z)$ by $a_{n+1}^+(z) = (1 - \zeta_0^{-1}z)a_n(z)$. It is interesting to compare this *extension operation* (from n to $n + 1$) with the *reduction operation* (from n to $n - 1$) reflected in (4.20).

As explained at the end of this section, in a more general setting, the result (5.9) can be established by straightforward verification, based on explicit expressions for $p_{n-1}^+(z)$ and $p_n^+(z)$. Here we use an alternative argument relying specifically on duality. Set the positive constant $r = \lambda_n^{\#}\mu_0^+/\mu_1^+$. In view of (5.6) and (4.3), the reduced version $\tilde{p}_n^+(z)$ of $p_{n+1}^+(z)$ is produced by the normalized mirror image $(r^{-1}\alpha_{n-1}^{\#}, r\alpha_{n-2}^{\#}, \cdots, r^{\varepsilon}\alpha_0^{\#})$, with $\varepsilon = (-1)^n$, of the sequence $(\alpha_0^{\#}, \alpha_1^{\#}, \cdots, \alpha_{n-1}^{\#})$ producing the polynomial $p_n^{\#}(z)$. Now it is a simple exercise to prove that the mirror image operation preserves the output polynomial of the recurrence (3.7). (See [12] for further details on this subject.) Therefore, $\tilde{p}_n^+(z)$ is proportional to $p_n^{\#}(z)$, so that the desired result (5.9) is nothing but (5.1). The main conclusion can be stated as follows (an application is given in the companion paper [12]).

PROPOSITION 6. *The distorted sequence of reflection coefficients* $(\rho_k^+)_{k=1}^n$, *defined in* (5.8), *produces a family of reduced symmetric polynomials* $\tilde{p}_k^+(z)$, *for* $k = 0, 1, \cdots, n$, *the last element of which is proportional to the predictor polynomial* $a_n(z)$.

Let us emphasize that the coefficients $\alpha_1^+, \alpha_2^+, \cdots, \alpha_n^+$ defined in (5.6) can be computed, together with the polynomials $p_k^+(z)$, by running the split Levinson algorithm of § 3 for the given Toeplitz matrix C_n; it suffices to set the parameter ω_1^+ equal to $-\zeta_0^{-1}$. (Note the property $\tau_0^+ = ic_0$.) The matrix identity (3.26) is satisfied up to $k = n$ (with $\mathbf{p}_l^+$ instead of $\mathbf{p}_l$, etc.), with $\tau_{n+1}^+ = 0$ and with C_{n+1} denoting the unique singular Toeplitz extension of order $n + 2$ of C_n. (Note that C_{n+1} is nonnegative definite and has nullity two.) Let us stress that all recurrence coefficients α_k^+ in this "extended" split Levinson algorithm are given by $\alpha_k^+ = \tau_{k-1}^+/\tau_k^+$, as in (3.27), including the last of them (for $k = n$). Further details on that subject can be found at the end of the paper.

As explained in the general theory of §§ 2 and 3, a family of second-kind antisymmetric polynomials $q_k^+(z)$ can be constructed from the same coefficient sequence $(\alpha_k^+)_{k=1}^n$, via the recurrence (3.21). It can be checked that the final polynomial, $q_{n+1}^+(z)$, is given by

$$q_{n+1}^+(z) = p_{n+1}^+(0)(1 - \zeta_0^{-1}z)r_n(z), \tag{5.10}$$

in terms of the classical second-kind predictor $r_n(z)$. The proof is basically the same as that of (5.9); details are omitted. In contrast with (5.10), note that none of the polynomials $q_k^+(z)$ with $1 \leqq k \leqq n$ vanishes at the point $z = \zeta_0$. The property $q_{n+1}^+(\zeta_0) = 0$ shows that the Jacobi parameter λ_{n+1}^+, defined formally as in (3.11), is equal to zero. In other words, we have $\lambda_n^+ = (\zeta_0^{-1/2}\alpha_n^+ + \zeta_0^{1/2}\bar{\alpha}_n^+)^{-1}$. Thus, the result (5.9) can be interpreted as an extension of the formula (3.32) where n is replaced by $n+1$ (and p_n by p_{n+1}^+). Note that the property $\lambda_{n+1}^+ = 0$ agrees with (3.10).

As alluded to above, the polynomials $p_k^+(z)$ can be introduced in an alternative manner, which is equally interesting. Assume that the available data are the entries of the Toeplitz matrix C_n, rather than the corresponding Schur–Szegö parameters $\rho_1, \cdots, \rho_n$. In this situation, the recurrence coefficients α_k can no longer be computed by way of (2.29) and (3.14). Choosing any complex number ε of unit modulus, let us apply the split Levinson algorithm of § 3 to the Toeplitz matrix C_n, with the initialization

$$p_0(z,\varepsilon) = \varepsilon^{-1/2}\zeta_0^{-1/2} + \varepsilon^{1/2}\zeta_0^{1/2}, \qquad p_1(z,\varepsilon) = \varepsilon^{-1/2} + \varepsilon^{1/2}z, \tag{5.11}$$

and $\tau_0 = c_0\varepsilon^{1/2}\zeta_0^{1/2}$. This is formally the same as (3.17), with $\omega_1(\varepsilon) = \varepsilon$ in place of ω_1. The algorithm produces some well-defined symmetric polynomials $p_k(z,\varepsilon)$ and recurrence coefficients $\alpha_k(\varepsilon)$, depending on the ε parameter. Of course, $p_k(z,\varepsilon)$ equals $p_k(z)$ for all k if and only if ε equals ω_1. Looking at the initial conditions, we see that the choice $\varepsilon = -\zeta_0^{-1}$ leads to the polynomials $p_k^+(z)$ examined above; we have

$$p_k(z, -\zeta_0^{-1}) = p_k^+(z), \tag{5.12}$$

for $k = 0, 1, \cdots, n$. As explained below, this identity extends to the case $k = n+1$, in a natural manner.

For any value of ε (with $|\varepsilon| = 1$), elementary computations give

$$p_{n-1}(z,\varepsilon) = \lambda_n^{-1}(\varepsilon)p_n(0,\varepsilon)\zeta_0^{-1/2}[a_{n-1}(z) + \omega_n(\varepsilon)\zeta_0\hat{a}_{n-1}(z)], \tag{5.13}$$

$$p_n(z,\varepsilon) = p_n(0,\varepsilon)[a_{n-1}(z) + \omega_n(\varepsilon)z\hat{a}_{n-1}(z)], \tag{5.14}$$

$$\tau_{n-1}(\varepsilon) = \sigma_{n-1}\lambda_n^{-1}(\varepsilon)p_n(0,\varepsilon)\omega_n(\varepsilon)\zeta_0^{1/2}, \tag{5.15}$$

$$\tau_n(\varepsilon) = \sigma_{n-1}p_n(0,\varepsilon)[\omega_n(\varepsilon) - \rho_n]. \tag{5.16}$$

As a consequence, provided we have $\varepsilon \neq \omega_1$, i.e., $\omega_n(\varepsilon) \neq \rho_n$, the split Levinson algorithm can be applied one step further to yield

$$p_{n+1}(z,\varepsilon) = p_{n+1}(0,\varepsilon)(1 - \zeta_0^{-1}z)a_n(z). \tag{5.17}$$

Comparing with (5.9), we see that (5.12) is actually valid for $0 \leqq k \leqq n+1$. Note that the special choice $\varepsilon = -\zeta_0^{-1}$ is characterized by the fact that $p_k(\zeta_0, \varepsilon)$ vanishes for all k, while this is true only for $k = n+1$ when we choose $\varepsilon \neq -\zeta_0^{-1}$. Thus, a family of reduced polynomials can be defined exclusively in the case $\varepsilon = -\zeta_0^{-1}$.

REFERENCES

[1] N. I. AKHIEZER, *The Classical Moment Problem*, Oliver and Boyd, London, 1965.

[2] Y. BISTRITZ, *Zero location with respect to the unit circle of discrete-time linear system polynomials*, Proc. IEEE, 72 (1984), pp. 1131–1142.

[3] A. BRUCKSTEIN AND T. KAILATH, *Some matrix factorization identities for discrete inverse scattering*, Linear Algebra Appl., 74 (1986), pp. 157–172.

[4] K. P. BUBE AND R. BURRIDGE, *The one-dimensional inverse problem of reflection seismology*, SIAM Rev., 25 (1983), pp. 497–559.

[5] P. DELSARTE AND Y. GENIN, *The split Levinson algorithm*, IEEE Trans. Acoust. Speech Signal Process., 34 (1986), pp. 470–478.

[6] ———, *On the splitting of classical algorithms in linear prediction theory*, IEEE Trans. Acoust. Speech Signal Process., 35 (1987), pp. 645–653.

[7] ———, *The tridiagonal approach to Szegö's orthogonal polynomials, Toeplitz linear systems, and related interpolation problems*, SIAM J. Math. Anal., 19 (1988), pp. 718–735.

[8] ———, *A survey of the split approach based techniques in digital signal processing applications*, Philips J. Res., 43 (1988), pp. 346–374.

[9] ———, *The multichannel split Levinson algorithm*, in Linear Circuits, Systems and Signal Processing: Theory and Applications, C. I. Byrnes, C. F. Martin, and R. E. Saeks, eds., North-Holland, Amsterdam, 1988, pp. 183–190.

[10] ———, *An introduction to the class of split Levinson algorithms*, in Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, G. H. Golub and P. Van Dooren, eds., NATO Advanced Study Institutes Series, Springer-Verlag, Berlin, New York, 1990, pp. 111–130.

[11] ———, *On the split approach based algorithms for* DSP *problems*, in Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, G. H. Golub and P. Van Dooren, eds., NATO Advanced Study Institutes Series, Springer-Verlag, Berlin, New York, 1990, pp. 131–148.

[12] ———, *Tridiagonal approach to the algebraic environment of Toeplitz matrices, part* II: *Zero and eigenvalue problems*, SIAM J. Matrix Anal. Appl., 12 (1991), to appear.

[13] P. DELSARTE, Y. GENIN, AND Y. KAMP, *On the class of positive definite matrices equivalent to Toeplitz matrices*, in Proc. Internat. Symposium on Mathematical Theory of Networks and Systems, N. Levan, ed., Western Periodicals, North Hollywood, CA, 1981, pp. 40–45.

[14] P. DELSARTE, Y. GENIN, Y. KAMP, AND P. VAN DOOREN, *Speech modelling and the trigonometric moment problem*, Philips J. Res., 37 (1982), pp. 277–292.

[15] J. GILEWICZ AND E. LEOPOLD, *Location of the zeros of polynomials satisfying three-term recurrence relations*. I. *General case with complex coefficients*, J. Approx. Theory, 43 (1985), pp. 1–14.

[16] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, North Oxford Academic, Oxford, U.K., 1983.

[17] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and Their Applications*, University of California Press, Berkeley, CA, 1958.

[18] T. KAILATH, A. VIEIRA, AND M. MORF, *Inverses of Toeplitz operators, innovations, and orthogonal polynomials*, SIAM Rev., 20 (1978), pp. 106–119.

[19] J. LE ROUX AND C. GUEGUEN, *A fixed point computation of partial correlation coefficients*, IEEE Trans. Acoust. Speech Signal Process., 25 (1977), pp. 257–259.

[20] J. MAKHOUL, *Linear prediction: A tutorial review*, Proc. IEEE, 63 (1975), pp. 561–580.

[21] M. MARDEN, *Geometry of Polynomials*, American Mathematical Society, Providence, RI, 1966.

[22] P. E. SAYLOR AND D. C. SMOLARSKI, *Computing the roots of complex orthogonal and kernel polynomials*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 1–13.

[23] G. SZEGÖ, *Orthogonal Polynomials*, American Mathematical Society, New York, 1959.

[24] J. L. WALSH, *Interpolation and Approximation by Rational Functions in the Complex Domain*, American Mathematical Society, Providence, RI, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 2, pp. 239–272, April 1991

005

ON RANDOM CORRELATION MATRICES*

R. B. HOLMES†

Abstract. This report contains a detailed study of random correlation matrices, including algebraic, statistical, and historical background. Such matrices are of particular interest because they serve to model "average signals" for simulation testing of signal processing algorithms. The statistical behavior of spectral functions of the two major types of random correlation matrices is extensively discussed in the latter half, from both theoretical and empirical aspects. The emphasis is on eigenvalue distribution and condition number behavior. Actual application to algorithm testing will be described in a subsequent report.

Key words. correlation matrix, random correlation matrix, random spectrum, Gram matrix, random Gram matrix, random eigenvalue, condition number, spacings, spectral distribution function

AMS(MOS) subject classifications. primary 15A52; secondary 15A12

1. Introduction.

This paper derives from a study of the relative efficacy of certain (group-theoretic) data transforms for various canonical signal processing tasks. Two such tasks are, in particular, data compression and decorrelation. For a given data transform, realized as a unitary matrix U, the extent of such activity can be measured from the transformed data covariance matrix. Thus if a data vector x has covariance C, its transform Ux has covariance $D = UCU^*$, and the data compression (respectively, decorrelation) efficiency of the transform U can be assessed by examination of the diagonal (respectively, off-diagonal) entries of D.

In order to make a serious statistical study of the efficiency of group transforms and filters for the various signal processing tasks, it is necessary to have an assortment of standardized signal models. These fall into two classes: parametric models and "purely random" models. The former determine, after sampling, structured covariance matrices with entries having a simple dependence on a few parameters. The simplest and most familiar example is the first-order Markov or autoregressive signal model, from which N samples generate the covariance matrix $[\rho^{|i-j|}]$, where $0 < |\rho| < 1$, and $1 \leqq i, j \leqq N$. On the other hand, it is somewhat less clear a priori what a "purely random" covariance structure might be. Clarification and discussion of this term is the primary object of the following sections. Speaking intuitively for the moment, it is evident that this term must be precisely defined if we are to be able to do any serious simulations of the action of the various transforms, and to eventually say that one or another of them, for fixed data dimension, is superior in the performance of a particular task "on the average."

This paper is written in a somewhat discursive style, with §§ 1 and 2, along with §§ 3.1 and 4.1, being essentially expository. Relevant definitions and aspects of numerical linear algebra are collected in Appendices A and B. Appendix A is primarily a review of known, if not "well-known," bounds on norms and eigenvalues. Appendix B focuses on the important concept of condition number, the behavior of which, under various conditions of randomness, is a major object of study of §§ 3 and 4. The main result in Appendix B is a sharp lower bound on the norm of the inverse of a correlation matrix.

1.1. Definitions.

In the background we have an N-dimensional real or complex-valued second-order random vector x. We will usually assume that x has zero mean

* Received by the editors November 12, 1988; accepted for publication (in revised form) November 29, 1989. This work was sponsored by the Lincoln Laboratory Innovative Research Program (IRP).

† Massachusetts Institute of Technology, Lincoln Laboratory, P.O. Box 73, Lexington, Massachusetts 02173 (davew@LL.LL.MIT.EDU).

$E(x) = \Theta$, the zero vector. The covariance matrix of x is the $N \times N$ matrix C_x defined by

$$C_x = [E(x_i \bar{x}_j)].$$

Such matrices are characterized as being Hermitian and positive semidefinite. We will, in fact, always assume that C_x is actually positive definite, so as to eliminate degenerate probability density functions. Hence the eigenvalues $\{\lambda_1, \cdots, \lambda_N\}$ of C_x are all positive; they constitute the *spectrum* $\sigma(C_x)$, of C_x, and their relative size will always be indicated by subscript: $\lambda_1 \geqq \lambda_2 \geqq \cdots \geqq \lambda_N > 0$.

We recall the statistical significance of these eigenvalues: letting $\{\phi_1, \cdots, \phi_N\}$ be the orthonormal set of eigenvectors corresponding to $\lambda_1, \cdots, \lambda_N$, we have

(a) $\lambda_i = \text{var}(\langle x, \phi_i \rangle)$;

(b) $\text{tr}(C_x) = \lambda_1 + \cdots + \lambda_N = E(\|x\|^2)$;

(c) $\lambda_{m+1} + \cdots + \lambda_N = \min_{S_m} E(d(x, S_m)^2)$, $m = 1, \cdots, N-1$.

The first assertion here is that λ_i is the variance of the ith *principal component* of x; these random variables occur as the coefficients in the expansion of x in the (Karhunen–Loève) basis $\{\phi_1, \cdots, \phi_N\}$. Statement (b) is a special case of (c) (take $m = 0$, there). The final assertion is that the best mean square approximation to x by m-dimensional subspaces S_m occurs when S_m is spanned by $\{\phi_1, \cdots, \phi_m\}$, with error as the indicated function of the eigenvalues. For applications of these and related formulas to multivariate statistics, pattern recognition, and signal processing (estimating x from noisy observations), see, respectively, [1], [18], and [33].

From now on we will make a slight specialization by assuming that all components of the random vector x have the same variance, which we take to be unity. It follows that the diagonal of C_x consists of ones, $\text{tr}(C_x) = N$, and the modulus of each off-diagonal entry c_{ij} satisfies $|c_{ij}| < 1$. These entries are, in fact, the correlation coefficients of the i and j components of x. Any such matrix is called a *correlation matrix*, and will be our primary object of study. Bounds and estimates for various quantities associated with such matrices are reviewed in Appendix A. Here we note that if C is any $N \times N$ correlation matrix, then $1 \leqq \|C\| = \lambda_1 \leqq \lambda_1 + \cdots + \lambda_N = N$, so that the set $\Gamma(N)$ of all such C is a bounded convex subset of the $N(N+1)/2$-dimensional real space of $N \times N$ Hermitian matrices. (If the scalars are complex, this latter space is of real dimension N^2.)

In general, it is difficult to tell by inspection whether a given symmetric or Hermitian matrix C with diagonal entries equal to one is positive definite, and hence a correlation matrix. Several nonlinear inequalities involving the off-diagonal entries must be satisfied; these correspond to the positivity of the leading principal minors of C.

Two simple sufficient conditions for positive definiteness are available, however. These are:

(a) C is diagonally dominant, so that the Gershgorin theorem can be applied; and

(b) C can be partitioned as

$$C = \begin{bmatrix} I_1 & F \\ F^* & I_2 \end{bmatrix}$$

where the I's are identity matrices, and F is a matrix whose (spectral) norm is less than one.

1.2. Notions of randomness. We now want to address the question of randomly selecting a correlation matrix of some fixed size. Our particular interest in this question has already been indicated in the introductory remarks above, and further motivation will be provided in the next section; in general, we may say simply that a satisfactory

answer to this question will permit generation of random test problems for a variety of statistical methods.

Roughly speaking, any method of generating random correlation matrices will begin by generating some number of pseudorandom variates uniformly distributed on the unit interval, and then performing certain deterministic mathematical steps to arrive at a correlation matrix. Four possible such methods will be described below, and two will be discussed at some length. But we have to acknowledge at the outset that no method is completely satisfactory. This is due to the lack of structure of the set $\Gamma(N)$, on the one hand, and to the presence of structure in the individual members of $\Gamma(N)$, on the other hand. That is, each element C of $\Gamma(N)$ has associated with it, as a matrix, entries, eigenvalues, and functions of these, such as norm, condition number, etc., all of which become random variables with their own distributions, which naturally depend on the manner in which C was produced. But the set $\Gamma(N)$ does not carry a natural invariant measure. This deficiency may be contrasted with the cases of the orthogonal or unitary groups, which carry a canonical (unimodular) Haar measure. Nor is $\Gamma(N)$ a homogeneous space, such as a sphere, on which a transformation group acts and leaves invariant some measure. Thus, while such phrases as "random orthogonal matrix" and "uniform distribution over the unit sphere S^{N-1}" have a clear conceptual meaning, and indeed there exist successful numerical procedures for generating such variates (see, in particular, [2], [57] for the former case), the situation remains murky for correlation matrices.

As a brief aside we offer two remarks. First, the topics of approximating and efficiently sampling from the uniform (Haar) distribution on finite or compact groups persist under current investigation. In addition to the references just given for the case of the orthogonal group, see recent articles by Takacs [60] for finite groups, and by Diaconis and Shahshahani and various coauthors ([11] and references) for an assortment of groups and applications. The basic approximation result, that the distributions of the successive terms in a random walk on a compact group converge vaguely to normalized Haar measure provided that the support of the common distribution of the individual terms is sufficiently diffuse, goes back at least to Grenander [24]. The condition on the support of the distribution can also be rephrased as a spectral property of its (operator-valued) Fourier transform. Second, although as noted above, $\Gamma(N)$ is not a homogeneous space, the cone $P(N)$ of all positive definite $N \times N$ Hermitian matrices is such a space. Namely, it is acted on by the general or special linear groups according to the rule

$$A \leftrightarrow TAT^*,$$

for $A \in P(N)$ and T nonsingular. It follows from general theory, including the fact that these linear groups are unimodular, that there is an invariant measure on $P(N)$ [44].

Returning now to the matter of random correlation matrices, we indicate four possible methods of generation; only the last two will be discussed in any detail, beginning in the next chapter.

Method 1. Direct acceptance–rejection. Here we must obtain symbolically the leading principal minors of the general symmetric $N \times N$ matrix with unit diagonal. This is possible for moderate size N via a computer algebra program. Requiring these minors to be positive then constitutes a set of $N - 2$ nonlinear constraints on the $N(N-1)/2$ off-diagonal entries. Then a set $C_{12}, \cdots, C_{1N}, C_{23}, \cdots, C_{2N}, \cdots, C_{N-1,N}$ of pseudorandom deviates uniformly distributed on $(-1, 1)$, or in the open unit disc, is generated and tested to see if the constraints are satisfied. If so, a correlation matrix C is obtained; if not, a new set of uniform deviates is generated, etc. To our knowledge, the distributional aspects of the spectral features of the resulting matrices are unknown.

As might be expected, this method is at best feasible for rather small values of N, say $N \leqq 6$. Indeed, based on 1,000 N trials, the empirical rejection rate $r_N = .195, .583, .875$, and .977, respectively, for $N = 3, 4, 5, 6$. Hence, the expected number of iterations until the algorithm succeeds in producing a correlation matrix is $(1 - r_N)^{-1} = 1.24, 2.4, 8.0$, and 45.5, for these values of N.

For the general theory of acceptance–rejection methods, reference may be made to Devroye [10].

Method 2. Perturbation about a mean. This method is discussed by Marsaglia and Olkin [40], which is generally the most current source of information about our subject. However, it is not of particular interest to us as, for our purposes, there is no reason to have in mind any a priori mean value.

Method 3. Random spectrum. As we know, the spectrum of an $N \times N$ correlation matrix consists of N positive numbers (not necessarily distinct) that sum to N. As will be recalled in the next section, every such set of N numbers occurs, so that the possible spectra fill out a simplex in real N-space. Since it is numerically feasible to generate pseudorandom uniform samples from this simplex, we can, by a succession of suitably chosen orthogonal or unitary transforms, arrive at a random correlation matrix. An automated procedure for doing this latter task is commercially available in the IMSL subroutine GGCOR. Statistical aspects of this method will be discussed at some length in § 3.

Method 4. Random Gram matrix. As is well known, every real positive-definite matrix A has a Cholesky factorization

$$A = TT^*,$$

where T is a uniquely defined lower triangular matrix with positive diagonal entries. Let the rows of T be denoted $t_1, \cdots, t_N$. Then

$$a_{ij} = \langle t_i, t_j \rangle,$$

and so A can be considered a Gram matrix defined by the vectors $\{t_1, \cdots, t_N\}$. If also A is a correlation matrix, then each vector t_i must have length one. Consequently, any procedure for generating pseudorandom unit vectors, with any distribution, will result in a random correlation matrix of Gram type. These vectors may or may not end in zeros, as in the Cholesky factorization, but naturally we do less work if they do. This method is the most efficient of the general methods, 1, 3, and 4; some of its statistical aspects will be discussed in § 4 (see also [40] again).

Whatever method we might eventually choose for a particular application will depend on the nature of the application and just which aspect of the random correlation matrices we wish to have an unambiguous uniform distribution.

1.3. Background and motivation. In terms of the preceding introductory material, and prior to the more technical developments of the remaining chapters, we will briefly review some of the relevant statistical literature. Specifically, we will comment on the contents of four articles, [4], [3], [31], [40], listed in chronological order.

In [4], Chalmers (1975) presents an algorithm which produces correlation matrices with a common spectrum. His motivation is the study of strongly structured data, that is, random vectors whose first two or three principal components explain much of the variability of the data. He wants to be able to distinguish between causes of the observed associations among different subsets of the components of the data, whether these are due to the physical nature of the variables themselves, or to some inherent structure in the data as captured by the underlying principal components. An empirical approach is

to generate other correlation matrices with eigenvalues identical to those observed, and to then compare results from these matrices with those from the original data. The algorithm itself is derived from a geometric lemma which asserts the existence of an infinite set of orthogonal generators to certain quadratic cones in real n-space. Normalizing these generators then leads to the columns of an orthogonal matrix which transforms a given diagonal matrix of eigenvalues into the desired correlation matrix.

In [3] Bendel and Mickey address the same problem as Chalmers, but more systematically, and with more concern for whether the resulting correlation matrices are truly "representative" of the entire class of correlation matrices with given spectrum, thought of as those which could arise from a given experiment. They note that parameterizing subsets of $\Gamma(N)$ by structure, e.g., equi-interclass correlation (constant off-diagonal entries) or first-order autoregressive (Markov-1 data) leads to very narrow classes of correlation matrices, unsuitable for many applications. Their approach is to treat the eigenvalues as parameters, especially when they, in turn, are functions of one or two parameters. For example, the eigenvalues might be required to form a geometric progression. In general, if the parameterized eigenvalues are roughly constant, and therefore approximately equal to one, the data variables are approximately independent, while a large spread in the range of the eigenvalues indicates strong interdependence between the variables.

Starting with a spectrum $\{\lambda_1, \cdots, \lambda_N\}$ and setting $D = \operatorname{diag}\{\lambda_1, \cdots, \lambda_N\}$, the method of Bendel and Mickey yields a correlation matrix C of the form

$$C = U^* D U,$$

where $U = V R_{N-1} \cdots R_2 R_1$. Here V is a randomly chosen orthogonal (or unitary) matrix and the R's are matrices representing Givens rotations, chosen successively to make one diagonal entry at a time of the product equal to one. The V's can be generated by various procedures (see references [2], [57] already mentioned in § 1.2). They go on to describe the application of their method to the problem of stopping rules in the statistical procedure known as stepwise regression. They also offer some comparisons between their method and that of Chalmers.

In [31], Johnson and Welch (1980) also emphasize the use of simulated data to test alternative selection procedures in stepwise regression, particularly to build confidence in the use of such procedures on real data with uncertain structure. If the joint distribution of the dependent and regressor variables is Gaussian, then it is standard to sample randomly from it by factoring the covariance matrix and using a string of pseudorandom $N(0, 1)$ variates. Thus only the structure of the distribution remains to be specified, and this, of course, is completely determined by the (mean and) covariance. If this is assumed, as it is, to be of correlation type, then it can be partitioned as

$$\begin{bmatrix} 1 & \rho^* \\ \rho & C \end{bmatrix},$$

where C is the intercorrelation matrix of the regressors, and ρ is the vector of correlation coefficients between the regressors and the dependent variable. So the emphasis is on generating such C's, and this is done by viewing C as a Gram matrix: $C = TT^*$, with the rows of T being unit vectors. They suggest generating each entry of T from a symmetric beta distribution, varying the free parameter from row to row. They note that a certain control over some aspects of the matrices so defined can be maintained, such as the degree of correlation between some regressors, and the coefficient of determination for the complete regressor set.

Finally, in [40], Marsaglia and Olkin give a rigorous mathematical description of Methods 2–4 described in the preceding section. Their major result is to obtain the explicit form of the distribution of the entries of a random Gram correlation matrix $C = TT^*$, when the entries of T are generated in a particular fashion. Some of this work will be reviewed later, in the appropriate context.

2. Two principal methods. As noted in § 1.2, only the methods labeled there as Methods 3 and 4 are discussed in any detail here. We begin this discussion now, setting the stage for the presentation of the new results later.

2.1. Random spectrum. As we know, the spectrum of a correlation matrix $C \in \Gamma(N)$ has a spectrum $\sigma(C) = \{\lambda_1, \cdots, \lambda_N\}$ consisting of N positive numbers of sum N. The set of all such N-tuples defines a simplex S_N, and we first want to observe that every point in S_N occurs in this fashion, that is,

$$U\{\sigma(C): C \in \Gamma(N)\} = S_N.$$

This is a consequence of a general characterization of Hermitian matrices due to several authors. Namely, if A is a Hermitian matrix of order N with eigenvalues $\lambda_1 \geqq \cdots \geqq \lambda_N$, and diagonal entries $d_1 \geqq \cdots \geqq d_N$, then

$$d_1 + \cdots + d_k \leqq \lambda_1 + \cdots + \lambda_k \tag{2.1}$$

for $1 \leqq k \leqq N - 1$, and

$$d_1 + \cdots + d_N = \lambda_1 + \cdots + \lambda_N = \operatorname{tr}(A); \tag{2.2}$$

(see [50]). Conversely, given real numbers $\{d_i, \lambda_i\}$ satisfying all these conditions, there exists a real symmetric matrix A with diagonal entries $d_1, \cdots, d_N$, and eigenvalues $\lambda_1, \cdots, \lambda_N$ (Horn [29], Mirsky [43]; in contemporary terminology, the vector of diagonal entries is *majorized* by the vector of eigenvalues of a Hermitian matrix [42]). In our case, however, the result follows more directly from a theorem of Fillmore [16], namely, that any matrix A is unitarily equivalent to one with a constant diagonal. This, in turn, is an easy consequence of the convexity of the numerical range $W(A)$, so that $\operatorname{tr}(A)/N \in W(A)$, and an induction argument.

The upshot of the above paragraph is that, given $(\lambda_1, \cdots, \lambda_N) \in S_N$, there is a correlation matrix C with these numbers as its spectrum. How, in practice, is such a matrix to be obtained? As already noted, answers have been given by Chalmers and Bendel and Mickey; there is also the paper by Chan and Li [5] which more generally provides an algorithm for constructing a real symmetric matrix with given diagonal entries and eigenvalues satisfying the conditions (2.1) and (2.2). It appears that for present purposes the most natural is that proposed by Bendel and Mickey, namely,

$$C = R_{N-1}^* \cdots R_2^* R_1^* D R_1 R_2 \cdots R_{N-1}, \tag{2.3}$$

where $D = \operatorname{diag}[\lambda_1, \cdots, \lambda_N]$, and R_k is a rotation in the plane spanned by the standard unit vectors l_k and l_{k+1}. The matrix R_k has the form

$$\begin{bmatrix} I_{k-1} & & & \\ & c & s & \\ & -s & c & \\ & & & I_{N-k-1} \end{bmatrix}$$

with $c^2 + s^2 = 1$. The rotation angle $\arccos(c)$ is chosen so that the kth diagonal entry of C is one.

We can strengthen the preceding remark by replacing the diagonal matrix D in (2.3) by $A = U^*DU$, U unitary. That is, A is an arbitrary positive-definite matrix with spectrum $\{\lambda_1, \cdots, \lambda_N\}$. Then A can still be transformed into a correlation matrix C, as before, and there are, in fact, exactly four choices for each R_k in (2.3).

To see this, consider first the specification of R_1. $R_1^* A R_1$ should have a diagonal entry equal to 1. Let A_p be a principal 2×2 submatrix of A, say

$$A_p = \begin{bmatrix} a & b \\ \bar{b} & d \end{bmatrix},$$

with $a, d > 0$. Let Q be a 2×2 orthogonal matrix, either a rotation

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix}$$

or a reflection

$$\begin{bmatrix} c & s \\ s & -c \end{bmatrix}.$$

Then the upper left entry of $Q^* A_p Q$ is $ac^2 + 2 \operatorname{Re}(b)cs + ds^2$, respectively. From the behavior of the Rayleigh quotient of A_p, we see that this quadratic form in (c, s) will somewhere assume the value one if and only if A_p has one eigenvalue less than or equal to one and the other greater than or equal to 1. Now since a, d are diagonal entries of A, and $\operatorname{tr}(A) = N$, A_p can be chosen so that one of a, d is less than or equal to 1, and the other greater than or equal to one. Its eigenvalues $\lambda_1 \geqq \lambda_2$ then satisfy

$$\lambda_2 = \min \langle A_p x, x \rangle \leqq \min(a, d)$$

$$\leqq 1 \leqq \max(a, d) \leqq \max \langle A_p x, x \rangle = \lambda_1.$$

So the condition on A_p is satisfied, and therefore four choices of Q exist to yield a 1 in the upper left corner of $Q^* A_p Q$. R_1 is then created immediately as a direct sum of Q and an identity matrix. The whole procedure can then be repeated, since the sum of the remaining diagonal entries of $R_1^* A R_2$ is $N - 1$.

The preceding remarks are a slight elaboration of some made at the end of § 4 of [40]. We note that in all the discussion of this section so far, there are no issues of randomness. These can be introduced in two stages. First, if a point $(\lambda_1, \cdots, \lambda) \in S_N$ is given, we can form the corresponding diagonal matrix D, select an orthogonal matrix V at random from the orthogonal group $O(N)$ with normalized Haar measure, and select a succession of orthogonal matrices, one of four choices at random at each step, so as to transform $V^* D V$ into a correlation matrix C. This C may fairly be said to be a random correlation matrix with specified spectrum. Second, the spectrum may itself be chosen from some probability distribution on S_N. The resulting matrices are said to have a *random spectrum*. The special case where the distribution of S_N is uniform will be discussed at some length in § 3. Some issues involved here are that this method is evidently rather computationally expensive, and that the distribution of the entries of the resulting correlation matrices is not well understood. However, we will be able to say something about the distribution of some global features of these matrices.

2.2. Random Gram matrix. We first quickly review the concept of Gram matrix. Let $x_1, \cdots, x_N$ be linearly independent vectors in any pre-Hilbert space. The corresponding *Gram matrix* is the $N \times N$ Hermitian matrix $G = G(x_1, \cdots, x_N)$ with (i, j)-entry $= \langle x_i, x_j \rangle$. The determinant $g(x_1, \cdots, x_N)$ is called the *Gramian* of the

set $\{x_1, \cdots, x_N\}$. Clearly, the covariance matrix of a set of normalized second-order random variables falls under this definition. In general, Gram matrices are positive semidefinite as follows from the formula

$$\langle G\alpha, \alpha \rangle = \| \sum \bar{\alpha}_i x_i \|^2 \geqq 0,$$

for any N complex numbers $\alpha_1, \cdots, \alpha_N$. Furthermore, as already noted in § 1.2, by Cholesky factorization, any real positive-definite matrix is a Gram matrix; more generally, any complex positive-semidefinite matrix has a positive-semidefinite square root, and so is a Gram matrix.

The Gramians are symmetric functions of their arguments, and obey the inequalities

$$0 \leqq g(x_1, \cdots, x_N) \leqq \| x_1 \|^2 \cdots \| x_N \|^2, \tag{2.4}$$

with equality on the left if and only if $\{x_i\}$ is linearly dependent, and equality on the right if and only if $\{x_i\}$ is orthogonal. To prove the right-hand inequality we first reduce to the case that each x_i is a unit vector, and then

$$\begin{aligned} g(x_1, \cdots, x_N)^{1/N} &= (\Pi \lambda_i)^{1/N} \\ &\leqq \frac{1}{N} \sum \lambda_i = \frac{1}{N} \operatorname{tr}(G) = 1, \end{aligned}$$

where $\{\lambda_i\} = \sigma(G)$.

We sense from this that the Gramian and other spectral features of the Gram matrix bear some relation to the relative orientation of the vectors $\{x_i\}$. Along this line we recall that if the vectors x_i belong to R^N, then

$$\operatorname{vol}\left(\left\{ \sum_1^N \lambda_i x_i : 0 \leqq \lambda_i \leqq \varepsilon_i \right\} \right) = g(x_1, \cdots, x_N)^{1/2} \prod_1^N \varepsilon_i, \tag{2.5}$$

so that, in particular, $g(x_1, \cdots, x_N)$ is the square of the volume of the parallelepiped spanned by the set $\{x_i\}$. If the x_i belong to some other space, (2.5) serves to define this volume.

In addition to the simple Hadamard inequality in (2.4), we have further

$$g(x_1, \cdots, x_m, y_1, \cdots, y_N) \leqq g(x_1, \cdots, x_m) g(y_1, \cdots, y_N),$$

and, in fact, the ratio of the left side to the right side is known to be $\sin^2 \alpha_1 \cdots \sin^2 \alpha_M$, where $\alpha_1, \cdots, \alpha_M$, $M \leqq N$, are the angles between span $\{x_i\}$ and span $\{y_j\}$.

Gram matrices occur naturally in all manner of least squares problems, such as Gram–Schmidt orthogonalization, linear regression, and pseudoinversion. Indeed, the basic problem of computing the orthogonal projection onto span $\{x_i\}$ requires the solution of a linear system with Gram coefficient matrix. It is familiar that as the basis vectors x_i deviate more from orthogonality, such problems become more ill-conditioned, and associated statistical procedures are said to suffer from "collinearity." For example, if $x_i(t)$ is the monomial t^i, and the inner product comes from some Lebesgue–Stieltjes measure with compact support, then the corresponding Gram matrices, indexed by N, have a condition number that grows at least as fast as 4^N; the classical Hilbert matrices are special cases ([61], the main result of this reference is reviewed in Appendix B; for a recent review of the collinearity problem with suggestions for its measurement by more subtle indicators than simply condition number, see [58]).

TABLE 1
Statistics for random $N \times N$ matrices.

	$N = 5$	10	$N = 5$	10
Mean c.n.	111.	553.	1.37E6	7.85E6
Standard dev.	754.	1.18E4	4.29E8	2.19E8
Median c.n.	15.8	40.5	114.	809.
Interquartile range	30.7	80.9	446.	3533.
Trimmed mean	39.2	95.4	1.57E3	1.57E4
Standard dev.	70.3	157.	5.48E3	6.65E4
Mean F norm	2.88	4.24	2.99	4.36
Mean norm	2.31	2.92	2.43	2.94
Mean min. e-value	.191	.100	.042	.009
Standard dev.	.161	.093	.054	.012
	Random spectrum		Random Gram	

As a reference point for later use, we record here a well-known distance formula involving Gramians. Let $M = \text{span}\ \{x_1, \cdots, x_N\}$, and let x be another point in the space containing M. Then

$$\text{dist}\,(x, M)^2 = \frac{g(x_1, \cdots, x_N, x)}{g(x_1, \cdots, x_N)}. \tag{2.6}$$

Recall that random Gram matrices were defined by Method 4 in § 1.2. In present notation the x_i are taken to be random vectors uniformly distributed over the sphere S^{N-1} in R^N. We now report some results from a small simulation, intended to compare such matrices with those of random spectrum (Method 3). We give here only the cases $N = 5$ and 10, as they appear typical of all cases considered. In each case, the summary statistics are based on 1,000 trials. In the left column of Table 1, c.n. means condition number, F norm mean Frobenius norm, and norm means spectral norm. Also, trimmed means that the largest one percent and smallest one percent of the samples have been deleted.

Probably the most striking contrast to be made on the basis of this numerical experiment is the much higher condition numbers of the random Gram matrices relative to those of the matrices with random spectrum. This aspect of the data persists even after trimming, and after passing to medians. It strongly suggests that random Gram matrices do not have a random spectrum. It also raises interesting questions about the relative orientation of a batch of two or more vectors drawn independently from the uniform distribution on the $(N - 1)$-sphere. Some of these will be considered in § 4 below.

3. Correlation matrices with random spectrum. Some background for the chapter was given in § 2.1. There, two essential steps in this process were recognized:

—Pick a point $\underline{\lambda} = (\lambda_1, \cdots, \lambda_N)$ "at random" from the simplex S_N, and form the diagonal matrix $D = \text{diag}\,(\lambda_1, \cdots, \lambda_N)$.

—Construct a matrix $C = R^*V^*DVR$, where V is drawn at random from the orthogonal group $O(N)$, and R is a product of randomly selected rotation/reflection matrices, chosen to successively put ones on the diagonal of C.

Naturally, the second step leaves $\sigma(C) = \underline{\lambda}$, and hence leaves all spectral functions of C unchanged. Among such functions are the spectral and Frobenius norms of C, and its condition number. (In general, any unitarily invariant norm of a Hermitian matrix is a

spectral function of that matrix.) Consequently, once a probability distribution is selected on S_N, so as to define the "at random" condition of the first step, the spectral functions of any correlation matrix defined by the second step may be studied directly. Note that this approach does not apply to other numerical attributes of C of possible interest, such as its individual entries; their distribution naturally depends in part on those of the V and R matrices.

3.1. Method of generation. From the preceding discussion we see that a probability distribution μ must be specified on the simplex S_N. Then a sample $\underline{\lambda}$ drawn from μ will be a random spectrum. Having no reason to weight any region of S_N more than another, we will take μ to be the uniform distribution on S_N, and denote by

$$\underline{\lambda} \sim U(S_N)$$

a point $\underline{\lambda}$ so chosen. Two tasks remain:

—Specify μ analytically,

—Specify μ operationally.

This latter task simply means to prescribe a method for a computer to make these draws in terms of an assumed capability to generate pseudorandom numbers $\sim U[0, 1]$.

The analytical specification of μ depends on (a special case of) the general theory of order statistics and spacings. Here we only need the case of independent samples from the uniform distribution. Thus let $u_{(1)} \leqq \cdots \leqq u_{(N-1)}$ be the order statistics of a random sample from $U([0, 1])$. Define $u_{(0)} = 0$, and $u_{(N)} = 1$. As shown by Wilks [66], the joint distribution of these order statistics is uniform over the simplex $\{0 \leqq x_1 \leqq \cdots \leqq x_{N-1} \leqq 1\}$ in R^{N-1}. Then the *spacings* of the sample are defined by

$$s_i = u_{(i)} - u_{(i-1)}, \qquad 1 \leqq i \leqq N.$$

Observe that for each sample, the spacings are positive numbers that sum to one. It was also shown by Wilks that the spacings vector $(s_1, \cdots, s_{N-1})$ is uniformly distributed over the simplex

$$\left\{0 \leqq x_i, \sum_1^{N-1} x_i \leqq 1\right\} \quad \text{in } R^{N-1}.$$

Now, for fixed N, the mapping

$$T(x_1, \cdots, x_{N-1}) = (x_1, \cdots, x_{N-1}, 1 - x_1 - \cdots - x_{N-1})$$

carries this simplex bijectively onto the simplex $\{0 \leqq x_i, \sum x_{i=1}\}$ in R^N, and carries over the uniform density (by an elementary change of variables).

The upshot of the preceding paragraph is that, for fixed N, the uniform density μ on the simplex S_N can be specified analytically as the distribution of the random vector

$$NT(s_1, \cdots, s_{N-1}). \tag{3.1}$$

And, of course, it follows that μ can be specified operationally in terms of pseudorandom number generator, and a sorting routine.

There is by now a fairly extensive literature of spacings (sometimes known as "gaps," "coverages," or "random division of an interval"). This topic can be traced back to the work of Whitworth [64] at the turn of the century on the distribution of the largest spacing. His result was utilized by Fisher [17] to give a significance test for the largest amplitude in a numerical harmonic analysis of a time series. (In fact, Fisher's test turns out to be the uniformly most powerful symmetric invariant decision procedure against simple periodicities. More recent work is concerned with compound periodicities, and

hence with the distribution of other functions of the spacings besides the maximum. It is not germane to detail any of this work here; the interested reader may consult some papers of Siegel, e.g., [51]. We merely want to draw attention to the unexpected link between the spacings concept and time-series analysis.)

In the later 1930s and then in the 1940s other work on distributions of functions of spacings was done by several authors: Levy, Greenwood, Moran, etc. Most of this originated as specific problems in applied statistics. The best review of all this is that of Pyke [46]. Other useful references are [8] and [55]. Among (many) other things, these references point out alternative analytical specifications of spacings. For instance, if $y_1, \cdots, y_N$ are independently exponentially distributed with arbitrary mean, and $z = y_1 + \cdots + y_N$, then the random vector

$$z^{-1}(y_1, \cdots, y_N)$$

is distributed as the spacings vector $T(s_1, \cdots, s_{N-1})$. Hence random spectra can also be generated by use of exponential variates. From this it follows that spacings can also be simulated from the (normalized) interarrival times of a Poisson process $\{N(t): t \geqq 0\}$ with $N(0) = 0$. Namely, if T_k is the elapsed time between the $(k-1)$st and the kth event, and $t > 0$ is fixed, then the conditional distribution, given $N(t) = n$, of

$$t^{-1}(T_1, \cdots, T_{n-1}, t - T_{N(t)})$$

is the same as the spacings vector. This is a classic construction of spacings with important modern applications to the limiting behavior of empirical processes [46].

The distribution of spacings and some functions thereof is also briefly discussed by Kendall and Moran in [34]. Naturally, geometric aspects are stressed. For instance, the joint distribution of the spacings is, with proper scale factor, exactly that of the lengths of the N perpendiculars from a random point inside the simplex S_N to the N sides. The authors go on to discuss some situations where probabilities can be computed from simplicial geometry.

3.2. Distribution of eigenvalues. Pursuant to the foregoing discussion we take as a random spectrum $\underline{\lambda} \in S_N$, N times the random vector of spacings defined by a random sample of size $N-1$ from the standard uniform distribution. We denote this vector as $\underline{\lambda} = (\lambda_1, \cdots, \lambda_N)$. In this short section we discuss the distribution of the λ_i, while in the next two sections we consider that of certain functions of the λ_i related to correlation matrices C with $\sigma(C) = \underline{\lambda}$.

We first note that the λ_i are exchangeable random variables since, because of the uniform distribution of $\underline{\lambda}$ on S_N, that distribution is unchanged under permutation of its components. It follows that the λ_i are identically distributed and, using the formula for the least order statistic, the distribution function F_N is given by

$$F_N(t) = 1 - \left(1 - \frac{t}{N}\right)^{N-1}. \tag{3.2}$$

Thus, for large N, λ_i is approximately exponentially distributed with mean one. From (3.2) we can conclude that

$$E(\lambda_i) = 1, \qquad \operatorname{var}(\lambda_i) = \frac{N-1}{N+1},$$

for each i.

Expressions for the joint distribution of the λ_i have been given by Steutel [55].

Namely,

$$\Pr(\lambda_1 > \alpha_1, \cdots, \lambda_N > \alpha_N) = \begin{cases} \left(1 - \frac{1}{N}\sum \alpha_i\right)^{N-1}, & \sum \alpha_i < N, \\ 0 \quad \text{if not} \end{cases}$$

and

$$\Pr(\lambda_1 \leqq \alpha_1, \cdots, \lambda_N \leqq \alpha_N) = 1 - \sum_{j=1}^{N} \left(1 - \frac{\alpha_j}{N}\right)^{N-1} + \sum_{1 \leqq j < k \leqq N} \left(1 - \frac{\alpha_j + \alpha_k}{N}\right)^{N-1} - + \cdots.$$

These formulae are derived by Laplace transform techniques and the relation, already alluded to in § 3.1, between the spacings distribution and that of certain exponential variates.

In similar fashion we could go on to describe the joint distribution of pairs (λ_i, λ_j), the associated covariance, etc. Here we will just note that

$$\operatorname{corr}(\lambda_i, \lambda_j) = \frac{-1}{N-1}.$$

But actually all such formulae of likely interest follow directly from the *multiple moments formula*

$$E(\lambda_1^{p_1} \cdots \lambda_N^{p_N}) = N^p \Gamma(N) \frac{\Gamma(p_1+1) \cdots \Gamma(p_N+1)}{\Gamma(p+N)}, \tag{3.3}$$

where $p = p_1 + \cdots + p_N$. This expression can either be derived by the Laplace transform method of Steutel [55], or, somewhat more directly and geometrically, as in Kendall and Moran [34, p. 34].

3.3. Distribution of norms. We continue with the assumption that we are dealing with a correlation matrix C whose spectrum $\underline{\lambda}$ has been chosen randomly according to $U(S_N)$. The issue now is the distribution of the norms $\|C\|$ and $\|C\|_F$, as defined in Appendix A.

Let us begin with $\|C\|_F^2 = \sum \lambda_i^2$ which, for both typographical and historical reasons, we will denote by $\mathrm{GM}(N)$. In the statistical literature this quantity is known as the Greenwood–Moran statistic, after the authors of [23] and [44]. It was originally proposed as a test for uniformity in response to a problem in epidemiology (time intervals between outbreaks of an infectious disease). Moran [44] derived a general formula for the moments of $\mathrm{GM}(N)$; it was rederived by Steutel [55]. For us, it is enough to use the moments formula (3.3) to obtain

$$E(\mathrm{GM}(N)) = \frac{2N^2}{N+1},$$

$$E(\mathrm{GM}(N))^2 = \frac{4N^4(N+5)}{(N+1)(N+2)(N+3)}$$

and hence

$$\operatorname{var}(\mathrm{GM}(N)) = \frac{4N^4(N-1)}{(N+1)^2(N+2)(N+3)}$$

$$= O(N).$$

A second point to be made about $\mathrm{GM}(N)$ is that it is (slowly!) asymptotically normal, a result due originally to Moran [44], and reproved by a more general method by Darling [7] (see also [46], [55]). As usual, Pyke has the most complete discussion of this topic. Once this asymptotic normality is established, the corresponding property of $(\mathrm{GM}(N))^{1/2} = \|C\|_F$ can be worked out by general theory concerning smooth functions of asymptotically normal variates. In fact, since $\mathrm{GM}(N)/2N$ has mean $N/(N+1) \approx 1$, and variance $\sigma_N^2 = O(1/N)$, its asymptotic normality implies that $(\mathrm{GM}(N))^{1/2}/2N$ is asymptotically normal with mean one, and variance $\sigma_N^2/4$. Hence $(\mathrm{GM}(N))^{1/2}$ is asymptotically normal with mean $\sqrt{2N}$, and variance $\frac{1}{2}N\sigma_N^2 = \frac{1}{2}$.

Next we consider the spectral norm $\|C\|$, for an $N \times N$ correlation matrix C with random spectrum as usual. Since $\|C\| = \lambda_{\max}$, the maximum eigenvalue of C, the distribution of $\|C\|$ is that of N times the maximum spacing determined by a random sample of $N-1$ points from the standard uniform distribution. We let V_N denote this maximum spacing, so that $\lambda_{\max} = \|C\| = NV_N$.

As already noted in § 3.1, the distribution of V_N goes back to Whitworth [64], and has a history of interesting statistical applications. A convenient source for this distribution is [7], wherein we can also find a derivation of the asymptotic behavior, due originally to Lévy [37]. We find that

$$\Pr(NV_N < x) = \sum_{k=0}^{N} (-1)^k \binom{N}{k} \left(1 - \frac{kx}{N}\right)_+^{N-1},$$

where $(t)_+ = \max(t, 0)$. From this we could derive the mean and higher moments, as needed. As a somewhat neater alternative, we can appeal to some well-known relations between the distribution of the spacings and certain exponential variates, as briefly reviewed in § 3.1. Now making use of the fact that the sum of exponential variates y_i is gamma-distributed, and the known distribution of the order statistics from the exponential distribution, we can obtain

$$NV_N \approx N \max\{y_i\}/z.$$

Also, a formula was given by Devroye [10]:

$$V_N = \left(\sum_1^N y_i/i\right) \Big/ z.$$

In both cases $z = y_1 + \cdots + y_N$. From all this we can deduce that

$$\begin{aligned} E(\lambda_{\max}) = E(NV_N) &= 1 + \frac{1}{2} + \cdots + \frac{1}{N} \\ &\approx \log N + \gamma, \end{aligned}$$

where $\gamma = .577\cdots$ is Euler's constant.

Finally, the Levy–Darling asymptotic formula for the maximum spacing, scaled to apply to the maximum eigenvalue of the matrix C is

$$\Pr(NV_N < \log N + x) \to \exp(-e^{-x}),$$

as $N \to \infty$. From this, it follows that

$$\operatorname{var}(NV_N) \to \pi^2/6,$$

as $N \to \infty$.

Thus we have obtained the exact means of the norm functions $\|C\|_F^2$ and $\|C\|$, and the asymptotic behavior of these, along with $\|C\|_F$, as $N \to \infty$. In particular, we have observed that the Frobenius norm tends to normality, while the spectral norm tends to obey an extreme value distribution.

3.4. Condition number expectation. We continue to study an $N \times N$ correlation matrix C with random spectrum. Now our focus is on the distribution of the condition number $\kappa(C)$, as defined by (B1). As we know from (B3), $\kappa(C) = \lambda_{\max}/\lambda_{\min}$, the ratio of the largest to the smallest eigenvalue of C. We have just described the distribution of $\lambda_{\max} = \|C\|$. In fact, the *joint* distribution of $(\lambda_{\max}, \lambda_{\min})$ can be inferred from the work of Darling [7], in the following form:

$$\Pr(\lambda_{\min} > x, \lambda_{\max} < y) = \sum_{j=0}^{N} \binom{N}{j} (-)^j \left(1 - x\left(\frac{N-j}{N}\right) - y\frac{j}{N}\right)_+^{N-1}.$$

From this, by letting $y \to N$, we obtain the distribution for the least eigenvalue:

$$\Pr(\lambda_{\min} > x) = (1-x)^{N-1}, \qquad 0 < x < 1. \tag{3.4}$$

This formula yields the moments of $\lambda_{\min}$ as

$$E(\lambda_{\min}) = \frac{1}{N}, \qquad \operatorname{var}(\lambda_{\min}) = \frac{N-1}{N^2(N+1)}.$$

We might pause here to collect the formulas giving the expected behavior of the eigenvalues, and their important functions, as a function of N, for $N \times N$ correlation matrices with random spectrum. Namely, we have seen that

$$E(\lambda_i) = 1, \qquad i \leqq \lambda_i \leqq N,$$

$$E(\lambda_{\max}) \approx \log N + \gamma,$$

$$E(\lambda_{\min}) = 1/N,$$

$$E(\textstyle\sum \lambda_i^2) \approx 2N.$$

Returning now to the joint distribution of $(\lambda_{\max}, \lambda_{\min})$, it can also be inferred from [7] that these quantities are asymptotically independent, as a consequence of the formula

$$\Pr\left(\lambda_{\min} > \frac{x}{N}, \lambda_{\max} < \log N - \log y\right) \to \exp(-x-y),$$

as $N \to \infty$. This permits us to write, for large N,

$$\begin{aligned} E(\kappa(C)) &= E(\lambda_{\max}/\lambda_{\min}) \\ &\approx E(\lambda_{\max}) E(1/\lambda_{\min}). \end{aligned}$$

However, although the first factor is finite, as we know, the second is not:

$$\begin{aligned} E\left(\frac{1}{\lambda_{\min}}\right) &= \int_0^1 \frac{1}{x}\frac{d}{dx}(1-(1-x)^{N-1})\,dx \\ &= (N-1)\int_0^1 \frac{(1-x)^{N-1}}{x}\,dx \\ &= (N-1)\int_0^1 \left(\frac{1}{x} + \cdots\right) dx = +\infty. \end{aligned}$$

This observation suggests that the condition number $\kappa(C)$ may not have a finite first moment. Additional grounds for such suspicion can be based on its validity at the other extreme case where $N = 2$. In this simple case the assertion goes as follows: if a single number s is drawn at random from the interval [0, 1], and U (respectively, V) is the min (respectively, max) of $\{s, 1 - s\}$, then the ratio V/U obeys the distribution

$$\Pr\left(\frac{V}{U} \leqq t\right) = \frac{t-1}{t+1},$$

and so has infinite expectation. This formula is derived by Feller [14, p. 24]. We now generalize this fact to the case of arbitrary N.

THEOREM. *Let C be a correlation matrix with random spectrum. Then the condition number $\kappa(C)$ has infinite expectation.*

Proof. In the notation of § 3.1 we let $0 < u_{(1)} < u_{(2)} < \cdots < u_{(N-1)} < 1$ be the order statistics of a random sample of size $N - 1$ from the standard uniform distribution. The joint distribution P of these statistics is the ordered $(N - 1)$-variate Dirichlet distribution [66, § 8.7], and is uniform over the region

$$\Omega = \{x: 0 < x_1 < x_2 < \cdots < x_{N-1} < 1\}$$

in R^N. Therefore,

$$E(\kappa(C)) = \int_{\Omega} \cdots \int \frac{\max\{u_{(1)}, u_{(2)} - u_{(1)}, \cdots, 1 - u_{(N-1)}\}}{\min\{\cdots \cdots\}} dP$$

$$= N(N-1)! \int_{T} \cdots \int \frac{\max\{\cdots\}}{u_{(1)}} du_{(1)} \cdots du_{(N-1)}$$

$$\geqq (N-1)! \int_{T} \cdots \int \frac{1}{u_{(1)}} du_{(1)} \cdots du_{(N-1)},$$

where T is the subregion of Ω defined by

$$x_1 \leqq \min\{x_2 - x_1, x_3 - x_2, \cdots, 1 - x_{N-1}\},$$

and we have used that the maximum spacing greater than or equal to $1/N$. Now, the last multiple integral over T exceeds, for sufficiently small $\varepsilon > 0$, the iterated integral

$$\int_0^{\varepsilon} \frac{dx_1}{x_1} \int_{2x_1}^{1-(N-2)x_1} dx_2 \int_{x_1+x_2}^{1-(N-3)x_1} dx_3 \cdots \int_{x_1+x_{N-2}}^{1-x_1} dx_{N-1}$$

$$= \int_0^{\varepsilon} \frac{1/(N-1)! + x_1 q(x_1)}{x_1} dx_1,$$

where q is a polynomial. This last integral is clearly divergent. □

This completes our discussion of correlation matrices with random spectrum except for the spectral distribution function, for which, see § 4.4.

4. Correlation matrices with random Gram structure. In this final chapter we discuss random Gram matrices, as defined in § 1.2, and briefly discussed in § 2.2, along with some simulation results. We will follow the same plan as in the preceding chapter, namely, generating such matrices and distribution of certain related random variables. Finally, we will make a few comparisons between the sample behavior of the two types of random matrices.

4.1. Method of generation. We recall from § 1.2 that an $N \times N$ random Gram matrix C has the form

$$C = TT^*, \tag{4.1}$$

where the rows of T are independently and identically distributed vectors distributed uniformly on the sphere S^{N-1} in R^N. That is, for each row $\underline{t}_i$ of T, we have

$$\underline{t}_i \sim U(S^{N-1}).$$

So, just as in § 3.1, the first question is how to express such random vectors in terms of standard univariate random variables.

Not surprisingly, this is a well-researched problem, with contributions dating back at least 30 years. The short paper by Marsaglia [39] has a review of this early literature, along with an improved method. More recent references are the pragmatic paper by Rubinstein [49], which also discusses the problem of § 3.1, and the extensive book of Devroye [9]. Again we distinguish between the analytic and the operational specification of $U(S^{N-1})$. The basic analytical result is that if X is a continuous radially symmetric N-dimensional random vector, then its projection on the sphere is uniformly distributed, that is,

$$X/\|X\| \sim U(S^{N-1}).$$

In particular, we can take $X \sim N(\theta, I)$, the standard spherical multivariate normal distribution. Operationally, the components of X can be generated by any of several standard pseudorandom normal variate routines. These eventually utilize pseudorandom uniform variates. The latter can also be used more directly to generate pseudorandom $U(S^{N-1})$ vectors, as is pointed out in [39], [9, Chap. V]. These are in addition to the brute force acceptance–rejection method, which tends to be very inefficient for large N ($N \geqq 5$, say). However, we will stick with the projected normally distributed random vectors.

Suppose now that we have a random vector $X \sim U(S^{N-1})$. It will be important to know how the components of X are distributed. It turns out that each

$$x_i^2 \sim \text{Beta}\left(\frac{1}{2}, \frac{N-1}{2}\right), \tag{4.2}$$

and that the density function of x_i is

$$C_N(1-t^2)^{(N-3)/2}, \qquad |t| < 1, \tag{4.3}$$

where $C_N = \Gamma(N/2)/(\Gamma(1/2)\Gamma((N-1)/2))$ is a normalizing constant. It is interesting to observe that these distributions vary considerably with dimension. In particular, we see that x_i follows an arc sine distribution when $(x_1, x_2) \sim U(S^1)$, while x_i is uniform on $(-1, 1)$ when $(x_1, x_2, x_3) \sim U(S^2)$. As N increases beyond three, the density is unimodal with an ever steeper peak at $t = 0$.

We might note here that the joint density of two or more of the x_i is also available, as a consequence of some work of Stam [54].

As a consequence of these facts we have the following geometrical lemmas: if X, Y are independently and uniformly distributed on S^{N-1}, then

$$E(\langle X, Y\rangle) = 0, \tag{4.4a}$$

$$E(\langle X, Y\rangle^2) = \frac{1}{N}, \tag{4.4b}$$

$$\text{var}\,(\langle X, Y\rangle^2) = \frac{2(N-1)}{N^2(N+2)}. \tag{4.4c}$$

Indeed, (4.4a) is a consequence of the so-called "formula of total expectation,"

$$E(f(X,Y)) = E(E(f(X,Y)\,|\,X)),$$

for scalar functions of two random vectors. The other two equations follow from realizing $\langle X, Y\rangle^2$ as the squared length of the projection of a random point in S^{N-1} on a random axis, along with standard properties of the beta distribution. This geometrical information will be used below in the next two sections.

4.2. Distribution of norms. As earlier, in § 3.3, we will study the distributional behavior of $\|C\|_F$ and $\|C\|$, but where now C is a random Gram matrix of the form of (4.1), with the rows of T uniformly distributed over the unit sphere of appropriate dimension.

The study of $\|C\|_F^2$ is greatly facilitated by the preceding results, since these imply that the square of each off-diagonal entry of C has the beta distribution of (4.2). It follows immediately that

$$\begin{aligned} E(\|C\|_F^2) &= N + 2\cdot\frac{1}{N}\cdot\frac{N(N-1)}{2} \\ &= 2N-1. \end{aligned} \tag{4.5}$$

However, a variance formula is not so immediate, as we indicate next.

We consider the second moment of $\|C\|_F^2$ about the origin, that is,

$$E(\|C\|_F^4) = E\left(\left(N + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_{ij}^2\right)^2\right). \tag{4.6}$$

Recalling that the first two moments of the beta distribution $B(a, b)$ are $a/(a+b)$ and $a(a+1)/(a+b)(a+b+1)$, respectively, we have, upon expansion of (4.6),

$$\begin{aligned} E(\|C\|_F^4) = N^2 &+ 4N\cdot\frac{1}{N}\cdot\frac{N(N-1)}{2} \\ &+ 4\cdot\frac{3}{N(N+2)}\cdot\frac{N(N-1)}{2} \\ &+ 4\cdot\frac{N(N-1)}{2}\left[\frac{N(N-1)}{2} - 1\right]\cdot x, \end{aligned} \tag{4.7}$$

where "x" is a generic notation for $E(c_{ij}^2 c_{kl}^2)$, for $i \neq k$ or $j \neq l$. Certainly if both $i \neq k$ and $j \neq l$, then $x = 1/N^2$, by independence.

In the remaining cases we are in the following situation. We have three random vectors u, v, w independently and identically distributed $U(S^{N-1})$ and we are considering the bivariate distribution of $(\langle u, v\rangle, \langle u, w\rangle)$. This distribution has also been considered by Stam [54], who gave a formula for the density of the trivariate distribution of $(\langle u, v\rangle, \langle u, w\rangle, \langle v, w\rangle)$. He also proved that this distribution converges in total variation to the standard normal distribution on R^3. In view of the complicated nature of the aforementioned density, and of the rather rapid approach to the normal, as shown by simulations, we will ignore possible weak dependencies for small N, and use the approximation $x = 1/N^2$ throughout (4.7). Therefore, after collecting terms there we arrive at the approximation

$$E(\|C\|_F^4) \approx 4N^2 - 4N - 1 + 6\frac{N-1}{N+2} + \frac{2}{N}. \tag{4.8}$$

Simulations show this to be actually very accurate for $N \geqq 5$. (In fact, extensive statistical testing never permits rejection of the hypothesis that the variates $\langle u, v \rangle$ and $\langle u, w \rangle$ are uncorrelated, for any N.) Finally, we see that

$$\operatorname{var}(\|C\|_F^2) \approx 6\frac{N-1}{N+2} + \frac{2}{N} - 2,$$

which, of course, is approximately four for large N.

These formulas for the first two moments of $\|C\|_F^2$ invite comparison with the corresponding formulas for correlation matrices with random spectra developed in the preceding chapter. While the means are very close, and asymptotically equivalent, there is a distinct difference in the behavior of the variances. Namely, the variance of $\|C\|_F^2$, when C has random spectrum, varies as $4N$, approximately, while that of $\|C\|_F^2$, when C is random Gram, is asymptotically constant.

Writing $\|C\|_F^2$ in the form used in (4.6), and appealing to the central limit theorem, the asymptotic normality follows readily:

$$\|C\|_F^2 \sim \text{Normal}\left(2N-1, 4\frac{N^2-2N+1}{N^2+N}\right),$$

for large N. As in the earlier section we could also establish the asymptotic normality of $\|C\|_F$, but at this point that can be left to the interested reader.

We now want to turn to the issue of the distribution of the spectral norm $\|C\|$ of a random Gram correlation matrix C. This particular topic brings us to the edge of a large and active field of research on the spectra of random matrices (see, for instance, Section II of the AMS conference proceedings [47]). This area has a long history, as indicated in the papers of Girko [20] and Geman [19] and their references, as well as the AMS volume. In turn it relates to many studies in the multivariate statistics field of spectral behavior of sample covariance matrices, for which see, for instance, Anderson [1].

The essential observation runs as follows. We have $C = TT^*$ as defined in (4.1). Then the columns of T^* are independent samples from the uniform distribution on S^{N-1}, and hence the matrix

$$S_N = \frac{1}{N}T^*T = \frac{1}{N}\sum_{k=1}^{N} t_k^* t_k$$

is the sample second moment matrix for such a distribution. (Here t_k is the kth row of the matrix T.) Now TT^* and T^*T always have the same eigenvalues, and hence, as a special case,

$$\|C\| = N\|S_N\|. \tag{4.9}$$

With this observation we can now refer to the considerable body of previous work mentioned in the preceding references, and also to Watson [63]. None of this seems to be exactly what we need. In particular, it is unlikely that we can ever know the precise distribution of $\|C\|$ for any fixed N. However, there are many asymptotic results. Here we will just take note of an improvement of Geman's theorem by Yin, Bai, and Krishnaiah, as referenced by Yin and Bai [67]. Namely, let X_p be a $p \times n$ random matrix with independently and identically distributed entries, $n = n(p)$ an increasing function of p with

$$\lim_{p\to\infty} \frac{p}{n} = y, \qquad 0 < y < \infty.$$

Suppose that the entries of X_p have mean zero, variance σ^2, and finite fourth moment. Then

$$\lim_{p\to\infty} \left\| \frac{1}{n} X_p X_p^* \right\| = (1+\sqrt{y})^2\sigma^2 \quad \text{a.s.} \tag{4.10}$$

At this point of the original version of this paper we made the conjecture that

$$\lim_{n\to\infty} \|C\| = 4 \quad \text{a.s.} \tag{4.11}$$

and advanced some reasons in its support. Here C is an $N \times N$ random Gram correlation matrix. First, we know from [19] that if $N \times N$ matrix G has independent entries, each a standard normal variate, then

$$\lim_{N\to\infty} \frac{1}{N} \|GG^*\| = 4 \quad \text{a.s.}$$

Next, let D be the $N \times N$ diagonal matrix with ith diagonal entry equal to $1/\|i\text{th column of } G\|$, and set $T^* = \text{GD}$. Then

$$\|C\| = \|T^*T\| = \|GD^2G^*\| = \frac{1}{N}\|G(ND^2)G^*\|,$$

and insofar as $ND^2 \approx I$ for large N, we may expect (4.11) to hold. We suggested that an order statistic analysis of the χ^2 distribution might be helpful here, but did not go further. However, one of the referees observed that this heuristic could be made rigorous, and the next paragraph is a slight paraphrase of his remarks.

The key fact is that not only is it true that

$$\lim_{n\to\infty} \frac{\chi^2(N)}{N} = 1 \quad \text{a.s.},$$

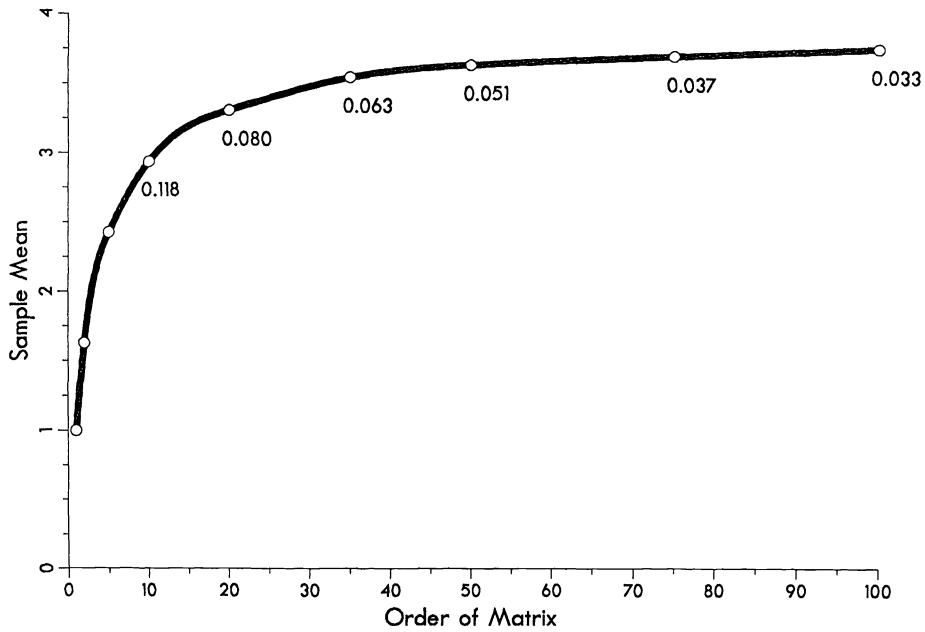


FIG. 1. *Empirical spectral norm of* $C = TT^*$.

but also the same is true for

$$\max\{\chi^2(N)/N : 1 \leqq i \leqq N\} \qquad \text{and} \qquad \min\{\chi^2(N)/N : 1 \leqq i \leqq N\}$$

[52]. Now with $H = \sqrt{ND}$, the eigenvalues of both H and H^{-1} tend to one almost surely, uniformly in i. So we have $\|(GH)\| \leqq \|(G)\|\|(H)\| \leqq \|(GH)\|\|(H)\|\|(H^{-1})\|$, which implies that $\|GG^*\|/\|G(ND^2)G^*\|$ tends to one as $N \to \infty$, and thereby validates (4.11).

Finally, a simulation for $N \leqq 100$ leads to the empirical curve of $E(\|C\|)$ against N shown in Fig. 1. Each data point for $N \leqq 50$ is based on 500 trials, while those for $N = 75, 100$ are based on 50 trials. The sample coefficient of variation is shown next to the data points.

4.3. Condition number expectation. In § 3.3 it was shown that the condition number of a correlation matrix with random spectrum has an infinite first moment. In the present section we will demonstrate the analogous fact for random correlation matrices of Gram type. The numerical results reported back in § 2.2 (refer to Table 1), along with their theoretical result just mentioned, certainly have prepared us for this fact. Recall that the key empirical difference between the two main types of random correlation matrices was, in fact, the comparatively ill conditioned nature of random Gram matrices. We will discuss some other aspects of the spectral behavior of such matrices in the next section.

THEOREM. *Let C be a random Gram correlation matrix. Then the condition number $\kappa(C)$ has infinite expectation.*

Proof. Making use of Taylor's inequality (B10) we have $C = TT^*$ and

$$\kappa(C) \equiv \|C\|\|C^{-1}\| \geqq 1/\min\, d_i^2$$

$$\geqq 1/d_1^2,$$

where $d_i \equiv \text{dist}\,(\underline{t}_i, M_i)$, $\underline{t}_i = i$th row of T, so each $\underline{t}_i$ is independently and identically distributed and $\sim U(S^{N-1})$, and $M_i = \text{span}\,\{\underline{t}_j : j \neq i\}$. Now since codim $(M_i) = 1$ in R^N, almost surely, d_1 is the magnitude of the projection of $\underline{t}_1$ on the line $M_1^{\perp}$. Let $\underline{u}_1$ be a unit vector in this subspace; then

$$d_1^2 = \langle \underline{t}_1, \underline{u}_1 \rangle^2.$$

That is, d_1^2 is the squared length of a random point on a random direction, and, as such, it has the distribution of (4.2), with moments given by (4.4). See also [40]. Therefore,

$$E(\kappa(C)) \geqq E(1/d_1^2)$$

$$= \int_0^1 \frac{t^{-1/2}(1-t)^{(N-3)/2}}{B(\frac{1}{2},(N-1)/2)} \frac{dt}{t},$$

and this integral clearly diverges at zero. □

4.4. Empirical spectral behavior. This final section addresses the question "How random is the spectrum of a random Gram matrix?". Now, in one sense, this question has already been answered by the results of §§ 3.3 and 4.2. Namely, in those sections we derived the behavior of $\|C\|_F^2$, $\|C\|_F$, and $\|C\|$ for both types of random correlation matrices. Expressing these functions of C in terms of the eigenvalues shows that, at least, not all spectral functions behave the same, and hence, in particular, that random Gram matrices do not have a random spectrum. Below, we will briefly discuss some other aspects of this question.

We begin by considering the behavior of the least eigenvalue λ_N of an $N \times N$ random Gram matrix C. In view of the boundedness of $\|C\|$ as $N \to \infty$ and the earlier observed

higher condition numbers of such matrices relative to those with random spectrum, we might expect λ_N to be much smaller than the least eigenvalue of a correlation matrix with random spectrum. Now, the same kind of argument as was used in § 4.3 leads to the conclusion that $\lim \lambda_N = 0$, almost surely. But the same is true for the least eigenvalue of a correlation matrix with random spectrum, as follows from the distribution function formula in (3.4). However, use of this formula in a Kolmogorov–Smirnov one-sample test of the hypothesis that λ_N obeys this distribution leads to its rejection, at the 99 percent level, at least for $N \leqq 20$.

Instead of dealing with the extreme eigenvalues λ_1, λ_N, of C we can also inquire about the behavior, in some sense, of the entire spectrum $\sigma(C)$. For example, we have already considered the statistic

$$\|C\|_F^2 = \sum_{i=1}^{N} \lambda_i^2,$$

and noted that asymptotically its mean behavior is that of a matrix with random spectrum, but its second moment behavior is quite different. But the most striking distinction between the two types of random correlation matrices can be made by considering their spectral distribution functions, and we turn to this topic next.

Given any square matrix A with real spectrum, its spectral distribution function F_A is defined by $F_A(x)$ = fraction of number of eigenvalues $\leqq x$. When A is random then, of course, so is $F_A(x)$. In this situation the asymptotic behavior of F_A has long been of interest (some useful surveys of this field are [62], [65], and [67]). Let us see what these distribution functions look like, for large N, for each of our two types of random correlation matrices. The situation is fairly simple for the case of random spectrum, thanks to the fact that each eigenvalue is identically distributed according to (3.2). It follows that the limiting spectral distribution function is $F_{RS}(x) \equiv 1 - e^{-x}$, $0 < x < \infty$.

By contrast, the situation for random Gram matrices is less immediate. The basic relevant fact is the "quarter-circle law for Gaussian matrices" [65]. This states that if the matrix G is defined as just after (4.11), then its spectral distribution function converges to the distribution function of a random variable S^2, where S has probability density function (pdf) of $= 1/2\pi((4 - t^2))^{1/2}$, $-2 \leqq t \leqq 2$ [62]. It routinely follows that S has a pdf of $1/2\pi((4/t - 1))^{1/2}$, $0 < t < 4$. But now the same analysis as validated (4.11) can be employed to show that the spectral distribution function of $N \times N$ random Gram matrices has the same limit. That is,

$$F_{\mathrm{RG}}(x) \equiv \frac{1}{2\pi} \int_0^x \sqrt{\left(\frac{4}{t} - 1\right)}\, dt, \qquad 0 < x < 4.$$

The fact that $F'_{\mathrm{RG}}(x)$ becomes infinite as $x \downarrow 0$ while $F'_{\mathrm{RG}}(x)$ does not is perhaps the most striking single distinction between our two types of random correlation matrices. But, of course, in principle it applies only in the limit as $N \to \infty$. What about the cases where N remains finite which is, after all, our primary concern? We offer two responses: one graphical and one statistical.

The first response is displayed in Figs. 2 and 3 for the case of random spectrum and random Gram matrices, respectively. In each case we consider in turn $N = 3$, 6, and 10, and from 500 trails for each N, we compute and plot the average of the spectral distribution functions. These averages may be compared with their limiting cases F_{RS} and F_{RG}, respectively.

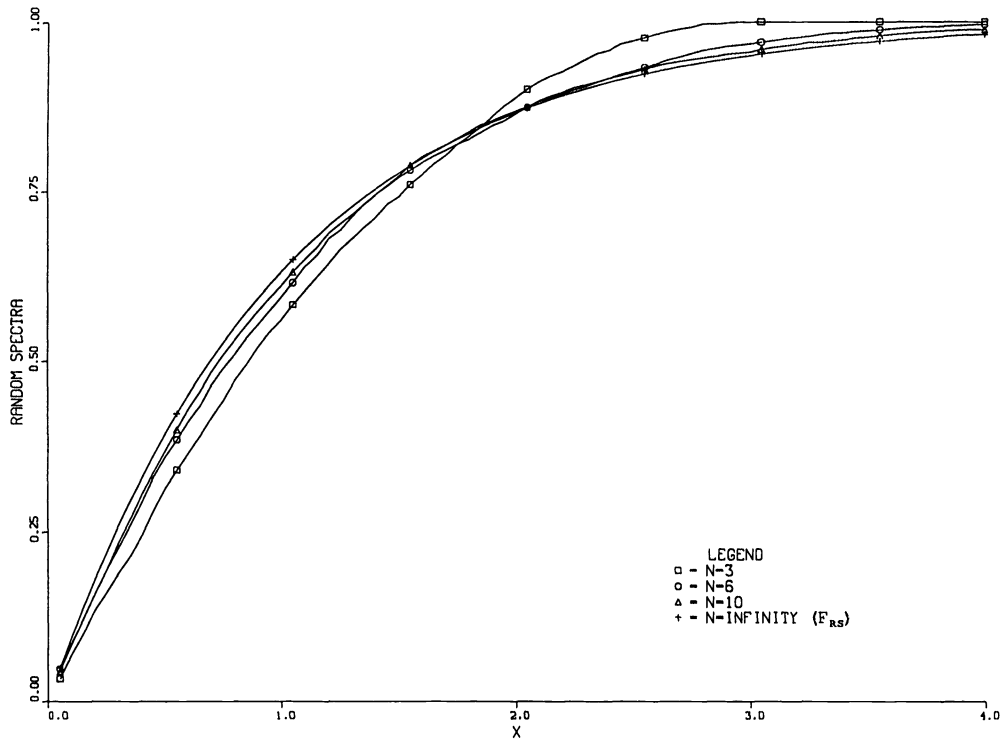


FIG. 2. *Limiting and expected spectral distribution functions for correlation matrices with random spectrum.*

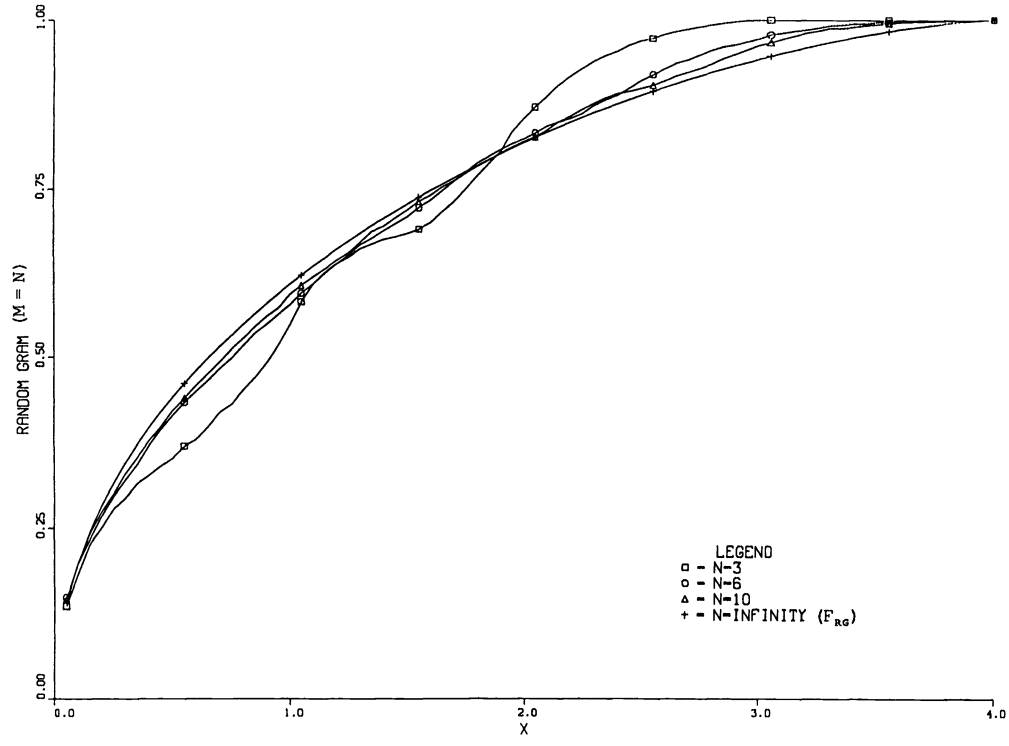


FIG. 3. *Limiting and expected spectral distribution functions for random Gram correlation matrices.*

The second response takes the form of a statistical test for uniformity based on the entire spectrum $\sigma(C) = \{\lambda_1, \cdots, \lambda_N\}$ of an $N \times N$ random Gram matrix, for $N \leqq 30$. Namely, the transformation

$$x_k = \frac{1}{N}(\lambda_{N-k+1} + \cdots + \lambda_{N-1} + \lambda_N)$$

replaces $\sigma(C)$ by a sample of points $\{x_1, \cdots, x_{N-1}\}$ in the unit interval which, under the null hypothesis of a uniform spectrum, is a sample from the uniform distribution $U([0, 1])$. In this case we applied Neyman's test [53] for uniformity, based on batches of 1,000 random Gram matrices for the various values of N. Here Neyman's statistic

$$N_2^2 = v_1^2 + v_2^2$$

was computed, where the v_j are the sample Fourier–Legendre coefficients when the density function f, from which the x's are drawn, is expanded in terms of Legendre polynomials:

$$f(x) = c \exp(1 + \sum c_j L_j(x)).$$

The motivation and theory of this test is discussed in the reference, and will not be given here. The distribution of N_2^2 is known approximately, and is asymptotically $\chi^2(2)$. The null hypothesis is to be rejected for large values of N_2^2. For each N we calculated the fraction of the 1,000 samples that exceeded various percentage points of the N_2^2 distribution, with the results indicated in Table 2. It is evident from these figures and the large number of trials that the null hypothesis of uniformity must be rejected. A closer examination of the data reveals that not only is there a very small eigenvalue λ_N, as noted above, but in fact there are enough small eigenvalues to pull the sample mean $\bar{x}$ far enough below 0.5 to greatly inflate the value of v_1 (precisely,

$$v_1 = \sqrt{12n}(\bar{x} - \tfrac{1}{2}),$$

where n = sample size = 1,000, here). Incidentally, the sample coefficient of variation of the Neyman statistics decreased steadily from 0.27 at $N = 5$ to 0.045 at $N = 30$, showing very little scatter about the increasingly large values of N_2^2.

Finally, we offer two comments about the empirical behavior of the condition number of random Gram matrices. First, for various N ($\leqq 20$) we generated batches of 1,000 each of random Gram matrices and correlation matrices with random spectra, and performed a Kolmogorov–Smirnov two-sample test on the respective condition numbers, to test the null hypothesis of a common distribution. This hypothesis was decisively rejected for all values of N, and this rejection continued when the samples were subjected to trimming.

TABLE 2

Fraction of Neyman statistics exceeding various percentage points, and sample average.

N %	50	90	95	Mean N_2^2
5	99.8	45	10.5	4.3
8	100	98.5	84.4	7.0
10	–	100	99.5	8.9
15	–	–	100	13.3
20	–	–	–	17.7
30	–	–	–	26.5

TABLE 3
Empirical ratio of condition number of collinearity measure for random Gram matrices.

N	Batch size	Sample mean ratio	Sample coeff. of var.
5	1,000	5.20	.21
10	1,000	8.82	.25
20	100	15.86	.24
35	100	22.65	.24
50	100	29.28	.22

Second, bearing in mind the condition number bounds established in Appendix B, we studied by simulation the tightness of the upper bound (B13). That is, for various N ($\leqq 50$) we generated batches of random Gram matrices, computed their condition numbers, the collinearity measure on the right-hand side of (B10), and then their ratio as in (B13). The results are displayed in Table 3. They suggest that the admittedly crude upper bound in (B13) can indeed be reduced, and perhaps even be replaced by a term that is of order $o(N)$.

5. Summary. Let us now, in conclusion, summarize not only the foregoing technicalities, but also the place of this material in a larger scheme. In addition, we will point out several issues that remain to be resolved.

As noted at the outset, our interest in random correlation matrices stems from their interpretation as covariance matrices of purely random or "average" (standardized) signals. A companion research project has as its goal the evaluation of the efficacy of various group-theoretic signal processing algorithms. One ingredient that must be specified before a well-defined question can be posed in this context is a definite signal model. As remarked in the Introduction, such models can either be defined by a few (typically $\leqq 2$) parameters, or they can be essentially nonparametric. A further possible subdivision of this latter class is into random stationary signals, or into purely random signals. The corresponding covariance matrices are then random correlation matrices with, in the first case, a Toeplitz structure. The question of generating such matrices, and the statistical behavior of the corresponding entries, spectral functions, etc., is interesting, and is being studied, with results to be reported elsewhere [28].

We therefore have chosen to concentrate on random correlation matrices of the two principal types defined in § 1.2, and studied in detail in §§ 2–4. We observed early on that random Gram matrices exhibited a more exaggerated spectral behavior relative to correlation matrices with random spectrum. As we discovered later, this is due to the very different small eigenvalue behavior as quantified by the respective spectral distribution functions.

Some other results, both theoretical and empirical, point out the very different spectral behavior of these two classes of random correlation matrices. For example, the spectral norm of a correlation matrix with random spectrum tends to grow logarithmically, while that of a random Gram matrix remains bounded, almost surely, as the dimension increases.

Some other new results pertain to condition number behavior. Specifically, in Appendix B we have extended earlier work of Taylor [61] on condition number lower bounds, and assessed their tightness. This result is strictly deterministic. We then used this bound to show that the condition number of random Gram matrices (of a fixed size) has an infinite first moment. In view of our earlier empirical observations, this conclusion was not a complete surprise. Yet it also turned out that correlation matrices

with random spectrum also have infinite first moment (for each fixed dimension $N \geqq 2$, the case $N = 2$ being due to Feller [14]).

We might offer an additional comment on the condition of random Gram matrices. Namely, referring back to the basic definition (§ 4.1), we could allow the row vectors t_i there to be drawn randomly from the unit sphere in a larger-dimensional space. Geometric intuition suggests that with more "room" in the sample space, collinearity should be less of a problem, with consequent improvement in conditioning. Numerical experiments show that, to an extent, this expectation is fulfilled. For example, in contrast with the data reported in Table 1, the mean (respectively, median) condition number of 500 5×5 random Gram matrices based on vectors drawn uniformly from the sphere S^9 is 11.6 (respectively, 8.7). The corresponding values for 500 10×10 random Gram matrices based on vectors drawn from S^{19} are 17.5 (respectively, 15.2). However, as long as $\dim(t_i)$ has the form $N + p$, where p is a fixed positive integer, then the limiting form of the spectral distribution function, as derived back in § 4.4, will remain the same.

Appendix A. Aspects of numerical linear algebra. This section contains a brief review of some quantitative aspects of linear algebra that are pertinent to the material that follows. For general background information on matrix theory we may refer to two recent volumes: Horn and Johnson [30] or Lancaster and Tismenetsky [35]. More specialized treatments of numerical linear algebra are given by Stewart [56] and Golub and Van Loan [22].

A.1. Bounds on norms and eigenvalues. Given an $N \times N$ matrix A we shall have occasion to use its operator or *spectral norm*

$$\|A\| = \max\{\|Ax\| / \|x\| : x \neq \theta\},$$

and its *Frobenius norm*

$$\|A\|_F = (\sum |a_{ij}|^2)^{1/2}.$$

In terms of the positive part $P = (AA^*)^{1/2}$ of A, we have

$$0 \leqq \|A\| = r_\sigma(P) \leqq \|A\|_F = \sqrt{\operatorname{tr}(P^2)}, \tag{A1}$$

where $r_\sigma(\cdot)$ means spectral radius. Bringing in the eigenvalues $\lambda_1 \geqq \cdots \geqq \lambda_N$ of A, and the singular values $s_1 \geqq \cdots \geqq s_N$ (these are the eigenvalues of P), we have $\|A\| = s_1$, and

$$\sum_1^N |\lambda_i|^2 \leqq \|A\|_F^2 = \sum_1^N s_i^2, \tag{A2}$$

with equality if and only if A is normal (theorem of Schur and Mirsky).

For general matrices A the singular values have many fascinating properties and applications, such as min-max characterizations, smooth dependence on A (which leads into perturbation theory), and geometric interpretations as distances from A to spaces of operators of lower rank. This latter property, on the one hand, leads into regularization techniques for least squares signal processing and, on the other, permits generalization to compact operators on infinite-dimensional spaces (s-number theory).

Let us now specialize to the case of primary interest here, namely, that where $A = C$ is a correlation matrix. Then we have

$$1 \leqq \|C\| = r_\sigma(C) = \inf |||C||| \leqq \max\{\|\text{row}_1\|_1, \cdots, \|\text{row}_N\|_1\} \leqq N, \tag{A3}$$

where $|||\cdot|||$ refers to a general matrix norm induced by some vector norm, and $\|\cdot\|_1$ is the l^1-vector norm. The expression "max $\{\cdots\}$" above is just the matrix norm induced

by the l^∞-vector norm. Its advantage, as with the bound $\|C\|_F$, is that it is immediately computable from the entries of C. Either of the extremes 1, N can be reached by some $C \in \Gamma(N)$. The second equality is true in much greater generality, in fact, for any operator on a Banach space [27].

If A is positive semidefinite, then

$$|a_{ij}| \leqq \sqrt{a_{ii}a_{jj}} \leqq \tfrac{1}{2}(a_{ii}+a_{jj}),$$

showing, in particular, that all off-diagonal entries of a correlation matrix have modulus less than or equal to one. Of course, such a matrix need not be diagonally dominant.

An improvement on the bound $\|C\| \leqq \|C\|_F$ has been noted by Leclerc [36], specifically for correlation matrices. Namely,

$$\|C\| \leqq 1 + \left(\frac{N-1}{N}\sum_0^2\right)^{1/2} \leqq \|C\|_F, \tag{A4}$$

where $\sum_0^2$ is the sum of squares of off-diagonal entries of C. The right-hand inequality here is strict unless all off-diagonal entries have modulus one. This bound on $\|C\|$ can be either larger or smaller than the "max" bound of (A3).

At this point we have given some upper bounds for $\|C\|$, and hence for all the (positive) eigenvalues of C. Upper bounds for $\|C^{-1}\|$ are equivalent to lower bounds on the eigenvalues of C, using $\|C^{-1}\| = r_\sigma(C^{-1})$; note that C^{-1}, while still positive definite, is no longer a correlation matrix in general. This kind of bound is not of particular interest to us here, but *lower* bounds on $\|C^{-1}\|$ are important, in connection with condition number estimates, and will be discussed later on. Here we will just recall an inequality of Kato [32], which gives a bound on $\|A^{-1}\|$, for any nonsingular A:

$$\|A^{-1}\| \leqq \|A\|^{N-1}/|\det(A)|.$$

There are innumerable inequalities pertaining to the eigenvalues of positive-definite matrices, and more generally to the singular values of arbitrary matrices. Here we mention just two. They are originally due to Fan [13], with a short proof now available [21], based on the characterization of the kth singular value $s_k(A)$ of a matrix (or operator) A as the distance from A to the set of matrices of rank $\leqq k-1$, in the spectral norm. Thus

$$s_{m+n-1}(A+B) \leqq s_m(A) + s_n(B),$$

$$s_{m+n-1}(AB) \leqq s_m(A)s_n(B),$$

for $m, n \geqq 1$.

Finally, we mention the concept of *spread* of a matrix A. This is the quantity

$$S(A) = \operatorname{diam}\sigma(A) \equiv \max|\lambda_i - \lambda_j|. \tag{A5}$$

When $A = C$, a correlation matrix, the following bounds on $S(C)$ can be derived:

$$2\max_{i\neq j}|c_{ij}| \leqq S(C) \leqq (2(\|C\|_F^2 - N))^{1/2}.$$

Since $\|C\|_F^2 = \sum \lambda_i^2$, the last inequality offers a lower bound on this quantity. But, in fact, a stronger two-sided inequality can be established, namely,

$$\frac{1}{2}S(C)^2 + N \leqq \|C\|_F^2 \leqq \frac{5}{4}NS(C)^2,$$

by working with the eigenvalues.

An alternate concept of spread, the condition number, is defined in Appendix B; it uses the maximum ratio as opposed to the maximum difference of eigenvalues, for positive-definite matrices (cf. (B3)).

Appendix B. Condition number estimates. The condition number $\kappa(A)$ of an arbitrary matrix A is defined by

$$\kappa(A) = \|A\|\|A^+\| = \|A\|\|A^{-1}\|, \tag{B1}$$

where the "+" means pseudoinverse, and the second equality is naturally only applicable if A is nonsingular. Note that this is the spectral condition number; other matrix norms might be used in (B1). In terms of the singular values of A we have

$$1 \leqq \kappa(A) = s_1/s_N, \tag{B2}$$

with equality if and only if A is a nonzero multiple of a unitary matrix. (Naturally, (B2) is restricted to nonsingular A.) Many kinds of singular matrices A can have $\kappa(A) = 1$; for instance, orthogonal projections and, more generally, partial isometries.

Condition numbers are widely used as measures of sensitivity of the solution of linear systems to inaccuracies in the data. Similarly, the condition number of the matrix of eigenvectors of a diagonalizable matrix measures the closeness of an approximate eigenvalue to the true spectrum. Roughly speaking, the percentage change in the (least squares) solution x of the system $Ax = b$ is bounded by the percentage variation in the data b times $\kappa(A)$, and this bound cannot be lowered. Thus $\kappa(A)$ is a measure of the inherent resistance of a particular system to accurate solution, and which does not depend on the particular numerical method employed. The larger the condition number, the more "ill conditioned" a particular system is, and the less we can infer a small error from a small residual.

We might also remark that $\kappa(A)$ can be characterized geometrically by the least angle ψ resulting as A is applied to all possible pairs of orthonormal vectors. Precisely,

$$\kappa(A) = \cot(\psi/2).$$

It is instinctive to want to measure ill-conditioning by some function of the eigenvalues, but this is only fruitful for normal matrices. For example, there is the $N \times N$ "Kahan matrix":

$$\begin{bmatrix} 1 & -1 & \cdot & \cdot & \cdot & -1 \\ 0 & 1 & \cdot & \cdot & & \\ \vdots & & & \cdot & \cdot & -1 \\ 0 & & \cdot & \cdot & \cdot & 1 \end{bmatrix},$$

which clearly has all eigenvalues equal to one, yet a condition number greater than $(N2^{(N-2)})^{1/2}$. However, when A is positive definite with eigenvalues $\lambda_1 \geqq \cdots \geqq \lambda_N$, then

$$\kappa(A) = \lambda_1/\lambda_N, \tag{B3}$$

and we have the inequality of Kato:

$$\kappa(A) \leqq \frac{4}{\det(A)}\left(\frac{\operatorname{tr}(A)}{N}\right)^N.$$

Thus, for correlation matrices C, $\kappa(C)\det(C)$ is a bounded function. Note that (B2) and (B3) together imply that

$$\kappa(A^*A) = \kappa(AA^*) = \kappa(A)^2,$$

showing how the familiar "normal equations" of many least squares procedures can become very ill conditioned (and, eventually, motivating the use of factorization methods which deal only with A, as an alternative).

In 1955, Riley [48] used the fact that

$$\kappa(A+\lambda I) \leqq \kappa(A) \tag{B4}$$

for any positive-definite A to suggest an iterative improvement procedure for solving an ill-conditioned linear system $Ax = b$. This was a forerunner of the ridge regression and regularization methods of statistics and signal processing, which trade off some bias for lowered mean square error. The inequality (B4) was greatly extended by Marshall and Olkin [41] who proved that

$$\kappa(A+B) \leqq \kappa(A)$$

whenever A and B are positive definite with $\kappa(B) \leqq \kappa(A)$.

We now turn to the matter of *lower* bounds for condition numbers. These will be of greatest interest for the case of Gram matrices, but we consider first, briefly, the general case. (We note, too, the considerable interest in recent years in numerical estimates—not bounds—for condition numbers, by estimating some norm of the inverse matrix [6], [12], [26].)

First, if A is any nonsingular matrix, with eigenvalues ordered by modulus: $|\lambda_1| \geqq \cdots \geqq |\lambda_N|$, then

$$|\lambda_1/\lambda_N| \leqq \kappa(A). \tag{B5}$$

This follows from the relations

$$\begin{aligned} \|A^{-1}\|^{-1} &= \inf\{\|Ax\|: \|x\| = 1\} \\ &\leqq \|Ae\| = |\lambda|, \end{aligned}$$

where e is any unit eigenvector associated with $\lambda \in \sigma(A)$. Of course, as the earlier example of the Kahan matrix illustrates, the left side of (B5) may severely underestimate the true condition number $\kappa(A)$, when A is not normal.

Now assume that A is positive definite. A variant of the well-known Kantorovich inequality [25] tells us that

$$\|x\|^2 \leqq \langle Ax, x\rangle\langle A^{-1}x, x\rangle \leqq \frac{(m_1+m_2)^2}{4m_1m_2}\|x\|^2, \tag{B6}$$

provided that

$$m_1 I \leqq A \leqq m_2 I,$$

for $0 < m_1 \leqq m_2$. Taking m_1 (respectively, m_2) to be the least (respectively, greatest) eigenvalue of A, and x any unit vector, we obtain

$$4\langle Ax, x\rangle\langle A^{-1}x, x\rangle \leqq \kappa + \frac{1}{\kappa} + 2 \leqq \kappa + 3,$$

yielding a lower bound for $\kappa = \kappa(A)$ for each x. Of course, an estimate involving A^{-1} is not of great practical value.

Another kind of inequality comes from the theory of Schur (or Hadamard) products of matrices. We will not review this concept in any detail here; see [59] for a nice survey. This product, for conformable matrices A, B, is defined by

$$[A \cdot B]_{i,j} = a_{ij} \cdot b_{ij}.$$

This multiplication, unlike the usual one, is commutative. The original result of Schur is that if A, B are positive semidefinite, then so is $A \cdot B$. An inequality of Fiedler [15] for positive-definite A reads

(B7) $$A \cdot A^{-1} \geqq I.$$

Note that, as a consequence of either this or the left side of the Kantorovich inequality, when C is a correlation matrix,

$$[C^{-1}]_{i,i} \geqq 1, \qquad i = 1, \cdots, N.$$

In 1982, Marcus [38] proved the matrix norm inequality

$$\|A \cdot B\| \leqq \|A\| \, \|B\|$$

for the Schur product. Taking $B = A^{-1}$ yields a lower bound for the condition number

$$\|A \cdot A^{-1}\| \leqq \kappa(A).$$

Finally, in preparation for part of our discussion in § 4, we want to specifically consider the case where A is a Gram matrix

$$A = G = G(x_1, \cdots, x_N),$$

in the notation of § 2.2, with each x_i a unit vector in some inner product space X. As already remarked in § 2.2 it has been empirically noted that many common Gram matrices tend to be ill conditioned, and an inequality derived in [61] can be used to quantify these observations by providing a lower bound for $\kappa(G)$ in terms of the relative orientations of the vectors $\{x_i\}$. By virtue of our own numerical experiments reported earlier, ill-conditioning is a prominent feature in random Gram matrices also. We now discuss an improved version of this inequality, and its sharpness. These results are purely deterministic; statistical implications are discussed in § 4.

We now work with a fixed Gram matrix $G = G(x_1, \cdots, x_N)$, $\|x_i\| = 1$, $i = 1, \cdots, N$. G is a correlation matrix, and $\|G\| = r_\sigma(G)$, so the real problem is to find a lower bound on $\|G^{-1}\|$ in terms of the vectors $\{x_i\}$. Let M (respectively, M_i) be the subspace of X spanned by $\{x_j\}_1^N$ (respectively, $\{x_j : j \neq i\}_1^N$). Let $\{v_j\}$ be the basis for M that is dual to $\{x_j\}$. Also, for an arbitrary real or complex unit vector e (according as X is real or complex), let $b = G^{-1}e$. Then

$$\begin{aligned} \|G^{-1}\| &\geqq \langle G^{-1}e, e \rangle = \langle b, Gb \rangle \\ &= \|\sum \bar{b}_j x_j\|^2 \equiv \|v\|^2. \end{aligned}$$

Now, with v as just defined, it is easily checked that

$$\bar{b} = \begin{bmatrix} \vdots \\ \langle v, v_i \rangle \\ \vdots \end{bmatrix},$$

so that if $v = v_i$, one of the dual basis vectors in M, then $Gb = e_i$, the standard unit basis vector.

We also observe that, since M_i is of codimension one in M, the duality formula for distance,

(B8) $$\operatorname{dist}(x, M_i) = \max \{ |\psi(x)| : \psi \in S(M_i^\perp) \},$$

for $x \in M$, implies that

$$d_i \equiv \operatorname{dist}(x_i, M_i) = \langle x_i, v_i / \|v_i\| \rangle$$
$$= 1 / \|v_i\|.$$

Putting all this together, we conclude that

$$\|G^{-1}\| \geqq \langle G^{-1} e_i, e_i \rangle = \|v_i\|^2$$
$$= \frac{1}{d_i^2} = \frac{g_i(x_1, \cdots, x_N)}{g(x_1, \cdots, x_N)}, \tag{B9}$$

where the last equality follows from the Gramian distance formula of (2.6), and "g_i" means the Gramian with x_i deleted.

The ensuing inequality

$$\|G^{-1}\| \geqq \max d_i^{-2} = \frac{1}{\min d_i^2} \tag{B10}$$

is due to Taylor [60, p. 46]. The major difference between his approach and the present one is that use of the duality formula (B8) strengthens the inequality by avoiding reliance on the Schwarz inequality. Thus the sole source of inequality in (B10) is the inequality appearing in (B9). This inequality is only a measure of the behavior of the Rayleigh quotient for G^{-1}, and does not explicitly involve the Gram structure of G. Hence the following theorem gives a measure of the tightness of Taylor's inequality (B10).

THEOREM.

$$\sup_{A \in \Gamma(N)} \frac{\|A^{-1}\|}{\max \langle A^{-1} e_i, e_i \rangle} = N.$$

Proof. For notational ease, we will replace A^{-1} by A, and then

$$\max \{ \langle A e_i, e_i \rangle : i = 1, \cdots, N \}$$

by $\mu(A)$. We first note that if A is any $N \times N$ positive-definite matrix,

$$1 \leqq \|A\| / \mu(A) \leqq N,$$

and that these bounds are sharp (within this larger class of matrices). The left inequality is trivial, and is achieved for diagonal matrices. The right inequality follows from

$$\|A\| = r_\sigma(A) = \lambda_1 \leqq \operatorname{tr}(A) \leqq N\mu(A).$$

To verify its sharpness, let $\varepsilon > 0$, and $D = \operatorname{diag}(1, \varepsilon, \cdots, \varepsilon)$, and apply the theory in § 2.1 to obtain A, unitarily equivalent to D, with constant diagonal. Then

$$1 = \|A\| \leqq \operatorname{tr}(A) = N\mu(A) = 1 + (N-1)\varepsilon;$$

now let $\varepsilon \downarrow 0$. So the point of the theorem is that if the A's are restricted to the class $\{A : A^{-1} \in \Gamma(N)\}$, the upper bound on $\|A\| / \mu(A)$ does not decrease.

To complete the proof, consider a special family of A's, namely, $\{A: A = aI_N + B, b_{ij} = (1 - \delta_{ij})b\}$, where $a > b > 0$. We have

$$\frac{\|A\|}{\mu(A)} = \frac{a+(N-1)b}{a} = 1+(N-1)\frac{b}{a}, \tag{B11}$$

and it will be shown that a and b can be chosen so that $A^{-1} \in \Gamma(N)$ and

$$\lim_{a \to \infty} \frac{b}{a} = 1. \tag{B12}$$

Let $\Delta = \det(A)$. We have

$$\Delta = (a-b)^{N-1}(a+(N-1)b),$$

and since the diagonal entries of A^{-1} are

$$\langle A^{-1}e_i, e_i\rangle = \frac{i\text{th-cofactor}}{\Delta},$$

it follows that

$$\langle A^{-1}e_i, e_i\rangle = \frac{(a+(N-2)b)(a-b)^{N-2}}{(a+(N-1)b)(a-b)^{N-1}}.$$

So, if A is to be a correlation matrix, a and b must satisfy the equation

$$a+(N-2)b = (a-b)(a+(N-1)b).$$

If we treat this as a (quadratic) equation for b and solve it, we obtain

$$b = \frac{(a-1)(N-2)+((N-2)^2(a-1)^2+4(N-1)(a^2-a))^{1/2}}{2(N-1)}.$$

After dividing both sides by a, and manipulating, we have

$$\frac{b}{a} = 1 - \frac{1}{a}\left(\frac{N-2}{2N-2}\right) + \frac{\sqrt{N^2+e}-N}{2N-2},$$

where

$$e = \frac{(N-2)^2}{a}\left(\frac{1}{a}-2\right) - \frac{4}{a}(N-1) < 0.$$

This shows that $b < a$ and that the limit in (B12) is one, as required. □

At this point we might justify an assertion made just after (A3), namely, that

$$\sup_{A \in \Gamma(N)} \|A\| = N.$$

We know from that equation that this supremum is at most N. That it is not less than N follows from consideration of the same family of matrices just used, and the value of the norms of such matrices given in (B11): just take $a = 1$ and let $b \to 1$.

To sum up, for a Gram matrix $G = G(x_1, \cdots, x_N)$, we have the lower bound on $\kappa(G)$, due to Taylor:

$$\kappa(G) \geqq \frac{1}{\min_i d_i^2},$$

an equivalent form

$$\kappa(G) \geqq \max_i \frac{g_i(x_1, \cdots, x_N)}{g(x_1, \cdots, x_N)},$$

and an upper bound on the tightness of this lower bound:

$$1 \leqq \frac{\kappa(G)}{\text{lower bound}} \leqq N^2. \tag{B13}$$

It is possible that this upper bound could be decreased, but we have not investigated this point. Some evidence was given earlier in § 4.4.

Acknowledgment. Mr. Thomas Loden very capably carried out the programming required for the foregoing numerical tests and simulations.

REFERENCES

[1] T. ANDERSON, *An Introduction to Multivariate Statistics*, John Wiley, New York, 1984.
[2] T. ANDERSON, I. OLKIN, AND L. UNDERHILL, *Generation of random orthogonal matrices*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 625–629.
[3] R. BENDEL AND R. MICKEY, *Population correlation matrices for sampling experiments*, Comm. Statist. B—Simulation Comput., 7 (1978), pp. 163–182.
[4] C. CHALMERS, *Generation of correlation matrices with a given eigen-structure*, J. Statist. Comput. Simulation, 4 (1975), pp. 133–139.
[5] N. CHAN AND K. LI, *Diagonal elements and eigenvalues of a real symmetric matrix*, J. Math. Anal. Appl., 91 (1983), pp. 562–566.
[6] A. CLINE, C. MOLER, G. STEWART, AND J. WILKINSON, *An estimate for the condition number of a matrix*, SIAM J. Numer. Anal., 16 (1979), pp. 368–375.
[7] D. DARLING, *On a class of problems related to the random division of an interval*, Ann. Math. Statist., 24 (1953), pp. 239–253.
[8] P. DEHEUVELS, *Spacings and applications*, in Proc. 4th Pannonian Symposium on Mathematical Statistics, D. Reidel, Dordrecht, Holland, 1983, pp. 1–30.
[9] L. DEVROYE, *Non-Uniform Random Variate Generation*, Springer-Verlag, New York, 1986.
[10] ———, *Laws of the iterated logarithm for order statistics of uniform spacings*, Ann Probab., 9 (1981), pp. 860–867.
[11] P. DIACONIS AND M. SHAHSHAHANI, *The subgroup algorithm for generating uniform random variables*, Prob. Engrg. Inform. Sci., 1 (1987), pp. 15–32.
[12] J. DIXON, *Estimating extremal eigenvalues and condition numbers of matrices*, SIAM J. Numer. Anal., 20 (1983), pp. 812–814.
[13] K. FAN, *Maximum properties and inequalities for the eigenvalues of completely continuous operations*, Proc. Nat. Acad. Sci., 37 (1951), pp. 760–766.
[14] W. FELLER, *An Introduction to Probability Theory and Its Applications*, Vol. 2, John Wiley, New York, 1971.
[15] M. FIEDLER, *Über eine Ungleichung für positiv definite Matrizen*, Math. Nachr., 23 (1961), pp. 197–199.
[16] P. FILLMORE, *On similarity and the diagonal of a matrix*, Amer. Math. Monthly, 76 (1969), pp. 167–168.
[17] R. FISHER, *Tests of significance in harmonic analysis*, in Proc. Roy. Soc. London Ser. A, 125 (1929), pp. 54–59.
[18] K. FUGUNAGA, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1972.
[19] S. GEMAN, *A limit theorem for the norm of random matrices*, Ann. Probab., 8 (1980), pp. 252–261.
[20] V. GIRKO, *Spectral theory of random matrices*, Russian Math Surveys, 40 (1984), pp. 77–120.

[21] I. GOHBERG AND M. KREIN, *Introduction to the Theory of Linear Nonselfadjoint Operators*, American Mathematical Society, Providence, RI, 1969.
[22] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.
[23] M. GREENWOOD, *The statistical study of infectious diseases*, J. Roy. Statist. Soc. Ser. A, 109 (1946), pp. 85–110.
[24] U. GRENANDER, *Probabilities on Algebraic Structures*, John Wiley, New York, 1963.
[25] W. GREUB AND W. RHEINBOLDT, *On a generalization of an inequality of L. V. Kantorovich*, Proc. Amer. Math. Soc., 10 (1959), pp. 407–415.
[26] W. HAGER, *Condition estimates*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 311–316.
[27] R. HOLMES, *A formula for the spectral radians of an operator*, Amer. Math. Monthly, 75 (1968), pp. 163–166.
[28] ———, *On random correlation matrices* II. *The Toeplitz case*, Tech. Report 816, M.I.T. Lincoln Laboratory, Lexington, MA, 1989; Comm. Statist. B—Simulation Comput., 18 (1989), pp. 1511–1537.
[29] A. HORN, *Doubly stochastic matrices and the diagonal of a rotation matrix*, Amer. J. Math., 76 (1954), pp. 620–630.
[30] R. HORN AND C. JOHNSON, *Matrix Analysis*, Cambridge University Press, New York, 1985.
[31] D. JOHNSON AND W. WELCH, *The generation of pseudo-random correlation matrices*, J. Statist. Comput. Simulation, 11 (1980), pp. 55–69.
[32] T. KATO, *Estimation of iterated matrices, with application to the von Neumann condition*, Numer. Math., 26 (1960), pp. 22–29.
[33] D. KAZAKOS, *Optimal constrained representation and filtering of signals*, Signal Process., 5 (1983), pp. 347–353.
[34] M. KENDALL AND P. MORAN, *Geometric Probability*, Griffin, London, 1963.
[35] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, Second Edition, Academic Press, New York, 1985.
[36] A. LECLERC, *Uni borne superieure pour les valeurs d'une matrice symetrique. Applications*, C. R. Acad. Sci. Paris, 287 (1978), pp. A553–555.
[37] P. LÉVY, *Sur la division d'un segment par des points chosis au hazard*, C. R. Acad. Sci. Paris, 208 (1939), pp. 147–149.
[38] M. MARCUS, *Eigenvalues, numerical ranges, stability analysis, and applications of number theory to computing*, Annual Scientific Report to Air Force Office of Scientific Research, Institute for Algebra and Combinatorics, University of California, Santa Barbara, CA, September 1982.
[39] G. MARSAGLIA, *Choosing a point from the surface of a sphere*, Ann. Math. Statist., 43 (1972), pp. 645–646.
[40] G. MARSAGLIA AND I. OLKIN, *Generating correlation matrices*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 470–475.
[41] A. MARSHALL AND I. OLKIN, *Norms and inequalities for condition numbers*, II, Linear Algebra Appl., 2 (1969), pp. 167–172.
[42] ———, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.
[43] L. MIRSKY, *Matrices with prescribed characteristic roots and diagonal elements*, J. London Math. Soc., 33 (1958), pp. 14–21.
[44] P. MORAN, *The random division of an interval*, Proc. Cambridge Philos. Soc., 89 (1947), pp. 92–98.
[45] L. NACHBIN, *The Haar Integral*, Van Nostrand, Princeton, 1965.
[46] R. PYKE, *Spacings*, J. Roy. Statist. Soc. Ser. B, 27 (1965), pp. 395–436.
[47] J. COHEN, H. KESTEN, AND C. NEWMAN, EDS., *Random Matrices and Their Applications*, Contemporary Mathematics 50, American Mathematical Society, Providence, RI, 1986.
[48] J. RILEY, *Solving systems of linear equations with a positive definite, symmetric, but possibly ill-conditioned matrix*, Math. Comp., 9 (1955), pp. 56–61.
[49] R. RUBINSTEIN, *Generating random vectors uniformly inside and on the surface of different regions*, European J. Oper. Res., 10 (1982), pp. 205–209.
[50] I. SCHUR, *Über eine Klasse von Mittelbildungen mit Anwendungen die Determinanten*, Theorie Sitzungsber. Berlin Math. Ges., 22 (1923), pp. 9–20.
[51] A. SIEGEL, *Testing for periodicity in a time series*, J. Amer. Statist. Assoc., 75 (1980), pp. 345–348.
[52] J. SILVERSTEIN, *The smallest eigenvalue of a large dimensional Wishart matrix*, Ann. Probab., 13 (1985), pp. 1364–1368.
[53] H. SOLOMON AND M. STEPHENS, *On Neyman's statistic for testing uniformity*, Comm. Statist. B—Simulation Comput., 12 (1983), pp. 127–134.
[54] A. STAM, *Limit theorems for uniform distributions on spheres in high-dimensional Euclidean spaces*, J. Appl. Probab., 19 (1982), pp. 221–228.

[55] F. STEUTEL, *Random division of an interval*, Statist. Neerlandica, 21 (1967), pp. 231–244.
[56] G. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.
[57] ———, *The efficient generation of random orthogonal matrices with an application to condition estimators*, SIAM J. Numer. Anal., 17 (1980), pp. 403–409.
[58] ———, *Collinearity and least squares regression*, Statist. Sci., 2 (1987), pp. 68–83.
[59] G. STYAN, *Hadamard products and multivariate statistical analysis*, Linear Algebra Appl., 6 (1973), pp. 217–240.
[60] L. TAKACS, *Harmonic analysis on Schur algebras and its applications in the theory of probability*, Probability Theory and Harmonic Analysis, J. A. Choi and W. Woyczynski, eds., Marcel Dekker, New York, 1986, pp. 227–283.
[61] J. TAYLOR, *The condition of Gram matrices and related problems*, Proc. Roy. Soc. Edinburgh Sect. A, 90 (1978), pp. 45–56.
[62] H. TROTTER, *Eigenvalue distributions of large hermitian matrices; Wigner's semicircle law and a theorem of Kac, Murdock, and Szego*, Adv. in Math., 54 (1984), pp. 67–82.
[63] G. WATSON, *Statistics on Spheres*, John Wiley, New York, 1983.
[64] W. WHITWORTH, *Choice and Chance*, Cambridge University Press, Cambridge, U.K., 1887.
[65] E. WIGNER, *Random matrices in physics*, SIAM Rev., 9 (1967), pp. 1–23.
[66] S. WILKS, *Mathematical Statistics*, John Wiley, New York, 1962.
[67] Y. YIN AND Z. BAI, *Spectra for large dimensional random matrices*, in Random Matrices and Their Applications, J. Cohen, H. Kesten, and C. Newman, eds., American Mathematical Society, Providence, RI, 1986, pp. 161–167.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 2, pp. 273–291, April 1991

006

RATIONAL ITERATIVE METHODS FOR THE MATRIX SIGN FUNCTION*

CHARLES KENNEY† AND ALAN J. LAUB†

Abstract. In this paper an analysis of rational iterations for the matrix sign function is presented. This analysis is based on Padé approximations of a certain hypergeometric function and it is shown that local convergence results for "multiplication-rich" polynomial iterations also apply to these rational methods. Multiplication-rich methods are of particular interest for many parallel and vector computing environments. The main diagonal Padé recursions, which include Newton's and Halley's methods as special cases, are globally convergent and can be implemented in a multiplication-rich fashion which is computationally competitive with the polynomial recursions (which are not globally convergent). Other rational iteration schemes are also discussed, including Laurent approximations, Cayley power methods, and globally convergent eigenvalue assignment methods.

Key words. Padé approximation, matrix sign function, Riccati equations, rational iterations

AMS(MOS) subject classifications. 15A24, 65D99, 65F99

1. Introduction. It is a classical result that the algebraic Riccati equation can be solved by using an invariant subspace of an associated Hamiltonian matrix. This motivated the introduction, by Roberts [21] in 1971, of the matrix sign function as a means of finding the positive and negative invariant subspaces of any matrix X which does not have eigenvalues on the imaginary axis. This and subsequent work [9] showed that the matrix sign function could be used to solve many problems in control theory.

The sign of X can be defined constructively as the limit of the Newton sequence

$$X_{n+1} = \tfrac{1}{2}(X_n + X_n^{-1}), \qquad X_0 = X, \tag{1.1}$$

$$\operatorname{sgn}(X) \equiv \lim_{n \to +\infty} X_n. \tag{1.2}$$

Newton's method has the pleasant feature that it is globally convergent; if X has no eigenvalues on the imaginary axis then the limit in (1.2) exists. As a definition, however, (1.2) does not reveal many of the important properties of the sign function. Because of this, it is useful to have an equivalent definition based on the Jordan canonical form of X (see [4], [7]). For a complex scalar z with $\operatorname{Re} z \neq 0$, define the sign of z by

$$\operatorname{sgn} z = \begin{cases} 1 & \text{if } \operatorname{Re} z > 0, \\ -1 & \text{if } \operatorname{Re} z < 0. \end{cases} \tag{1.3}$$

For a complex matrix X such that $\Lambda(X) \subset \mathbb{C}^+ \cup \mathbb{C}^-$ (i.e., X has no eigenvalues on the imaginary axis) let T take X to Jordan form:

$$X = T^{-1} \begin{bmatrix} P & 0 \\ 0 & N \end{bmatrix} T, \tag{1.4}$$

* Received by the editors September 28, 1989; accepted for publication (in revised form) November 15, 1989. This research was supported by National Science Foundation (and Air Force Office of Scientific Research) grant ECS87-18897, National Science Foundation grant DMS88-00817, and Air Force Office of Scientific Research contract AFOSR-89-0167.

† Department of Electrical and Computer Engineering, University of California, Santa Barbara, California 93106 (laub%lanczos@hub.ucsb.edu).

where P and N are in block diagonal Jordan form with, respectively, positive and negative real part eigenvalues. Then the sign of X is given by

$$\operatorname{sgn}(X) = T^{-1}\begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} T, \tag{1.5}$$

where I and $-I$ in (1.5) have the same dimensions as P and N in (1.4). This shows immediately that the sign of X is a square root of the identity which commutes with X:

$$S^2 = I, \qquad XS = SX, \tag{1.6}$$

where $S = \operatorname{sgn}(X)$.

Using (1.4) in (1.1), shows that the eigenvalues $\lambda_j^{(n)}$ of X_n are decoupled from each other and obey the scalar recursions

$$\lambda_j^{(n+1)} = \frac{1}{2}\left(\lambda_j^{(n)} + \frac{1}{\lambda_j^{(n)}}\right), \qquad \lambda_j^{(0)} = \lambda_j(X), \tag{1.7}$$

with $\lim_{n \to +\infty} \lambda_j^{(n)} = \operatorname{sgn}(\lambda_j)$. This decoupling greatly simplifies the analysis of methods like (1.1).

Because of the need for pivoting, matrix inversions are sometimes not as amenable to parallel or vector implementation as matrix multiplications. Thus, a current trend in evaluating $\operatorname{sgn}(X)$ and related functions such as the polar decomposition [5], [11], [12] is to favor algorithms which are "multiplication-rich," such as the Newton–Schulz iteration

$$X_{n+1} = \tfrac{1}{2} X_n (3I - X_n^2). \tag{1.8}$$

(The recursion (1.8) is obtained from (1.1) by using Schulz's approximation $X_n^{-1} \cong X_n + (I - X_n^2)X_n$ as suggested in [12].) This method avoids the matrix inversion in (1.1) and is quadratically convergent provided

$$\|I - X^2\| < 1, \tag{1.9}$$

where $\|\cdot\|$ is any reasonable matrix norm (see Theorem 5.2). If (1.9) is not satisfied then a starter method such as (1.1) must be used until $\|I - X_n^2\| < 1$.

Higher-order polynomial recursions for the polar decomposition of a nonsingular matrix were developed independently by Kovarik [17] and Leipnik [18] and are applicable to the matrix sign function. These methods are based on polynomial approximations of the hypergeometric function

$$(1-\xi)^{-1/2} = 1 + \frac{1}{2}\xi + \frac{3}{8}\xi^2 + \cdots, \tag{1.10}$$

and generate convergent matrix sequences provided that (1.9) is satisfied. The motivation for studying this function is that for nonzero real x, $\operatorname{sgn} x = x/|x| = x/(1-\xi)^{1/2}$ where $\xi = 1 - x^2$. In § 3, we show that the sufficient condition (1.9) actually provides a rather good approximation to the true region of convergence for these methods. Consequently, we might feel that loss of global convergence is the price that must be paid in order to use multiplication-rich algorithms. Rather surprisingly, this is not the case.

For example, recursions based on rational (Padé) approximations of $(1-\xi)^{-1/2}$ have much larger regions of convergence. In fact, the main diagonal approximations (those for which the degree m of the denominator is equal to or one greater than the

degree k of the numerator) lead to globally convergent iterations that satisfy an elegant error formula:

$$(S-X_n)(S+X_n)^{-1}=(S-X_0)^{\gamma^n}(S+X_0)^{-\gamma^n}, \tag{1.11}$$

where $\gamma = k + m + 1$ is the order of the approximation. (For Newton's method, a similar result was proved by Balzer [3, eq. (39)] and by Roberts [21, § 1.3].) These methods are easily modified to allow exact one-step convergence of specified eigenvalues (much like the eigenvalue assignment schemes of Balzer in [3]) while still remaining globally convergent. An analysis of the Halley family of algorithms of Gander [10] for the polar decomposition shows that these methods belong to this class of assignment procedures. The work in [10] can also be adapted to give a local convergence theory for general sign function iterations of the form $X_{n+1} = F(X_n)$.

A second family of globally convergent multiplication-rich methods is based on the Cayley transform

$$Y=(I-X)(I+X)^{-1}, \tag{1.12}$$

which takes the positive real part eigenvalues of X inside the unit circle and the negative real part eigenvalues of X outside the unit circle. If Y is multiplied by itself repeatedly, then these eigenvalues move toward zero and infinity, respectively. Transforming back to get $\tilde{X}_\nu$,

$$\tilde{X}_\nu=(I-Y^\nu)(I+Y^\nu)^{-1} \tag{1.13}$$

moves these eigenvalues very near one and minus one, respectively. (If X has -1 as an eigenvalue, then $I + X$ is singular and a modified version of (1.12), (1.13) must be used.) A fascinating correspondence between the Cayley power method and the Padé approximation method is that if the power ν in (1.13) is equal to γ^n in (1.11), then X_n is equal to $\tilde{X}_\nu$! This does not mean, however, that these two methods should be viewed as identical because in this case the Padé method requires n matrix inversions while the Cayley method requires only two. Similar equivalency results for different members of the Padé method can also be proved (see Theorem 3.4). An interesting sidelight on the Cayley power method is that (1.12) can be replaced by any transformation which is a rational or analytic function of X that takes the right- and left-half complex planes inside and outside the unit disk, respectively. For example, if $Y = e^{-X}$ then Y^ν is just the fundamental solution matrix to $\dot{Y} = -XY$ at time ν: $Y^\nu = e^{-\nu X}$ and $\tilde{X}_\nu = (I - e^{-\nu X})(I + e^{-\nu X})^{-1}$. Note in this case that $I + e^{-\nu X}$ is never singular, since the eigenvalues of X are not on the imaginary axis.

In the next section we present the theory of the Padé approximants of $(1 - \xi)^{-1/2}$ for $k \geqq m - 1$, which is based on well-known results for hypergeometric functions. This theory is then used to analyze scalar sign function recursions in § 3, where we also show how it can be adapted to give globally convergent eigenvalue assignment iterations. In § 4 we consider other rational iterations including Laurent methods. These scalar results are useful because matrix convergence is predicated on the scalar convergence of the eigenvalues of X (§ 5). This leads to local convergence results for $k \geqq m - 1$, and global convergence for the main diagonal approximants $k = m$ and $k = m - 1$.

2. Padé approximations to $(1 - \xi)^{-1/2}$. Let $(\alpha)_n = (\alpha)(\alpha + 1)\cdots(\alpha + n - 1)$ with $(\alpha)_0 = 1$, and define the family of hypergeometric functions

$$ {}_2F_1(\alpha,\beta,\gamma,\xi)=\sum_{n=0}^{+\infty}\frac{(\alpha)_n(\beta)_n}{n!(\gamma)_n}\xi^n. \tag{2.1}$$

From [1],

$$(1-\xi)^{-1/2} = {}_2F_1(\tfrac{1}{2}, 1, 1, \xi) \equiv f(\xi). \tag{2.2}$$

In general, the $[k/m]$ Padé approximant to f is a rational function P_{km}/Q_{km} where $\deg(P_{km}) = k$, $\deg(Q_{km}) = m$, and

$$f(\xi) - \frac{P_{km}(\xi)}{Q_{km}(\xi)} = O(\xi^{k+m+1}). \tag{2.3}$$

Because f is a hypergeometric function, a great deal is known about P_{km} and Q_{km} [1]. First of all [13], Q_{km} is related to the set of orthogonal polynomials over [0, 1] defined with respect to the weight function $\omega(\xi) = (\xi^{-1/2}/\pi)(1-\xi)^{-1/2}\xi^{k+1-m}$ for $k \geqq m-1$. If ψ_m is the mth such polynomial with $\psi_m(1) = 1$, then

$$Q_{km}(\xi) = \xi^m \psi_m(\xi^{-1}), \tag{2.4}$$

and $Q_{km}(0) = 1$. From (2.4), the zeros of Q_{km} are just the reciprocals of the zeros of ψ_m. Since the zeros of ψ_m are simple [22] and lie in (0, 1), the zeros of Q_{km} are also simple and lie in $(1, \infty)$. (This result could have been anticipated from another point of view since $(1-\xi)^{-1/2}$ has a natural branch cut along $(1, \infty)$ and, as noted in [2, pp. 51–57], the zeros and poles of a Padé approximant tend to fall along the branchcuts of the functions they approximate.) Denoting the zeros of Q_{km} by $1 < z_1 < z_2 < \cdots < z_m$, we may write

$$Q_{km}(\xi) = \prod_{i=1}^{m} (z_i - \xi)/z_i. \tag{2.5}$$

This identity is useful for convergence analysis, but a more convenient form [1] is

$$\begin{aligned} Q_{km}(\xi) &= {}_2F_1(-m, -\tfrac{1}{2}-k, -k-m, \xi) \\ &= \sum_{n=0}^{m} \frac{(-m)_n(-\frac{1}{2}-k)_n \xi^n}{n!(-k-m)_n} \\ &\equiv \sum_{n=0}^{m} q_n^{km} \xi^n. \end{aligned} \tag{2.6}$$

From [13], P_{km} is given by

$$\begin{aligned} P_{km}(\xi) &= \sum_{n=0}^{k} \frac{(\frac{1}{2})_n(\frac{1}{2}-m)_m(n-k-m)_m}{n!(-k-m)_m(n+\frac{1}{2}-m)_m} \xi^n \\ &\equiv \sum_{n=0}^{k} P_n^{km} \xi^n. \end{aligned} \tag{2.7}$$

The key to the local error analysis of Padé recursions is the following theorem, which was proved by Leipnik [18, Thm. 1] and stated by Kovarik [17, lemma following Thm. 2] for the polynomial case $m = 0$.

THEOREM 2.1. *For $k \geqq m-1$,*

$$Q_{km}^2(\xi) - (1-\xi)P_{km}^2(\xi) = \xi^{k+m+1}\left(\sum_{i=1}^{\mu} c_i \xi^i\right), \tag{2.8}$$

where $c_i = c_i(k, m) > 0$ *for* $0 \leqq i \leqq \mu \equiv \max(2k+1, 2m) - (k+m+1)$, *and*

$$Q_{km}^2(1) = \sum_{i=1}^{\mu} c_i. \tag{2.9}$$

Proof. From (2.3) and the fact that $Q_{km}^2(\xi) - (1-\xi)P_{km}^2(\xi)$ is a polynomial of order $\mu + k + m + 1$,

$$Q_{km}^2(\xi) - (1-\xi)P_{km}^2(\xi) = \frac{Q_{km}^2}{f^2}\left(f - \frac{P_{km}}{Q_{km}}\right)\left(f + \frac{P_{km}}{Q_{km}}\right)$$

$$= \xi^{k+m+1}\left(\sum_{i=1}^{\mu} c_i \xi^i\right),$$

for some constants $c_0, c_1, \cdots, c_\mu$. Setting $\xi = 1$ gives (2.9). It remains to show that the coefficients c_i are positive. The idea of the proof is best illustrated by considering the diagonals, $m = k - t$, for $t = -1, 0, 1, \cdots, k$ in the Padé table. (For example, see Table 1.) For the first main diagonal, $t = -1$, $\mu = 0$ and multiplying out the left side of (2.8) gives $c_0 = (q_{k+1}^{kk+1})^2 > 0$. For the second main diagonal, $t = 0$, $\mu = 0$, and $c_0 = (P_k^{kk})^2 > 0$. For the first superdiagonal, $t = 1$, $\mu = 1$, and

$$c_0 = (P_k^{kk-1})^2 > 0, \qquad c_1 = P_k^{kk-1}(P_{k-1}^{kk-1} - P_k^{kk-1}) + P_k^{kk-1}P_{k-1}^{kk-1}.$$

In general, for $t \geqq 0$, $\mu = t$, and the coefficients, c_s can be written as the sum of terms of the form

$$P_{k-r}^{kk-t}(P_{k+r-s}^{kk-t} - P_{k+r-s+1}^{kk-t}), \tag{2.10}$$

and

$$P_{k-r}^{kk-t}P_{k+r-s}^{kk-t}, \tag{2.11}$$

where

$$0 \leqq r \leqq s \leqq t \leqq k. \tag{2.12}$$

We complete the proof of the theorem by showing that each term of the type (2.10) or (2.11) is positive. From (2.7),

$$P_{k+r-s}^{kk-t} = \frac{(\frac{1}{2})_{k+r-s}(\frac{1}{2}+t-k)_{k-t}(t+r-s-k)_{k-t}}{(k+r-s)!(t-2k)_{k-t}(\frac{1}{2}+t+r-s)_{k-t}}. \tag{2.13}$$

Since both P_{k+r-s}^{kk-t} and P_{k-r}^{kk-t} have sign $(-1)^{k-t}$,

$$P_{k-r}^{kk-t}P_{k+r-s}^{kk-t} > 0. \tag{2.14}$$

Using (2.13),

$$P_{k-r}^{kk-t}(P_{k+r-s}^{kk-t} - P_{k+r-s+1}^{kk-t}) = P_{k-r}^{kk-t}P_{k+r-s}^{kk-t}\left(1 - \frac{(s-r)(t-s+r+\frac{1}{2})}{(s-r+k-t)(k+r-s+1)}\right) > 0$$

by (2.14) and (2.12) because $(s-r)/(s-r+k-t) \leqq 1$ and

$$(t-s+r+\tfrac{1}{2})/(k-s+r+1) < 1.$$

(Note that the degenerate case $k = t = s = r$ does not cause a problem because (2.10) then reduces to $P_0^{k0}P_k^{k0}$, which is positive by (2.14).) □

3. Scalar Padé recursions. As we show in § 5, the convergence of the matrix sequence $\{X_n\}$ is determined by the convergence of the scalar sequences for the eigenvalues of X_0. The scalar Padé recursions have the form

$$x_{n+1} = x_n \frac{P_{km}(1-x_n^2)}{Q_{km}(1-x_n^2)}, \tag{3.1}$$

where P_{km}/Q_{km} is the $[k/m]$ Padé approximant to $(1-\xi)^{-1/2}$. Table 1 gives the expressions for the right-hand side of (3.1) for k and m between zero and three. For example, the case $k = 0$, $m = 1$ gives

$$x_{n+1} = \frac{2x_n}{1+x_n^2}, \tag{3.2}$$

which might be called the "inverse" Newton method for solving the equation $x^2 - 1 = 0$ since the values $x_1, x_2, \cdots$ generated by (3.2) are the inverses of those generated by the "regular" Newton method

$$x_{n+1} = \frac{1}{2}\left(x_n + \frac{1}{x_n}\right). \tag{3.3}$$

The case $k = 1$, $m = 1$ gives Halley's method (see [10] for a related application). The next theorem generalizes the local convergence results of Leipnik [18] and Kovarik [17].

THEOREM 3.1. *Let* $|1 - x_0^2| < 1$ *for* $x_0 \in \mathbb{C}$ *and define* $\{x_n\}$ *by* (3.1) *for* $k \geqq m - 1$. *Then*

$$|1-x_n^2| \leqq |1-x_0^2|^{(k+m+1)^n}, \tag{3.4}$$

and

$$\lim_{n \to +\infty} x_n = \operatorname{sgn}(x_0). \tag{3.5}$$

Proof. By (3.1),

$$1 - x_1^2 = (Q_{km}^2(\xi) - (1-\xi)P_{km}^2(\xi))/Q_{km}^2(\xi), \tag{3.6}$$

where $\xi = 1 - x_0^2$. But Q_{km} has zeros $z_1, \cdots, z_m$ in $(1, +\infty)$, so by (2.5)

$$|Q_{km}(\xi)| = \prod_{i=1}^{m} \frac{|z_i - \xi|}{z_i} \geqq \prod_{i=1}^{m} \frac{z_i - |\xi|}{z_i} > \prod_{i=1}^{m} \frac{z_i - 1}{z_i} = Q_{km}(1). \tag{3.7}$$

TABLE 1
Padé recursions for the matrix sign function.

	$k = 0$	$k = 1$	$k = 2$	$k = 3$
$m = 0$	x	$\frac{x}{2}(3-x^2)$	$\frac{x}{8}(15-10x^2+3x^4)$	$\frac{x}{16}(35-35x^2+21x^4-5x^6)$
$m = 1$	$\frac{2x}{1+x^2}$	$\frac{x(3+x^2)}{1+3x^2}$	$\frac{x}{4}\frac{(15+10x^2-x^4)}{(1+5x^2)}$	$\frac{x}{8}\frac{(35+35x^2-7x^4+x^6)}{(1+7x^2)}$
$m = 2$	$\frac{8x}{3+6x^2-x^4}$	$\frac{4x(1+x^2)}{1+6x^2+x^4}$	$\frac{x(5+10x^2+x^4)}{1+10x^2+5x^4}$	$\frac{x}{2}\frac{(35+105x^2+21x^4-x^6)}{(3+42x^2+35x^4)}$
$m = 3$	$\frac{16x}{5+15x^2-5x^4+x^6}$	$\frac{8x(3+5x^2)}{5+45x^2+15x^4-x^6}$	$\frac{2x(3+10x^2+3x^4)}{1+15x^2+15x^4+x^6}$	$\frac{x(7+35x^2+21x^4+x^6)}{1+21x^2+35x^4+7x^6}$

Using Theorem 2.1 in (3.6) gives

$$\begin{aligned} |1-x_1^2| &\leqq |\xi|^{k+m+1}\left(\sum_{i=1}^{\mu} c_i|\xi|^i\right)\Big/|Q_{km}(\xi)|^2 \\ &\leqq |1-x_0^2|^{k+m+1}\left(\sum_{i=1}^{\mu} c_i\right)\Big/|Q_{km}(\xi)|^2 \\ &\leqq |1-x_0^2|^{k+m+1}Q_{km}^2(1)/|Q_{km}(\xi)|^2 \\ &\leqq |1-x_0^2|^{k+m+1} \end{aligned}$$

by (2.9) and (3.7). Repeating this argument gives (3.4). From (3.4), $x_n^2 \to 1$. To see that $x_n \to \operatorname{sgn}(x_0)$, let $h(x) = xP_{km}(1-x^2)/Q_{km}(1-x^2)$. Since the only poles of h lie on the imaginary axis, h is continuous on the set

$$S \equiv \{x : |1-x^2| < 1\} \equiv S_+ \cup S_-, \tag{3.8}$$

where $S_+ \equiv \{x \in S : \operatorname{Re} x > 0\}$, $S_- \equiv \{x \in S : \operatorname{Re} x < 0\}$. By (3.4), h takes S into S. Since $S_+ \cap S_- = \varnothing$ and each is a connected set, $h(S_+)$ must lie entirely in S_+ or S_-, because the continuous image of a connected set is connected. But $1 \in S_+$ and $h(1) = 1 \in S_+$, so $h(S_+) \subset S_+$. Similarly, $h(S_-) \subset S_-$. Thus if $x_0 \in S_\pm$ then $x_n \in S_\pm$ for all n, and by (3.4),

$$\lim_{n \to +\infty} x_n = \operatorname{sgn}(x_0). \qquad \square$$

In order to assess how well the set S in (3.8) approximates the region of convergence for the recursions in (3.1), we define the basins of attraction for the fixed points ± 1 of h:

$$B_+ \equiv \{x : \lim_{n \to +\infty} x_n = 1\}, \qquad B_- \equiv \{x : \lim_{n \to +\infty} x_n = -1\}. \tag{3.9}$$

The Julia set [6], [19] for the recursion (3.1) is the boundary of the basin of attraction of $+1$:

$$J_{km} = \partial B_+. \tag{3.10}$$

Because of the unusual properties associated with Julia sets, J_{km} is also the boundary of the basin of attraction for -1:

$$J_{km} = \partial B_-. \tag{3.11}$$

(See [19] for a very readable introduction to Julia sets and the properties of rational recursions such as (3.1); for a deeper study, see [6].)

Computationally, J_{km} can be approximated by starting with (almost) any point $z_0 \in \mathbb{C}$ and then reversing (3.1) to solve for the predecessors of z_0:

$$z_n = z_{n+1}P_{km}(1-z_{n+1}^2)/Q_{km}(1-z_{n+1}^2), \tag{3.12}$$

where $z_n = x_{-n}$ in (3.1). Since (3.12) can be written as a polynomial in z_{n+1} of order $\mu_1 = \max(2k+1, 2m)$, there are μ_1 solutions $z_{n+1}^{(i)}$ to (3.12), one of which is selected at random to continue the iteration. This scheme takes advantage of the fact that for the forward recursion (3.1), the Julia set is repulsive; points near J_{km} move to ± 1. In reverse, under (3.12), the Julia set becomes attractive and nearly all orbits of points are dense in J_{km} (see [6, Thm. 2.5]). Thus by plotting $\{z_n^{(i)}\}$ for $n > 30$ (to allow the initial points time to approach the Julia set) we obtain a good graphical approximation of J_{km} and

thus can assess easily the real region of convergence of (3.1) as compared to the set $|1 - x^2| < 1$. This was done for each of the recursions given in Table 1 (excluding the globally convergent main diagonal recursions), and the results are displayed in Figs. 1–9, along with the set $|1 - x^2| = 1$ for comparison (this set looks like an "infinity" symbol centered at zero). In each of these figures, the principal domains of attraction of ± 1 are the largest connected regions, inside the Julia set, which contain ± 1, respectively. The other connected regions nested within the Julia set map onto these principal domains after a finite number of steps in (3.1). For the multiplication-rich polynomial recursions ($m = 0$), the set $|1 - x^2| < 1$ provides a rather good approximation to the actual region of convergence. However, as m increases toward k, that is, as we move toward the main diagonals $k = m$ or $k = m - 1$, the region of convergence becomes much larger than $|1 - x^2| < 1$.

We now show that along the main diagonals, the regions of convergence are as large as possible and we have, in fact, global convergence. That is, if x_0 is not on the imaginary axis then $\lim_{n \to +\infty} x_n = \operatorname{sgn}(x_0)$.

First note a rather remarkable property of (3.1) for $k = m$ and $k = m - 1$: the polynomials $-xP_{km}(1 - x^2)$ and $Q_{km}(1 - x^2)$ are, respectively, the odd and even parts of $(1 - x)^{k+m+1}$. This makes it very easy to write down the appropriate recursion. For example, if $k = m = 2$, then

$$(1-x)^{k+m+1} = 1 - 5x + 10x^2 - 10x^3 + 5x^4 - x^5,$$

so $-xP_{22}(1 - x^2) = -5x - 10x^3 - x^5$, and $Q_{22}(1 - x^2) = 1 + 10x^2 + 5x^4$. Thus the [2/2] recursion is $x_{n+1} = x_n(5 + 10x_n^2 + x_n^4)/(1 + 10x_n^2 + 5x_n^4)$. This property can be proved either by manipulating the series (2.6), (2.7) or by starting with $-xP_{km}(1 - x^2)$ and $Q_{km}(1 - x^2)$ as the odd and even parts of $(1 - x)^{k+m+1}$ and then showing that (2.3) is satisfied.

THEOREM 3.2. *Let $x_0 \in \mathbb{C}^+ \cup \mathbb{C}^-$ and let $\{x_n\}$ be defined by (3.1) for $k = m$ or $k = m - 1$. Then*

$$\frac{1-x_n}{1+x_n} = \left(\frac{1-x_0}{1+x_0}\right)^{(k+m+1)^n} \quad \text{for } x_0 \in \mathbb{C}^+ \tag{3.13}$$

and

$$\frac{1+x_n}{1-x_n} = \left(\frac{1+x_0}{1-x_0}\right)^{(k+m+1)^n} \quad \text{for } x_0 \in \mathbb{C}^-. \tag{3.14}$$

In either case, for $s = \operatorname{sgn}(x_0)$,

$$\frac{s-x_n}{s+x_n} = \left(\frac{s-x_0}{s+x_0}\right)^{(k+m+1)^n}. \tag{3.15}$$

Proof. Equations (3.13) and (3.14) are identical except for being inverses of each other to avoid division by zero when $x_0 = \pm 1$. Let $x_0 \in \mathbb{C}^+$ for convenience and set $x_1 = x_0 P_{km}(1 - x_0^2)/Q_{km}(1 - x_0^2)$. By the preceding remarks, for any x, and $k = m$ or $m - 1$,

$$Q_{km}(1-x^2) - xP_{km}(1-x^2) = (1-x)^{k+m+1}. \tag{3.16}$$

Replacing x by $-x$ gives

$$Q_{km}(1-x^2) + xP_{km}(1-x^2) = (1+x)^{k+m+1}. \tag{3.17}$$

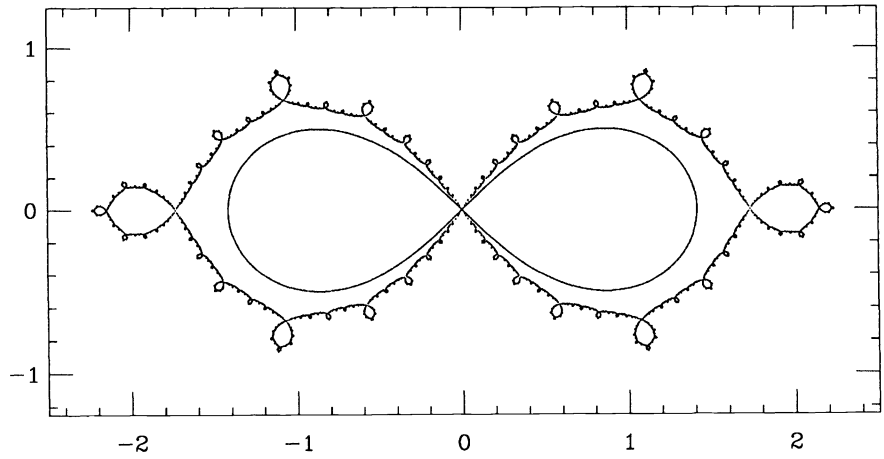


FIG. 1. *Padé convergence region for* $k = 1$, $m = 0$.

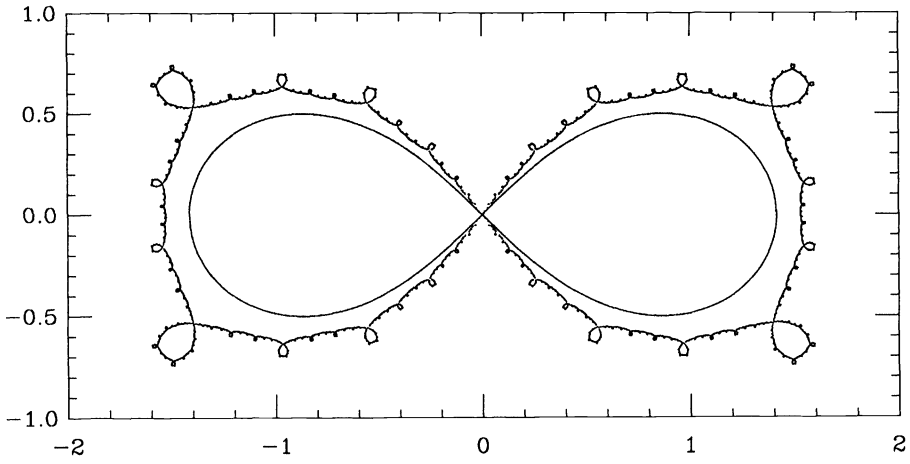


FIG. 2. *Padé convergence region for* $k = 2$, $m = 0$.

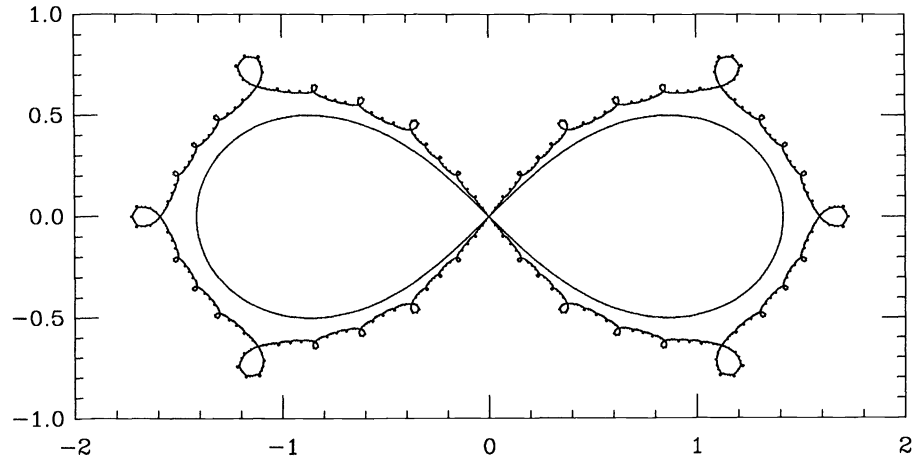


FIG. 3. *Padé convergence region for* $k = 3$, $m = 0$.

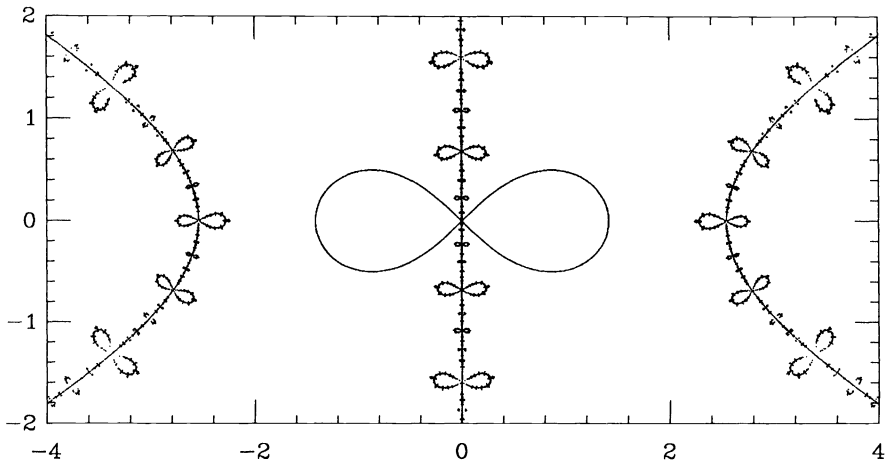


FIG. 4. *Padé convergence region for* $k = 0$, $m = 2$.

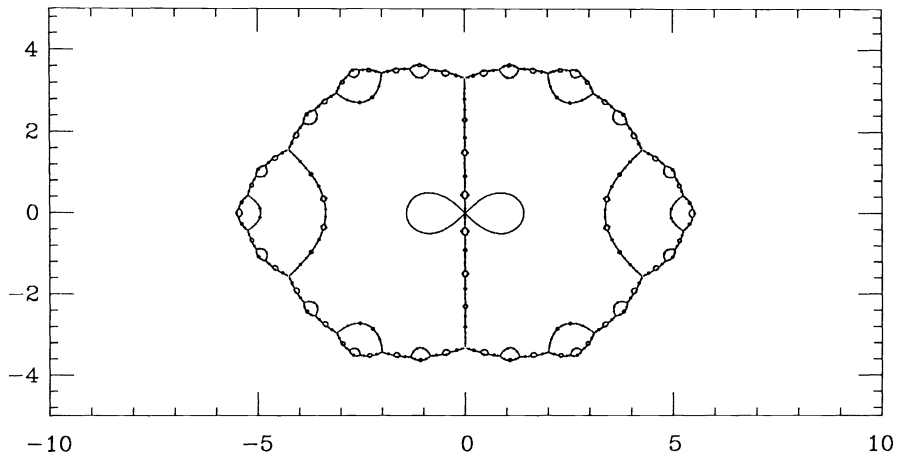


FIG. 5. *Padé convergence region for* $k = 2$, $m = 1$.

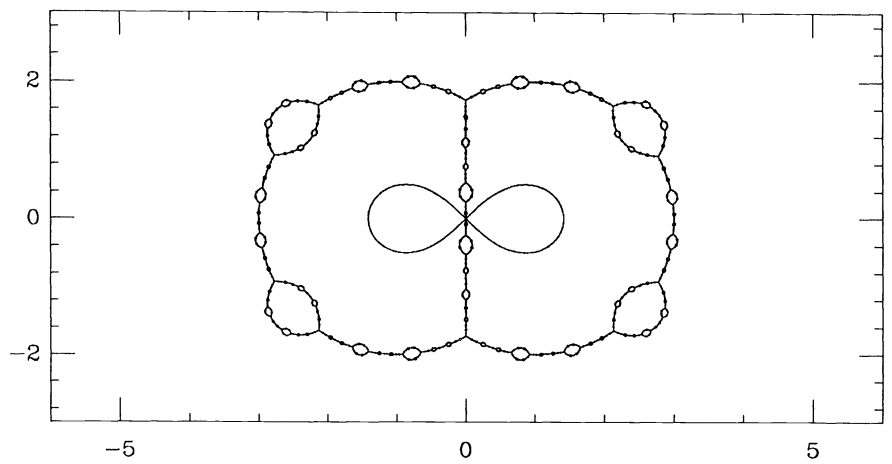


FIG. 6. *Padé convergence region for* $k = 3$, $m = 1$.

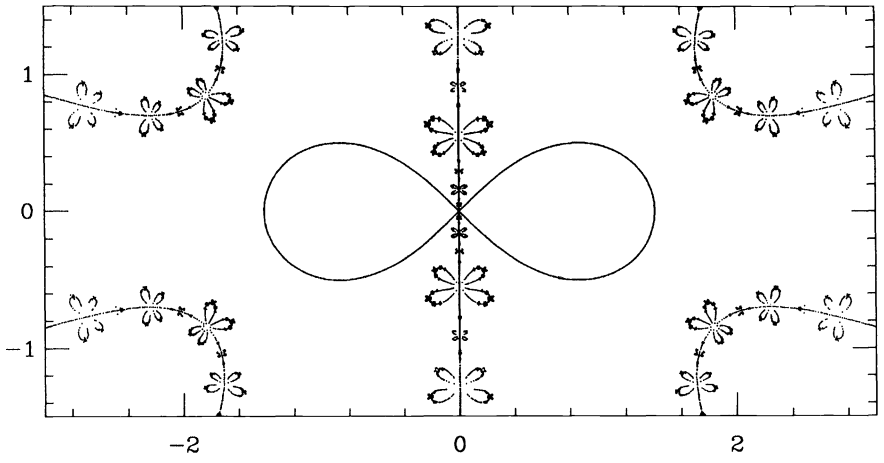


FIG. 7. *Padé convergence region for* $k = 0$, $m = 3$.

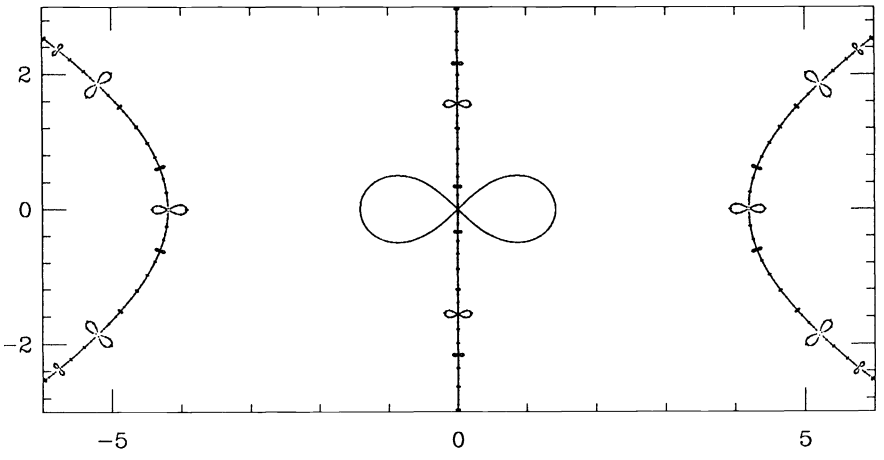


FIG. 8. *Padé convergence region for* $k = 1$, $m = 3$.

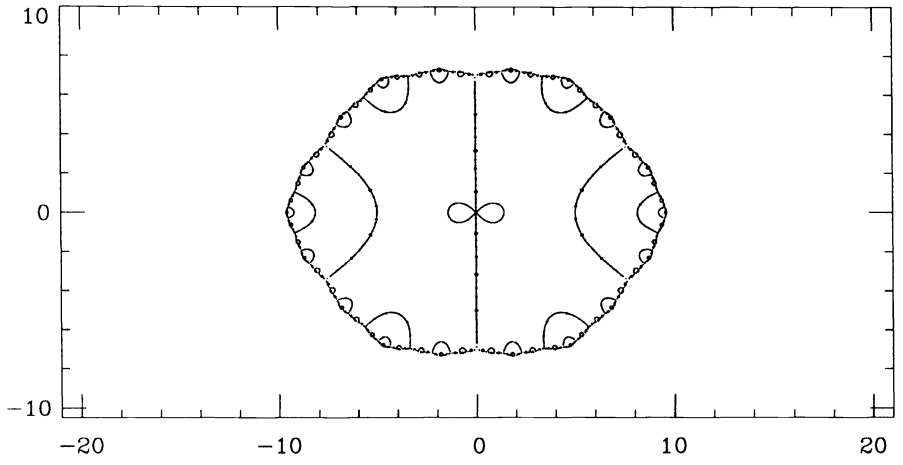


FIG. 9. *Padé convergence region for* $k = 3$, $m = 2$.

Thus,

$$1-x_1=(Q_{km}(1-x_0^2)-x_0P_{km}(1-x_0^2))/Q_{km}(1-x_0^2)$$
$$=(1-x_0)^{k+m+1}/Q_{km}(1-x_0^2),$$

and

$$1+x_1=(1+x_0)^{k+m+1}/Q_{km}(1-x_0^2).$$

Dividing, we obtain (3.13) for $n = 1$. Repeat to get the general statement. □

From Theorem 3.2, we immediately get Theorem 3.3.

THEOREM 3.3 (Global Convergence). *If $x_0 \in \mathbb{C}^+ \cup \mathbb{C}^-$, then for $k = m$ or $k = m-1$, with $m \geqq 1$ $\lim_{n\to+\infty} x_n = \operatorname{sgn}(x_0)$.*

Proof. By Theorem 3.2, we need only show that $|(1-x_0)/(1+x_0)| < 1$ for $x_0 \in \mathbb{C}^+$ or $|(1+x_0)/(1-x_0)| < 1$ for $x_0 \in \mathbb{C}^-$. Let $x_0 = \rho e^{i\theta} \in \mathbb{C}^+$ with $-\pi/2 < \theta < \pi/2$. Then $|(1-x_0)/(1+x_0)|^2 = (1-2\rho\cos\theta+\rho^2)/(1+2\rho\cos\theta+\rho^2) < 1$, and similarly for $x_0 \in \mathbb{C}^-$. □

From Theorem 3.2, we see that the distance measure from x_0 to 1 given by $d_+(x_0) \equiv |(1-x_0)/(1+x_0)|$ and its counterpart for -1, $d_-(x_0) \equiv |(1+x_0)/(1-x_0)|$, are more natural than $|1-x|$ and $|1+x|$, respectively. For example, $x_0 = 10^{-6}$ and $1/x_0 = 10^6$ are equidistant from 1 under the regular Newton method (since (1.2) is symmetric with respect to x_0 and $1/x_0$) but $|1-x_0| \cong 1$ while $|1-1/x_0| \cong 10^6$. (See [3] and [15].)

Theorem 3.2 is also useful in establishing the equivalence of certain methods in the Padé table. For example, if $x_0 \in \mathbb{C}^+$, then two steps of the inverse Newton method ($k = 0$, $m = 1$) give

$$\frac{1-x_2}{1+x_2}=\left(\frac{1-x_0}{1+x_0}\right)^4. \tag{3.18}$$

However, if $\tilde{x}_1$ denotes the result of taking one step from x_0 with the recursion ($k = 1$, $m = 2$), then

$$\frac{1-\tilde{x}_1}{1+\tilde{x}_1}=\left(\frac{1-x_0}{1+x_0}\right)^4. \tag{3.19}$$

Solving for x_2 and $\tilde{x}_1$, we find $x_2 = \tilde{x}_1$. Similarly, if we take one step with ($k = 0$, $m = 1$) followed by a step with ($k = 3$, $m = 3$) the result would be the same as one step with ($k = 6$, $m = 7$).

THEOREM 3.4 (Equivalency). *Let $x_0 \in \mathbb{C}^+ \cup \mathbb{C}^-$ and let x_r be the result of applying r steps of the (possibly different) main diagonal Padé recursions $[k_1/m_1], \cdots, [k_r/m_r]$. Then $x_r = \tilde{x}_{\tilde{r}}$, where $\tilde{x}_{\tilde{r}}$ is obtained by $\tilde{r}$ main diagonal steps $[\tilde{k}_1/\tilde{m}_1], \cdots, [\tilde{k}_{\tilde{r}}/\tilde{m}_{\tilde{r}}]$, provided that both are of the same order, i.e.,*

$$\prod_{i=1}^{r}(k_i+m_i+1)=\prod_{i=1}^{\tilde{r}}(\tilde{k}_i+\tilde{m}_i+1)\equiv\rho. \tag{3.20}$$

Proof. Applying Theorem 3.2 for each individual step,

$$\left(\frac{1-x_r}{1+x_r}\right)=\left(\frac{1-x_0}{1+x_0}\right)^{\rho}=\left(\frac{1-\tilde{x}_{\tilde{r}}}{1+\tilde{x}_{\tilde{r}}}\right).$$

Solving for x_r and $\tilde{x}_{\tilde{r}}$ gives $x_r = \tilde{x}_{\tilde{r}}$. If $x_0 \in \mathbb{C}^-$, use (3.14). □

4. Other rational methods. In this section we consider other rational iterations, including eigenvalue assignment methods, Cayley transform methods, and Laurent series methods. Eigenvalue assignment methods were introduced by Balzer [3], in the form of scaled Newton methods which move specified real eigenvalues to $x = 1$ in one step. These methods were shown to be globally but not quadratically convergent. By using the methods of Theorem 3.2, it is easy to construct globally convergent methods of arbitrarily high order that will move any selected set $\{\lambda_i\}$ of real or complex conjugate eigenvalues to $x = 1$ in one step.

For example, if we want a fourth-order method which assigns $\lambda_1 = 2$, $\lambda_2 = 1 + i$, and $\lambda_3 = 1 - i$ to $x = 1$, then we let $-xp(x^2)$ and $q(x^2)$ be, respectively, the odd and even terms in the expansion of $(1-x)^4(2-x)(1+i-x)(1-i-x)$:

$$(1-x)^4(2-x)(1+i-x)(1-i-x)$$
$$= 4 - 22x + 52x^2 - 69x^3 + 56x^4 - 28x^5 + 8x^6 - x^7.$$

Then

$$xp(x^2) = 22x + 69x^3 + 28x^5 + x^7 = x(22 + 69x^2 + 28x^4 + x^6),$$
$$q(x^2) = 4 + 52x^2 + 56x^4 + 8x^6 = 4(1 + 13x^2 + 14x^4 + 2x^6),$$

and the desired iteration is

$$x_{n+1} = \frac{x_n(22 + 69x_n^2 + 28x_n^4 + x_n^6)}{4(1 + 13x^2 + 14x^4 + 2x^6)}.$$

In order to prove global convergence for these assignment methods, we need the following lemma.

LEMMA 4.1. *Let* $\operatorname{Re} z > 0$, $\operatorname{Re} \lambda > 0$, *and* $r > 0$. *Then*

$$\left|\frac{r-z}{r+z}\right| < 1, \tag{4.1}$$

and

$$\left|\left(\frac{\lambda - z}{\lambda + z}\right)\left(\frac{\bar{\lambda} - z}{\bar{\lambda} + z}\right)\right| < 1. \tag{4.2}$$

Proof. If we set $x = z/r$, then $\operatorname{Re} x > 0$ and

$$\left|\frac{r-z}{r+z}\right| = \left|\frac{1-x}{1+x}\right| < 1,$$

as in the proof of Theorem 3.3. Now say $\lambda = re^{i\theta}$, $z = \rho e^{i\phi}$ where $\phi, \theta \in (-\pi/2, \pi/2)$. Then

$$\begin{aligned}|(\lambda - z)(\bar{\lambda} - z)|^2 &= (r^2 - 2\rho r \cos\theta \cos\phi + \rho^2 \cos 2\phi)^2 \\ &\quad + \sin^2\phi(2\rho^2 \cos\phi - 2\rho r \cos\theta)^2 \\ &< (r^2 + 2\rho r \cos\theta \cos\phi + \rho^2 \cos 2\phi)^2 \\ &\quad + \sin^2\phi(2\rho^2 \cos\phi + 2\rho r \cos\theta)^2 \\ &= |(\lambda + z)(\bar{\lambda} + z)|^2.\end{aligned}$$

□

THEOREM 4.2. *Let* $\{\lambda_1, \lambda_2, \cdots, \lambda_\mu\}$ *be a conjugate symmetric set in the open right-half plane and* $-xp(x^2)$ *and* $q(x^2)$ *be, respectively, the odd and even parts of* $(1-x)^\gamma(\lambda_1 - x)\cdots(\lambda_\mu - x)$. *Then the iterative method*

$$x_{n+1} = \frac{x_n p(x_n^2)}{q(x_n^2)} \tag{4.3}$$

is globally convergent of order γ *and takes* $\{\lambda_1, \lambda_2, \cdots, \lambda_\mu\}$ *to* $x = 1$ *in one step. Moreover, for* $s = \operatorname{sgn}(x_0)$,

$$\frac{s - x_{n+1}}{s + x_{n+1}} = \left(\frac{s - x_n}{s + x_n}\right)^\gamma \left(\frac{\lambda_1 s - x}{\lambda_1 s + x}\right)\cdots\left(\frac{\lambda_\mu s - x}{\lambda_\mu s + x}\right), \tag{4.4}$$

and

$$\left|\frac{s - x_{n+1}}{s + x_{n+1}}\right| \leqq \left|\frac{s - x_n}{s + x_n}\right|^\gamma. \tag{4.5}$$

Proof. We shall prove (4.4) and (4.5) for the case $s = 1$, since the case $s = -1$ follows immediately. From (4.3),

$$\frac{1 - x_{n+1}}{1 + x_{n+1}} = \frac{q(x_n^2) - x_n p(x_n^2)}{q(x_n^2) + x_n p(x_n^2)}$$

$$= \frac{(1-x)^\gamma(\lambda_1 - x)\cdots(\lambda_\mu - x)}{(1+x)^\gamma(\lambda_1 + x)\cdots(\lambda_\mu + x)},$$

which proves (4.4). Inequality (4.5) then follows from (4.4) and Lemma 4.1. □

Remark 1. Since $xp(x^2)/q(x^2)$ in (4.3) is an odd function, it also moves $\{-\lambda_1, -\lambda_2, \cdots, -\lambda_\mu\}$ to -1 in one step.

Remark 2. In [10], Gander gives a family of quadratically convergent methods which depend on a parameter f:

$$x_{n+1} = x_n \frac{2f - 3 + x_n^2}{f - 2 + f x_n^2}. \tag{4.6}$$

In Theorem 2 of [10], it is shown that (4.6) is globally convergent for $f > 2$ and for $f = 3$ gives Halley's method, which is cubically convergent. For $f < 2$, prescaling must be done to ensure convergence. We can interpret Gander's method as a second-order method which makes one real eigenvalue assignment. Expand

$$(1-x)^2(\lambda - x) = \lambda - (2\lambda + 1)x + (2 + \lambda)x^2 - x^3,$$

and use the method of Theorem 4.2 to obtain the iteration

$$x_{n+1} = \frac{x_n(2\lambda + 1 + x_n^2)}{\lambda + (2 + \lambda)x_n^2}. \tag{4.7}$$

This is the same as (4.6) for $\lambda = f - 2$. Thus the condition $f > 2$ for global convergence in (4.6) is just the requirement that the real eigenvalue λ, which gets mapped to $x = 1$, must be in the right-half plane as in Theorem 4.2. Moreover, $f = 3$ corresponds to $\lambda = 1$ being triply assigned to $x = 1$, so that the iteration is cubically convergent (Halley's method).

Remark 3. Allowing some of the eigenvalues λ in Theorem 4.2 to be multiple results in methods in which λ is mapped to $x = 1$ and points near λ are taken at least quadratically to $x = 1$. For example, expanding $(1-x)^2(2-x)^2$ gives the second-order method in which $\lambda = 2$ is doubly assigned to one:

$$x_{n+1} = \frac{x_n(12+6x_n^2)}{4+13x_n^2+x_n^4}.$$

If $x_0 = 2.1$, then $x_1 = .99985\cdots$.

As indicated in the Introduction, another family of methods can be based on the Cayley transform. For $x \neq -1$, let

$$y = \frac{1-x}{1+x}. \tag{4.8}$$

Let $\tilde{x}_\nu$ denote the result of multiplying y by itself ν times and then transforming back:

$$\tilde{x}_\nu = \frac{1-y^\nu}{1+y^\nu}. \tag{4.9}$$

From this we see that

$$\frac{(1-\tilde{x}_\nu)}{(1+\tilde{x}_\nu)} = y^\nu = \left(\frac{1-x}{1+x}\right)^\nu. \tag{4.10}$$

Now suppose that x_n is defined by (3.1) for one of the main diagonal ($k = m$ or $k = m-1$) Padé recursions where $\nu = (k+m+1)^n$. By the Equivalency Theorem 3.4, we must have

$$\tilde{x}_\nu = x_n. \tag{4.11}$$

Thus the Cayley transform method and the Padé recursions produce exactly the same results, except that the arithmetic operations of inversion and multiplication have been rearranged. It was pointed out earlier that this can have a significant effect in the matrix case, since the Cayley transform approach is multiplication-rich compared to the Padé methods. We now extend the Cayley transform method to the case where $x = -1$ or where -1 is an eigenvalue of X in the matrix case.

From (4.8) and (4.9),

$$\tilde{x}_\nu = \frac{1-\left(\frac{1-x}{1+x}\right)^\nu}{1+\left(\frac{1-x}{1+x}\right)^\nu} = \frac{(1+x)^\nu-(1-x)^\nu}{(1+x)^\nu+(1-x)^\nu}. \tag{4.12}$$

The next lemma shows that the right-hand side of (4.12) is well defined for any x which is not on the imaginary axis.

LEMMA 4.3. *Let* $x \in \mathbb{C}^+ \cup \mathbb{C}^-$. *Then* $(1+x)^\nu + (1-x)^\nu \neq 0$ *for any positive integer* ν.

Proof. Suppose to the contrary that $(1+x)^\nu + (1-x)^\nu = 0$. Then $x \neq 1$, so $(1+x)^\nu/(1-x)^\nu = -1$. This means that $(1+x)/(1-x)$ is a νth root of -1: $(1+x)/(1-x) = e^{i\theta}$ where θ is not an odd multiple of π (else $x = +\infty$). Solving for x we find $x = (\sin\theta/(1+\cos\theta))i \notin \mathbb{C}^+ \cup \mathbb{C}^-$, which is a contradiction. □

We end this section with a short discussion of Laurent methods, which are polynomial iterations in x and x^{-1} of the form

$$x_{n+1} = \sum_{j=-\nu}^{\nu} b_{j\nu} x_n^j.$$

These methods are motivated by a desire to generate a "multiplication-rich" iteration once X^{-1} has been computed. For example, Newton's method is of this form with $\nu = 1$, $b_{-11} = \frac{1}{2} = b_{11}$. If we let

$$L(x) = \sum_{j=-\nu}^{\nu} b_{j\nu} x^j,$$

then the coefficients $b_{j\nu}$ can be determined from $L(1) = 1$, $L'(1) = 0, \cdots, L^{(2\nu-1)}(1) = 0$. (Other conditions which assign specified eigenvalues to $x = 1$ can be used as well.)

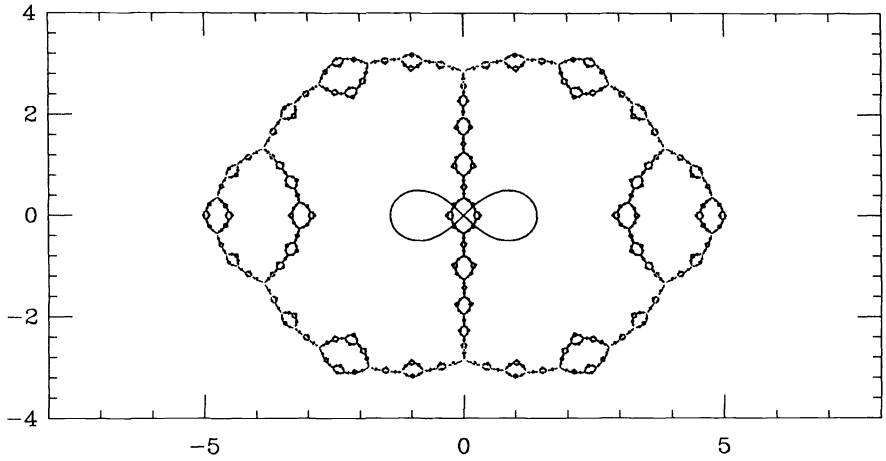


FIG. 10. *Laurent convergence region for* $\nu = 3$.

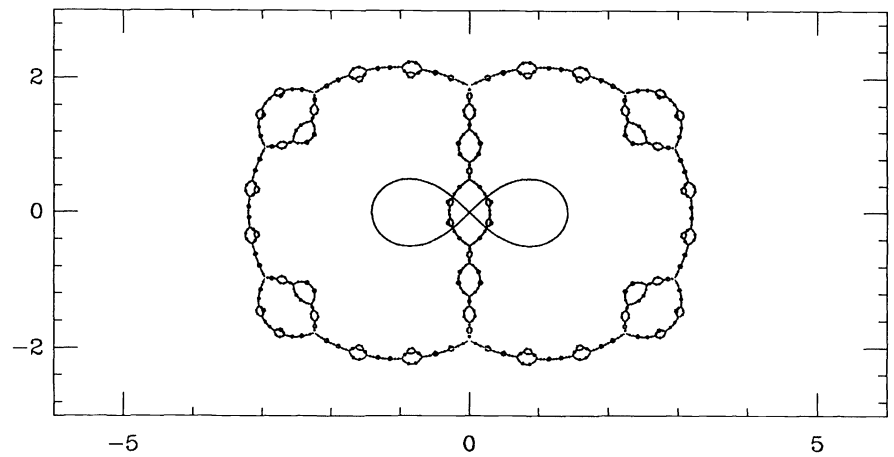


FIG. 11. *Laurent convergence region for* $\nu = 5$.

Because of symmetry reasons we generally want L to be an odd function, $L(-x) = -L(x)$, so that ν should be odd and $b_{j\nu} = 0$ whenever j is even. After Newton's method ($\nu = 1$) the next two methods ($\nu = 3$ and $\nu = 5$) are of order four and six, respectively, and take the form

$$x_{n+1} = \frac{1}{16}\left(-\frac{1}{x_n^3} + \frac{9}{x_n} + 9x_n - x_n^3\right) \quad \text{for } \nu = 3,$$

$$x_{n+1} = \frac{1}{7552}\left(\frac{73}{x_n^5} - \frac{660}{x_n^3} + \frac{4270}{x_n} + 4580x_n - 815x_n^3 + 104x_n^5\right) \quad \text{for } \nu = 5.$$

These methods are multiplication-rich in the sense that they require one matrix inversion and $\nu + 1$ multiplies per step. However, they are not globally convergent and, in fact, the region of convergence for these two methods does not even include the set $|x^2 - 1| < 1$, as do the Padé methods. This is illustrated in Figs. 10 and 11, where the set $|x^2 - 1| = 1$ is included for comparison.

5. Matrix convergence. In this section we show that convergence in the matrix case is determined by the scalar convergence of the eigenvalues. This allows us to apply the scalar convergence results of the previous sections to the matrix case.

The following general result is the key to this process.

LEMMA 5.1. *Let $R = R(x)$ be an odd rational function such that $R(1) = 1$ and $R'(1) = 0$. Let $x_0 \in \mathbb{C}^+ \cup \mathbb{C}^-$ such that $\lim_{n\to+\infty} x_n = \operatorname{sgn}(x_0)$, where $x_{n+1} = R(x_n)$. Let X_0 be a Jordan block of the form*

$$X_0 = \begin{bmatrix} x_0 & 1 & & \\ & x_0 & \ddots & 0 \\ & & \ddots & 1 \\ 0 & & & x_0 \end{bmatrix}.$$

Then the matrix sequence defined by $X_{n+1} = R(X_n)$ satisfies $\lim_{n\to+\infty} X_n = \operatorname{sgn}(X_0)$.

Proof. Let $R_1(x) = R(x)$, $R_2(x) = R(R(x))$, and in general $R_{n+1}(x) = R(R_n(x))$. Because X_0 is a Jordan block,

$$X_n = R_n(X_0) = \begin{bmatrix} a_1 & a_2 & \cdots & a_\nu \\ & a_1 & \ddots & \vdots \\ & & \ddots & a_2 \\ 0 & & & a_1 \end{bmatrix}, \tag{5.1}$$

where ν is the order of X_0 and

$$a_j = a_j(n) = \frac{1}{(j-1)!}\frac{d^{j-1}}{dx^{j-1}}R_n\Big|_{x_0}.$$

Thus $a_1(n) = R_n(x_0) = x_n \to \operatorname{sgn}(x_0)$ by assumption. For $j = 2$,

$$a_2(n) = \frac{dR_n}{dx}\Big|_{x_0} = \frac{dR}{dx}\Big|_{R_{n-1}(x_0)} \frac{dR_{n-1}}{dx}\Big|_{x_0} = \frac{dR}{dx}\Big|_{x_{n-1}} a_2(n-1),$$

by the chain rule. But $\lim_{n\to+\infty} dR/dx|_{x_{n-1}} = dR/dx|_{\operatorname{sgn}(x_0)} = 0$ since $\operatorname{sgn}(x_0) = \pm 1$ and $dR/dx(\pm 1) = 0$ by assumption. Thus $a_2(n) \to 0$.

As an induction hypothesis suppose that $a_j(n) \to 0$ for $2 \leqq j \leqq i-1$. Then by the chain rule

$$a_i(n) = \left.\frac{dR}{dx}\right|_{x_{n-1}} a_i(n-1) + r_n,$$

where r_n has a fixed form, independent of n, involving sums and products of a_j for $2 \leqq j \leqq i-1$. Thus $r_n \to 0$ by the induction hypothesis. Since $dR/dx|_{x_{n-1}}$ also tends to zero we have $\lim_{n\to+\infty} a_i(n) = 0$. This means that $\lim_{n\to+\infty} X_n = \operatorname{sgn}(x_0)I = \operatorname{sgn}(X_0)$. □

Using Lemma 5.1, we obtain the matrix analogues of Theorems 3.1–3.4 and the Cayley power method.

THEOREM 5.2. *Let $k \geqq m-1$ and assume that the eigenvalues of X_0 lie in $\mathbb{C}^+ \cup \mathbb{C}^-$. Assume that $\|I - X_0^2\| < 1$ and define*

$$X_{n+1} = -X_n P_{km}(I - X_n^2) Q_{km}^{-1}(I - X_n^2).$$

Then

$$\|I - X_n^2\| < \|I - X_0^2\|^{(k+m+1)^n}, \tag{5.2}$$

and

$$\lim_{n\to+\infty} X_n = \operatorname{sgn}(X_0). \tag{5.3}$$

Proof. The condition $\|I - X_0^2\| < 1$ ensures that $|1 - \lambda^2| < 1$ for any eigenvalue λ of X_0. Hence by Theorem 3.1, the eigenvalues $\lambda_n^{(i)}$ for X_n converge to $\operatorname{sgn}(\lambda_0^{(i)})$. By Lemma 5.1 and the definition of $\operatorname{sgn}(X_0)$ in terms of its Jordan form, (5.3) is true. The matrix inequality (5.2) can be obtained by using the matrix analogue of the arguments in the proof of Theorem 3.1. □

THEOREM 5.3. *Let $\Lambda(X_0) \subset \mathbb{C}^+ \cup \mathbb{C}^-$ and assume that $k = m$ or $k = m-1$ in (4.3). Then for $\gamma = k + m + 1$*

$$\lim_{n\to+\infty} X_n = \operatorname{sgn} X_0 \equiv S, \tag{5.4}$$

$$(S - X_n)(S + X_n)^{-1} = [(S - X_0)(S + X_0)^{-1}]^{\gamma^n}, \tag{5.5}$$

and

$$X_n = (A^{\gamma^n} - B^{\gamma^n})(A^{\gamma^n} + B^{\gamma^n})^{-1}, \tag{5.6}$$

where

$$A = I + X_0 \quad \text{and} \quad B = I - X_0. \tag{5.7}$$

Proof. By Theorem 3.3 the eigenvalues of X_0 converge under (3.1) to the appropriate value of ± 1. By Lemma 5.1, this means that $\lim_{n\to\infty} X_n = \operatorname{sgn}(X_0)$. Equation (5.5) is obtained by considering the individual Jordan blocks and using (3.15). Similarly, use Lemma 4.3 to see that (5.6) is true for each Jordan block and hence for X_n itself. □

6. Conclusion. In this paper, we have presented a theory of rational recursions for the matrix sign function, including Padé, Laurent, Cayley transform, and eigenvalue assignment methods. Of particular interest are the globally convergent main diagonal Padé iterations and their multiplication-rich Cayley transform equivalents.

Several important aspects concerning the numerical evaluation of sign function iterations have been treated elsewhere and so have not been discussed here. For example, scaling can significantly increase the speed of convergence of X_n to sgn (X) as noted in [3] and [4]; for scaling related to the polar decomposition, see [11]. The choice of optimal and nearly optimal scaling constants for Newton's method is discussed at length in [15] and it is not hard to adapt these results to the main diagonal Padé recursions. Similarly, the problem of estimating the sensitivity of the sign of a matrix is considered in [16], based on the work in [8], [14], and [20].

REFERENCES

[1] G. BAKER, *Essentials of Padé Approximants*, Academic Press, New York, 1975.

[2] G. BAKER AND P. GRAVES-MORRIS, *Padé approximants*, Vol. 13, Encyclopedia of Mathematics and Its Applications, Gian-Carlo Rota, ed., Addison-Wesley, London, 1981.

[3] L. A. BALZER, *Accelerated convergence of the matrix sign function method of solving Lyapunov, Riccati and other matrix equations*, Internat. J. Control, 32 (1980), pp. 1057–1078.

[4] G. BIERMAN, *Computational aspects of the matrix sign function solution to the* ARE, in Proc. 23rd Conference on Decision and Control, Las Vegas, NV, 1984, pp. 514–519.

[5] Å BJÖRCK AND C. BOWIE, *An iterative algorithm for computing the best estimate of an orthogonal matrix*, SIAM J. Numer. Anal., 8 (1971), pp. 358–364.

[6] H. BROLIN, *Invariant sets under iteration of rational functions*, Ark. Mat., 6 (1967), pp. 103–144.

[7] R. BYERS, *Solving the algebraic Riccati equation with the matrix sign function*, Linear Algebra Appl., 85 (1987), pp. 267–279.

[8] A. CLINE, C. B. MOLER, G. W. STEWART, AND J. H. WILKINSON, *An estimate for the condition number of a matrix*, SIAM J. Numer. Anal., 16 (1979), pp. 368–375.

[9] E. D. DENMAN AND A. N. BEAVERS, *The matrix sign function and computation in systems*, Appl. Math. Comput., 2 (1976), pp. 63–94.

[10] W. GANDER, *Algorithms for the polar decomposition*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 1102–1115.

[11] N. J. HIGHAM, *Computing the polar decomposition—with applications*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1160–1174.

[12] N. J. HIGHAM AND R. S. SCHREIBER, *Fast polar decomposition of an arbitrary matrix*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 648–655.

[13] C. KENNEY AND A. J. LAUB, *Padé error estimates for the logarithm of a matrix*, Internat. J. Control, 50 (1989), pp. 707–730.

[14] ———, *Condition estimates for matrix functions*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 191–209.

[15] ———, *On scaling Newton's method for polar decomposition and the matrix sign function*, SIAM J. Matrix Anal. Appl., 13 (1992), to appear.

[16] ———, *Polar decomposition and matrix sign function condition estimates*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 488–504.

[17] Z. KOVARIK, *Some iterative methods for improving orthonormality*, SIAM J. Numer. Anal., 7 (1970), pp. 386–389.

[18] R. B. LEIPNIK, *Rapidly convergent recursive solution of quadratic operator equations*, Numer. Math., 17 (1971), pp. 1–16.

[19] H. O. PEITGEN, D. SAUPE, AND F. V. HAESELER, *Cayley's problem and Julia sets*, Math. Intelligencer, 6 (1984), pp. 11–20.

[20] J. R. RICE, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287–310.

[21] J. D. ROBERTS, *Linear model reduction and solution of the algebraic Riccati equation by use of the sign function*, Internat. J. Control, 32 (1980), pp. 677–687.

[22] G. SZEGÖ, *Orthogonal Polynomials*, American Mathematical Society, Providence, RI, 1939.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 2, pp. 292–309, April 1991

007

THE RESTRICTED TOTAL LEAST SQUARES PROBLEM: FORMULATION, ALGORITHM, AND PROPERTIES*

SABINE VAN HUFFEL† AND HONGYUAN ZHA‡

Abstract. The restricted total least squares (RTLS) problem, presented in this paper, is devised for solving overdetermined sets of linear equations $AX \approx B$ in which the data $[A; B]$ are perturbed by errors of the form $E^* = DEC$. D and C are known matrices and E is an arbitrary but bounded matrix. By choosing D and C appropriately, the RTLS problem formulation can handle any weighted least squares (LS), generalized LS, total LS, and generalized total LS problem. Also, equality constraints can be imposed.

In order to solve these problems, a computationally efficient and numerically reliable restricted TLS algorithm, based on the restricted singular value decomposition (RSVD), of the matrix triplet $([A; B], D, C)$, is developed. This RSVD is a generalization of the ordinary SVD for triple matrix products. The matrices involved may be rank-deficient and the explicit formation of matrix inverses and products is avoided. Using the RSVD, some properties of the RTLS problem are proven.

Key words. generalized total least squares, generalized least squares, restricted singular value decomposition, numerical linear algebra

AMS(MOS) subject classifications. 15A18, 65F20

C.R. classification. G1.3

1. Introduction.

Every linear parameter estimation problem gives rise to an overdetermined set of linear equations $AX \approx B$. Usually, $R(B) \not\subseteq R(A)$ and hence this set does not have an exact solution. In these cases, a best estimate $\hat{X}$ of X is found by fitting a best subspace S to the data $[A; B]$ such that rank $(\tilde{A})$ = rank $([\tilde{A}; \tilde{B}])$, where $[\tilde{A}; \tilde{B}]$ is the projection of $[A; B]$ into S. By solving this adjusted set $\tilde{A}X = \tilde{B}$, $\hat{X}$ is obtained.

In the ordinary least squares (LS) approach the measurements A are assumed to be free of error ($\tilde{A} = A$) and hence, all errors are confined to the right-hand side matrix B. However, this assumption is frequently unrealistic: sampling errors, human errors, modelling errors, and instrument errors may imply inaccuracies in A as well. For those cases, the total least squares (TLS) approach has been devised and amounts to fitting a "best" subspace to $[A; B]$ when the errors in the measurements A and B are uncorrelated with zero mean and equal variance. As proven in [9], this TLS solution $\hat{X}$ is a strongly consistent estimate of the true solution X of the corresponding unperturbed set $A_0X = B_0$ provided $\lim_{m\to\infty} A_0^TA_0/m$ exists and is positive definite, i.e., $\hat{X}$ converges to X with probability one as the number of equations m is going to infinity.

However, in many linear parameter estimation problems some columns of A may be error-free. Moreover, the errors in the remaining data may be correlated and not equally sized. In order to maintain consistency of the result when solving these problems, the ordinary TLS problem has been generalized [29], [27]. The generalized TLS (GTLS) problem assumes that the first columns A_1 of $A = [A_1; A_2]$ are error-free and that square nonsingular, error equilibration matrices D and C are known such that the errors in $D^{-1}[A_2; B]C^{-1}$ are equilibrated, i.e., uncorrelated with zero mean and equal variance.

* Received by the editors April 5, 1989; accepted for publication (in revised form) February 23, 1990.

† Electronics, Systems, Automation and Technology (ESAT) Laboratory, Department of Electrical Engineering, Katholieke Universiteit Leuven, Kardinaal Mercierlaan 94, B-3001 Heverlee, Belgium, and Belgian National Fund of Scientific Research (N.F.W.O.) (vanhuffel@esat.kuleuven.ac.be). This author's research was sponsored by the Belgian National Fund of Scientific Research (N.F.W.O.).

‡ Department of Computer Science, Stanford University, Stanford, California 94305 (zha@na-net.stanford.edu).

This GTLS problem can now be further generalized. Before defining this so-called *restricted* TLS problem, we introduce our notation used throughout this paper.

A matrix is always represented by a capital letter, e.g., A. The corresponding lowercase letter with the subscript i and ij refers to the ith column and (i, j)th entry respectively, e.g., a_i, a_{ij}.

The superscript T denotes the transpose of a vector or matrix.

The $m \times m$ identity matrix is denoted by I_m.

The notation $\operatorname{diag}(\alpha_1, \cdots, \alpha_p)$, $p = \min\{m, n\}$, is used to denote an $m \times n$ matrix A, defined by $a_{ij} = 0$ whenever $i \neq j$ and $a_{ii} = \alpha_i$ for $i = 1, \cdots, p$.

$R(M)$, $\operatorname{Null}(M)$, $M^\dagger$, $\|M\|_F$, and $\|M\|_2$ denote, respectively, the range, null space, pseudoinverse, Frobenius norm, and 2-norm of a matrix M.

$\mathscr{E}(\cdot)$ is the expected value operator.

Restricted TLS formulation. Consider the following set of m linear equations in $n \times d$ unknowns X:

$$AX \approx B \quad A \in \mathscr{R}^{m \times n}, \quad B \in \mathscr{R}^{m \times d}, \quad X \in \mathscr{R}^{n \times d}, \tag{1}$$

where the data matrix $[A; B] = [A_0; B_0] + E^*$. A_0, B_0 are the error-free data and E^* represents a "restricted" perturbation matrix of the form

$$E^* = DEC. \tag{2}$$

$D \in \mathscr{R}^{m \times p}$ and $C \in \mathscr{R}^{q \times (n+d)}$ are known matrices while $E \in \mathscr{R}^{p \times q}$ is unknown and arbitrary but bounded.

Then, the problem of finding a matrix $[\Delta\hat{A}; \Delta\hat{B}]$ of the form $[\Delta\hat{A}; \Delta\hat{B}] = D\hat{E}C$ such that

$$R(B - \Delta\hat{B}) \subseteq R(A - \Delta\hat{A}), \tag{3}$$

and

$$\|\hat{E}\|_F \quad \text{is minimal,} \tag{4}$$

is referred to as the **restricted TLS** (RTLS) **problem** and any X satisfying

$$(A - \Delta\hat{A})X = B - \Delta\hat{B} \tag{5}$$

is called a **RTLS solution**.

Whenever the RTLS solution is **not unique,** a **weighted minimum norm** solution, denoted by $\hat{X}$, is singled out in the sense that $\|F_1\hat{X}F_2\|_F$ is minimized with $F_1 \in \mathscr{R}^{n \times n}$ and $F_2 \in \mathscr{R}^{d \times d}$ appropriately chosen nonsingular weighting matrices.

The term "*restricted* TLS" originates in its applications. We try to find a matrix $\hat{E}$ of minimal (unitarily invariant) norm that reduces the rank of $[A; B] - D\hat{E}C$ with given $[A; B]$, D, and C sufficiently such that the approximate set $([A; B] - D\hat{E}C)\begin{bmatrix} X \\ -I \end{bmatrix} = 0$ is solvable. Hence, we attempt to reduce the rank of $[A; B]$ by *restricting* the modifications to the column space of D and the row space of C.

Observe that this RTLS formulation is more general than the GTLS formulation given in [29] and [27]. Indeed, D and C may be rectangular, even rank-deficient, and error-free columns in A, if any, need not be stored in the first columns of A. Using an appropriately chosen D and C, even columns in B may be error-free and equality constraints can be imposed (see § 2). Finally, if the RTLS solution is not unique, then not only the solution $\hat{X}$ with minimal $\|\hat{X}\|_F$ can be singled out but any other $\hat{X}$ with an appropriately "weighted" minimal norm $\|F_1\hat{X}F_2\|_F$.

The primary goal of this paper is to present a general problem formulation which includes a whole variety of well-known problems, together with an efficient and numer-

ically reliable algorithm which solves these problems. In fact, the RTLS formulation can handle any LS, generalized LS, TLS, and generalized TLS problem, as well as any variant of these problems, and even more general problems which are not yet fully investigated. Also equality constraints and weighting matrices can be included. In § 2 it is shown how the aforementioned problems are converted to RTLS problems, illustrating the generality of the RTLS problem. Not only does this unified approach allow for an elegant problem formulation but at the same time it provides a deeper geometrical and algebraic insight in the connections between the different problems.

In order to solve the RTLS problem, the restricted singular value decomposition (RSVD), introduced by Zha [31], [32], can readily be applied to (3)–(4) to yield the solution of the RTLS problem. In essence, the RSVD applies to a given triplet of (possibly complex) matrices T, D, C of compatible dimensions and provides a factorization of the matrix T, relative to the matrices D and C. It can be considered as the ordinary SVD of the matrix T, but with different (possibly nonnegative definite) inner products applied in its column and in its row space. The properties and structure of the RSVD are investigated in detail in [5] and [32], as well as its connection to generalized eigenvalue problems, canonical correlation analysis and other generalizations of the SVD. Additionally in [5], a lot of applications are discussed. Just as the SVD (respectively, generalized SVD) is a valuable tool for the solution and analysis of the TLS (respectively, generalized TLS) problem, so the RSVD plays the same role of the more general RTLS problem, as pointed out in § 3.

Based on this RSVD, a powerful and numerically reliable RTLS algorithm is outlined in § 4 and can be used in practice to compute the RTLS solution. Its main difference with respect to the ordinary TLS algorithm is the use of the RSVD instead of the ordinary SVD. Its greatest advantage is the fact that only *one* algorithm is needed to solve a wide variety of problems. Of course, there are much faster direct ways of solving special RTLS problems (see, e.g., [12]). Nevertheless, if the problem is not too large and since computing power is generally so cheap, it will often make sense to use the RTLS algorithm at least when code for it becomes part of widely available and reliable subroutine packages. This is because the added information on structure and sensitivity that it provides, the ease of switching from one problem to another and analyzing the impact of any restriction on the sensitivity of the solution, can be very helpful in understanding the problem and finding the best problem formulation.

Although RTLS problems in their most general form are not yet fully understood and investigated, it is the authors' belief that the RTLS problem and the RTLS algorithm will become an important tool in the analysis and numerical solution of numerous problems.

Finally, § 5 gives the conclusions.

2. Special RTLS problems. It is easy to see that the RTLS formulation can handle any LS and generalized LS problem, as well as every TLS and generalized TLS problem. Indeed:

If D equals $m \times m$ identity matrix, denoted by I_m, and $C = I_{n+d}$ (respectively, $C = [0_{d\times n}; I_d]$), then we have the **ordinary TLS** (respectively, **LS**) formulation [11], [12], [26]. LS and TLS problems arise in a broad class of scientific disciplines such as signal processing [28], system identification [2], [13], automatic control or in general engineering, statistics [9], economics, medicine [23], etc.

If $D = I_m$ and $C = [0_{q\times n_1}; I_q]$ with $q = n + d - n_1$, a **mixed LS-TLS** problem is obtained: this is an extension of the ordinary TLS and LS problems which assumes the

first n_1 columns of A to be error-free [10], [23], [5]. For instance, in regression analysis, e.g., in curve fitting and intercept models [9], we often encounter such problems, as well as in system identification [22] and signal processing applications [28] whenever some signals can be observed without error while the other ones are disturbed by zero-mean white noise.

The **generalized TLS (GTLS)** problem also assumes that the first n_1 columns A_1 of $A = [A_1; A_2]$ are error-free but moreover allows for correlations between the errors E_2^* in the noisy submatrix $[A_2; B]$ provided square, nonsingular matrices D and C_2 are known such that the elements of $D^{-1}E_2^*C_2^{-1}$ are equilibrated, i.e., uncorrelated with equal variance. By taking $C = [0; C_2]$, C_2, and D as defined above, we have $\hat{E} = [0_{m\times n_1}; D^{-1}[\Delta\hat{A}_2; \Delta\hat{B}]C_2^{-1}]$, which shows the correspondence between the RTLS formulation and the more restrictive GTLS formulation used in [27] and [24]. If $D = I_m$, then C_2 is, up to a factor of proportionality, given by the square root of the error covariance matrix $\mathcal{E}(E_2^{*T}E_2^*)$ which defines the correlations between the errors in each row of $[A_2; B]$. These GTLS problems frequently occur in regression analysis [8], as well as in transfer function modelling [28] and the identification of multi-input multi-output systems whose outputs, and possibly the inputs, are disturbed by zero-mean correlated noise [13], [15], [16].

If rank $(D) = p$ and $C = [0_{d\times n}; I_d]$, we obtain the **generalized LS (GLS)** problem formulated as follows [19]:

$$\min \|E\|_F \quad \text{such that } B = AX + DE.$$

D represents, up to a factor of proportionality σ, the square root of the symmetric nonnegative-definite covariance matrix $\mathcal{E}(E_2^*E_2^{*T}) = \sigma^2 DD^T$, where E_2^* represents the errors in the right-hand side matrix B. Many estimation problems are led to the solution of a GLS problem. For instance, in control engineering the Kalman filter can be viewed as a GLS problem [14]. In regression analysis, the one-dimensional ($d = 1$) GLS solution provides the best linear unbiased estimate of X in all cases of the general Gauss–Markov model, given by $b = Ax + e^*$, where e^* is a zero-mean random vector with covariance matrix $\sigma^2 DD^T$ [19]. Whenever the coefficient matrix A is ill conditioned, this GLS solution is very sensitive to perturbations. In order to stabilize the solution and decrease its variance, Zha and Hansen [33] suggested adding Tikhonov regularization and proposed the following **regularized** Gauss–Markov model: $\min\{\|e\|_2^2 + \lambda^2\|Cx\|_2^2\}$ such that $Ax + De = b$. Just as the generalized SVD is the main tool for the analysis and solution of the general Gauss–Markov model, so the restricted SVD plays the same role of the regularized Gauss–Markov model, as pointed out in [33].

The RTLS formulation can also handle **any variant** of the GTLS and GLS problem allowing for error-free columns in A_2 and B as well, e.g., the RTLS problem:

$$(6) \qquad AX \approx B, \quad E^* = [DEC_1; 0_{m\times d}] \quad \text{with } C = [C_1; 0_{q\times d}] \quad \text{and } D, C_1 \text{ known,}$$

defines a GTLS problem which assumes B to be error-free. If $C_1 = [0_{1\times(i-1)}, 1, 0_{1\times(n-i)}]$, $q = d = 1$, then (6) defines a GLS problem in which A and B are error-free except for the ith column of A. Furthermore, the RTLS formulation also allows to solve any **(under)determined** set of linear equations (i.e., $m \leqq n$), defined as follows:

$$\min_{\hat{X}} \|F_1\hat{X}F_2\|_F \quad \text{such that } A\hat{X} = B.$$

This set always has one exact solution $\hat{X}$. Hence, E^* must be set to zero in the RTLS formulation, e.g., by taking $D = 0$ or $C = 0$.

Finally, observe that **equality constraints** can be imposed, given by the error-free rows of $[A; B]$, e.g., if the first m_1 rows of $[A; B]$ represent equality constraints, then

$$D = \begin{bmatrix} 0_{m_1 \times p} \\ D_2 \end{bmatrix}.$$

By adding the assumption that also n_1 columns of $[A; B]$ are error-free, a whole class of RTLS problems $AX \approx B$ can be defined, characterized by the fact that only one submatrix of $[A; B]$ is perturbed and may be changed by the RTLS algorithm, e.g.,

$$D = \begin{bmatrix} D_1 \\ 0_{m_1 \times p} \end{bmatrix} \quad \text{and} \quad C = [C_1; 0_{q \times n_1}].$$

These structured perturbation problems have been treated by Demmel [3], [4] and arise in the design of control systems based on H^∞ optimization [7] and in stability analysis of various problems in linear algebra. In practice, engineering design problems are very often formulated as optimization problems using LS (or TLS) approaches. As pointed out in [1], many of these problems involve equality constraints which represent some physical laws, e.g., in inverse kinematics of redundant and parallel manipulators, robot trajectory planning, mechanical systems design, etc. As shown above, constrained linear LS problems are easily transformed into RTLS problems but also constrained nonlinear LS problems can be solved with RTLS provided the solution method alters the problem in each iteration step to a constrained linear LS problem of the form:

$$\min_x (f - Ex)^T W (f - Ex) \quad \text{or} \quad \min_x \| W^{1/2} (f - Ex) \|_2 \tag{7}$$

$$\text{subject to} \quad Gx = h. \tag{8}$$

W is a positive-definite weight matrix and $W^{1/2}$ is its square root. By defining

$$A = \begin{bmatrix} G \\ E \end{bmatrix}^{m_1}, \qquad B = \begin{bmatrix} h \\ f \end{bmatrix},$$

$$D = \begin{bmatrix} 0 \\ W^{1/2} \end{bmatrix}^{m_1}, \qquad C = [0, \cdots, 0, 1]$$

in (1)–(2), (7)–(8) can be reformulated as an RTLS problem. An example of a constrained TLS problem, discussed in [21], consists in estimating the granulometry of minerals in a separator. Here, the equality constraints represent mass balance equations of the system in equilibrium.

If $F_1 \neq I_m$ or $F_2 \neq I_d$, the corresponding **"weighted"** problems are considered, e.g., the GLS problem given in [30] and [6] ($d = 1$):

$$\min_{e,x} \|e\|_2^2 \quad \text{such that } b = Ax + De \quad \text{with } \|F_1 x\|_2 \text{ minimal},$$

is a special RTLS problem with $C = [0_{1 \times n}; 1]$, $F_2 = 1$, and F_1, D as defined above. This problem is encountered in many important applications, e.g., aircraft wing flutter analysis [17].

Finally, observe that the RTLS problem is **not always solvable**. Two kinds of unsolvable RTLS problems must be distinguished. First of all, the RTLS problem is un-

solvable if no matrix $\hat{E}$ exists such that rank $([A; B] - D\hat{E}C) \leqq n$, e.g., the following RTLS problem described in [4]:

$$[A;B]=\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} \quad \text{and} \quad [\Delta\hat{A};\Delta\hat{B}]=D\hat{E}C \quad \text{given by} \begin{bmatrix}0\\ 1\end{bmatrix}\hat{e}[0\ 1]=\begin{bmatrix}0 & 0\\ 0 & \hat{e}\end{bmatrix}.$$

Since the corrections $[\Delta\hat{A}; \Delta\hat{B}]$, applied to the data $[A; B]$, may only affect the perturbed entries in $[A; B]$ (specified by D and C), the rank of $[A; B]$ cannot be sufficiently reduced in these problems in order to obtain a solvable set $(A - \Delta\hat{A})X = B - \Delta\hat{B}$. Second, condition (4) may be satisfied but not condition (3), i.e., there exists an $\hat{E}$ with minimal $\|\hat{E}\|_F$ such that rank $([A; B] - D\hat{E}C) \leqq n$ but $R(B - \Delta\hat{B}) \not\subseteq R(A - \Delta\hat{A})$. In this case, many $\mathcal{O}(\varepsilon)$ perturbations $\hat{E}_\varepsilon$ of $\hat{E}$ can make $R(B - \Delta\tilde{B}) \subseteq R(A - \Delta\tilde{A})$, where $[\Delta\tilde{A}; \Delta\tilde{B}] = D\hat{E}_\varepsilon C$, but there is no smallest value of $\|\hat{E}_\varepsilon\|_F$ for which (3) is satisfied, e.g., [12, p. 420]:

$$[A;B]=\begin{bmatrix}1 & 0 & 1\\ 0 & 0 & 1\\ 0 & 0 & 1\end{bmatrix}, \quad C=I, \quad D=I, \quad \hat{E}_\varepsilon=\begin{bmatrix}0 & 0 & 0\\ 0 & \varepsilon & 0\\ 0 & \varepsilon & 0\end{bmatrix}.$$

This kind of unsolvability only occurs when A is (nearly) rank-deficient or when the set of equations (1) is highly conflicting. By imposing additional constraints on the solution space, i.e., for all $\begin{bmatrix}w\\ 0\end{bmatrix} \in \text{Null}(D\tilde{E}C)$, for all $\tilde{E} \in \{E \mid \|E\|_F$ minimal and rank $([A; B] - DEC) \leqq n\}$:

$$[\Delta\hat{A};\Delta\hat{B}]\begin{bmatrix}w\\ 0\end{bmatrix}\begin{matrix}n\\ d\end{matrix}=0 \Leftrightarrow D\hat{E}C\begin{bmatrix}w\\ 0\end{bmatrix}=0, \tag{9}$$

these RTLS problems can be made solvable and are referred to as *nongeneric* RTLS problems. We say that these RTLS problems are **not** solvable in the **generic** sense (i.e., condition (4), subject to (3), is not satisfied) but may be solvable in the **nongeneric** sense (i.e., condition (4), subject to (3) and (9), is satisfied). See [26] for more details.

Note that the nongeneric RTLS problem can also be considered as a slightly changed generic RTLS problem $AX \approx B$, $E^* = DE\tilde{C}$ in which $\tilde{C}$ is an appropriately chosen projection of C into $R(C)$ such that

$$R(\tilde{C}) \perp R\left(\left\{\begin{bmatrix}w\\ 0\end{bmatrix}\begin{matrix}n\\ d\end{matrix} \middle| \begin{bmatrix}w\\ 0\end{bmatrix} \text{ satisfies (9)}\right\}\right).$$

3. Relation with the restricted SVD and properties. While the ordinary TLS algorithm is strongly based on the ordinary SVD [12], the RTLS algorithm is strongly based on the restricted SVD (RSVD) as shown in § 4. The RSVD, introduced by Zha [31], [32] as the implicit SVD of a triple matrix product (see also [5]) and therefore closely related to the S, T-SVD [30] and the HK-SVD [6], can be described as follows.

DEFINITION. **Restricted singular values (RSVs)** of (T, D, C). Let $T \in \mathbb{R}^{m\times n}$ be perturbed by a matrix of the form $E^* = DEC$ where $D \in \mathbb{R}^{m\times p}$, $E \in \mathbb{R}^{p\times q}$, and $C \in \mathbb{R}^{q\times n}$. Then the RSVs of the triplet (T, D, C) are defined as follows:

$$\sigma_i(T,D,C)=\min_{E\in\mathbb{R}^{p\times q}}\{\|E\|_2 \mid \text{rank}\,(T+DEC)\leqq i-1\}, \qquad i=1,\cdots,n.$$

The RSVs are arranged in nondecreasing order of magnitude, i.e., $\sigma_i \geqq \sigma_{i+1}$. Observe that the RSVs are generalizations of the well-known ordinary singular values and generalized singular values. Indeed, the ordinary singular values are given by $\sigma_i(T, I_m, I_n)$

while the generalized singular values are given by $\sigma_i(T, I_m, C)$ or $\sigma_i(T, D, I_n)$ [32]. If, for some i, no matrix E exists such that $\operatorname{rank}(T + DEC) \leqq i - 1$, we simply define $\sigma_i(T, D, C) = \infty$. Also, define $\sigma_i(T, D, C) = 0$ if $m < n$ and $m + 1 \leqq i \leqq n$.

These RSVs are computed from the RSVD of (T, D, C), as follows.

THEOREM 1. **Restricted singular value decomposition (RSVD)** of (T, D, C). *If* $T \in \mathbb{R}^{m\times n}$, $D \in \mathbb{R}^{m\times p}$, *and* $C \in \mathbb{R}^{q\times n}$, *then there exist orthonormal* $U \in \mathbb{R}^{p\times p}$ *and* $V \in \mathbb{R}^{q\times q}$ *and nonsingular* $P \in \mathbb{R}^{m\times m}$ *and* $Q \in \mathbb{R}^{n\times n}$ *such that*

$$PTQ = \begin{bmatrix} \Sigma_T & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{matrix} s_2, \\ t_2 \end{matrix} \qquad PDU = \begin{bmatrix} \Sigma_D \\ 0 \end{bmatrix} t_2, \qquad V^T CQ = [\Sigma_C; 0]$$

$$\begin{matrix} & s_1 & t_1 \end{matrix} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad t_1$$

where

$$\Sigma_T = \begin{bmatrix} I_j & 0 & 0 & 0 \\ 0 & I_k & 0 & 0 \\ 0 & 0 & I_l & 0 \\ 0 & 0 & 0 & \tilde{S}_T \end{bmatrix}, \qquad \Sigma_D = \begin{bmatrix} I_j & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \tilde{S}_D & 0 \\ 0 & 0 & 0 & I_{s_2} \end{bmatrix} \begin{matrix} \\ k+l \\ \\ \end{matrix},$$

$$\Sigma_C = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & I_l & 0 & 0 \\ 0 & 0 & \tilde{S}_C & 0 \\ 0 & 0 & 0 & I_{s_1} \end{bmatrix},$$

$$j+k$$

and $\tilde{S}_T = \operatorname{diag}(\tilde{\tau}_1, \cdots, \tilde{\tau}_s) \in \mathbb{R}^{s\times s}$, $\tilde{S}_D = \operatorname{diag}(\tilde{\delta}_1, \cdots, \tilde{\delta}_s) \in \mathbb{R}^{s\times s}$ *and* $\tilde{S}_C = \operatorname{diag}(\tilde{\gamma}_1, \cdots, \tilde{\gamma}_s) \in \mathbb{R}^{s\times s}$, $\tilde{\tau}_g > 0$, $\delta_g > 0$, $\tilde{\gamma}_g > 0$, *for all* $g = 1, \cdots, s$.

Let $ljk = l + j + k$ *and define*

$$\begin{aligned}
\tau_i &= 1, & \delta_i &= 1, & \gamma_i &= 0, & i &= 1, \cdots, j, \\
&= 1, & &= 0, & &= 0, & i &= j+1, \cdots, j+k, \\
&= 1, & &= 0, & &= 1, & i &= j+k+1, \cdots, ljk, \\
&= \tilde{\tau}_g, & &= \tilde{\delta}_g, & &= \tilde{\gamma}_g, & i &= ljk+g \quad \textit{and} \quad g = 1, \cdots, s, \\
&= 0, & &= 1, & &= 1, & i &= ljk+s+1, \cdots, ljk+s+\min(s_1, s_2),
\end{aligned}$$

then the RSVs *of* (T, D, C) *are given by*

$$\begin{aligned}
\sigma_i(T, D, C) &= \frac{\tau_i}{\delta_i \gamma_i}, & i &= 1, \cdots, \operatorname{rank}(T) \\
&= 0, & i &= \operatorname{rank}(T) + 1, \cdots, n.
\end{aligned}$$

Trivial RSVs, *defined by* $\frac{0}{0}$, *are set to zero*.

This theorem is proven in [32]. The integer indices are determined by the rank of the given matrices. Defining

$$r_{tdc} = \operatorname{rank}\left(\begin{bmatrix} T & D \\ C & 0 \end{bmatrix}\right), \qquad r_{tc} = \operatorname{rank}\left(\begin{bmatrix} T \\ C \end{bmatrix}\right),$$

$$r_{td} = \operatorname{rank}([T; D]),$$

we have

$$j = r_{tc} + \operatorname{rank}(D) - r_{tdc},$$

$$k = r_{tdc} - \operatorname{rank}(D) - \operatorname{rank}(C),$$

$$l = r_{td} + \operatorname{rank}(C) - r_{tdc},$$

$$r = r_{tdc} + \operatorname{rank}(T) - r_{td} - r_{tc},$$

$$s_1 = r_{tc} - \operatorname{rank}(T) \quad \text{and} \quad t_1 = n - r_{tc},$$

$$s_2 = r_{td} - \operatorname{rank}(T) \quad \text{and} \quad t_2 = m - r_{td}.$$

The properties and structure of the RSVD are investigated in detail in [5] and [32]. In particular, it is shown that the RSVD not only allows for an elegant treatment of algebraic and geometric problems in a wide variety of applications, but that its structure provides a powerful tool in simplifying proofs and derivations that are algebraically rather complicated. Additionally in [5], many applications are discussed including the analysis of the extended shorted operator, unitarily invariant norm minimization with rank constraints, rank minimization in matrix balls, the analysis and solution of linear matrix equations. In particular, it is shown how the RSVD allows us to solve a special class of RTLS problems, namely, constrained TLS problems $AX \approx B$ with some error-free rows and columns in $[A; B]$ and the close connection to Carlson's generalized Schur complement is emphasized. Finally, generalized Gauss–Markov models with and without constraints are discussed: it is shown how these models convert to special RTLS problems and how the RSVD simplifies the solution of these generalized LS problems with constraints.

As pointed out in [32], RSVs play an important role in the **rank determination** of a matrix T_0 under a restricted perturbation of the form $E^* = DEC$ where E satisfies $\|E\|_2 \leqq \varepsilon$ and ε is known. It is easy to see that the best estimate of the rank of $T_0 = T - DEC$, (T, D, C) known, is given by

$$\operatorname{rank}_\varepsilon(T) = \max\{i \mid \sigma_i(T, D, C) > \varepsilon\},$$

i.e., there exists a matrix $\hat{E}$ satisfying $\|\hat{E}\|_2 \leqq \varepsilon$ such that $\operatorname{rank}(T - D\hat{E}C) = i$ but no $\tilde{E}$ exists satisfying $\|\tilde{E}\|_2 \leqq \varepsilon$ such that $\operatorname{rank}(T - D\tilde{E}C) < i$. This matrix $\hat{E}$ can be easily computed from the RSVD of (T, D, C) and allows us to compute the RTLS correction matrix $[\Delta\hat{A}; \Delta\hat{B}] = D\hat{E}C$ satisfying (3)–(4) (see § 4). Indeed, comparing the definition of the RSVs with the RTLS condition (4), the following theorem can be proven.

THEOREM 2. *Assume that the* RTLS *problem* $A_{m\times n}X \approx B_{m\times d}$ *with given* $D_{m\times p}$, $C_{q\times(n+d)}$ *is solvable and* $\sigma_n([A; B], D, C) > \sigma_{n+1}([A; B], D, C)$, *then*:

The RSVD *of the* RTLS *approximation* $([A; B] - [\Delta\hat{A}; \Delta\hat{B}], D, C)$ *is given by the* RSVD *of* $([A; B], D, C)$, *in which the smallest* d RSVs *are equal to zero*.

Proof. Since the RTLS problem is solvable, there exists an E such that $R(B - \Delta B) \subseteq R(A - \Delta A)$, $[\Delta A; \Delta B] = DEC$. This implies that there exist $X \in \mathbb{R}^{n\times d}$ such that $(A - \Delta A)X = B - \Delta B$, i.e.,

$$\{[A;B] - DEC\}\begin{bmatrix} X \\ -I_d \end{bmatrix} = 0.$$

Thus, the matrix in curly brackets has at most rank n. By following the argument in

Theorem 4.4 of [32], it can be shown that

$$\|E\|_F^2 \geqq \|\hat{E}\|_F^2 = \sum_{i=n+1}^{n+d} \sigma_i^2([A;B],D,C) = \min_{E \in \mathscr{R}^{p\times q}} \{\|E\|_F^2 \mid \operatorname{rank}([A;B]-DEC) \leqq n\}$$

and that equality results by setting $E = \hat{E}$. The condition $\sigma_n > \sigma_{n+1}$ ensures that $\hat{E}$ is the unique minimizer. Hence, the unique RTLS approximation $[\hat{A};\ \hat{B}] = [A;\ B] - [\Delta\hat{A};\ \Delta\hat{B}]$ of rank n satisfying (3)–(4) is obtained by setting $[\Delta\hat{A};\ \Delta\hat{B}] = D\hat{E}C$, i.e., the RSVD of $([\hat{A};\ \hat{B}],\ D,\ C)$ is given by the RSVD of $([A;\ B],\ D,\ C)$ by setting $\sigma_{n+1}([A;\ B],\ D,\ C) = \cdots = \sigma_{n+d}([A;\ B],\ D,\ C) = 0$. □

The concept of RSVs also allows us to define the **conditions of solvability** of RTLS problems (see § 2).

- If $\sigma_{n+1}([A;\ B],\ D,\ C) < \infty$ but $R(B - \Delta\hat{B}) \not\subseteq R(A - \Delta\hat{A})$ for $\sigma_{n+1}([A;\ B],\ D,\ C) = \cdots = \sigma_{n+d}([A;\ B],\ D,\ C) = 0$, then the RTLS problem is not solvable in the generic sense (but may be solvable in the nongeneric sense).
- If $\sigma_{n+1}([A;\ B],\ D,\ C) = \infty$, then the RTLS problem is not solvable in the (non)generic sense. Finally, the following theorem allows us to derive the **consistency** of the solution of most RTLS problems described in § 2.

THEOREM 3. *Consider the* RTLS *problem, defined by* (1)–(2), *where C is of the form*

$$C = \begin{bmatrix} 0 & 0 \\ 0 & C_2 \end{bmatrix} = \begin{matrix} \\ n_1 \end{matrix}\!\!\begin{matrix} \\ n+d-n_1 \end{matrix}\!\!\begin{bmatrix} C_a & C_{ab} \\ C_{ba} & C_b \end{bmatrix}\!\!\begin{matrix} n \\ d \end{matrix}\!\!\begin{matrix} n & d \end{matrix}$$

and C_2, D square, known, and nonsingular. Denote by σ' (respectively, σ) the minimal restricted singular value of the matrix triplet $(A,\ D,\ C_a)$ (respectively, $([A;\ B],\ D,\ C)$). Let σ have multiplicity d and denote $\mathscr{D} = D^TD$ and

$$\mathscr{C} = C^TC = \begin{bmatrix} \mathscr{C}_a & \mathscr{C}_{ab} \\ \mathscr{C}_{ba} & \mathscr{C}_b \end{bmatrix}\!\!\begin{matrix} n \\ d \end{matrix}\!\!\begin{matrix} n & d \end{matrix}.$$

If $\sigma' > \sigma$, the RTLS *solution is given by*

$$\hat{X} = (A^T\mathscr{D}^{-1}A - \sigma^2\mathscr{C}_a)^{-1}(A^T\mathscr{D}^{-1}B - \sigma^2\mathscr{C}_{ab}). \tag{10}$$

Proof. Since D is nonsingular, the solution $\hat{X}$ of the RTLS problem, defined above, equals the solution of the generalized TLS problem $A^*X \approx B^*$ where $[A^*;\ B^*] = D^{-1}[A;\ B]$ and the perturbations of $[A^*;\ B^*]$ have the form

$$E^* = [0;\ EC_2]\!\!\begin{matrix} \\ n_1 \end{matrix}.$$

Using the definition of the RSVs, it is easy to verify that $\sigma = \min_i \{\sigma_i([A^*;\ B^*],\ I_m,\ C)\}$ (respectively, $\sigma' = \min_i \{\sigma_i(A^*,\ I_m,\ C_a)\}$) is the minimal generalized singular value of the matrix pair $([A^*;\ B^*],\ C)$ (respectively, $(A^*,\ C_a)$). Equation (10) then follows immediately from Theorem 4 of [29], applied to this generalized TLS problem. □

Theorem 3 proves the correspondence between the RTLS solution and well-known expressions of consistent estimates in statistics, as well as in system identification. If $D = I_m$, (10) is a well-known expression in **linear regression** analysis. Its consistency and other statistical properties have been investigated by Gallo [8] and Fuller [34]. Gleser [9] studied the special case that $C = I_{n+d}$ and $n_1 = 0$, corresponding to the

ordinary TLS problem. Using their results and assuming that $\lim_{m\to\infty} A_0^T A_0/m$ (defined below) exists and is positive definite, it can be concluded that the RTLS solution is a **strongly consistent** estimate of the true parameters X of the **general errors-in-variables model**, defined as

$$(11)\qquad B_0 = (A_0)_{m\times n} X_{n\times d} = A_1 X_1 + (A_2)_0 X_2, \quad A_2 = (A_2)_0 + \Delta A_2, \quad B = B_0 + \Delta B.$$

A_1 is known but $(A_2)_0$ and B_0 are not. The observations A_2 and B of the unknown values $(A_2)_0$ and B_0 contain measurement errors ΔA_2 and ΔB such that the rows of $[\Delta A_2; \Delta B]$ are independently and identically distributed (i.i.d.) with zero mean and known positive-definite covariance matrix $\mathscr{C}_2 = C_2^T C_2$, up to a factor of proportionality. Not only in statistics, but also in system identification, expression (10) (with $D = I_m$) is well known (see, e.g., [22], [15]). In particular, (10) arises as a consistent estimator in transfer function modelling. These models are given by

$$(12)\qquad y(t) + a_1 y(t-1) + \cdots + a_{n_a} y(t-n_a) = b_1 u(t-1) + \cdots + b_{n_b} u(t-n_b),$$

where the $\{u(t)\}$ and $\{y(t)\}$ are the input and output sequences, respectively, and $\{a_j\}$ and $\{b_j\}$ are the unknown constant parameters of the system. If sufficient observations are taken, (12) gives rise to an overdetermined Toeplitz-like set of equations. If the observed in- and outputs (respectively, observed outputs) are disturbed by mutually independent stationary zero-mean white noise sequences of equal variance, i.e., $C_2 = I$ and $n_1 = 0$ (respectively, $= n_b$), the RTLS solution of this set, given by (10), coincides with the Koopmans–Levin estimate [2] (respectively, the compensated LS estimate [22]) and is hence strongly consistent under the same conditions as described in [2] and [22]. Several authors, e.g., [13], [15], [16], extended these results in order to prove consistency for multi-input multi-output systems, modelled by (12), in which the disturbances are not necessarily white provided the covariance matrix $\mathscr{C}_2 = C_2^T C_2$ of the correlated noise in the input-output data is known, up to a factor of proportionality. See [29] for a complete list of references and a detailed description of the consistency conditions.

The condition $D = I_m$ is not a real restriction for consistency. Indeed, if D is non-singular, the same consistency conditions apply to the transformed data $[A^*; B^*] = D^{-1}[A; B]$. Also, the assumption that the first n_1 columns of A are error-free is not a restriction. In fact, we can always find an appropriate permutation of the columns of A such that the consistency conditions are satisfied.

Summarizing, these results prove that the solution $\hat{X}$ of RTLS problems, given by (1)–(4), in which $C = [0; C_2]$, C_2 and D are square and nonsingular, is consistent under mild conditions whenever $\mathscr{E}(E^T E) \sim I$, and E has zero mean. Consistency of the solution of more general RTLS problems has not yet been proven and needs to be further analyzed.

4. Algorithm. The basic problem is to find a matrix $[\Delta\hat{A}; \Delta\hat{B}] = D\hat{E}C$ with minimal $\|\hat{E}\|_F$ such that rank $([A; B] - [\Delta\hat{A}; \Delta\hat{B}]) \leqq n$ and (3) is satisfied. Heretofore, the RSVD of the matrix triplet (T, D, C), $T = [A; B]$, is computed. This is done as follows. First of all, the *regular* submatrix triplet $(T_{33}^{(5)}, D_{31}^{(5)}, C_{13}^{(5)})$, which only contains the nontrivial finite RSVs of the general matrix triplet (T, D, C), is extracted from (T, D, C) by means of orthogonal transformations. Since $D_{31}^{(5)}$ and $C_{13}^{(5)}$ are nonsingular, the implicit SVD algorithm of Ewerbring and Luk [6] or Zha [31] can now be applied (Steps 2.1–2.2) in order to compute the SVD of the triple matrix product $D_{31}^{(5)-1} T_{33}^{(5)} C_{13}^{(5)-1}$ without explicitly forming the products and without inverting $D_{31}^{(5)}$ or $C_{13}^{(5)}$. This guarantees its better numerical performance. Moreover, by first performing orthogonal transformations (Step 1), the RTLS algorithm only needs to compute the implicit three-product SVD of

a smaller submatrix, hereby improving its computational efficiency. $\hat{E}$ satisfying (3)–(4) can now be computed and also a basis of the null space of $[A - \Delta\hat{A}; B - \Delta\hat{B}]$ of dimension at least d (Steps 2.3–2.6). From the latter, the RTLS solution $\hat{X}$ can be deduced (Step 3). If the RTLS solution is not unique, the solution $\hat{X}$ with minimal $\|F_1\hat{X}F_2\|_F$ is singled out.

ALGORITHM **RTLS**.

Given

- the data matrix $T_{m\times(n+d)} = [A_{m\times n}; B_{m\times d}]$.
- the matrices $D_{m\times p}$ and $C_{q\times(n+d)}$, as defined in the RTLS formulation.
- nonsingular weighting matrices $(F_1)_{n\times n}$ and $(F_2)_{d\times d}$ such that, in case of non-uniqueness, the RTLS solution $\hat{X}$ with minimal $\|F_1\hat{X}F_2\|_F$ is singled out.

Step 1. **Reduction of $T - D\hat{E}C$ to $\hat{T} - \hat{D}\tilde{E}\hat{C}$ by orthogonal transformations.** Five orthogonal transformations, with row or column pivoting wherever possible, are performed in order to separate the *regular* submatrix triplet from (T, D, C). Denote the kth transformation by:

$$\begin{bmatrix} P_1^{(k)} & 0 \\ 0 & P_2^{(k)} \end{bmatrix}\begin{bmatrix} T^{(k-1)} & D^{(k-1)} \\ C^{(k-1)} & 0 \end{bmatrix}\begin{bmatrix} Q_1^{(k)} & 0 \\ 0 & Q_2^{(k)} \end{bmatrix} = \begin{bmatrix} T^{(k)} & D^{(k)} \\ C^{(k)} & 0 \end{bmatrix},$$

where $P_i^{(k)}$ and $Q_i^{(k)}$ $(i = 1, 2)$ are orthogonal matrices defining the kth transformation or the row or column permutation or simply the identity matrix. $T^{(k)}$, $C^{(k)}$, and $D^{(k)}$ are the transformed T, D, and C after k transformations. All the submatrices are conformally partitioned. Let $T^{(0)} = T$, $C^{(0)} = C$, and $D^{(0)} = D$.

1.1. Transform $C^{(0)}$ to $[0; C_2^{(1)}]$ such that $C_2^{(1)}$ has full column rank:

$$\begin{bmatrix} [T_1^{(1)} & T_2^{(1)}] & D \\ [0 & C_2^{(1)}] & 0 \end{bmatrix}$$

1.2. Transform $T_1^{(1)}$ to $\begin{bmatrix} T_{11}^{(2)} \\ 0 \end{bmatrix}$ such that $T_{11}^{(2)}$ has full row rank:

$$\begin{bmatrix} \begin{bmatrix} T_{11}^{(2)} & T_{12}^{(2)} \\ 0 & T_{22}^{(2)} \end{bmatrix} & \begin{bmatrix} D_1^{(2)} \\ D_2^{(2)} \end{bmatrix} \\ [\,0 \quad C_2^{(2)}\,] & 0 \end{bmatrix}$$

1.3. Transform $D_2^{(2)}$ to $\begin{bmatrix} 0 \\ D_3^{(3)} \end{bmatrix}$ such that $D_3^{(3)}$ has full row rank:

$$\begin{bmatrix} \begin{bmatrix} \begin{bmatrix} T_{11}^{(3)} & T_{12}^{(3)} \\ 0 & T_{22}^{(3)} \end{bmatrix} \\ \begin{bmatrix} 0 & T_{32}^{(3)} \end{bmatrix} \end{bmatrix} & \begin{bmatrix} D_1^{(3)} \\ 0 \\ D_3^{(3)} \end{bmatrix} \\ [\,0 \quad C_2^{(3)}\,] & 0 \end{bmatrix}$$

1.4. Transform $T_{22}^{(3)}$ to $[T_{22}^{(4)}; 0]$ such that $T_{22}^{(4)}$ has full column rank:

$$\begin{bmatrix} \begin{bmatrix} \begin{bmatrix} T_{11}^{(4)} & T_{12}^{(4)} & T_{13}^{(4)} \\ 0 & T_{22}^{(4)} & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & T_{32}^{(4)} & T_{33}^{(4)} \end{bmatrix} \end{bmatrix} & \begin{bmatrix} D_1^{(4)} \\ 0 \\ D_3^{(4)} \end{bmatrix} \\ [\,0 \quad C_2^{(4)} \quad C_3^{(4)}\,] & 0 \end{bmatrix}$$

1.5. Transform $D_3^{(4)}$ to $[D_{31}^{(5)};0]$ and $C_3^{(4)}$ to $\begin{bmatrix} C_{13}^{(5)} \\ 0 \end{bmatrix}$ such that $D_{31}^{(5)}$ and $C_{13}^{(5)}$ are nonsingular and upper triangular:

$$\begin{bmatrix} \hat{T} & \hat{D} \\ \hat{C} & 0 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} T_{11}^{(5)} & T_{12}^{(5)} & T_{13}^{(5)} \\ 0 & T_{22}^{(5)} & 0 \\ 0 & T_{32}^{(5)} & T_{33}^{(5)} \end{bmatrix} & \begin{bmatrix} D_{11}^{(5)} & D_{12}^{(5)} \\ 0 & 0 \\ D_{31}^{(5)} & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & C_{12}^{(5)} & C_{13}^{(5)} \\ 0 & C_{22}^{(5)} & 0 \end{bmatrix} & 0 \end{bmatrix} \begin{matrix} m_1 \\ m_2 \\ m_3 \\ n_3+d \\ q-n_3-d \end{matrix}$$

$$\begin{matrix} n_1 & n_2 & n_3+d & m_3 & p-m_3 \end{matrix}$$

$m_3 \leftarrow m - m_1 - m_2$; $\quad n_3 \leftarrow n - n_1 - n_2$

1.6. { *check solvability of* RTLS *problem* }
If $n - m_1 - n_2 < 0$ then stop { RTLS *problem not solvable.* $\nexists \hat{E}$ *such that* rank $(T - D\hat{E}C) \leqq n$ }

Step 2. **Combination of $\tilde{E}$ and Null($T - D\hat{E}C$).**
If $m_3 = 0$ then begin $\tilde{Z}_2 \leftarrow \mathrm{I}_{n_3+d}$; $r \leftarrow 0$; go to Step 2.5 end
If $n_3 + d = 0$ then begin $r \leftarrow 0$; go to Step 2.5 end

2.1. { *Reduction of* $D_{31}^{(5)}, T_{33}^{(5)}$, *and* $C_{13}^{(5)}$ *to upper triangular form* }

If $m_3 \geqq n_3 + d$ then begin

$$T_{33}^{(5)} = Q_T \begin{bmatrix} R_T \\ 0 \end{bmatrix} \qquad \text{(QR factorization)}$$

$$H \leftarrow D_{31}^{(5)T} Q_T$$

$$H = Q_D \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{matrix} n_3+d \\ \\ n_3+d \end{matrix} \qquad \text{(QL factorization)}$$

$$E \leftarrow L_{11}^T; \quad F \leftarrow R_T; \quad G \leftarrow C_{13}^{(5)}$$

end

else begin

$$T_{33}^{(5)} = [R_T; 0] Q_T^T \qquad \text{(RQ factorization)}$$

$$H \leftarrow C_{13}^{(5)} Q_T$$

$$H = Q_C \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} \begin{matrix} m_3 \\ \\ m_3 \end{matrix} \qquad \text{(QR factorization)}$$

$$E \leftarrow D_{31}^{(5)}; \quad F \leftarrow R_T; \quad G \leftarrow R_{11}$$

end

2.2. { *Compute the implicit* 3*-product* SVD : $D_{31}^{(5)-1} T_{33}^{(5)} C_{13}^{(5)-1} = U \Sigma V^T$ }
$\tilde{U}^T E^{-1} F G^{-1} \tilde{V} = \mathrm{diag}\,(\sigma_1, \cdots, \sigma_s)$ $\quad \sigma_{i-1} \geqq \sigma_i \quad i = 2, \cdots, s = \min\{m_3, n_3 + d\}$
$\tilde{U}, \tilde{V} \in \mathcal{R}^{s \times s}$ orthonormal

$$\text{If } m_3 \geqq n_3 + d \text{ then } U \leftarrow Q_D \begin{bmatrix} \tilde{U} & 0 \\ 0 & I_{m_2 - n_3 - d} \end{bmatrix}; \qquad V \leftarrow \tilde{V}$$

$$\text{else } U \leftarrow \tilde{U}; V \leftarrow Q_C \begin{bmatrix} \tilde{V} & 0 \\ 0 & I_{n_3+d-m_3} \end{bmatrix}; \qquad \sigma_i \leftarrow 0 \quad \text{for } i = m_3 + 1,$$

$\cdots, n_3 + d$

2.3. { *Compute* $\tilde{E}$ }
If not user determined, compute r by means of a user-defined rank determinator R_0 such that rank $([A - \Delta\hat{A}; B - \Delta\hat{B}]) = m_1 + n_2 + r \leqq n$:

$$\sigma_1 \geqq \cdots \geqq \sigma_r > R_0 \geqq \sigma_{r+1} \geqq \cdots \geqq \sigma_{n_3+d}$$

$$\left\{ \textit{then } \tilde{E}_{p\times q} = \begin{bmatrix} \tilde{E}_{11} & 0 \\ 0 & 0 \end{bmatrix} \textit{ with } \tilde{E}_{11} = \sum_{i=r+1}^{s} \sigma_i u_i v_i^T \right\}$$

2.4. { *Compute a basis* $\tilde{Z}_2$ *of* Null $(T_{33}^{(5)} - D_{31}^{(5)}\tilde{E}_{11}C_{13}^{(5)})$ }
Solve $C_{13}^{(5)}\tilde{Z}_2 = [v_{r+1}, \cdots, v_{n_3+d}]$ (back substitution)

2.5. { *Compute a basis* $\tilde{Z}$ *of* Null $(\hat{T} - \hat{D}\tilde{E}\hat{C})$ }
$H \leftarrow D_{11}^{(5)}[u_{r+1}, \cdots, u_s]\,\mathrm{diag}\,(\sigma_{r+1}, \cdots, \sigma_s) - T_{13}^{(5)}\tilde{Z}_2$

If $m_1 = n_1 > 0$ then solve $T_{11}^{(5)}\tilde{Z}_1 = H$ (back substitution)
else { *compute minimum norm solution of* $T_{11}^{(5)}\tilde{Z}_1 = H$ *and* Null $(T_{22}^{(2)})$ }
if $m_1 > 0$ then begin

$$T_{11}^{(5)} = [R_{T_{11}}; 0] \begin{matrix} m_1 \\ \begin{bmatrix} Q_{71}^T \\ Q_{72}^T \end{bmatrix} \end{matrix} \qquad \text{(RQ factorization)}$$

solve $R_{T_{11}}\tilde{Z}_{11} = H$ (back substitution)
$\tilde{Z}_1 \leftarrow Q_{71}\tilde{Z}_{11}$
end
else $Q_{72} \leftarrow I_{n_1}$

$$\tilde{Z} \leftarrow \begin{matrix} \\ n_3+d-r \end{matrix} \begin{matrix} \begin{bmatrix} \tilde{Z}_1 & Q_{72} \\ 0 & 0 \\ \tilde{Z}_2 & 0 \end{bmatrix} \\ n_1-m_1 \end{matrix} \begin{matrix} n_1 \\ n_2 \\ n_3+d \end{matrix}$$

2.6. { *Compute a basis* Z of Null $(T - D\hat{E}C)$ }

$$Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \begin{matrix} n \\ d \end{matrix} \leftarrow Q_1^{(1)}Q_1^{(2)}Q_1^{(4)}\tilde{Z} \qquad \text{(back transformation)}$$

Step 3. **RTLS solution** $\hat{X}$.

3.1. If $m_1 + n_2 + r < n$ then begin
$Z_1 \leftarrow F_1 Z_1$; $Z_2 \leftarrow F_2^{-1} Z_2$
if $(F_1 \neq I$ or $F_2 \neq I$ or $C \neq I_{n+d})$ and $d > 1$,
orthonormalize:

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = Q_z R_z \quad \text{with } Q_z^T Q_z = I_{n_3+d-r+n-m_1};$$

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \leftarrow Q_z$$

3.2. Perform Householder transformations Q such that:

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} Q = \begin{bmatrix} W & Y \\ 0 & \Gamma \end{bmatrix} \begin{matrix} n \\ d \end{matrix} \quad \text{and} \quad \Gamma_{d\times d} \text{ upper triangular}$$

If Γ nonsingular then begin { RTLS *problem generic* }
solve $\hat{X}\Gamma = -Y$ (back substitution)
if $m_1 + n_2 + r < n$ then $\hat{X} \leftarrow F_1^{-1}\hat{X}F_2^{-1}$
end
else begin { RTLS *problem nongeneric* }
if $r \leqq 0$ then stop { *nongeneric* RTLS *problem not solvable* }
$r \leftarrow r - \rho$ ρ multiplicity of σ_r
if $r < 0$ then stop { *nongeneric* RTLS *problem not solvable* }
go back to Step 2.4
end

END

The following comments are in order:

- Submatrices with zero row or column dimension may be annihilated, e.g., if $m_1 = 0$, then $T_{11}^{(5)}$ does not exist, nor do $\tilde{Z}_1$ in Step 2.5 and Z_1 in Step 3.
- Step 2.1 of the RTLS algorithm, based on the canonical correlation computation procedure of [6], reduces all three matrices involved in the three-product SVD to upper triangular form of equal dimension. In Step 2.2, the algorithm PSVD-2 of [6] can readily be applied to find the implicit three-product SVD. For more details and an analysis of the computational complexity, see [6]. An alternative method for computing the implicit three-product SVD is described in [31]. The special case where $D_{31}^{(5)} = I_{m_3}$ reduces PSVD-2 to the well-known generalized SVD algorithms [18], [20] for computing the SVD of the product FG^{-1} implicitly. These algorithms are all based on an implicit Kogbetliantz approach and are suitable for **parallel implementation.** If $C_{13}^{(5)} = I$ and $D_{31}^{(5)} = I$, the SVD in Step 2.2 is simply the ordinary SVD.
- The case $m_1 < n_1$ implies that linear dependencies exist between the columns of $[A; B]$. If only columns of A are involved in the linear dependency, then the corresponding rows in X cannot be uniquely determined. Statistically, this means that some linear relations (or functionals) of the model parameters cannot be estimated. In this case, A is rank-deficient and the RTLS problem will be nongeneric [26], [25].
- If the RTLS problem is solvable and has a unique solution, the RTLS approximation is given by ($\tilde{E}$ is defined in Step 2.3):

$$[A - \Delta\hat{A}; B - \Delta\hat{B}] = [A; B] - D\hat{E}C \quad \text{with } \hat{E} = Q_2^{(3)}Q_2^{(5)}\tilde{E}P_2^{(5)}P_2^{(1)}.$$

- The $n + d$ RSVs of (T, D, C) are equal to the $n + d$ RSVs of $(\hat{T}, \hat{D}, \hat{C})$ because of the invariance of RSVs with respect to orthogonal transformations [32]. Hence, the $n + d$ RSVs of (T, D, C) are given by the $n_2 + m_1$ infinite RSVs, the $n_3 + d$ RSVs of $(T_{33}^{(5)}, D_{31}^{(5)}, C_{13}^{(5)})$ and $n_1 - m_1$ trivial RSVs, to be considered as zeros ($n_1 > m_1$ if Null $(T) \cap$ Null $(C) \neq \{0\}$). The d smallest RSVs of (T, D, C) are thus given by $n_1 - m_1$ trivial RSVs and the $d - n_1 + m_1$ smallest RSVs of $(T_{33}^{(5)}, D_{31}^{(5)}, C_{13}^{(5)})$. The latter are (made) zero in the RTLS approximation $[A - \Delta\hat{A}; B - \Delta\hat{B}]$ (Step 2.2).
- If the RTLS problem (3)–(4) has a unique solution $\hat{E}$ or is one-dimensional ($d = 1$) or the d smallest RSVs of $([A; B], D, C)$ coincide, then $\hat{E}$ always has minimal $\|\hat{E}\|_F$ and minimal $\|\hat{E}\|_2$. Only if these conditions are not satisfied (i.e., $d > 1$ and $\sigma_n([A; B], D, C) = \sigma_{n+1}([A; B], D, C) > \sigma_{n+d}([A; B], D, C)$), the RTLS algorithm singles out a weighted minimum norm solution $\hat{X}$ which still satisfies (3) but (4) is only satisfied for the two-norm, i.e., the correction matrix $\hat{E}$, corresponding to the computed minimum norm solution $\hat{X}$, has minimal $\|\hat{E}\|_2$ but it is not guaranteed that $\|\hat{E}\|_F$ remains minimal (see also [23, p. 31]).

The following theorem proves that the RTLS algorithm indeed solves the RTLS problem.

THEOREM 4. *If the solution of the* RTLS *problem* (3)–(4) *exists and is unique, the* RTLS *algorithm solves the* RTLS *problem.*

Proof. We use the notation used in the RTLS definition and the RTLS algorithm. Formula (3) implies that there exist $\hat{X} \in \mathscr{R}^{n\times d}$ such that $(A - \Delta\hat{A})\hat{X} = B - \Delta\hat{B}$, i.e.,

$$\{[A;B]-[\Delta\hat{A};\Delta\hat{B}]\}\begin{bmatrix}\hat{X}\\ -I_d\end{bmatrix}=0 \Leftrightarrow \{T-D\hat{E}C\}\begin{bmatrix}\hat{X}\\ -I_d\end{bmatrix}=0. \tag{13}$$

Thus the matrix in curly brackets has at most rank n. If rank $(T) \leqq n$ then $\hat{E} = 0$ satisfies (3)–(4) and $[\hat{X}^T; -I_d]^T \in \text{Null}\,(T)$, which is of dimension at least d. If rank $(T) > n$, this must be accomplished by finding a matrix $\hat{E} \neq 0$, reducing the rank of $T - D\hat{E}C$ such that $\|\hat{E}\|_F$ is minimal and $T - D\hat{E}C$ has at most rank n. This $\hat{E}$ is found as follows. In Step 1, using only orthogonal transformations $T - D\hat{E}C$ is reduced to $\hat{T} - \hat{D}\tilde{E}\hat{C}$ given by

(14)

$$\hat{T}-\hat{D}\tilde{E}\hat{C}=\begin{bmatrix} T_{11}^{(5)} & T_{12}^{(5)}-[D_{11}^{(5)};D_{12}^{(5)}]\tilde{E}\begin{bmatrix}C_{12}^{(5)}\\ C_{22}^{(5)}\end{bmatrix} & T_{13}^{(5)}-[D_{11}^{(5)};D_{12}^{(5)}]\tilde{E}\begin{bmatrix}C_{13}^{(5)}\\ 0\end{bmatrix}\\ 0 & T_{22}^{(5)} & 0\\ 0 & T_{32}^{(5)}-[D_{31}^{(5)};0]\tilde{E}\begin{bmatrix}C_{12}^{(5)}\\ C_{22}^{(5)}\end{bmatrix} & T_{33}^{(5)}-D_{31}^{(5)}\tilde{E}_{11}C_{13}^{(5)}\end{bmatrix},$$

$$\begin{aligned}\text{rank}\,(T-D\hat{E}C) &= \text{rank}\,(\hat{T}-\hat{D}\tilde{E}\hat{C})\\ &= \text{rank}\,(T_{11}^{(5)})+\text{rank}\,(T_{22}^{(5)})+\text{rank}\,(T_{33}^{(5)}-D_{31}^{(5)}\tilde{E}_{11}C_{13}^{(5)})\\ &= m_1+n_2+r.\end{aligned}$$

The reduction of $T - D\hat{E}C$ of rank greater than n to rank less than or equal to n can only be accomplished by reducing the rank of $T_{33}^{(5)} - D_{31}^{(5)}\tilde{E}_{11}C_{13}^{(5)}$ to $r = n - m_1 - n_2$. Since $D_{31}^{(5)}$ and $C_{13}^{(5)}$ are nonsingular, the matrix $\tilde{E}_{11}$ with minimal $\|\tilde{E}_{11}\|_F$ is directly given from the SVD of $D_{31}^{(5)-1}T_{33}^{(5)}C_{13}^{(5)-1} = U\,\text{diag}\,(\sigma_1, \cdots, \sigma_s)V^T$ by applying the Eckart–Young theorem [12], i.e., the best lower rank r approximation $D_{31}^{(5)-1}T_{33}^{(5)}C_{13}^{(5)-1} - \tilde{E}_{11}$ that minimizes $\|\tilde{E}_{11}\|_F$, is given by $\tilde{E}_{11} = \sum_{i=r+1}^{s}\sigma_i u_i v_i^T$.

The RTLS problem has a unique solution if $m_1 + n_2 + r = n$ and $\sigma_r > \sigma_{r+1}$. Then $\tilde{E}_{11}$ is the unique minimizer. Since $\|\tilde{E}\|_F$ must be minimal, $\tilde{E}$ is given by $\begin{bmatrix}\tilde{E}_{11} & 0\\ 0 & 0\end{bmatrix}$. From this, $[\Delta\hat{A};\ \Delta\hat{B}] = D\hat{E}C = DQ_2^{(3)}Q_2^{(5)}\tilde{E}P_2^{(5)}P_2^{(1)}C$ follows immediately. Since $\|\hat{E}\|_F = \|\tilde{E}\|_F$, $\hat{E}$ also satisfies (3). Since $\hat{E}$ is known, the RTLS solution can now be computed from Null $(T - D\hat{E}C)$. Indeed from (13), $R([\hat{X}^T; -I]^T) \subseteq \text{Null}\,(T - D\hat{E}C)$. Hereto, we first compute Null $(\hat{T} - \hat{D}\tilde{E}\hat{C})$ and then use the relation $\tilde{Z} \in \text{Null}\,(\hat{T} - \hat{D}\tilde{E}\hat{C}) \Leftrightarrow Z = Q_1^{(1)}Q_1^{(2)}Q_1^{(4)}\tilde{Z} \in \text{Null}\,(T - D\hat{E}C)$. Now, using (14) and partitioning

$$\tilde{Z}=\begin{bmatrix}\tilde{Z}_1\\ \tilde{Z}_0\\ \tilde{Z}_2\end{bmatrix}\begin{matrix}n_1\\ n_2\\ n_3+d\end{matrix},$$

$(\hat{T} - \hat{D}\tilde{E}\hat{C})\tilde{Z} = 0$ equals

$$T_{11}^{(5)}\tilde{Z}_1+(T_{12}^{(5)}-D_{11}^{(5)}\tilde{E}_{11}C_{12}^{(5)})\tilde{Z}_0+(T_{13}^{(5)}-D_{11}^{(5)}\tilde{E}_{11}C_{13}^{(5)})\tilde{Z}_2=0, \tag{15}$$

$$T_{22}^{(5)}\tilde{Z}_0=0,$$

$$(T_{32}^{(5)}-D_{31}^{(5)}\tilde{E}_{11}C_{12}^{(5)})\tilde{Z}_0+(T_{33}^{(5)}-D_{31}^{(5)}\tilde{E}_{11}C_{13}^{(5)})\tilde{Z}_2=0. \tag{16}$$

Since $T_{22}^{(5)}$ is nonsingular, $\tilde{Z}_0 = 0$. Since $D_{31}^{(5)}$, $C_{13}^{(5)}$ are nonsingular and using the SVD of $D_{31}^{(5)-1} T_{33}^{(5)} C_{13}^{(5)-1}$ and $\tilde{E}_{11}$, (16) also yields

$$(D_{31}^{(5)-1} T_{33}^{(5)} C_{13}^{(5)-1} - \tilde{E}_{11}) C_{13}^{(5)} \tilde{Z}_2 = 0 \Leftrightarrow \sum_{i=1}^{r} \sigma_i u_i v_i^T C_{13}^{(5)} \tilde{Z}_2 = 0.$$

Since the $(n_3 + d - r)$-dimensional space $R([v_{r+1}, \cdots, v_{n_3+d}]) \perp R([v_1, \cdots, v_r])$, (16) is satisfied if $C_{13}^{(5)} \tilde{Z}_2 \in R([v_{r+1}, \cdots, v_{n_3+d}])$, or equivalently, $\tilde{Z}_2 \in R(C_{13}^{(5)-1}[v_{r+1}, \cdots, v_{n_3+d}])$. Hence, the $(n_3 + d - r)$ basis vectors of Null $(\hat{T} - \hat{D}\tilde{E}\hat{C})$ are given by

$$\begin{bmatrix} \tilde{Z}_1 \\ \tilde{Z}_0 \\ \tilde{Z}_2 \end{bmatrix} = \begin{bmatrix} \tilde{Z}_1 \\ 0 \\ \tilde{Z}_2 \end{bmatrix} = \begin{bmatrix} -T_{11}^{(5)\dagger}(T_{13}^{(5)} - D_{11}^{(5)} \tilde{E}_{11} C_{13}^{(5)}) \tilde{Z}_2 \\ 0 \\ \tilde{Z}_2 \end{bmatrix},$$

as computed in Step 2.5 of the RTLS algorithm. These are the only basis vectors of Null $(\hat{T} - \hat{D}\tilde{E}\hat{C})$ if $n_1 = m_1$. If $n_1 > m_1$, then $n_1 - m_1$ extra basis vectors of Null $(\hat{T} - \hat{D}\tilde{E}\hat{C})$, corresponding to trivial restricted singular values, can be computed from the RQ factorization of $T_{11}^{(5)}$, as follows:

$$T_{11}^{(5)} \tilde{Z}_{11} = 0 \Leftrightarrow [R_{T_{11}}; 0] \begin{bmatrix} Q_{71}^T \\ Q_{72}^T \end{bmatrix} \tilde{Z}_{11} = 0 \Rightarrow R(\tilde{Z}_{11}) \subseteq R(Q_{72}).$$

Hence, the $n_1 - m_1$ extra basis vectors are given by $[Q_{72}^T; 0]^T$, as computed in Step 2.5. We then have

$$\text{Null}\,(\hat{T} - \hat{D}\tilde{E}\hat{C}) = R(\tilde{Z}) \Leftrightarrow \text{Null}\,(T - D\hat{E}C) = R(Z) \quad \text{with } Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \begin{matrix} n \\ d \end{matrix} = Q_1^{(1)} Q_1^{(2)} Q_1^{(4)} \tilde{Z}.$$

Since $\hat{X}$ satisfies (13) and the problem has a unique solution, it follows that the d-dimensional solution space $R([\hat{X}^T; -I]^T) = R(Z) = \text{Null}\,(T - D\hat{E}C)$. Since the problem is solvable, Z_2 is nonsingular and then $\hat{X} = -Z_1 Z_2^{-1}$ follows immediately (as computed in Step 3.2). □

If $\sigma_{n-m_1-n_2} = \sigma_{n-m_1-n_2+1}$, the solution is not unique. In this case, the dimension of Null $(T - D\hat{E}C) > d$ and $r < n - m_1 - n_2$ with r the largest value such that $\sigma_r > \sigma_{r+1}$. This implies that any $\tilde{E}_{11} = \sum_{i=1}^{d} \tilde{\sigma}_i \tilde{u}_i \tilde{v}_i^T$ with $\tilde{u}_i \in R([u_{r+1}, \cdots, u_s])$, $\sigma_{r+1} \leqq \tilde{\sigma}_i \leqq \sigma_s$ and $\tilde{v}_i \in R([v_{r+1}, \cdots, v_{n_3+d}])$, $s = \min\{m_3, n_3 + d\}$, will have minimal $\|E_{11}\|_2 = \sigma_{n-m_1-n_2}$ (but no longer minimal $\|\tilde{E}_{11}\|_F$ if $d > 1$ and the d smallest RSVs of $([A; B], D, C)$ are not equal, i.e., $\|\tilde{E}_{11}\|_F \geqq (\sum_{i=n-m_1-n_2+1}^{s} \sigma_i^2)^{1/2}$). The weighted minimum norm RTLS solution, computed in Step 3, is obtained as follows. From any basis

$$\begin{bmatrix} G_1 \\ G_2 \end{bmatrix} \begin{matrix} n \\ d \end{matrix}$$

of a d-dimensional subspace in $R([\frac{Z_1}{Z_2}])$, a RTLS solution $\hat{X} = -G_1 G_2^{-1}$ can be computed. We have to find that $\hat{X}$ such that $\|F_1 G_1 G_2^{-1} F_2\|_F$ is minimized, i.e., find a basis

$$\begin{bmatrix} \tilde{G}_1 \\ \tilde{G}_2 \end{bmatrix}$$

of a d-dimensional subspace in

$$R\left(\begin{bmatrix} F_1 Z_1 \\ F_2^{-1} Z_2 \end{bmatrix}\right)$$

with minimal $\|\tilde{G}_1\tilde{G}_2^{-1}\|_F$. As proven in [29], this minimum is found by performing orthogonal transformations Q onto an orthonormal basis Q_z (computed in Step 3.1) of the considered range such that Q_z is reduced to $\begin{bmatrix} W & Y \\ 0 & \Gamma \end{bmatrix}$. If the problem is solvable, then Γ is nonsingular and $\tilde{G}_1\tilde{G}_2^{-1} = Y\Gamma^{-1}$ has minimal $\|\tilde{G}_1\tilde{G}_2^{-1}\|_F = \|F_1\hat{X}F_2\|_F$. Hence, $F_1\hat{X}F_2 = -Y\Gamma^{-1}$ or $\hat{X} = F_1^{-1}(-Y\Gamma^{-1})F_2^{-1}$, as computed in Step 3.2.

5. Conclusions. Consider an overdetermined set of linear equations $AX \approx B$, in which the data $[A; B]$ are assumed to be perturbed by errors of the form $E^* = DEC$, D and C are known matrices, and E is an arbitrary but bounded matrix. The problem of finding an estimate $\hat{E}$ of E with minimal $\|\hat{E}\|_F$ such that $([A; B] - D\hat{E}C)\begin{bmatrix} X \\ -I \end{bmatrix} = 0$ is solvable, is referred to as the restricted total least squares (RTLS) problem. If the solution of the RTLS problem is not unique, a weighted minimum norm solution is singled out. By choosing D and C appropriately, the RTLS problem formulation can handle any weighted least squares (LS), generalized LS, TLS and generalized TLS problem. Also, equality constraints can be imposed. Moreover, RTLS allows to declare any column or row of $[A; B]$ to be error-free and allows for correlations between the errors in the remaining data provided the corresponding error covariance matrices are known, up to a factor of proportionality. The RTLS algorithm, which solves the RTLS problem, is based on the restricted singular value decomposition (SVD), a generalization of the SVD for triple matrix products. This restricted SVD allows $[A; B]$, D, and C to be rank-deficient and avoids to invert or multiply the matrices explicitly. This guarantees the better numerical performance of the RTLS algorithm with respect to the explicit transformation procedures. Moreover, by first performing orthogonal transformations, the RTLS algorithm only needs to compute the implicit three-product SVD of a smaller submatrix, hereby improving its computational efficiency. Whenever the elements of E have zero mean and equal variance, i.e., $\mathcal{E}(E^TE) \sim I$, consistency of the solution of RTLS problems where $C = [0; C_2]$, C_2 and D square and nonsingular, can be proven under mild conditions. More general consistency results are not yet proven and need to be further analyzed.

REFERENCES

[1] J. ANGELES, K. ANDERSON, AND C. GOSSELIN, *An orthogonal-decomposition algorithm for constrained least-square optimization*, in Proc. 13th American Society of Mechanical Engineers Design and Automation Conference, Boston, MA, 1982, pp. 215–220.

[2] M. AOKI AND P. C. YUE, *On a priori error estimates of some identification methods*, IEEE Trans. Automat. Control, 15 (1970), pp. 541–548.

[3] J. W. DEMMEL, *The smallest perturbation of a submatrix which lowers the rank and constrained total least squares problems*, SIAM J. Numer. Anal., 24 (1987), pp. 199–206.

[4] ———, *On structured singular values*, Tech. Report, Courant Institute of Mathematical Sciences, New York University, New York, 1989; IEEE Trans. Automat. Control, submitted.

[5] B. L. R. DE MOOR AND G. H. GOLUB, *The restricted singular value decomposition: Properties and applications*, Numerical Analysis Project NA-89-03, Department of Computer Science, Stanford University, Stanford, CA, 1989; SIAM J. Matrix Anal. Appl., 12 (1991), to appear.

[6] L. M. EWERBRING AND F. T. LUK, *Canonical correlations and generalized* SVD: *Applications and new algorithms*, J. Comput. Appl. Math., 27 (1989), pp. 37–52.

[7] M. FAN AND A. TITS, *Characterization and efficient computation of the structured singular value*, IEEE Trans. Automat. Control, 31 (1986), pp. 734–743.

[8] P. P. GALLO, *Consistency of regression estimates when some variables are subject to error*, Comm. Statist. B—Theory Methods, 11 (1982), pp. 973–983.

[9] L. J. GLESER, *Estimation in a multivariate "errors in variables" regression model: Large sample results*, Ann. Statist., 9 (1981), pp. 24–44.

[10] G. H. GOLUB, A. HOFFMAN, AND G. W. STEWART, *A generalization of the Eckart–Young–Mirsky matrix approximation theorem*, Linear Algebra Appl., 88/89 (1987), pp. 317–327.

[11] G. H. GOLUB AND C. F. VAN LOAN, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.

[12] ———, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[13] A. GROSJEAN AND C. FOULARD, *Extensions of the Levin's method (or eigenvector method) for the identification of discrete, linear, multivariable, stochastic, time invariant, dynamic systems*, in Proc. 4th International Federation of Automatic Control Symposium on Identification and System Parameter Estimation, Tbilisi, USSR, 1976, pp. 2003–2010.

[14] S. J. HAMMARLING, *The numerical solution of the Kalman filtering problem*, NAG Tech. Report TR1/85, Numerical Algorithms Group Central Office, Wilkinson House, Oxford OX2 8DR, U.K., 1985.

[15] P. N. JAMES, P. SOUTER, AND D. C. DIXON, *Suboptimal estimation of the parameters of discrete systems in the presence of correlated noise*, Electronics Lett., 8 (1972), pp. 411–412.

[16] Ü. KOTTA, *Structure and parameter estimation of multivariable systems using the eigenvector method*, in Proc. 5th International Federation of Automatic Control Symposium on Identification and System Parameter Estimation, Darmstadt, FRG, 1979, pp. 453–458.

[17] W. E. LARIMORE AND F. T. LUK, *System identification and control using* SVDs *of systolic arrays*, in Proc. High Speed Computing, Society of Photo-Optical Instrumentation Engineers, San Diego, CA, Vol. 880, 1988, pp. 37–48.

[18] F. T. LUK, *A parallel method for computing the generalized singular value decomposition*, J. Parallel Distrib. Comput., 2 (1985), pp. 250–260.

[19] C. C. PAIGE, *The general linear model and the generalized singular value decomposition*, Linear Algebra Appl., 70 (1985), pp. 269–284.

[20] ———, *Computing the generalized singular value decomposition*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1126–1146.

[21] J. RAGOT AND M. AUBRUN, *Application de la régression orthogonale sous contrainte linéaire à un problème d'équilibrage de bilan-matière*, Rev. Statist. Appl., 30 (1982), pp. 45–56.

[22] P. STOICA AND T. SÖDERSTRÖM, *Bias correction in least squares identification*, Internat. J. Control, 35 (1982), pp. 449–457.

[23] S. VAN HUFFEL, *Analysis of the total least squares problem and its use in parameter estimation*, Ph.D. thesis, Department of Electrical Engineering, Katholieke Universiteit, Leuven, Belgium, June 1987.

[24] ———, *The generalized total least squares problem: formulation, algorithm and properties*, in Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, G. Golub and P. Van Dooren, eds., NATO ASI Series F, Springer-Verlag, Berlin, 1990, pp. 651–660.

[25] S. VAN HUFFEL AND J. VANDEWALLE, *Algebraic relationships between classical regression and total least-squares estimation*, Linear Algebra Appl., 93 (1987), pp. 149–162.

[26] ———, *Analysis and solution of the nongeneric total least squares problem*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 360–372.

[27] ———, *Reliable and efficient techniques based on total least squares for computing consistent estimators in models with errors in the variables*, in Mathematics in Signal Processing, II, J. G. McWhirter, ed., Clarendon Press, Oxford, U.K., 1990, pp. 593–603.

[28] ———, *Comparison of total least squares and instrumental variable methods for parameter estimation of transfer function models*, Internat. J. Control, 50 (1989), pp. 1039–1056.

[29] ———, *Analysis and properties of the generalized total least squares problem $AX \approx B$ when some or all columns in A are subject to error*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 294–315.

[30] C. F. VAN LOAN, *Generalizing the singular value decomposition*, SIAM J. Numer. Anal., 13 (1976), pp. 76–83.

[31] H. ZHA, *A numerical algorithm for computing the restricted singular value decomposition of matrix triplets*, Scientific Report 89-1, Konrad-Zuse Zentrum für Informationstechnik, Berlin, FRG, 1989; Linear Algebra Appl., submitted.

[32] ———, *Restricted singular value decomposition of matrix triplets*, Scientific Report 89-2, Konrad-Zuse Zentrum für Informationstechnik, Berlin, FRG, 1989; SIAM J. Matrix Anal. Appl., 12 (1991), pp. 172–194.

[33] H. ZHA AND P. C. HANSEN, *Regularization and the general Gauss–Markov linear model*, Tech. Report 89-18, Department of Mathematics, University of California, Los Angeles, CA, 1989; Math. Comp., to appear.

[34] W. A. FULLER, *Error Measurement Models*, John Wiley, New York, 1987.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 2, pp. 310–320, April 1991

008

ON THE INVERSE M-MATRIX PROBLEM FOR REAL SYMMETRIC POSITIVE-DEFINITE TOEPLITZ MATRICES*

I. KOLTRACHT† AND M. NEUMANN†

Abstract. Necessary and sufficient conditions are obtained for a real symmetric positive-definite Toeplitz matrix R to be an inverse of an M-matrix in terms of its Schur coefficients. Related problems are also considered, such as when such a matrix R can be extended to a higher-dimensional real symmetric positive-definite Toeplitz matrix whose inverse is an M-matrix or, under less restrictive conditions on R, when only its Cholesky factors are inverses of M-matrices. The proofs are constructive and allow the generation of such R's with the various aforementioned properties.

Key words. Toeplitz matrices, M-matrices, inverse M-matrix problem, reflection coefficients

AMS(MOS) subject classifications. 15A57, 15A48

1. Introduction. Real symmetric positive-definite (RSPD) Toeplitz matrices, M-matrices, and inverse M-matrices appear in a large variety of applications (see, for example, Bunch [B], Willoughby [W], Berman and Plemmons [BP], Varga [V], and Young [Y]). In this paper we consider the question of when an RSPD Toeplitz matrix is an inverse M-matrix. In the study of RSPD Toeplitz matrices, certain parameters, often called the *reflection* or *Schur coefficients*, appear in a natural way. These coefficients turn out not only to have a physical meaning which we shall explain later, but to also be useful in characterizing sign properties of the entries of inverses of RSPD Toeplitz matrices.

An RSPD Toeplitz matrix has the form $R = \{r_{|i-j|}\}_{i,j=0}^{N}$, where, without any loss of generality, it is assumed that $r_0 = 1$. The corresponding Schur coefficients $c_1, c_2, \cdots, c_N$ can be obtained, for example, via the Levinson recursion procedure as follows (see [L]).

LEVINSON ALGORITHM.

1. Set

$$g_0(0) := 1. \tag{1}$$

2. For $k = 1, 2, \cdots, N$ compute

$$c_k = -\sum_{j=0}^{k-1} r_{j+1} g_{k-1}(j) \tag{2}$$

and compute

$$g_k = \frac{1}{1-c_k^2}\left\{\begin{pmatrix} 0 \\ g_{k-1}(0) \\ \vdots \\ g_{k-1}(k-1) \end{pmatrix} + c_k \begin{pmatrix} g_{k-1}(k-1) \\ \vdots \\ g_{k-1}(0) \\ 0 \end{pmatrix}\right\}. \tag{3}$$

The vectors g_k, $k = 0, \cdots, N$ are the last columns of the inverses of $R_k =$

* Received by the editors March 30, 1989; accepted for publication (in revised form) December 14, 1989.
† Department of Mathematics, University of Connecticut, Storrs, Connecticut 06268 (KOLTRACH@UCONNVM.BITNET). The work of the first author was supported by National Science Foundation grant DMS-8801961. The work of the second author was supported by Air Force Office of Scientific Research grant 88-0047.

$\{r_{|i-j|}\}_{i,j=0}^{k}$ and yield the UDU^T factorization of R_N^{-1} with

$$U=\begin{pmatrix} 1 & g_1(0)/g_1(1) & \cdots & g_N(0)/g_N(N) \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & g_N(N-1)/g_N(N) \\ 0 & 0 & \cdots & 1 \end{pmatrix} \tag{4}$$

and with

$$D=\operatorname{diag}(g_0(0), g_1(1), \cdots, g_N(N)). \tag{5}$$

As we are assuming that R is positive definite, clearly all the diagonal entries of D must be positive.

Recall now that an $n \times n$ real matrix A is a nonsingular M-matrix if A has the representation

$$A = sI - B,$$

where B is an $n \times n$ matrix whose entries are nonnegative numbers and $s > \rho(B)$ with $\rho(\cdot)$ denoting the spectral radius of a matrix. Of the many characterizations that can be found in the literature (e.g., see [BP], [FP], [V]) for a real $n \times n$ matrix A with nonpositive off-diagonal entries to be a nonsingular M-matrix, two characterizations are important for our work here. The first is that A be *monotone*, that is, it must possess an inverse whose entries are all nonnegative. The second characterization is that A possess an LU-factorization with both triangular factors being nonsingular M-matrices themselves.

The inverse M-matrix problem can be stated as follows: *Characterize those $n \times n$ nonsingular nonnegative matrices whose inverses are* M-*matrices*. For several subclasses of the nonsingular nonnegative matrices the inverse M-matrix problem has been partially or completely solved. Willoughby [W] has given tight sufficient conditions on $0 < y < x < 1$ such that if all the off-diagonal elements of a unit diagonal positive matrix lie in (y, x), then the matrix is an inverse M-matrix. Markham [Ma] has characterized the totally nonnegative matrices whose inverses are M-matrices. Lewin and Neumann [LN] have determined a graph-theoretic characterization for those (0, 1)-matrices which are inverses of M-matrices. Finally, we mention that in Johnson [J] an overview and a survey of the inverse M-matrix problem is given.

Returning to our present problem, namely, the characterization of the RSPD Toeplitz matrices which are inverses of M-matrices, we see from (3) and (5) that a necessary condition for R to be an inverse of an M-matrix is that U is an M-matrix or, equivalently, that $g_k(i) \leqq 0$ for all $k = 1, \cdots, N$ and $i = 0, 1, \cdots, k-1$.

In § 2 (cf. Theorem 1) we give necessary and sufficient conditions for U to be an M-matrix in terms of the Schur coefficients. The proof is constructive and allows efficient recursive generation of all RSPD Toeplitz matrices R whose Cholesky factor is an inverse of an M-matrix.

For $N \leqq 3$ (i.e., for matrices of size at most 4×4) the conditions of Theorem 1 also turn out to be necessary and sufficient for R itself to be an inverse of an M-matrix. For $N > 3$ these conditions are no longer sufficient as a 5×5 example at the beginning of § 3 illustrates. We go on in this section (cf. Theorem 2) and give necessary and sufficient conditions, again in terms of Schur coefficients, for an RSPD Toeplitz matrix R to be an inverse of an M-matrix. The manner of our proof is recursive and thus is typical of many other algorithms for Toeplitz matrices. It shows when a $k \times k$ RSPD Toeplitz matrix whose inverse is an M-matrix can be embedded as a leading $k \times k$

principal submatrix of a $(k+1)\times(k+1)$ RSPD Toeplitz matrix whose inverse is an M-matrix. This then gives us tools of extending, when possible, RSPD matrices which are inverses of M-matrices to matrices with same properties, but of higher dimensions.

In § 4 we consider a method (cf. Theorem 4) of transforming a given RSPD Toeplitz matrix whose Cholesky factors are inverses of M-matrices into an RSPD Toeplitz matrix whose inverse is an M-matrix. The method is based on an observation that if R_k is an RSPD Toeplitz matrix whose Cholesky factor is an inverse of an M-matrix and g_k is the last column of its inverse, then the vector $g_{k,d} = [g_k(0), \cdots, g_k(k-1), dg_k(k)]^T$ is the last column of the inverse of a unique RSPD Toeplitz matrix $R_k(d)$ whose Cholesky factor is an inverse of an M-matrix. If, in addition, $g_k(k-(i-j))$ is strictly less than zero for those indices $[k/2] > i > j > 0$ for which $(R_k^{-1})_{i,j}$ are positive, then there exists a minimal d, call it $d_{\min}$, such that for any $d \geqq d_{\min}$ the matrix $R_k(d)$ is RSPD Toeplitz whose inverse is an M-matrix. In a sense then it is shown how to construct an RSPD Toeplitz matrix which is an inverse to an M-matrix and which is nearest to a given nonnegative Toeplitz matrix.

The proof of Theorem 4 is, similar to the proofs of previous theorems, constructive and leads to an algorithm for the generation of various RSPD Toeplitz matrices whose inverse is an M-matrix of increasing sizes.

Finally, some consequences and possible applications of our results here are as follows: (i) Due to a well-known relation between the last columns of inverses of RSPD Toeplitz matrices and monic polynomials whose roots all lie in the interior of the unit disk (see Theorem 3 and preceding references), Algorithm 1 shows how to generate all polynomials whose coefficients (other than the leading one) are nonpositive. Such polynomials can be useful in the analysis of stationary time series (see, for example, Robinson [R] and Makhoul [M]), as they correspond to monotonic time series. (ii) In Propositions 1 and 2 and Theorem 1 we show that a necessary condition for an RSPD Toeplitz matrix to be an inverse of an M-matrix is that the reflection coefficients are all nonpositive but bounded away from -1. RSPD Toeplitz matrices play a role in the pressure wave propagation in layered media (see, for example, [R] and Koltracht and Lancaster [KL2]), and in the electromagnetic wave propagation in transmission lines (see, for example, Kuznetsov and Stratonovich [KS]), where the reflection coefficient between two layers characterizes the change of wave velocity in adjacent layers. The fact that all reflection coefficients are nonpositive corresponds to a situation where the velocity increases with depth monotonically. The boundedness away from -1 characterizes media which allow higher energy transmission into deeper layers. The important problem of the design of layered media with prescribed properties at the boundaries which maximize transmitted energy (solved for the continuous media in [KS, Chap. 4]) is an open problem for the discrete case. Since the optimal medium has to be monotone, our results indicate that the solution to this problem in the discrete case must be found among RSPD Toeplitz matrices whose Cholesky factor is an inverse M-matrix.

2. RSPD Toeplitz matrices whose Cholesky factor is an inverse of an M-matrix. In this section we characterize all RSPD Toeplitz matrices $R = \{r_{|i-j|}\}_{i,j=0}^{N}$ whose factor G in the Cholesky factorization $R = GG^T$ is an inverse of an M-matrix. For this purpose we mention that it is well known that R is RSPD if and only if the Schur coefficients $c_1, \cdots, c_N$ computed via the Levinson recursion (1)–(3) are less than one in absolute value. Using this recursion it is also possible to generate RSPD Toeplitz matrices by choosing appropriate coefficients c_k (see Koltracht and Lancaster [KL1]) as follows.

Given $r_0 = 1, r_1, \cdots, r_{k-1}$ such that $R_{k-1} = \{r_{|i-j|}\}_{i,j=0}^{k-1}$ is RSPD and g_{k-1} which satisfies $R_{k-1}g_{k-1} = [0, \cdots, 0, 1]^T$, choose any $|c_k| < 1$ and set

$$r_k = -\left[c_k + \sum_{j=0}^{k-2} r_{j+1} g_{k-1}(j)\right] \Big/ g_{k-1}(k-1). \tag{6}$$

Then $R_k = \{r_{|i-j|}\}_{i,j=0}^{k}$ is also RSPD. The vector g_k can be computed via (3) and the iteration can be repeated.

PROPOSITION 1. *If R is an* RSPD *Toeplitz matrix and for $k = 1, 2, \cdots, N$, $g_k(j) \leqq 0$, $j = 0, 1, \cdots, k-1$, then $r_k \geqq 0$ for all $k = 1, 2, \cdots, N$. If, in addition, $r_1 > 0$, then $r_k > 0$ for $k = 2, \cdots, N$.*

Proof. Since $c_1 = -r_1$ it follows that $c_1 \leqq 0$. Suppose that for $k > 1$ $r_1 \geqq 0, \cdots, r_{k-1} \geqq 0$. From (3), $g_k(0) = (c_k/(1 - c_k^2))g_k(k)$. But then, as $g_k(0) \leqq 0$ and $g_k(k) > 0$, it follows that $c_k \leqq 0$. Thus $r_k \geqq 0$.

Assume now that $r_1 > 0$ and, by induction, suppose that $r_2 > 0, \cdots, r_{k-1} > 0$. This together with $R_{k-1}g_{k-1} = e_{k-1}$ imply that $g_{k-1}(i) < 0$ for some $i < k-1$. From (6) we obtain now that $r_k > 0$. This proves the proposition. □

PROPOSITION 2. *Let R be an* RSPD *Toeplitz matrix. Then $g_k(j) \leqq 0$ for all $k = 1, 2, \cdots, N$ and $j = 0, 1, \cdots, k-1$ if and only if*

$$B_k \leqq c_k \leqq 0, \qquad k = 1, 2, \cdots, N, \tag{7}$$

where

$$B_k = \max_{j=1,\cdots,k-1} \left\{ -\frac{g_{k-1}(j-1)}{g_{k-1}(k-j-1)}, -\frac{g_{k-1}(k-j-1)}{g_{k-1}(j-1)} \right\} \tag{8}$$

over all $g_{k-1}(j-1) \neq 0$ or $g_{k-1}(k-j-1) \neq 0$.

Proof. It follows from the Levinson recursion algorithm, in particular (3), that $g_k(0) = [c_k/(1 - c_k^2)]g_k(k)$ and hence $c_k \leqq 0$. For $j = 1, \cdots, k-1$ we know that

$$\begin{pmatrix} g_k(j) \\ g_k(k-j) \end{pmatrix} = \frac{1}{1-c_k^2} \begin{pmatrix} 1 & c_k \\ c_k & 1 \end{pmatrix} \begin{pmatrix} g_{k-1}(j-1) \\ g_{k-1}(k-j-1) \end{pmatrix}.$$

Thus the proposition holds if and only if the conditions $c_k \leqq 0$ and the conditions

$$g_{k-1}(j-1) + c_k g_{k-1}(k-j-1) \leqq 0 \tag{9}$$

and

$$g_{k-1}(k-j-1) + c_k g_{k-1}(j-1) \leqq 0 \tag{10}$$

hold for all $j = 1, \cdots, k-1$. Finally, conditions (9) and (10) are equivalent to (8). The proposition is thus proved. □

We remark that in the definition of B_1 the set of indices is empty which is interpreted as $B_1 = -\infty$. Proposition 2 can be restated as follows.

THEOREM 1. *Let R be an* RSPD *Toeplitz matrix. Then the Cholesky factor of R is an inverse of an* M*-matrix if and only if for $k = 1, 2, \cdots, N$,* (7) *and* (8) *hold.*

The constructive proof of Proposition 2 shows how we can generate RSPD Toeplitz matrices whose Cholesky factor is an inverse of an M-matrix from such matrices of a smaller size.

ALGORITHM 1.

1. Start with $r_0 = 1$ and $g_0(0) = 1$. Choose any $-1 < c_1 \leqq 0$. Set $r_1 = -c_1$ and compute

$$g_1 = \frac{1}{1-c_1^2}\begin{pmatrix} c_1 \\ 1 \end{pmatrix}.$$

2. For $k = 2, \cdots, N$ compute

$$B_k = \max_{j=1,\cdots,k-1}\left\{-\frac{g_{k-1}(j-1)}{g_{k-1}(k-j-1)}, -\frac{g_{k-1}(k-j-1)}{g_{k-1}(j-1)}\right\}$$

over all $g_{k-1}(j-1) \neq 0$ or $g_{k-1}(k-j-1) \neq 0$. Choose any c_k such that $B_k \leqq c_k \leqq 0$ and $-1 < c_k$. Set

$$r_k = -\left[c_k + \sum_{j=0}^{k-2} r_{j+1} g_{k-1}(j)\right] \Big/ g_{k-1}(k-1)$$

and compute

$$g_k = \frac{1}{1-c_k^2}\left\{\begin{pmatrix} 0 \\ g_{k-1}(0) \\ \vdots \\ g_{k-1}(k-1) \end{pmatrix} + c_k \begin{pmatrix} g_{k-1}(k-1) \\ \vdots \\ g_{k-1}(0) \\ 0 \end{pmatrix}\right\}.$$

3. RSPD Toeplitz matrices whose inverses are M-matrices. A well-known result due to Fiedler and Pták [FP] tells us that a necessary and sufficient condition for a real matrix with nonpositive off-diagonal entries to be a nonsingular M-matrix is that it has an LU-factorization with both triangular factors being nonsingular M-matrices. Thus the conditions of Theorem 1 are obviously necessary for the matrix R itself to be an inverse of an M-matrix. For $N \leqq 3$, that is, for RSPD matrices whose order does not exceed four, these conditions are also sufficient. For $N < 3$ this immediately follows from the fact that any real symmetric Toeplitz matrix satisfies the equality

$$JRJ = R, \tag{11}$$

where

$$J = \begin{pmatrix} 0 \cdots 0 & 0 \\ 0 \cdots 1 & 1 \\ \vdots \quad \vdots & \vdots \\ 1 \cdots 0 & 0 \end{pmatrix}.$$

For $N = 3$ the only entry to check is $x = (R_3^{-1})_{2,1}$. Indeed

$$R_3^{-1} = \begin{pmatrix} g_3(3) & g_3(2) & g_3(1) & g_3(0) \\ g_3(2) & y & x & g_3(1) \\ g_3(1) & x & y & g_3(2) \\ g_3(0) & g_3(1) & g_3(2) & g_3(3) \end{pmatrix}.$$

Now x can be represented in terms of the $g_3(j)$'s only by the use of the Trench recursion (see Trench [T]) or its matrix equivalent, the Gohberg–Semencul formula (Gohberg

and Feldman [GF]), which for the 4×4 case is given by

(12)

$$R_3^{-1} = \frac{1}{g_3(3)} \begin{pmatrix} g_3(3) & 0 & 0 & 0 \\ g_3(2) & g_3(3) & 0 & 0 \\ g_3(1) & g_3(2) & g_3(3) & 0 \\ g_3(0) & g_3(1) & g_3(2) & g_3(3) \end{pmatrix} \begin{pmatrix} g_3(3) & g_3(2) & g_3(1) & g_3(0) \\ 0 & g_3(3) & g_3(2) & g_3(1) \\ 0 & 0 & g_3(3) & g_3(2) \\ 0 & 0 & 0 & g_3(3) \end{pmatrix}$$
$$- \begin{pmatrix} 0 & 0 & 0 & 0 \\ g_3(0) & 0 & 0 & 0 \\ g_3(1) & g_3(0) & 0 & 0 \\ g_3(2) & g_3(1) & g_3(0) & 0 \end{pmatrix} \begin{pmatrix} 0 & g_3(0) & g_3(1) & g_3(2) \\ 0 & 0 & g_3(0) & g_3(1) \\ 0 & 0 & 0 & g_3(0) \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

It now follows that

$$x = g_3(1)g_3(2) + g_3(2)g_3(3) - g_3(1)g_3(0).$$

Since $g_3(1)g_3(0) \geqq 0$ it will follow that $x \leqq 0$ if we can show that $g_3(3) \geqq |g_3(1)|$. But this follows from the equality

$$r_1 g_3(0) + 1 g_3(1) + r_1 g_3(2) + r_2 g_3(3) = 0,$$

in which the first three terms are nonpositive, the last term is nonnegative, and $0 \leqq r_2 < 1$. In fact, for any N we immediately deduce the following proposition.

PROPOSITION 3. *Under the conditions of Theorem* 1, $g_k(k) > |g_k(j)|$ *for* $j = 0, 1, \cdots, k-1$ *and* $k = 1, 2, \cdots, N$.

We remark that the Proposition 3 does not hold for general RSPD Toeplitz matrices.

For $N > 3$ the conditions of Theorem 1 do not necessarily imply that R is an inverse of an M-matrix. For $N = 4$ we can see this from the following example which we calculated using MATLAB on the SUN-3/50 Workstation and which we generated by Algorithm 1. The 4×4 Toeplitz matrix is defined by $r_0 = 1$, $r_1 = 0.21131896972656$, $r_2 = 0.12334449623931$, $r_3 = 0.34758334703019$, and $r_4 = 0.13265249942545$. Then

$$G^{-1} = \begin{pmatrix} 1.0000 & 0 & 0 & 0 & 0 \\ -0.2162 & 1.0231 & 0 & 0 & 0 \\ -0.0846 & -0.1991 & 1.0266 & 0 & 0 \\ -0.3501 & -0.0215 & -0.1815 & 1.0846 & 0 \\ -0.0049 & -0.3493 & -0.0214 & -0.1799 & 1.0847 \end{pmatrix}$$

while

$$R^{-1} = \begin{pmatrix} 1.1765 & -0.1951 & -0.0232 & -0.3788 & -0.0053 \\ -0.1951 & 1.2088 & -0.1930 & \boxed{0.0396} & -0.3788 \\ -0.0232 & -0.1930 & 1.0873 & -0.1930 & -0.0232 \\ -0.3788 & \boxed{0.0396} & -0.1930 & 1.2088 & -0.1951 \\ -0.0053 & -0.3788 & -0.0232 & -0.1951 & 1.1765 \end{pmatrix},$$

where the entries of matrices are rounded to four decimal places for easier visualization which does not affect the signs of the entries.

Consider now the case $N \geqq 4$. Here we easily see that if a nonboundary entry of R_{k-1}^{-1} is positive, then the entry of R_k^{-1} with the same indices is also positive. Indeed,

recall that if R_{k-1} and R_k are real symmetric and invertible, then

$$R_k^{-1} = \begin{pmatrix} R_{k-1}^{-1} + \dfrac{1}{g_k(k)}\gamma_k\gamma_k^T & \gamma_k \\ \gamma_k^T & g_k(k) \end{pmatrix}, \tag{13}$$

where $\gamma_k^T = [g_k(0), \cdots, g_k(k-1)]$. Thus if the i,jth entry of R_{k-1}^{-1} becomes positive, then $(R_k^{-1})_{i,j} = (R_{k-1}^{-1})_{i,j} + (1/g_k(k))g_k(i)g_k(j)$ is also positive.

Using (13) we can, however, formulate necessary and sufficient conditions for R_k^{-1} to be an inverse of an M-matrix in terms of entries of $R_{k-1}^{-1}, \cdots, R_0^{-1}$.

THEOREM 2. *An* RSPD *Toeplitz matrix R is an inverse of an* M*-matrix if and only if for $k = 1, \cdots, N$,*

$$c_k \in \Delta_k := \left(\bigcap_{[k/2] > i > j > 0} \Delta_{i,j}^k \right) \cap \Delta_0^k, \tag{14}$$

where $\Delta_0^k = [B_k, 0]$ with B_k defined as in Proposition 2 *and where*

$$\Delta_{i,j}^k := [\rho_{i,j}^{(1)}, \rho_{i,j}^{(2)}], \tag{15}$$

with $\rho_{i,j}^{(1)} \leqq \rho_{i,j}^{(2)}$ being the roots of the quadratic

$$\begin{aligned} &[g_{k-1}(k-i-1)g_{k-1}(k-j-1) - g_{k-1}(k-1)(R_{k-1}^{-1})_{i,j}]x^2 \\ &\quad + [g_{k-1}(i-1)g_{k-1}(k-j-1) + g_{k-1}(j-1)g_{k-1}(k-i-1)]x \\ &\quad + g_{k-1}(k-1)(R_{k-1}^{-1})_{i,j} + g_{k-1}(i-1)g_{k-1}(j-1) = 0, \end{aligned} \tag{16}$$

provided they are real and where $\Delta_{i,j}^k := \phi$ if the roots of the quadratic are complex.

Proof. Observe that the condition $c_k \in \Delta_0^k$ is necessary and sufficient for $g_k(j) \leqq 0$, $j = 0, 1, \cdots, k-1$. It now follows from (13) that R_k^{-1} is an M-matrix if and only if for all $i, j = 1, \cdots, k-1$,

$$(R_{k-1}^{-1})_{i,j} + \frac{g_k(i)g_k(j)}{g_k(k)} \leqq 0.$$

Using the Levinson recursion procedure (in particular, see (3)), this is equivalent to

$$\begin{aligned} &(1-c_k^2)g_{k-1}(k-1)(R_{k-1}^{-1})_{i,j} \\ &\quad + [g_{k-1}(i-1) + c_k g_{k-1}(k-i-1)][g_{k-1}(j-1) + c_k g_{k-1}(k-j-1)] \leqq 0, \end{aligned}$$

which is equivalent to having the left-hand side of (16) less than or equal to zero when $x = c_k$. Since the leading term of (16) is nonnegative this will happen if and only if (16) has real roots $\rho_{i,j}^{(1)} \leqq \rho_{i,j}^{(2)}$ and $\rho_{i,j}^{(1)} \leqq c_k \leqq \rho_{i,j}^{(2)}$.

Since $R_k^{-1} = R_k^{-T}$ and $JR_k^{-1} = R_k^{-1}J$ we must test only those indices i, j for which $[k/2] > i > j > 0$ and the proof is concluded. □

We remark that the verification of the conditions of Theorem 2 is quite complicated and certainly is not much simpler than just inverting the matrix R_k. However, Theorem 2 shows how to obtain all existing extensions of a given RSPD Toeplitz matrix R_{k-1}, which is an inverse M-matrix, to R_k with the same properties as by choosing arbitrary $c_k \in \Delta_k$. We further comment that if for some index i, $B_k = -g_{k-1}(i-1)/g_{k-1}(k-i-1)$, then the choice of $c_k = B_k$ implies that $(R_k^{-1})_{i,j} = (R_{k-1}^{-1})_{i,j}$, which is the best possible in the sense that any other choice of c_k will make $(R_k^{-1})_{i,j}$ closer to zero (which can be seen from (9) and (13)). On the other hand, for the boundary entries of

R_k^{-1} we have that

$$g_k(i) = \frac{1}{1-c_k^2}(g_{k-1}(i-1) + c_k g_{k-1}(k-i-1))$$

and hence the choice of

$$c_k = -\frac{g_{k-1}(i-1)}{g_{k-1}(k-i-1)}$$

implies that $g_k(i) = 0$, and hence $B_{k+1} = 0$, which leaves only the choice $c_{k+1} = 0$, and so also $c_{k+m} = 0$ for $m = 1, 2, \cdots$. Thus if for some m, Δ_{k+m} does not contain zero, then no continuation of R_k is possible.

The remarks indicate the unlikelihood for an RSPD Toeplitz matrix, which is an inverse of an M-matrix, to have high-dimensional extensions with the same properties. This observation is supported by a large number of numerical experiments which we have carried out with randomly generated RSPD Toeplitz matrices. It is also consistent with the observation in Willoughby [W] that the likelihood for a nonnegative matrix to be an inverse of an M-matrix decreases as the dimension grows. Large RSPD Toeplitz matrices whose inverse is an inverse of an M-matrix do exist, of course, as the following example illustrates. Let

$$R = \begin{pmatrix} 1 & a & \cdots & a^N \\ a & 1 & \cdots & a^{N-1} \\ \vdots & \vdots & \vdots & \vdots \\ a^N & \cdots & a & 1 \end{pmatrix}$$

be the Kac–Murdock matrix with $0 < a < 1$. Then

$$R^{-1} = \frac{1}{1-a^2}\begin{pmatrix} 1 & -a & \cdots & 0 \\ -a & 1+a^2-a & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & -a & 1 \end{pmatrix}.$$

Note that this example does not satisfy the sufficient condition in [W].

4. Construction of M-inverse RSPD Toeplitz matrices of increasing size. In this section we consider the question of how to modify a given RSPD Toeplitz matrix whose Cholesky factor is an inverse of an M-matrix into an RSPD Toeplitz matrix which is an inverse of an M-matrix and which is related in some sense to R. If $r_j > 0$, $j = 0, \cdots, N$, then it follows from a result of Johnson [J] that one way to do this is to add to R a scalar matrix dI with $d > 0$ sufficiently large. An interesting open question now is: *what is the minimal d for which it is so*? We suggest here a different sort of modification. For that purpose we first introduce the following fact, known in the signal processing literature as the Schur–Kohn criterion (see Robinson [R] and references there), which is also a consequence of a more general result of Krein ([K]) concerning Hermitian Toeplitz matrices.

THEOREM 3. *Let* R_k *be an* rspd *Toeplitz matrix and let the vector* $g_k = [g_k(0), \cdots, g_k(k-1), g_k(k)]$ *be the last column of its inverse. Then the polynomial* $g_k(k)z^k + \cdots + g_k(1)z + g_k(0)$ *has all its zeros inside the unit circle. Conversely, if* $b_k z^k + \cdots + b_1 z + b_0$ *is a polynomial with real coefficients and all its zeros inside the unit circle, then there exists a unique* rspd *Toeplitz matrix* R_k *such that the vector* $[b_0, \cdots, b_{k-1}, b_k]^T$ *is the last column of* R_k^{-1}.

We shall also require the following fact.

PROPOSITION 4. *Let the polynomial* $p(z) = z^k + a_{k-1}z^{k-1} + \cdots + a_0$ *have the nonpositive coefficients* $a_{k-1}, \cdots, a_0$ *and suppose that all its roots lie in the interior of the unit disc. Then for any* $0 \leqq \alpha \leqq 1$, *the polynomial* $p_\alpha(z) = z^k + \alpha[a_{k-1}z^{k-1} + \cdots + a_0]$ *also has all its zeros inside the unit circle.*

Proof. The zeros of $p_\alpha(z)$ coincide with eigenvalues of its companion matrix

$$A_\alpha = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 1 \\ -\alpha a_0 & -\alpha a_1 & \cdots & -\alpha a_{k-1} \end{pmatrix}.$$

Note that entrywise $0 \leqq A_\alpha \leqq A_1 = A$. Therefore, by a well-known result (see [BP], for example),

$$\rho(A_\alpha) \leqq \rho(A),$$

where $\rho(A)$ denotes the spectral radius of A. The proposition is proved. □

Note that the above proof does not generalize in a straightforward way for the case when the coefficients have different signs. (It seems plausible, however, that Proposition 4 holds under more general conditions.) Combining Theorem 3 and Proposition 4 we get the following corollary.

COROLLARY. *Let* R_k *be an* RSPD *Toeplitz whose Cholesky factor is an inverse of an* M*-matrix and let* $g_k = [g_k(0), \cdots, g_k(k)]^T$ *be the last column of its inverse. Then for any* $d \geqq 1$ *there exists a unique* RSPD *Toeplitz matrix* $R_k(d)$ *whose Cholesky factor is an inverse of an* M*-matrix such that the vector* $g_{k,d} = [g_k(0), \cdots, g_k(k-1), dg_k(k)]^T$ *is the last column of* $R_k^{-1}(d)$.

The natural question to address now is the following. Given an RSPD Toeplitz matrix R_k whose Cholesky factor is an inverse of an M-matrix, is there a $d \geqq 1$ such that $R_k(d)$ is an RSPD Toeplitz matrix which is an inverse of an M-matrix? If yes, then characterize all such d's.

THEOREM 4. *Let* R_k *be an* RSPD *Toeplitz whose Cholesky factor is an inverse of an* M*-matrix. Let*

$$\Omega = \left\{ (i,j) \middle| \left[\frac{k}{2}\right] > i > j > 0 \text{ and } (R_k^{-1})_{i,j} > 0 \right\}.$$

Then there exists a $d \geqq 1$ *such that* $R_k(d)$ *is an* RSPD *Toeplitz matrix which is an inverse to an* M*-matrix if and only if* $g_k(k-(i-j)) < 0$ *for all* $(i,j) \in \Omega$. *In this case* $R_k(d)$ *is an* RSPD *Toeplitz whose inverse is an* M*-matrix if and only if*

$$d \geqq \max_{\Omega} \left\{ 1 - \frac{(R_k^{-1})_{i,j}}{g_k(k-(i-j))g_k(k)} \right\}. \tag{17}$$

Proof. The proof is based on the Gohberg–Semencul formula for R_k^{-1} [GF], which is similar to (12) (with k instead of 3) and in which entries of R_k^{-1} are expressed in terms of entries of g_k. In the symmetric case we have for $i > j$,

$$\begin{aligned} (R_k^{-1})_{i,j} = {} & [g_k(k-i)g_k(k-j) + \cdots + g_k(k-(i-j)-1)g_k(k-1)] \\ & + g_k(k-(i-j))g_k(k) - [g_k(i)g_k(j) + \cdots + g_k(i-j)g_k(0)]. \end{aligned} \tag{18}$$

If $(R_k^{-1})_{i,j} > 0$ and $g_k(k-(i-j)) = 0$ for some $i > j$ then, clearly, $(R_k^{-1}(d))_{i,j} = (R_k^{-1})_{i,j} > 0$ for any d. Suppose now that $g_k(k-(i-j)) < 0$ whenever $(R_k^{-1})_{i,j} > 0$. It

follows from (18) that

$$(R_k^{-1}(d))_{i,j}=(d-1)g_k(k-(i-j))g_k(k)+(R_k^{-1})_{i,j}.$$

Thus $R_k^{-1}(d)$ has nonpositive off-diagonal entries if and only if

$$(d-1)g_k(k-(i-j))g_k(k)<-(R_k^{-1})_{i,j}$$

for all $i>j$ such that $(R_k^{-1})_{i,j}>0$. Since $J_kR_k^{-1}(d)=R_k^{-1}(d)J_k$ it follows that only the indices $(i,j)\in\Omega$ have to be tested. □

The result of Theorem 4 in conjunction with Algorithm 1 can be used to generate various M-inverse rspd Toeplitz matrices of increasing size. Indeed, the condition $g_k(j)>0, j=0, 1, \cdots, k-1$ can be easily incorporated in Algorithm 1. We obtain the following procedure.

ALGORITHM 2. Let R_{k-1} be an RSPD Toeplitz matrix which is an inverse of an M-matrix.

1. Compute g_{k-1} using the Levinson algorithm.
2. Compute B_k via (8) and choose arbitrary c_k such that $B_k<c_k<0$ and $-1<c_k$.
3. Compute g_k using Levinson recursion procedure (see (3)). Thus $g_k(k)>0$ and $g_k(j)<0$ for $j=0, 1, \cdots, k-1$.
4. For $[k/2]>i>j>0$ compute $R_k^{-1}(i,j)$ using the Gohberg–Semencul formula (18) and find
$$d_{\min}=\max_{\Omega}\left\{1-\frac{(R_k^{-1})_{i,j}}{g_k(k-(i-j))g_k(k)}\right\}.$$
5. Choose any $d\geqq d_{\min}$ and define $g_{k,d}=[g_k(0), \cdots, g_k(k-1), dg_k(k)]$. Then $R_k(d)$ is an RSPD Toeplitz M-inverse matrix and the vector $g_{k,d}$ is the last column of its inverse.

We remark that Algorithm 2 is efficient and its computational complexity is of order $O(k^2)$. Indeed, the Levinson algorithm for computing g_{k-1} and g_k is of order $O(k^2)$ while the computation of entries of R_k^{-1} can be done recursively by the use of (18), namely,

$$R_k^{-1}(i,j)=R_k^{-1}(i-1,j-1)+g_k(k-i)g_k(k-j)-g_k(i)g_k(j),$$

which also can be done in $O(k^2)$ operations. Finally, the inversion of $R_k^{-1}(d)$, if necessary, is similarly of order $O(k^2)$ since the inverse of a Toeplitz matrix belongs to the class of *close to Toeplitz matrices* for which $O(k^2)$ algorithms are well known (see Gohberg, Kailath, and Koltracht [GKK] and references there).

Acknowledgment. We thank Dr. A. Bruckstein for useful discussions.

REFERENCES

[BP] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in Mathematical Sciences*, Academic Press, New York, 1979.

[B] J. BUNCH, *Stability of methods for solving Toeplitz systems of equations*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 349–364.

[FP] M. FIEDLER AND V. PTÁK, *On matrices with nonpositive off-diagonal elements and positive principal minors*, Czech. Math. J., 12 (1962), pp. 382–400.

[GF] I. GOHBERG AND I. M. FELDMAN, *Convolution Equations and Projection Methods for Their Solution*, in Transl. Math. Monographs, Vol. 41, American Mathematical Society, Providence, RI, 1974.

[GKK] I. GOHBERG, T. KAILATH, AND I. KOLTRACHT, *Efficient solution of linear systems of equations with recursive structure*, Linear Algebra Appl., 80 (1986), pp. 81–113.

[J] C. R. JOHNSON, *Inverse* M-*matrices*, Linear Algebra Appl., 47 (1982), pp. 195–216.

[KL1] I. KOLTRACHT AND P. LANCASTER, *Condition Numbers of Toeplitz and Block Toeplitz Matrices*, in Operator Theory: Advances and Applications, Vol. 18, Birkhäuser-Verlag, Basel, Switzerland, 1986, pp. 271–300.

[KL2] ———, *Threshold algorithms for the prediction of reflection coefficients in a layered medium*, Geophys., 53 (1988), pp. 908–920.

[K] M. G. KREIN, *Distribution of roots of polynomials orthogonal on the unit circle with respect to a sign alternating weight*, Theoret. Funkcii Functinal Anal. i Prilozen., 2 (1966), pp. 131–137. (In Russian.)

[KS] P. I. KUZNETSOV AND R. L. STRATONOVICH, *The Propagation of Electromagnetic Waves in Multiconductor Transmission Lines*, Pergamon Press, Macmillan, New York, 1964.

[LN] M. LEWIN AND M. NEUMANN, *On the inverse* M-*matrix problem for* (0, 1)-*matrices*, Linear Algebra Appl., 30 (1980), pp. 41–50.

[L] N. LEVINSON, *The Wiener* RMS *error criterion in filter design and prediction*, J. Math. Phys., 25 (1947), pp. 261–278.

[M] J. MAKHOUL, *Linear prediction: A tutorial review*, Proc. IEE-E, 63 (1975), pp. 571–580.

[Ma] T. L. MARKHAM, *Nonnegative matrices whose inverses are* M-*matrices*, Proc. Amer. Math. Soc., 36 (1971), pp. 326–330.

[R] E. A. ROBINSON, *Seismic Inversion and Deconvolution*, Seismic Exploration Vol. 4a, Geophysical Press, London, Amsterdam, 1984.

[T] W. K. TRENCH, *An algorithm for the inversion of finite Toeplitz matrices*, J. Soc. Indust. Appl. Math., 12 (1964), pp. 512–522.

[V] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1961.

[W] R. A. WILLOUGHBY, *The inverse* M-*matrix problem*, Linear Algebra Appl., 18 (1977), pp. 75–94.

[Y] D. M. YOUNG, *Iterative Solution to Large Linear Systems*, Academic Press, New York, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 2, pp. 321–326, April 1991

009

TOEPLITZ MATRICES AND COMMUTING TRIDIAGONAL MATRICES*

RONALD PERLINE†

Abstract. A new proof is presented of the existence of commuting tridiagonal matrices for a particular family of Toeplitz matrices.

Key word. Toeplitz matrices

AMS(MOS) subject classification. 33A70

1. Introduction. Let M denote an $N \times N$ real symmetric Toeplitz matrix, that is, a matrix with entries given by

$$M(i,j) = r_{|i-j|}, \qquad r_k \text{ real}, \qquad 1 \leqq i \leqq j \leqq N.$$

Such matrices occur in many applications. As a consequence, efforts have been made to take advantage of their special structure in the study of their linear algebraic properties. For example, fast algorithms for solving matrix systems of the form $Mx = b$ exist [GY], [K].

In [G1] and [G2], Grünbaum addresses the eigenvalue problem $Mv = \lambda v$ for Toeplitz matrices. His approach is as follows. Suppose M has the property that M *commutes* with some symmetric tridiagonal matrix D which has simple spectrum (we will call such tridiagonal matrices nontrivial). This implies that the eigenvectors of D are shared by M. This reduces the problem to finding eigenvectors for symmetric tridiagonal D, a problem which has been studied extensively and for which there exists an efficient technology [Pa], [WR].

The difficulty, of course, is finding the commuting matrix D. Not all Toeplitz matrices M have such an associated D. In fact, in [G2], Grünbaum classifies all such Toeplitz matrices: up to an addition of some scalar multiple of the identity and multiplication by an overall constant, they are of the form

$$M(i,j) = r_{|i-j|} = \frac{\sin(\alpha)(i-j)}{\sin(\beta)(i-j)}, \tag{1}$$

where α, β are free parameters. The associated nontrivial commuting tridiagonal matrix $D = D(\alpha, \beta)$ is also given explicitly in [G2].

Let us remark on the form of the matrices given by (1). Upon multiplication by β, then letting $\beta \to 0$, we obtain Toeplitz matrices $M(i,j) = \sin(\alpha)(i-j)/(i-j)$. Thus M looks to be a discrete analogue of the finite convolution operator with kernel $K(x,y) = \sin(\alpha)(x-y)/(x-y)$ studied by Slepian and Landau and Pollak in [S1]–[S5]. The analysis of this operator is facilitated by the existence of a commuting second-order differential operator.

In fact, the general matrix given by (1) can be seen to be a discrete analogue of the finite convolution kernel $\sin c(x-y)/\sinh b(x-y)$ characterized in [G6] and [M] as having a commuting second-order differential operator.

* Received by the editors April 10 1989; accepted for publication (in revised form) December 15, 1989. This research was partially supported by a National Science Foundation/North Atlantic Treaty Organization Postdoctoral Fellowship.

† Department of Mathematics and Computer Science, Drexel University, Philadelphia, Pennsylvania 19104. This work was done while the author was a visitor at the Eidgenössische Technische Hochscule, Zürich, Switzerland.

This idea of finding a commuting operator to facilitate the spectral analysis of a given operator seems to be a fruitful one and has been extended to many other contexts (see, for example, [G3]–[G5], [P1], [P2], [Pe1], [Pe2], [DS]).

In the process of trying to better understand the commutativity property of the matrices given by (1), we discovered a remarkable feature from which we can derive the commutativity property described above, in a manner which we find conceptually simpler than the calculations in [G2]. Consider the *semi-infinite* matrices M given by (1). Then the following result is true.

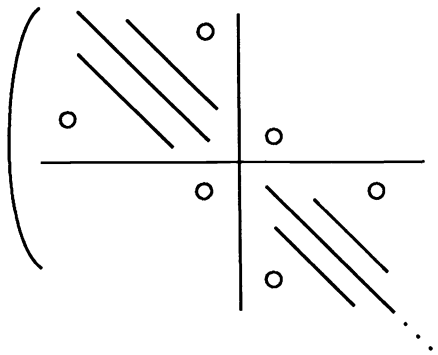
THEOREM. *There exist two linearly independent, nontrivial symmetric tridiagonal (semi-infinite) matrices A and B commuting with M.*

The details appear below. But let us immediately describe how this theorem allows us to prove the desired commutativity result for *finite* Toeplitz matrices. Let P_N denote the semi-infinite diagonal matrix with $P_N(i, i) = 1$, $1 \leqq i \leqq N$, zero otherwise. Then $M_N \equiv P_N M P_N$ is a semi-infinite matrix whose only nonzero entries lie in the upper $N \times N$ block; and this is exactly the finite form of the matrices given in (1).

Now consider matrices of the form $aA + bB$, a, b scalars. Nonzero a_N, b_N can certainly be chosen so that

$$(a_N A + b_N B)(N, N+1) = (a_N A + b_N B)(N+1, N) = 0.$$

The resulting matrix has the form shown below.



Thus $D_N \equiv a_N A + b_N B$ commutes with P_N. By linearity of commutators, it also commutes with M, hence with $M_N = P_N M P_N$. From the commutativity of M_N and D_N, it follows that their upper $N \times N$ blocks commute, giving the desired finite commutativity result. We remark that the idea of "cutting off" a tridiagonal matrix into blocks by removing corner elements has been used by us before in a similar context (see [Pe1]).

After some preliminaries on properties of Chebyshev polynomials, we give explicit formulas for the matrices A, B and a_N, b_N. In Lemmas 3 and 4 we demonstrate the commutativity of M with A and B, respectively.

2. Chebyshev polynomials. We will be considering the Chebyshev polynomials T_n and U_n of the first and second kind. A useful reference for their properties is [MOS]. We recall the following facts:

(I) $T_0(X) = 1$, $T_1(X) = X$; $U_0(X) = 1$, $U_1(X) = 2X$; and both T_n and U_n satisfy the three-term recursion relation

$$2XP_n(X) = P_{n+1}(X) + P_{n-1}(X), \qquad n \geqq 1.$$

(II) If $X = \cos\theta$, then T_j and U_j can be written as

$$T_j(X) = \cos(j\theta), \qquad U_j(X) = \frac{\sin(j+1)\theta}{\sin\theta}.$$

We will consider Toeplitz matrices of the form $M(i, j) = r_{|i-j|}$, $r_l = U_{l-1}(KX)/U_{l-1}(X)$, K some real parameter ($r_0 = 0$). Note that r_l can be rewritten as

$$r_l = \frac{U_{l-1}(K\cos\theta)}{U_{l-1}(\cos\theta)} = \frac{U_{l-1}(\cos\phi)}{U_{l-1}(\cos\theta)} = \frac{\sin\theta}{\sin\phi}\frac{\sin(l\phi)}{\sin(l\theta)},$$

where $\cos\phi = KX$.

Thus these matrices are indeed of the form given by (1).

We will need two lemmas about Chebyshev polynomials.

LEMMA 1. *Define*

$$C_l = \left[\frac{U_{l+2}(KX)}{U_{l+2}(X)} - \frac{KU_{l+1}(KX)}{U_{l+1}(X)}\right],$$

$$D_l = \left[\frac{U_l(KX)}{U_l(X)} - \frac{KU_{l+1}(KX)}{U_{l+1}(X)}\right].$$

Then

$$\frac{tD_l + C_l}{tC_l + D_l} = -\frac{tD'_l - C'_l}{tC'_l - D'_l}, \tag{2}$$

where t is an indeterminant, and

$$D'_l = U_{l+2}(X), \qquad C'_l = U_l(X).$$

Proof. The desired identity follows from the single statement

$$C_l D'_l + C'_l D_l = 0$$

and this, in turn, can be derived using the three-term recursion relation for Chebyshev polynomials.

LEMMA 2. *Define*

$$E_l = \left[\frac{KU_{l+1}(KX)}{U_{l+1}(X)} - \frac{U_{l+2}(KX)}{U_{l+2}(X)}\right],$$

$$F_l = \left[\frac{U_{l+2}(KX)}{U_{l+2}(X)} - \frac{U_l(KX)}{U_l(X)}\right],$$

$$G_l = \left[\frac{U_l(KX)}{U_l(X)} - \frac{KU_{l+1}(KX)}{U_{l+1}(X)}\right].$$

Then

$$\frac{t^2G_l + tF_l + E_l}{t^2E_l + tF_l + G_l} = \frac{tF'_l - E'_l}{tE'_l - F'_l},$$

$F' = U_{l+2}(X)$, $E'_l = U_l(X)$, and t is an indeterminant.

The proof follows from the two identities $G_lE'_l - E_lF'_l = 0$ and $E_l + F_l + G_l = 0$.

3. The commuting matrices. We can now prove Lemma 3.

LEMMA 3. *Let A be the semi-infinite symmetric tridiagonal matrix with entries*

$$A(j,j) = 2KT_{2j}(X) - 2,$$

$$A(j,j+1) = A(j+1,j) = -\frac{T_{2j+1}(X)}{X} + 1.$$

Then A commutes with M, $M(i,j) = r_{|i-j|} = U_{i-j-1}(KX)/U_{i-j-1}(X)$.

Proof. Let $Z \equiv [A,M] = AM - MA$. Then Z is obviously skew symmetric. Consider $Z(i,j)$, $j > i$. We have

$$\begin{aligned} Z(i,j) = {} & A(i,i-1)M(i-1,j) + A(i,i)M(i,j) + A(i,i+1)M(i+1,j) \\ & - M(i,j-1)A(j-1,j) - M(i,j)A(j,j) - M(i,j+1)A(j+1,j), \end{aligned}$$

which equals, after rearrangement,

$$\begin{aligned} & -\frac{U_{j-i-2}(KX)}{U_{j-i-2}(X)}\left[\frac{T_{2i+1}(X)}{X} - \frac{T_{2j-1}(X)}{X}\right] \\ & \quad + \frac{U_{j-i-1}(KX)}{U_{j-i-1}(X)}[2KT_{2i}(X) - 2KT_{2j}(X)] \\ & \quad + \frac{U_{j-i}(KX)}{U_{j-i}(X)}\left[\frac{T_{2j+1}(X)}{X} - \frac{T_{2i-1}(X)}{X}\right]. \end{aligned}$$

We want to show that this expression is equal to zero. Using the recursion relation $2XT_i(X) = T_{i+1}(X) + T_{i-1}(X)$ we obtain (after rearrangement and introducing the new index $l = j - i - 2$)

$$\begin{aligned} XZ(i,j) = {} & T_{2i+2l+5}\left[\frac{U_{l+2}(KX)}{U_{l+2}(X)} - \frac{KU_{l+1}(KX)}{U_{l+1}(X)}\right] \\ & + T_{2i+2l+3}\left[\frac{U_l(KX)}{U_l(X)} - \frac{KU_{l+1}(KX)}{U_{l+1}(X)}\right] \\ & + T_{2i+1}\left[\frac{KU_{l+1}(KX)}{U_{l+1}(X)} - \frac{U_l(KX)}{U_l(X)}\right] \\ & + T_{2i-1}\left[\frac{KU_{l+1}(KX)}{U_{l+1}(X)} - \frac{U_{l+2}(KX)}{U_{l+2}(X)}\right] \\ = {} & C_l T_{2i+2l+5} + D_l T_{2i+2l+3} - D_l T_{2i+1} - C_l T_{2i-1} = (*), \end{aligned}$$

where C_l, D_l are defined in Lemma 1.

We require $(*)$ to be zero. Observe that for fixed K, l, X, this is a linear constant coefficient difference equation, with characteristic polynomial

$$C_l S^{2i+2l+5} + D_l S^{2i+2l+3} - D_l S^{2i+1} - C_l S^{2i-1} = 0$$

or, letting $t = S^2$,

$$t^{l+2}[tC_l + D_l] - [tD_l + C_l] = 0,$$

thus

$$t^{l+2} = \frac{(tD_l + C_l)}{(tC_l + D_l)} = -\frac{(tD'_l - C'_l)}{(tC'_l - D'_l)}$$

by Lemma 1. Hence the characteristic polynomial can be rewritten:

$$t^{l+2}[tC'_l - D'_l] + [tD'_l - C'_l] = C'_l[t^{l+3} - 1] + D'_l[t - t^{l+2}] = 0.$$

Multiplying by $\sin\theta$, where $X = \cos\theta$, and using the trigonometric formula for Chebyshev polynomials finally results in

$$(t^{l+3} - 1)\sin(l+1)\theta + (t - t^{l+2})\sin(l+3)\theta = 0.$$

It is easy to check that $t = e^{2i\theta}$ is a solution. Thus $S = e^{i\theta}$; so $\tilde{T}_y = e^{ij\theta}$ is a solution to the recurrence equation $(*) = 0$. Also, since the coefficients in $(*)$ are real, $\mathrm{Re}(\tilde{T}_j) = \cos(j\theta) = T_j$ is also a solution. This proves $Z(i, j) = 0$.

LEMMA 4. *Let B be the symmetric tridiagonal matrix given by*

$$B(j,j) = K[T_{2j-2} - T_{2j}],$$

$$B(j,j+1) = B(j+1,j) = (T_{2j+1} - T_{2j-1})/2x.$$

Then B commutes with M.

The proof parallels that of Lemma 3; we just record the relevant computations, which appear:

(i) By multiplying through by $2X$, it suffices to look at $[B', M]$, B' given by

$$B'(j,j) = K[T_{2j-3} - T_{2j+1}],$$

$$B'(j,j+1) = B'(j+1,j) = T_{2j+1} - T_{2j-1};$$

(ii) The commutativity condition $Z(i,j) = [B', M](i,j) = 0$ is equivalent to the following recursion relation for Chebyshev polynomials (again, $l = j - i - 2$):

$$T_{2l+2i+5}E_l + T_{2l+2i+3}F_l + T_{2l+2i+1}G_l + T_{2i+1}G_l + T_{2i-1}F_l + T_{2i-3}E_l = 0,$$

where E_l, F_l, G_l are defined in Lemma 2;

(iii) The recursion relation has characteristic polynomial

$$S^{2l+4}[S^4E_l + S^2F_l + G_l] + [S^4G_l + S^2F_l + E_l] = 0.$$

Letting $t = S^2$, and using Lemma 2, this is seen to be equivalent to

$$t^{l+2}[tE'_l - F'_l] + [tF'_l - E'_l] = 0,$$

an equation for which $t = e^{2i\theta}$ is a solution. The desired commutativity result is thus obtained.

LEMMA 5. *Let $a_N = T_N(X)$, $b_N = U_N(X)$. Then $(a_NA + b_NB)(N, N+1) = 0$.*

Proof. We simply must show that

$$\left(-\frac{T_{2N+1}(X)}{X} + 1\right)T_N(X) + \left(\frac{T_{2N+1}(X) - T_{2N-1}(X)}{2X}\right)U_N(X) = 0.$$

But by using the trigonometric form of the Chebyshev polynomials, this reduces to a trigonometric identity that is easy to verify directly.

COROLLARY. *$D(N) = a_NA + b_NB$ commutes with $M_N = P_NMP_N$. In fact, as the reader can verify, the upper $N \times N$ block of $D(N)$ is exactly the matrix given in* [G2, p. 30].

Acknowledgments. This paper was written while the author was visiting the Forschungsinstitut für Mathematik at the ETH, Zürich, whose hospitality he is happy to acknowledge. The author benefited from the reviews of the referees, in particular, from comments concerning the proofs of Lemmas 1 and 2, for which he is pleased to thank them. Finally, thanks go to Rahel Boller for typing the manuscript.

REFERENCES

[DS] B. DICKINSON AND K. STEIGLITZ, *Eigenvectors and functions of the discrete Fourier transform*, IEEE Trans. Acoust. Speech Signal Process., 30 (1982), pp. 25–31.

[G1] F. A. GRÜNBAUM, *Eigenvectors of a Toeplitz matrix: discrete version of the prolate spheroidal wave functions*, SIAM J. Algebraic Discrete Methods, 2 (1981), pp. 136–141.

[G2] ———, *Toeplitz matrices commuting with tridiagonal matrices*, Linear Algebra Appl., 40 (1981), pp. 25–36.

[G3] F. A. GRÜNBAUM, L. LONGHI, AND M. PERLSTADT, *Differential operators commuting with finite convolution operators: some nonabelian examples*, SIAM J. Appl. Math., 42 (1982), pp. 941–952.

[G4] F. A. GRÜNBAUM, *A new property of reproducing kernels for classical orthogonal polynomials*, J. Math. Anal. Appl., 95 (1983), pp. 491–500.

[G5] ———, *The eigenvectors of the discrete Fourier transform: a version of the Hermite functions*, J. Math. Anal. Appl., 88 (1982), pp. 355–363.

[G6] ———, *Second order differential operators commuting with convolution integral operators*, LBL Tech. Report 9298, Lawrence Berkeley Laboratory, Berkeley, CA, 1979.

[GY] F. GUSTAVSON AND D. YON, *Fast computation for Toeplitz systems*, IBM RC 7551, IBM Research Center, Yorktown Heights, NY, March 1979.

[K] T. KAILATH, *Inverses of Toeplitz operators, innovations and orthogonal polynomials*, SIAM Rev., 20 (1978), pp. 106–119.

[M] J. MORRISON, *On the commutation of finite integral operators with difference kernels and linear selfadjoint differential operators*, Abstract, Notices American Mathematical Society, Providence, RI, 1962, p. 119.

[MOS] W. MAGNUS, F. OBERHETTINGER, AND R. SONI, *Formulas and Theorems for the Special Functions of Mathematical Physics*, Springer-Verlag, Berlin, New York, 1966.

[Pa] B. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[P1] M. PERLSTADT, *Chopped orthogonal polynomials expansions—some discrete cases*, SIAM J. Algebraic Discrete Methods, 4 (1983), pp. 94–100.

[P2] ———, *A property of orthogonal polynomial families with polynomial duals*, SIAM J. Math. Anal., 15 (1984), pp. 1043–1054.

[Pe1] R. PERLINE, *Discrete time-band limiting operators and commuting tridiagonal matrices*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 192–195.

[Pe2] ———, *Self-dual polynomials, Hermite matrices, and Heisenberg functionals*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 373–377.

[S1] D. SLEPIAN AND H. O. POLLAK, *Prolate spheroidal wave functions, Fourier analysis and uncertainty* I, Bell System Tech. J., 40 (1961), pp. 43–64.

[S2] H. J. LANDAU AND H. O. POLLAK, *Prolate spheroidal wave functions, Fourier analysis and uncertainty* II, Bell System Tech. J., 40 (1961), pp. 65–84.

[S3] ———, *Prolate spheroidal wave functions, Fourier analysis and uncertainty* III, Bell System Tech. J., 41 (1962), pp. 1295–1336.

[S4] D. SLEPIAN, *Prolate spheroidal wave functions, Fourier analysis and uncertainty* IV, Bell System Tech. J., 43 (1964), pp. 3009–3058.

[S5] ———, *Prolate spheroidal wave functions, Fourier analysis and uncertainty* V, Bell System Tech. J., 57 (1978), pp. 1371–1430.

[WR] J. WILKINSON AND C. REINSCH, *Handbook for Automatic Computation*, Vol. 2, Springer-Verlag, Berlin, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 2, pp. 327–353, April 1991

010

UPDATING AND DOWNDATING OF ORTHOGONAL POLYNOMIALS WITH DATA FITTING APPLICATIONS*

SYLVAN ELHAY†, GENE H. GOLUB‡, AND JAROSLAV KAUTSKY§

Abstract. New methods for updating and downdating least squares polynomial fits to discrete data are derived and assessed using polynomials orthogonal on *all* the data points being used. Rather than fixing on one basis throughout, the methods adaptively update and downdate both the least squares fit and the polynomial basis. This is achieved by performing similarity transformations on the tridiagonal Jacobi matrices representing the basis. Although downdating is potentially unstable, experimental results show that the methods give satisfactory results for low degree fits. Details of new algorithms implementing the methods are given, the most economical of which needs $14n + O(1)$ flops and $2n$ square roots to update a fit of order n.

Key words. updating, downdating, least squares, polynomial fits, discrete inner products, orthogonal polynomials, Jacobi matrices

AMS(MOS) subject classifications. primary: 65D10, 65F30; secondary: 65D30, 65D32

1. Introduction.

Updating and downdating are commonly applied to the decomposition of an $m \times n$ matrix $X = QR$, $m > n$ into orthogonal factor Q and upper triangular factor R, when new rows of X are added or old ones removed. While the original factoring requires $O(mn^2)$ operations, the updates need only $O(n^2)$ operations. In some applications the $m \times m$ factor Q need not be stored but if it is needed it can be saved in product form using $O(mn)$ locations; Q and R therefore represent a powerful data compression device. The up/downdating thus leads to savings in both computer operations and storage.

Updating is usually done by perfectly stable orthogonal rotations. Downdating a row of X can be achieved by rotations if we know the corresponding row of Q (see [7]). Downdating can also be done by hyperbolic rotations [6], the condition of which depends on the data. Hyperbolic rotations (called *hyper-rotations* here) and the stability of downdating have attracted some attention in the literature [1], [14], [2], and [11].

In the context of polynomial least squares data fitting, the techniques mentioned above apply if we use a fixed polynomial basis to represent the required fits. The matrix X is then a Vandermonde-like matrix of the values of the basis polynomials at the data points.

The choice of the polynomial basis strongly influences the stability of the computation. Thus a basis chosen to suit some given input data may be inappropriate once more data points are added or existing data points are removed. For every data set there is, however, a natural polynomial basis—the polynomials orthogonal on that data. This then leads to the problem of adaptively up/downdating the orthogonal polynomial basis together with the coefficients of the least squares fit. The solution of this problem, rather

* Received by the editors June 1, 1989; accepted for publication (in revised form) November 30, 1989.

† Computer Science Department, University of Adelaide, Adelaide, South Australia 5000 (elhay@cs.adelaide.edu.au).

‡ Computer Science Department, Stanford University, Stanford, California 94305 (na.golub@na-net.stanford.edu). The work of this author was supported in part by U.S. Army grant DAAL03-87-K-0095, the John Simon Guggenheim Memorial Foundation, and the Flinders University of South Australia Research Grant. Some preliminary work was also done while he was a guest worker at the Bell Telephone Laboratories in the summer of 1987.

§ School of Mathematical Sciences, Flinders University, Bedford Park, South Australia, 5042 (j.kautsky@cc.flinders.edu.au).

than the up/downdating of the QR factorization, is the central task of our paper, which is organised as follows.

In § 2 we define the problem in detail and in § 3 we derive some well-known properties of orthogonal polynomials expressed in matrix notation. The presentation is essentially self-contained: we find it easier—and possibly instructive—to prove the properties in question rather than to quote the facts and to translate them. The inspiration for this approach stems from the works of Wilf [15] and Golub and Welsch [8] and has been already used by the authors extensively (see, e.g., [10] for further references). Some preliminary work in this area was also done by Parker [16].

Section 4 deals with updating methods. We begin with the details of the updating and downdating problems in terms of the chosen representation.

Theorem 4.1 in § 4.2 gives a constructive characterization of the solution based on orthogonal similarity. This is a special case of the technique, widely used in numerical linear algebra, which transforms a symmetric matrix into a similar tridiagonal matrix. In this instance the transformation is achieved more economically by rotations than by elementary Householder matrices because of the structure of the problem. This idea is due to Rutishauser [13] and more recently has been treated by Gragg and Harrod [9] and Boley and Golub [3].

In § 4.3 we present methods which bear the same relation to those in § 4.2 as the LR factorization does to the QR factorization, i.e., the orthogonal similarity is replaced by a triangular similarity. We call the methods *Lanczos-type methods* because the construction of the similarity follows the same lines as the Lanczos method. Two special Lanczos-type methods are derived in §§ 4.4 and 4.5. The first, based on determinantal relations of Rutishauser [12], is the most computationally economical algorithm for updating the basis but is not well suited to updating the least squares fit. The second, based on the special structure of the similarity matrix observed in [5], leads to the most economical of all the algorithms here.

All the methods mentioned so far produce a representation of the least squares polynomials of all possible degrees, including the maximal degree polynomial fit that interpolates the given data. Limiting the calculation to produce all fits up to some prescribed lower degree introduces new aspects of the problem which are dealt with in § 4.6. For example, updating by a data point which is new to the full data set may be equivalent to updating by a point which already belongs to the (different) data set associated with the currently held representation. Effectively this means *changing* the weight of an existing data point. Theorem 4.2 in § 4.7 characterizes the existence of the solution to the updating problem for this situation.

Section 5.1 defines the downdating problem and § 5.2 introduces downdating methods which are closely related to the updating methods of § 4. In addition we give in Theorem 5.1 of § 5.3 a downdating which is based on the reversal of the orthogonal rotations method. This leads naturally in Theorem 5.2 of § 5.4 to conditions for the existence of the downdate. Essentially, downdating is possible if the weight involved is sufficiently small; for larger weights it may be necessary to reduce the degree of the polynomial fit.

Algorithmic implementations of the up- and downdates by the four methods are given in § 6 together with their computational complexity. In addition a version of one of the algorithms, scaled to avoid overflow when adding a data point outside the current data, is given in § 6.3.2.

Finally, § 7 demonstrates the methods by computing a moving least squares polynomial fit to noisy data. Both the computed least squares polynomial and its represen-

tation, as generated by the different methods, are briefly examined and compared with accurately computed results. These limited tests indicate the following:

(a) Although the accuracy deteriorates as the degree of the fit increases, the methods are satisfactory for low degree polynomials.

(b) Computational complexity and convenience, rather than other considerations, are the main determinants in choosing between the four methods.

2. The problem.

PROBLEM 1 (weighted least squares data fitting). Given the discrete data

$$Y_N = \{x_j, w_j, y_j\}_{j=1}^N,$$

find polynomial q of degree n such that

$$\sum_{j=1}^{N} w_j^2 (y_j - q(x_j))^2 \tag{2.1}$$

is minimised.

Suppose Y_{N+1} is obtained from Y_N by augmenting a new triplet of data. Solving Problem 1 for Y_{N+1} *assuming* the knowledge of its solution q for Y_N is called *updating* the least squares fit. Conversely, solving Problem 1 for Y_N assuming the knowledge of its solution for Y_{N+1} is called *downdating*. In other words, updating and downdating means modifying the solutions of Problem 1 when one data set (triple) is added to or removed from the data.

The importance of updating and downdating is threefold:

- Problem 1 can be solved by updating, starting from the trivial solution for $N = 1$ (taking into account obvious restrictions on discreteness of x_j's and relations between n and N).
- Using downdating and updating we can calculate "sliding" least squares fits—include new measurements and discard old ones or even change the weights of selected data sets.
- Using suitable representation of the solutions we do not have to store—or refer to—data already used.

The key word here is representation. We choose to represent the least squares fit q by the coefficients of its expansion in terms of the polynomials orthonormal with respect to the discrete inner product

$$[f,g]_{W_N} = \sum_{j=1}^{N} w_j^2 f(x_j) g(x_j) \tag{2.2}$$

where we have denoted $W_N = \{x_j, w_j\}_{j=1}^N$. We need therefore to consider also the following problem.

PROBLEM 2 (discrete orthonormal polynomials). Given the discrete data W_N find polynomials p_k of exact degree k, $k = 0, 1, \cdots, n$ such that

$$[p_i, p_k]_{W_N} = \delta_{ik}, \qquad i,k = 0,1,\cdots,n \tag{2.3}$$

where $i \neq k$ implies $\delta_{ik} = 0$ and $\delta_{ii} = 1$.

Here we choose to represent the orthonormal polynomials by the coefficients of the three-term relation which they satisfy. This representation has several advantages, say in comparison with standard powers expansions:

- It allows for stable evaluation of the values of the polynomials and other manipulations with them.
- It is concise—for Problem 2 we have $2n$ values representing $\frac{1}{2}n^2$ power coefficients of $p_0, p_1, \cdots, p_n$. For Problem 1 it appears generous to use $3n$ values to represent one polynomial but these $3n$ values give *all* least squares fits up to degree n.
- Using matrix notation the Problems 1 and 2 can be re-formulated, in the terms of the chosen representation, as problems in numerical linear algebra. This gives a new insight into the existing methods and leads to simple derivation of both old and new algorithms.

We note that the comments on the importance of updating and downdating made above with reference to Problem 1 apply equally to the Problem 2.

3. Matrix notation and basic relations. For any vector of functions $\mathbf{u} = (u_1, \cdots, u_n)^T$, $\mathbf{v} = (v_1, \cdots, v_m)^T$ and scalar product $[\cdot, \cdot]$, discrete as in (2.2) or a continuous one, we denote by $[\mathbf{u}, \mathbf{v}^T]$ the constant $n \times m$ matrix with elements $[u_i, v_j]$. Some obvious rules apply—for example, if A and B are constant matrices of suitable sizes then

$$[A\mathbf{u}, \mathbf{v}^T B] = A[\mathbf{u}, \mathbf{v}^T]B,$$

which we shall use freely. Furthermore, we denote by $\|\cdot\|$ the Euclidean vector 2-norm.

For any polynomials p_j of exact degree j, $j = 0, 1, \cdots, n$, there exists a unique, constant, lower Hessenberg matrix J and scalar β_n such that, for any t,

$$t\mathbf{p}_n(t) = J\mathbf{p}_n(t) + \beta_n p_n(t)\mathbf{e}_n \tag{3.1}$$

where we have denoted $\mathbf{p}_n = (p_0, p_1, \cdots, p_{n-1})^T$. Here, as elsewhere, $\mathbf{e}_j$ is the jth column of an identity matrix with appropriate dimension. The matrix J is unreduced, i.e., its superdiagonal elements $\beta_j = \mathbf{e}_{j+1}^T J \mathbf{e}_j, j = 1, \cdots, n-1$, together with β_n, do not vanish.

Given an unreduced lower Hessenberg matrix J and nonzero scalars p_0 and β_n, the polynomials $p_1, p_2, \cdots, p_n$ can be determined recursively from (3.1). We call J the *recurrence matrix* for polynomials $\mathbf{p}_n$, p_n; because of the one-to-one correspondence it can be used to represent the polynomials.

The polynomials $\mathbf{p}_n$, p_n will be orthonormal with respect to a scalar product $[\cdot, \cdot]$ (such as, for example, the scalar product $[\cdot, \cdot]_{W_N}$ in (2.2) corresponding to the data W_N) if

$$[\mathbf{p}_n, \mathbf{p}_n^T] = I_n, \qquad [\mathbf{p}_n, p_n] = \mathbf{0}, \qquad [p_n, p_n] = 1. \tag{3.2}$$

Combining (3.2) and (3.1) we now have

$$[t\mathbf{p}_n, \mathbf{p}_n^T] = [J\mathbf{p}_n, \mathbf{p}_n^T] + \beta_n \mathbf{e}_n [p_n, \mathbf{p}_n^T] = J,$$

from which it follows that, the matrix on the left being symmetric, the recurrence matrix J for orthonormal polynomials must also be symmetric and thus tridiagonal.

The *Jacobi* (symmetric, tridiagonal, unreduced) *matrix*

$$J = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & \cdots & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \cdots & 0 \\ 0 & \beta_2 & \alpha_3 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha_n \end{pmatrix}$$

has real distinct eigenvalues, say $\lambda_1, \cdots, \lambda_n$. From (3.1) it is immediate that all eigenvalues λ_j are roots of p_n and $\mathbf{p}_n(\lambda_j)$ are the corresponding eigenvectors. So denoting $P =$

$(\mathbf{p}_n(\lambda_1), \cdots, \mathbf{p}_n(\lambda_n))$ and $\Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$, we have

$$JP = P\Lambda.$$

As J is symmetric, P has orthogonal columns ($P^TP\Lambda = \Lambda P^TP$ implies P^TP is diagonal). Denoting $D := \operatorname{diag}(\nu_1, \nu_2, \cdots, \nu_n)$ where

$$\nu_j = (\mathbf{p}_n(\lambda_j)^T\mathbf{p}_n(\lambda_j))^{-1/2}, \qquad j = 1, 2, \cdots, n$$

we have

$$P^TP = D^{-2}. \tag{3.3}$$

Thus PD is orthogonal and

$$PD^2P^T = I. \tag{3.4}$$

This shows that the polynomials $\mathbf{p}_n$, orthonormal with respect to some scalar product (such as the one in (2.2)), are also orthonormal with respect to the discrete scalar product $[\cdot, \cdot]_{\tilde{W}_n}$ given by the data $\tilde{W}_n = \{\lambda_j, \nu_j\}_{j=1}^n$. Finally we note that

$$\mathbf{e}_1^TPD = (\nu_1, \cdots, \nu_n)p_0, \tag{3.5}$$

which allows for the recovery (apart from signs, which are not relevant) of the scaled weights ν_j from any set of normalised eigenvectors of J.

Turn now to the least squares fit Problem 1 with data Y_N. Define $\mathbf{y}_N := (y_1, \cdots, y_N)^T$. Let $\mathbf{p}_N$ be polynomials orthonormal with respect to the inner product (2.2) and denote by J_N, P_N, D_N the matrices J, P, and D above. For any polynomial $q = \mathbf{p}_N^T\mathbf{d}_N$ with coefficients $\mathbf{d}_N = (d_0, \cdots, d_{N-1})^T$ the quantity (2.1) to be minimised is (we note that for $\mathbf{x}_N = (x_1, \cdots, x_N)^T$ we have $q(\mathbf{x}_N) = P_N^T\mathbf{d}$)

$$\begin{aligned}[y-q, y-q]_{W_N} &= \| D_N(\mathbf{y}_N - q(\mathbf{x}_N)) \|^2 \\ &= \| D_N(\mathbf{y}_N - P_N^T\mathbf{d}) \|^2 \\ &= \| P_ND_N^2\mathbf{y}_N - \mathbf{d} \|^2,\end{aligned} \tag{3.6}$$

the last equality being a consequence of the above observation that P_ND_N is orthogonal. Now choosing $\mathbf{d}_N = P_ND_N^2\mathbf{y}_N$ gives the solution to the *interpolating* problem for the data Y_N as the expansion in the polynomials orthonormal on W_N (the weights are not important in that case). However, from (3.6) it follows that for any $n < N$, $\hat{\mathbf{d}}_n = (d_0, \cdots, d_{n-1}, 0, 0, \cdots, 0)^T$ gives the solution to the least squares fit problem by polynomials of degree less than n. The shortened vector $\mathbf{d}_n = (d_0, \cdots, d_{n-1})^T$ is now given by

$$\mathbf{d}_n = P_{n,N}D_N^2\mathbf{y}_N \tag{3.7}$$

where $P_{n,N}$ is the matrix comprising the first n rows of P_N.

The solution to the Problem 2 is now given by the first n diagonal elements $\alpha_1, \cdots, \alpha_n$ and $n-1$ off-diagonal elements $\beta_1, \cdots, \beta_{n-1}$ of the matrix J_N (we may denote the submatrix J_n) which together with the zeroth moment $\mu_0 = \sum_{j=1}^N w_j^2 = [1, 1]_{W_N}$ represent the first n orthonormal polynomials $\mathbf{p}_n$. The solution to the Problem 1 is then given by μ_0, J_n, and $\mathbf{d}_n$, for which $q = \mathbf{d}_n^T\mathbf{p}_n$.

The evaluation of $q(t)$ can conveniently be done in two ways. Denote by

$$K(t)\mathbf{p}_n(t) = p_0(t)\mathbf{e}_1 \tag{3.8}$$

the system of equations defined by the first n rows of the $n+1$ equations in

$$\begin{pmatrix} \mathbf{e}_1^T \\ (J_n - tI) \end{pmatrix} \mathbf{p}_n(t) = \begin{pmatrix} p_0(t) \\ -\beta_n p_n(t) \mathbf{e}_n \end{pmatrix}.$$

Given the vector $\mathbf{d}_n$, the approximating polynomial $q(t)$ can be evaluated at a point by a forward-substitution on (3.8) for $\mathbf{p}_n$ and then an inner product with $\mathbf{d}_n$.

A second more direct way is the Clenshaw recursion in which we have

$$q(t) = p_0 \mathbf{e}_1^T \mathbf{y}$$

where $\mathbf{y}$ is the solution obtained by a back-substitution on the system

$$K(t)^T \mathbf{y} = \mathbf{d}_n.$$

4. Updating methods.

4.1. A reformulation of the updating problem. The problem of updating is now, with respect to the representation of the solutions introduced in § 3, as follows.

Given

(a) $\mu_0 > 0$,

(b) J_n a Jacobi matrix of size n,

(c) a vector $\mathbf{d}_n$,

(d) an integer $\tilde{n} = n$ or $n+1$, and

(e) a triplet $\{x, w, y\}$

((a), (b), and (c) representing the solution of Problem 1 for some data set Y_N), find $\tilde{\mu}_0$, $\tilde{\mathbf{d}}_{\tilde{n}}$, and $\tilde{J}_{\tilde{n}}$ representing the solution for the data set $Y_N \cup \{x, w, y\}$.

We have formulated the updating problem only for Problem 1—the corresponding version for Problem 2 is obtained by omitting the coefficients $\mathbf{d}$ and the function values.

As we shall see, the data set Y_N does not enter into the calculations; it is of course important for the formulation of the problem and for the following brief discussion of the existence and meaningfulness of the solutions.

The solution of the updating problem always exists for $\tilde{n} = n$. Requiring that with an update we can increase our matrix size to $\tilde{n} = n+1$ implies that $N = n$, that is that the points x_j of Y_N are the eigenvalues λ_j of J_n. The increase is then possible if and only if x is not one of the λ_j's.

Finally, as the solution for the one-point set $Y_1 = \{x, w, y\}$ is trivially

$$\mu_0 = w^2, \qquad J_1 = (x), \qquad d_0 = y|w|,$$

we can indeed use updating (with $\tilde{n} = n+1$) to build up the solution to Problems 1 and 2 for any values of n and N without storing the data set Y_N—as long as we increase n only in the beginning of the process.

4.2. Rotation method. To describe the solution of the updating problem, we will temporarily deal with the case $n = N$ and $\tilde{n} = N+1$. We denote here $\mathbf{w}_N = (w_1, \cdots, w_N)^T$ the vector of weights and $\Lambda_N = \operatorname{diag}(x_1, \cdots, x_N)$ the diagonal matrix of the points of our data set. We also introduce

$$\sigma_N = \sqrt{\mu_0^{(N)}} = (w_1^2 + \cdots + w_N^2)^{1/2} = \|\mathbf{w}_N\|.$$

The following theorem now gives the solution of the updating problem.

THEOREM 4.1. Given σ_N, J_N, *and* $\mathbf{d}_N$ *for the data* Y_N *and also the triple* $\{x, w, y\}$ *the solution for* $Y_{N+1} = Y_N \cup \{x, w, y\}$ *is*

$$\tilde{J}_{N+1} = Q\begin{pmatrix} J_N & \mathbf{0} \\ \mathbf{0}^T & x \end{pmatrix} Q^T, \tag{4.1}$$

$$\tilde{\sigma}_{N+1} = (\sigma_N^2 + w^2)^{1/2}, \tag{4.2}$$

$$\tilde{\mathbf{d}}_{N+1} = Q\begin{pmatrix} \mathbf{d}_N \\ wy \end{pmatrix}, \tag{4.3}$$

where the orthogonal similarity matrix Q *is uniquely determined by requiring* $\tilde{J}_{N+1}$ *to be tridiagonal and* Q *to satisfy*

$$Q(\sigma_N \mathbf{e}_1 + w\mathbf{e}_{N+1}) = \tilde{\sigma}_{N+1}\mathbf{e}_1. \tag{4.4}$$

Proof. Denoting $Q_N = P_N D_N$ of § 3, we have $J_N = Q_N \Lambda_N Q_N^T$ and, from (3.5),

$$\sigma_N \mathbf{e}_1^T Q_N = \mathbf{w}_N^T.$$

Similarly, $\tilde{J}_{N+1} = \tilde{Q}_{N+1}\Lambda_{N+1}\tilde{Q}_{N+1}^T$ and $\tilde{\sigma}_{N+1}\mathbf{e}_1^T\tilde{Q}_{N+1} = \mathbf{w}_{N+1}^T$. Noting that $\Lambda_{N+1} = \operatorname{diag}(\Lambda_N, x)$ and $\mathbf{w}_{N+1}^T = (\mathbf{w}_N^T, w)$ (Λ and $\mathbf{w}$ escalate), we see that substituting for Q with

$$Q = \tilde{Q}_{N+1}\begin{pmatrix} Q_N^T & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix} \tag{4.5}$$

verifies (4.1) and (4.4). The norm-preserving property of (4.4) leads to (4.2), which is consistent with the definition of σ_N.

Turning to the Fourier coefficients we have from (3.7) that

$$\mathbf{d}_N = Q_N D_N \mathbf{y}_N \quad \text{and} \quad \tilde{\mathbf{d}}_{N+1} = \tilde{Q}_{N+1} D_{N+1}\tilde{\mathbf{y}}_{N+1},$$

where again $D_{N+1} = \operatorname{diag}(\mathbf{w}_{N+1}) = \operatorname{diag}(D_N, w)$ and $\tilde{\mathbf{y}}_{N+1}^T = (y_1, \cdots, y_N, y) = (\mathbf{y}_N^T, y)$ escalate. Thus

$$\tilde{\mathbf{d}}_{N+1} = \tilde{Q}_{N+1}\begin{pmatrix} D_N\mathbf{y}_N \\ wy \end{pmatrix} = \tilde{Q}_{N+1}\begin{pmatrix} Q_N^T\mathbf{d}_N \\ wy \end{pmatrix} = Q\begin{pmatrix} \mathbf{d}_N \\ wy \end{pmatrix}, \tag{4.6}$$

proving (4.3). It is well known that the orthogonal similarity Q transforming the given matrix in (4.1) into an upper Hessenberg matrix $\tilde{J}_{N+1}$ is fully determined by its first row $Q^T\mathbf{e}_1 = 1/\tilde{\sigma}_{N+1}(\sigma_N\mathbf{e}_1 + w\mathbf{e}_{N+1})$, as specified in (4.4). □

For algorithmic implementation Q is constructed as

$$Q = R_N R_{N-1} \cdots R_1 \tag{4.7}$$

where R_j is a rotation between the jth and $(N+1)$st rows, $j = 1, 2, \cdots, N$. The first rotation R_1 achieves (4.4) which, by definition, will not be affected by the subsequent rotations; these are determined so that in the intermediate matrices

$$K_n = R_n R_{n-1} \cdots R_2 R_1 \begin{pmatrix} J_N & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix} R_1^T R_2^T \cdots R_{n-1}^T R_n^T$$

the first $n-1$ elements of the last (i.e., $(N+1)$st) row vanish. An inspection shows that K_n is tridiagonal up to the last row and column which has nonzero elements only in nth,

$(n+1)$st, and $(N+1)$st positions (so that $J_{N+1} = K_N$ is indeed tridiagonal). Furthermore, the step $K_{n+1} = R_{n+1}K_nR_{n+1}^T$ involves, besides the last row and column, only:

- The nth off-diagonal element of K_n to determine the rotation R_{n+1} and its new value in K_{n+1},
- The $(n+1)$st diagonal and off-diagonal elements to determine their new values.

These observations lead to the following corollary.

COROLLARY 4.1. *The size n leading submatrix $\tilde{J}_n$ of $\tilde{J}_{N+1}$ is the leading submatrix of K_n and only the size n leading submatrix J_n of J_N and the triple $\{x, w, y\}$ are needed to determine rotations $R_1, \cdots, R_n$ and $\tilde{J}_n$. Similarly, only the first n elements of $\mathbf{d}_N$ and the rotations $R_1, \cdots, R_n$ are sufficient to obtain the first n elements of $\mathbf{d}_{N+1}$.*

4.3. Lanczos-type methods. The method we wish to describe now is a modification of the Lanczos method where the orthogonal matrix Q is essentially replaced by a lower triangular matrix (it has been described and called LTL—the lower triangular Lanczos method—in [10]). We shall describe it here again for two reasons. First, unlike the case in [10], we deal here with polynomials orthogonal with respect to a *finite* discrete scalar product. Second, we use it as a starting point from which to derive two new methods.

Let us assume, as before, that the polynomials $\mathbf{p}_N$, p_N, orthonormal with respect to $[\cdot, \cdot]_{w_N}$, satisfy the three-term relation

$$t\mathbf{p}_N(t) = J_N\mathbf{p}_N(t) + \beta_N\mathbf{e}_N p_N(t). \tag{4.8}$$

We note that p_N cannot be normalised by the scalar product (recall from above that $x_1, \cdots, x_N$ are eigenvalues of J_N and thus roots of p_N) because $[p_N, f]_{w_N} = 0$ for any f. The β_N can be chosen in such a way that p_N is, say, monic. This is not necessarily a good choice for numerical calculation (which is in fact independent of β_N) but is convenient for the present discussion. We are now seeking the Jacobi matrix $\tilde{J}_{N+1}$ for the polynomials $\tilde{\mathbf{p}}_{N+1}, \tilde{p}_{N+1}$ orthonormal with respect to $[\cdot, \cdot]_{w_{N+1}}$. They must satisfy

$$t\tilde{\mathbf{p}}_{N+1}(t) = \tilde{J}_{N+1}\tilde{\mathbf{p}}_{N+1}(t) + \tilde{\beta}_{N+1}\mathbf{e}_{N+1}\tilde{p}_{N+1}(t). \tag{4.9}$$

Obviously ($\tilde{\beta}_{N+1}$ being again chosen to make $\tilde{p}_{N+1}$ monic),

$$\tilde{p}_{N+1}(t) = (t - x_{N+1})p_N(t),$$

so that by denoting $\mathbf{p}_{N+1}^T = (\mathbf{p}_N^T, p_N)$ we obtain

$$t\mathbf{p}_{N+1}(t) = J_{N+1}\mathbf{p}_{N+1}(t) + \mathbf{e}_{N+1}\tilde{p}_{N+1}(t) \tag{4.10}$$

where

$$J_{N+1} = \begin{pmatrix} J_N & \beta_N\mathbf{e}_N \\ \mathbf{0}^T & x_{N+1} \end{pmatrix}.$$

There exists a constant nonsingular lower triangular matrix L such that

$$\mathbf{p}_{N+1} = L\tilde{\mathbf{p}}_{N+1}. \tag{4.11}$$

By substituting (4.11) into (4.10) and comparing it with (4.9), premultiplied by L, we obtain immediately

$$J_{N+1}L = L\tilde{J}_{N+1}, \qquad \tilde{\beta}_{N+1}\mathbf{e}_{N+1}^T L\mathbf{e}_{N+1} = 1. \tag{4.12}$$

Denoting the elements of $\tilde{J}_{N+1}$ similarly as in J_N and $\mathbf{v}_j = L\mathbf{e}_j$ the columns of L, we have from (4.12) the recurrence

$$(1-\delta_{j,N+1})\tilde{\beta}_j\mathbf{v}_{j+1} + \tilde{\alpha}_j\mathbf{v}_j + \tilde{\beta}_{j-1}\mathbf{v}_{j-1}(1-\delta_{j1}) = J_{N+1}\mathbf{v}_j, \qquad j = 1, 2, \cdots, N+1. \tag{4.13}$$

As discussed in detail in [10], $\tilde{J}_{N+1}$ can be explicitly constructed by evaluating alternatively its elements and the columns of L using the fact that $\mathbf{e}_k^T \mathbf{v}_j = 0$ for $k < j$ (clearly analogous to the Lanczos method) *and* from the knowledge of $\mathbf{v}_1$.

It remains to show how to find $\mathbf{v}_1 = L\mathbf{e}_1$ for this particular case. We have

$$M_{N+1} := [\mathbf{p}_{N+1}, \mathbf{p}_{N+1}^T]_{w_{N+1}} = LL^T \tag{4.14}$$

so that, with $\rho_1 = \mathbf{e}_1^T L \mathbf{e}_1 = \mathbf{e}_1^T \mathbf{v}_1$, we have

$$\rho_1 \mathbf{v}_1 = LL^T \mathbf{e}_1 = [\mathbf{p}_{N+1}, p_0]_{w_{N+1}} \tag{4.15}$$

$$= \mathbf{e}_1 + p_0 w_{N+1}^2 \mathbf{p}_{N+1}(x_{N+1}), \tag{4.16}$$

which determines $\mathbf{v}_1$ explicitly.

We remark that this LTL method, described and devised here in matrix notation, is equivalent to the modified Chebyshev's algorithm discussed in [4, p. 295] and suffers the instabilities mentioned there, particularly when the knots x_j are uniformly spaced. As the rotations method provides stable answers, it is not likely, however, that the underlying map itself is ill conditioned (as suggested in [4]) but only the LTL method is then unsuitable. We note that the elements σ_{kl} of (2.12) in [4] are indeed those of L^T as

$$L^T = [\tilde{\mathbf{p}}_{N+1}, \tilde{\mathbf{p}}_{N+1}^T]_{w_{N+1}} L^T = [\tilde{\mathbf{p}}_{N+1}, \mathbf{p}_{N+1}^T]_{w_{N+1}}.$$

We will now show that we can proceed with the recurrence constructing $\tilde{J}_{N+1}$ using the diagonal and subdiagonal elements of L only.

Denoting

$$\mathbf{v}_j = (\underbrace{0, \cdots, 0}_{j-1}, \rho_j, \tau_j, \cdots)^T$$

we obtain, for tridiagonal $\tilde{J}_{N+1}$, from (4.13) taken elementwise,

$$\tilde{\beta}_{j-1}\rho_{j-1} = \mathbf{e}_{j-1}^T \tilde{J}_{N+1} \mathbf{v}_j = \beta_{j-1}\rho_j, \qquad j = 2, 3, \cdots, N+1, \tag{4.17}$$

$$\tilde{\alpha}_j \rho_j + \tilde{\beta}_{j-1}\tau_{j-1} = \mathbf{e}_j^T J_{N+1} \mathbf{v}_j = \begin{cases} \alpha_j \rho_j + \beta_j \tau_j, & 1 \leqq j \leqq N, \\ x\rho_{N+1}, & j = N+1, \end{cases} \tag{4.18}$$

$$\tilde{\beta}_{N+1} = 1/\rho_{N+1},$$

the last equation coming from (4.12).

Unfortunately, we cannot generate the ρ_j and τ_j using (4.13) without calculating the other elements of L, too. In the next two sections we present two other ways to determine ρ_j, τ_j without using (4.13).

Turning now to the updating of the Fourier coefficients $\mathbf{d}$, we have

$$L\tilde{\mathbf{d}}_{N+1} = \begin{pmatrix} \mathbf{d}_N + w^2 y \mathbf{p}_N(x) \\ w^2 y p_N(x) \end{pmatrix} \tag{4.19}$$

by a calculation which parallels (4.6). The updated coefficients can be found by a forward substitution.

4.4. Solution by determinants. In this section we derive formulae (see [12]) for the ρ_j and τ_j from subdeterminants of the Gram matrix of modified moments M_{N+1} of (4.14).

Let us denote (Choleski decomposition escalates)

$$M_k = [\mathbf{p}_k, \mathbf{p}_k^T]_{w_{N+1}} = \begin{pmatrix} M_{k-1} & \mathbf{m}_k \\ \mathbf{m}_k^T & \mu_k \end{pmatrix} = L_k L_k^T$$

and

$$L_k = \begin{pmatrix} L_{k-1} & \mathbf{0} \\ \mathbf{r}_k^T & \rho_k \end{pmatrix}.$$

We have, generally,

$$\det(M_k) = \det(L_k)^2 = \rho_1^2 \rho_2^2 \cdots \rho_k^2$$

so that

$$\rho_k = (\det(M_k)/\det(M_{k-1}))^{1/2}, \qquad k = 2, 3, \cdots \tag{4.20}$$

while

$$\rho_1 = M_1^{1/2}.$$

We now derive an explicit expression for τ_k. Defining two $k \times k+1$ selection matrices

$$S = (I_k \quad \mathbf{0}), \qquad \tilde{S} = \begin{pmatrix} I_{k-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & 0 & 1 \end{pmatrix},$$

we have

$$\tilde{M}_k = \tilde{S} M_{k+1} S^T = \begin{pmatrix} M_{k-1} & \mathbf{m}_k \\ \mathbf{m}_{k+1}^T & \end{pmatrix}$$

and

$$\tilde{S} L_{k+1} L_{k+1}^T S^T = \begin{pmatrix} L_{k-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{r}_{k+1} & & \rho_{k+1} \end{pmatrix} \begin{pmatrix} L_{k-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{r}_k^T & \rho_k & 0 \end{pmatrix}^T$$

in which, the last columns not contributing, the right-hand side is a product of two triangular matrices. Taking determinants, we have

$$\det(\tilde{M}_k) = \rho_1^2 \rho_2^2 \cdots \rho_{k-1}^2 \rho_k \tau_k,$$

from which

$$\tau_k = \det(\tilde{M}_k)(\det(M_k)\det(M_{k-1}))^{-1/2}. \tag{4.21}$$

So far we worked with general M_k, that is the Gram matrix of the "old" polynomials $\mathbf{p}$ with respect to the inner product determining the "new" orthogonal polynomials. In this sense, the above formulae could be used to improve the LTL method of [10] whenever the above determinants are obtainable. This is the case in the present situation as

$$[f, g]_{w_{N+1}} = [f, g]_{w_N} + w_{N+1}^2 f(x_{N+1}) g(x_{N+1})$$

so that

$$M_k = I_k + \mathbf{u}_k \mathbf{u}_k^T \tag{4.22}$$

where $\mathbf{u}_k = w_{N+1} \mathbf{p}_k(x_{N+1})$. Denoting $\psi_j = w_{N+1} p_{j-1}(x_{N+1})$, the jth element of $\mathbf{u}_k$,

$j \leqq k$ and independent of k, we have also

$$\mathbf{m}_k = \psi_k \mathbf{u}_{k-1}, \qquad \mu_k = 1 + \psi_k^2$$

and

$$\tilde{M}_k = \begin{pmatrix} I + \mathbf{u}_{k-1}\mathbf{u}_{k-1}^T & \psi_k \mathbf{u}_{k-1} \\ \psi_{k+1}\mathbf{u}_k^T & \end{pmatrix} = \begin{pmatrix} I + \mathbf{u}_{k-1}\mathbf{u}_{k-1}^T & \psi_k \mathbf{u}_{k-1} \\ \psi_{k+1}\mathbf{u}_{k-1}^T & \psi_{k+1}\psi_k \end{pmatrix}.$$

By a well-known formula

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(D) \det(A - BD^{-1}C),$$

so that we now have

$$\det(\tilde{M}_k) = \psi_k \psi_{k+1} \det(I + \mathbf{u}_{k-1}\mathbf{u}_{k-1}^T - \psi_k \mathbf{u}_{k-1}\psi_{k+1}\mathbf{u}_{k-1}^T/\psi_k\psi_{k+1}) = \psi_k\psi_{k+1}. \tag{4.23}$$

Of course, the other determinant follows easily from (4.22):

$$\det(M_k) = 1 + \mathbf{u}_k^T \mathbf{u}_k = 1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_k^2. \tag{4.24}$$

To update the Fourier coefficients we need to solve (4.19). The right-hand side is known because $w^2 y \mathbf{p}_{N+1}(x) = wy\mathbf{u}_{N+1}$. By transposing (4.12), one can derive a recurrence for the rows of L which is similar to that in (4.13) for the columns of L but with J and $\tilde{J}$ interchanged. The forward substitution can thus be performed as the elements of $\tilde{J}$ are obtained (see algorithm TLD in § 6.3.1 for details).

4.5. Solution using the special form of L. Although the calculation of the new Jacobi matrix by the method in the previous section is convenient, the computation of the rows of the L matrix requires $O(n^2)$ operations while other methods need $O(n)$. As pointed out in [5], the matrix L, being the Cholesky factor of a rank-one update to a diagonal matrix, has a special form. We exploit this property to calculate the updated Fourier coefficients more efficiently. In fact, the special form of L leads to a new method which we now describe.

From (4.14) and from (4.22), we have

$$LL^T = M_{N+1} = \operatorname{diag}(I_N, 0) + \mathbf{u}_{N+1}\mathbf{u}_{N+1}^T. \tag{4.25}$$

The form referred to in [5] is special in that the subdiagonal elements of the factor L are the elements of the strictly lower triangular part of a rank-one matrix. Comparing the elements of (4.25) we have for $j < k$

$$\mathbf{e}_k^T L \mathbf{e}_j = \psi_k q_j,$$

where the diagonal elements ρ_j of L and q_j must satisfy

$$\rho_j^2 = 1 - \delta_{j,N+1} + \psi_j^2(1 - (q_1^2 + q_2^2 + \cdots q_{j-1}^2)), \tag{4.26}$$

$$\rho_j q_j = \psi_j(1 - (q_1^2 + q_2^2 + \cdots + q_{j-1}^2)). \tag{4.27}$$

We can therefore evaluate ρ_j and τ_j alternatively and this is sufficient to proceed with the recurrence in (4.17) and (4.18) for which we need ρ_j and the ratio

$$\frac{\tau_j}{\rho_j} = \frac{\psi_{j+1}\psi_j}{\psi_j^2 + 1/(1 - (q_1^2 + q_2^2 + \cdots + q_{j-1}^2))}.$$

The forward substitution for the Fourier coefficients now simplifies significantly to

$$\tilde{d}_j = \left(d_j + wy\psi_j - \psi_j \sum_{k=1}^{j-1} \tilde{d}_k q_k \right) \Big/ \rho_j$$

where the sum is accumulated through the process. In the algorithmic implementation shown later, we save some operations by rearranging the recurrences in (4.26) and (4.27) to compute $z_j := 1 - (q_1^2 + q_2^2 + \cdots + q_{j-1}^2)$ rather than q_j.

4.6. Updating a partial matrix. The four methods in §§ 4.2–4.5 were described for the case when $n = N$ (see the formulation in § 4.1). For practical purposes we are mostly interested in the situation where $n \ll N$. We have already pointed out in Corollary 4.1 (§ 4.2) that when updating by the rotations method it is sufficient to store and use the submatrix J_n of J_N (corresponding to W_N) of size $n < N$ to calculate the submatrix $\tilde{J}_n$ of $\tilde{J}_{N+1}$ (corresponding to $W_{N+1} = W_N \cup \{x_{N+1}, w_{N+1}\}$). The same observation applies to the other three methods of §§ 4.3–4.5 because of their forward recurrence character.

It is worth noting that, once a restriction to some size n has been adopted, further increase is not meaningfully possible, i.e., to represent the full data set Y_N. In fact the partial solution of size $n < N$ may be visualized as representing a class of data sets, the smallest of which is the Gauss quadrature corresponding to the matrix J_n (its eigenvalues as knots, etc.) and the function values of the least squares fit polynomial at the quadrature knots. Any increase in the size of J_n (i.e., any increase in the degree of the least squares polynomial fit) then corresponds to augmenting the new data points to this minimal data set.

4.7. Changing the weight of an existing data point. Theorem 4.1 was derived under the assumption that the knots x_j in Y_{N+1} were distinct. We now discuss what happens when x is equal to one of the knots in Y_N, say $x = x_N$. Let us consider (as in Theorem 4.1) the size N solution $\tilde{J}_N$, $\tilde{\mathbf{d}}_N$ for the data $Y_{N+1} = \{x_j, w_j, y_j\}_{j=1}^{N+1}$ where $\{x_{N+1}, w_{N+1}, y_{N+1}\} \to \{x_N, w, y\}$ for some w and y. The limit of such a change will correspond to the solution for the data set $Y_N = \{x_j, \tilde{w}_j, \tilde{y}_j\}_{j=1}^{N}$ where

$$\begin{aligned} &\tilde{w}_j = w_j, \qquad \tilde{y}_j = y_j, \qquad j = 1, \cdots, N-1, \\ &\tilde{w}_N = \sqrt{w_N^2 + w^2}, \end{aligned} \tag{4.28}$$

$$\tilde{y}_N = (w_N^2 y_N + w^2 y)/(w_N^2 + w^2). \tag{4.29}$$

To see this note that (4.28) is obvious from the definition of $\tilde{\sigma}_N$ and implies that the limiting $\tilde{J}_N$ corresponds to $\tilde{W}_N = \{x_j, \tilde{w}_j\}_{j=1}^{N}$ independently of $\tilde{y}_j$. Denoting (as in (3.7) for Y_{N+1})

$$P_{N,N+1} = (P \quad \mathbf{p}_N(x_N) \quad \mathbf{p}_N(x_{N+1})),$$

we now have

$$\begin{aligned} \mathbf{d}_N &= P_{N,N+1} D_{N+1}^2 \mathbf{y}_{N+1} \\ &= P D_{N-1}^2 \mathbf{y}_{N-1} + w_N^2 y_N \mathbf{p}_N(x_N) + w_{N+1}^2 y_{N+1} \mathbf{p}_N(x_{N+1}), \end{aligned}$$

so that in the limit,

$$\tilde{\mathbf{d}}_N = P D_{N-1}^2 \mathbf{y}_{N-1} + (w_N^2 y_N + w^2 y) \mathbf{p}_N(x_N),$$

from which (4.29) follows directly.

We thus have a situation similar to the case discussed in § 4.2.

THEOREM 4.2. *Given* σ_N, J_N, *and* $\mathbf{d}_N$ *for the data* Y_N *and also* $\{x, w, y\}$, *where* $x = x_N$, *the solution for* $\tilde{Y}_N$ *is*

$$\begin{aligned} \tilde{J}_N &= QJ_NQ^T, \\ \tilde{\sigma}_N &= (\sigma_N^2 + w^2)^{1/2}, \\ \tilde{\mathbf{d}}_N &= Q(\mathbf{d}_N + (\tilde{w}_N\tilde{y}_N - w_Ny_N)\mathbf{q}_N), \end{aligned} \tag{4.30}$$

where the orthogonal similarity matrix Q *is uniquely determined by requiring* $\tilde{J}_N$ *to be tridiagonal and*

$$Q(\sigma_N\mathbf{e}_1 + (\tilde{w}_N - w_N)\mathbf{q}_N) = \tilde{\sigma}_N\mathbf{e}_1.$$

Here $\mathbf{q}_N$ *is the eigenvector of* J_N *corresponding to* x_N, *scaled to have norm one and a positive first element.*

Proof. Denoting again $J_N = Q_N\Lambda_NQ_N^T$ with $\sigma_NQ_N^T\mathbf{e}_1 = w_N$ we require that $\tilde{J}_N = \tilde{Q}_N\Lambda_N\tilde{Q}_N^T$ with $\tilde{\sigma}_N\tilde{Q}_N^T\mathbf{e}_1 = \tilde{\mathbf{w}}_N = \mathbf{w}_N + (\tilde{w}_N - w_N)\mathbf{e}_N$. Thus $\tilde{J}_N = QJ_NQ^T$ with $Q = \tilde{Q}_NQ_N^T$ so that

$$\tilde{\sigma}_NQ^T\mathbf{e}_1 = Q_N\tilde{\mathbf{w}}_N = \sigma_N\mathbf{e}_1 + (\tilde{w}_N - w_N)Q_N\mathbf{e}_N,$$

as required. The normalization of the eigenvector $\mathbf{q}_N = Q_N\mathbf{e}_N$ must be such as to force $\tilde{\sigma}_N > \sigma_N$ for $\tilde{w}_N > w_N$. For the Fourier coefficients we then have ($\tilde{D}_N = \operatorname{diag}(\tilde{\mathbf{w}}_N) = D_N + (\tilde{w}_N - w_N)\mathbf{e}_N\mathbf{e}_N^T$)

$$\begin{aligned} \tilde{\mathbf{d}}_N &= \tilde{Q}_N\tilde{D}_N\tilde{\mathbf{y}}_N \\ &= QQ_N(D_N + (\tilde{w}_N - w_N)\mathbf{e}_N\mathbf{e}_N^T)(\mathbf{y}_N + (\tilde{y} - y_N)\mathbf{e}_N), \end{aligned}$$

which gives (4.30) after cancellation. □

This theorem shows that it is possible to modify the weight of a chosen knot by orthogonal similarity as in § 4.2. However, the eigenvector $\mathbf{q}_N$ would then have to be evaluated. On the other hand, the four updating methods discussed so far can be applied to the present situation—the only difference is that the resulting matrix $\tilde{J}_{N+1}$ cannot then be unreduced because it has two equal eigenvalues. It is then a consequence of Theorem 4.2 that $\tilde{\beta}_N = 0$.

We note that the proof of Theorem 4.1 depends on the matrix $\tilde{J}_{N+1}$ having a uniquely determined set of eigenvectors. Here this is not the case as the two equal eigenvalues have a two-dimensional invariant subspace. In fact, condition (4.4) determining the first row of Q is based on one choice of eigenvectors in this subspace while the required modification of the weight in the resulting matrix corresponds to another choice of these eigenvectors.

5. Downdating methods.

5.1. A reformulation of the downdating problem. The problem of downdating is now, with respect to the representation of the solutions introduced in § 3, as follows:

Given

(a) $\mu_0 > 0$,

(b) J_n, a Jacobi matrix of size n,

(c) a vector $\mathbf{d}_n$,

(d) a triple $\{x, w, y\}$, and

(e) $\{x, w, y\} \in Y_{N+1}$

((a), (b), and (c) representing the solution of Problem 1 for some data set Y_N), find $\tilde{n}$, $\tilde{\mu}_0$, $\tilde{\mathbf{d}}_{\tilde{n}}$, and $\tilde{J}_{\tilde{n}}$ representing the solution for the data set $\tilde{Y}_N = Y_{N+1} \setminus \{x, w, y\}$ where $\tilde{n}$ is the largest integer for which a solution exists.

As before, we have formulated the downdating problem only for Problem 1—the corresponding version for Problem 2 is obtained by omitting the coefficients $\mathbf{d}$ and values y.

The problem as formulated here always has a solution with $\tilde{n} = \min\{n, N\}$. In practical cases the set Y_{N+1} may have been discarded and so one would not know if the given triple belonged to the set. We are thus interested in solving the problem without assuming (e) above, in which case $\tilde{n}$ may decrease further.

The rest of this section is therefore set out as follows. We first derive downdating methods closely related to the four updating methods of § 4 for the case $n = N + 1$, i.e., when the solution (of size $\tilde{n} = N$) is known to exist. Next we derive another downdating method based on reversing the rotations method. We then turn to the problem of downdating a partial matrix ($n < N + 1$) and characterize the existence of the solution in that case—the downdating may, in general, decrease the size $\tilde{n}$ of the solution or the solution may not exist at all. This leads to stopping criteria for all the methods discussed.

5.2. Methods based on a complex weight. As formulated, Theorem 4.2 describes only an increase of the weight of the data triple $\{x_N, w_N, y_N\}$, i.e., $w^2 > 0$ implies $\tilde{w}_N > w_N$. However, replacing w^2 in Theorem 4.2 by $-w^2$ (note that only w^2 is used there, not w) leads to a solution $\tilde{J}_N$ (with $\tilde{w}_N < w_N$) which remains real and unreduced of size N as long as $w^2 < w_N^2$. In the limit, when $\tilde{w}_N = 0$, we will have $\tilde{\beta}_{N-1} = 0$, $\tilde{\alpha}_N = x_N$ and the size $N-1$ submatrix of $\tilde{J}_N$ will be the result of the downdating problem of § 5.1 (although for $n = N$ rather than $N + 1$).

We have pointed out in § 4.7 that the modification of a weight can be achieved by any of the three methods for updating. For downdating this simply means replacing w^2 by $-w^2$ and thus, if necessary, w by iw, $i^2 = -1$.

Inspection of each of these methods shows that, throughout the calculation, the $N \times N$ submatrix which is required remains real and those elements of the calculations which are complex quantities remain pure imaginary. Thus the whole calculation can be done very conveniently in real arithmetic (see the algorithms in § 6 for details).

In the case of the rotation method the updating similarity matrices are products of (orthogonal) plane rotations: a $\left(\begin{smallmatrix} c & s \\ -s & c \end{smallmatrix}\right)$, $c^2 + s^2 = 1$ matrix imbedded in an identity. The similarity matrices here are products of elementary matrices which are of the form $\left(\begin{smallmatrix} c & is \\ -is & c \end{smallmatrix}\right)$, $c^2 - s^2 = 1$ imbedded in an identity. These matrices, which may be called hyper-rotations because the sines and cosines are replaced by hyperbolic sines and cosines, are (complex) orthogonal but not unitary. While plane rotations always have condition unity the condition of a hyper-rotation is $(|c| + |s|)/(|c| - |s|)$ which can be arbitrarily large. The inherent instability of downdating is manifest here in the possible ill-conditioning of these hyper-rotation matrices.

5.3. An eigenvector method for downdating. The downdating problem in § 5.1 is a converse of the updating problem as formulated in § 4.1 (up to exchanging J and $\tilde{J}$, etc.). We can therefore attempt to solve it by reversing one of the methods for updating—this is particularly suitable for the rotations method. The analogue of Theorem 4.1 now is Theorem 5.1.

THEOREM 5.1. *Given σ_{N+1}, J_{N+1}, and $\mathbf{d}_{N+1}$ for the data $Y_{N+1} = Y_N \cup \{x, w, y\}$ the solution for Y_N is*

$$\begin{pmatrix} J_N & \mathbf{0} \\ \mathbf{0}^T & x \end{pmatrix} = Q J_{N+1} Q^T,$$

$$\tilde{\sigma}_N = (\sigma_N^2 - w^2)^{1/2}, \tag{5.1}$$

$$\begin{pmatrix} \tilde{\mathbf{d}}_N \\ wy \end{pmatrix} = Q d_{N+1}$$

where the (real) orthogonal similarity matrix Q is uniquely determined by requiring $\tilde{J}_N$ to be tridiagonal and

$$Q(\sigma_{N+1}\mathbf{e}_1 - w\mathbf{q}) = \tilde{\sigma}_N\mathbf{e}_1$$

where $\mathbf{q}$ is the eigenvector of J_{N+1} corresponding to x, scaled to have norm one and a positive first element.

The similarity Q is now the product of the same rotations as in (4.7): $Q^T = R_N R_{N-1}\cdots R_1$. In fact these can again be calculated implicitly after finding R_N first and applying it to J_{N+1}. For R_N, only the ratio of the last two elements of the eigenvector $\mathbf{q}$ is required, so $\mathbf{q}$ need not even be scaled. These unscaled last two elements could, in theory, be obtained by one step of the back-substitution on $J_{N+1} - xI$; however, as we shall see in the next section, this is not practical.

5.4. Downdating a partial matrix and the existence of the solution. When we have available only a submatrix J_n of the whole matrix J_{N+1}, $n < N+1$ and the triple to be downdated $\{x, w, y\}$, the question of the existence of the downdating solution is equivalent to the existence of some matrix $J_{\hat{N}}$, $\hat{N} > n$ with the following properties:

- The given J_n is a submatrix of $J_{\hat{N}}$.
- x is an eigenvalue of $J_{\hat{N}}$ with the weight w scaled to the current zeroth moment σ_{N+1}^2.

The smallest matrix with these properties will have dimension $\hat{N} = n+1$ and will be of the form

$$\begin{pmatrix} J_n & \beta\mathbf{e}_n \\ \beta\mathbf{e}_n^T & \alpha \end{pmatrix},$$

where α and β are to be chosen so that $J_{\hat{N}}$ has eigenvalue x with weight w. Denoting the corresponding normalized eigenvector $\binom{\mathbf{q}}{\eta}$, this leads to equations

$$\begin{aligned} \mathbf{e}_1^T\mathbf{q} &= w/\sigma_{N+1}, \\ (J_n - xI)\mathbf{q} &= -\beta\eta\mathbf{e}_n, \\ \beta\mathbf{e}_n^T\mathbf{q} + \alpha\eta &= x\eta. \end{aligned} \tag{5.2}$$

The first n of these $n+2$ equations form a lower triangular system for $\mathbf{q}$ (independent of α and β) which can thus be determined uniquely. The solution $\mathbf{q}$ is proportional to w, so for w not too large we have $\|\mathbf{q}\| < 1$ and a real $\eta = (1 - \mathbf{q}^T\mathbf{q})^{1/2}$ exists. Otherwise, for larger w, we might need to decrease the size of the solution to $\tilde{n} < n$ so that the norm of the shorter vector $\mathbf{q}$ is less than 1—this will imply the existence of the solution of the downdating problem for that size submatrix $J_{\tilde{n}}$. The unknown α and β are now easily obtained from the last two equations of (5.2); the sign of η may be chosen so that β is positive. The solution of the downdating problem can now be obtained as discussed in § 5.3 from the last two elements of the eigenvector $(\mathbf{q}^T\ \eta)^T$. We note here that, contrary to the case discussed in § 5.3, the complete forward substitution is indeed necessary as we do not know the whole matrix J_{N+1} needed in order to do only one step of back-substitution as suggested there.

We have thus established the following result.

THEOREM 5.2. *Given σ_{N+1} and J_n, $n \leqq N$, the solution for some Y_{N+1} and the downdating pair $\{x, w\}$, denote $\mathbf{q}$ the solution of just the first $n-1$ equations of $(J_n - xI)\mathbf{q} = \mathbf{0}$ satisfying $\mathbf{e}_1^T\mathbf{q} = w/\sigma_{N+1}$. Let $\tilde{n}$ be the largest integer such that $\tilde{\mathbf{q}}^T\tilde{\mathbf{q}} < 1$ where $\tilde{\mathbf{q}}$ is the vector of the first $\tilde{n}$ elements of $\mathbf{q}$. Then $\tilde{n}$ is the largest size of the leading submatrix $J_{\tilde{n}}$ of J_n for which it is possible to downdate by the pair $\{x, w\}$.*

Obviously, if $w \geqq \sigma_{N+1}$ then $\tilde{n} = 0$ and no downdating is possible. Furthermore, all three downdating methods in § 5.2 must produce a solution of size $\tilde{n}$; suitable criteria for recognizing this critical size are given in the algorithms. They correspond to identifying the depletion of the total *mass*, i.e., the current zeroth moment, during the process.

The constructive nature of Theorem 5.2, for downdating a Jacobi matrix, means that a practical method can be based on it. The vector $\mathbf{d}_{\tilde{n}}$ of given Fourier coefficients has also to be extended by an unknown element, say $\mathbf{d}_{\tilde{n}+1}$, so that after the transformation (5.1) the last element of the left-hand side is wy. We achieve this by the following method. Note first that elements $\mathbf{e}_j^T \mathbf{d}_{\tilde{n}+1}$ are known for the range $j = 1, 2, \cdots, \tilde{n}$ and wy is known. We may rewrite (5.1) as

$$\begin{pmatrix} \tilde{\mathbf{d}}_{\tilde{n}} \\ 0 \end{pmatrix} + wy\mathbf{e}_{\tilde{n}+1} = Q\begin{pmatrix} \hat{\mathbf{d}}_{\tilde{n}} \\ 0 \end{pmatrix} + \gamma Q\mathbf{e}_{\tilde{n}+1}, \tag{5.3}$$

for some scalar γ. As the $R_{\tilde{n}}$, $R_{\tilde{n}-1}, \cdots, R_1$ become available we apply them to the known $\hat{\mathbf{d}}_{\tilde{n}}$ and $\mathbf{e}_{\tilde{n}+1}$. After all rotations have been applied we use the last row of (5.3) to solve for γ. The right-hand side of (5.3) then gives the solution.

The elementary matrices used are real (orthogonal) rotations. So the stability of the process depends on the accuracy with which the first rotation used (R_n) is determined. This again is limited by the condition of the forward substitution phase of the process and the size of η. This must be taken into account when applying the above-mentioned stopping criterion.

6. The algorithms. The algorithms here avoid complex arithmetic by exploiting the fact, mentioned in § 5.2, that whenever some computed quantities are complex, they are pure imaginary. Thus in most cases, the updating and downdating algorithms for the same method have been combined into a single version in which the upper sign of a $\pm$ or $\mp$ pair is to be read for updates and the lower sign for downdates. Consequently, a test such as "if $\sigma^2 \pm w^2 < 0 \cdots$" is not necessary if only the + sign is used.

The algorithms have the following input and output.

Input:

$n \geqq 1$ the size of the current solution

the Jacobi matrix J_n in the form of two vectors $\{\alpha_1, \alpha_2, \cdots, \alpha_n\}$ and $\{\beta_1, \beta_2, \cdots, \beta_{n-1}\}$

the current zeroth moment σ

coefficients $d_1, d_2, \cdots, d_n$ of the least squares fit polynomial

a triple $\{x, w, y\}$

Output:

$\tilde{n}$ the size of the solution

the new zeroth moment $\tilde{\sigma}$

the Jacobi matrix $\tilde{J}_{\tilde{n}}$, in the form of two vectors $\{\alpha_1, \alpha_2, \cdots, \alpha_{\tilde{n}}\}$ and $\{\beta_1, \beta_2, \cdots, \beta_{\tilde{n}-1}\}$, which overwrite J

coefficients $\tilde{d}_1, \tilde{d}_2, \cdots, \tilde{d}_{\tilde{n}}$ of the up/down dated least squares fit polynomial, which overwrite the d_i's

The new size $\tilde{n}$ will satisfy

$$\tilde{n} = \begin{cases} n+1 \quad \text{or} \quad n & \text{if } J \text{ has been updated,} \\ \leqq n & \text{if } J \text{ has been downdated.} \end{cases}$$

Of course if a downdate is attempted which it is impossible to perform $\tilde{n} = 0$ will result and no solution will be returned.

These algorithms can be used to build up a solution of any required size starting with the solution to the problem for $Y_1 = \{x, w, y\}$ which, as was pointed out in § 4.1, is

$$\alpha_1 = x, \quad \sigma = |w|, \quad d_1 = \sigma y.$$

We use comments (marked by #) to connect the variables in the algorithms to their derivation in the text of the previous sections. The calculations related to the up/downdating of the Fourier coefficients d_i are identified by indentation. These steps can be omitted if only $\tilde{J}$ is required.

Symbols in bold type represent vector quantities.

6.1. RHR—up/downdating algorithm using rotations/hyper-rotations. This algorithm is an implementation of the update method described in § 4.2 and its downdating variant based on § 5.2. In the case of the updating version of this algorithm, the first two tests are trivially false and the fourth will be true whenever x is an eigenvalue of J_n.

$\tilde{n} = n + 1$
if $\sigma^2 \pm w^2 \leqq 0$ then set $\tilde{n} = 0$ and exit. endif:
$\tilde{\sigma} = \sqrt{\sigma^2 \pm w^2}$
$c = \sigma/\tilde{\sigma}$ # set up (hyper)rotation for implicit step
$s = w/\tilde{\sigma}$
$\sigma = \tilde{\sigma}$
$\theta_1 = cs(x - \alpha_1)$ #θ_1, θ_2 are nonzero elements-
$\theta_2 = -s\beta_1$ #introduced by the implicit step
$\alpha_{\tilde{n}} = c^2 x \pm s^2 \alpha_1$ # apply (hyper)rotation
$\alpha_1 = c^2 \alpha_1 \pm s^2 x$
if $n > 1$ then set $\beta_1 = c\beta_1$ endif:
    $d_{\tilde{n}} = cwy - sd_1$
    $d_1 = cd_1 \pm swy$
for $i = 2, 3, \cdots, \tilde{n} - 1$ # chase matrix back to tridiagonal form
    if $\beta_{i-1}^2 \pm \theta_1^2 \leqq 0$ then set $\tilde{n} = i - 1$ and exit. endif:
    $\zeta = \sqrt{\beta_{i-1}^2 \pm \theta_1^2}$ # set up (hyper)rotation
    $c = \beta_{i-1}/\zeta$
    $s = \theta_1/\zeta$
    $\beta_{i-1} = \zeta$
    $t = c^2 \alpha_i \pm 2cs\theta_2 \pm s^2 \alpha_{\tilde{n}}$
    $r = c^2 \alpha_{\tilde{n}} \mp 2cs\theta_2 \pm s^2 \alpha_i$
    $\alpha_i = t$
    $\alpha_{\tilde{n}} = r$
    $\theta_1 = cs(\alpha_{\tilde{n}} - \alpha_i) + \theta_2(c^2 \mp s^2)$
    $\theta_2 = -s\beta_i$
    $\beta_i = c\beta_i$
        $t = d_i$
        $d_i = ct \pm sd_{\tilde{n}}$
        $d_{\tilde{n}} = cd_{\tilde{n}} - st$
endfor i:
if downdate or $\theta_1 = 0$ then
    set $\tilde{n} = n$ and exit.
else
    $\beta_{\tilde{n}-1} = |\theta_1|$ # adjustment for larger J
    $d_{\tilde{n}} = \text{sign}(\theta_1) d_{\tilde{n}}$
endif:

6.2. REV—rotations eigenvector downdate method. This is an implementation of the method derived in § 5.4, which is a reversal of the update method by rotations.

$\tilde{n} = n$
$u_{\text{old}} = 0$
$u = w/\sigma$
$u_{\text{new}} = (x - \alpha_1)u/\beta_1$
if $\sigma^2 - w^2 < 0$ then set $\tilde{n} = 0$ and exit. endif:
$\tilde{\sigma} = \sqrt{\sigma^2 - w^2}$
$m = 1 - u^2$
if $m - u_{\text{new}}^2 < 0$ then set $\tilde{n} = 1$ and goto label1: endif:
$m = m - u_{\text{new}}^2$
for $j = 2, 3, \cdots, \tilde{n} - 1$ #forward substitution for the eigenvector
  $u_{\text{old}} = u\ ;\ u = u_{\text{new}}$
  $u_{\text{new}} = ((x - \alpha_j)u - \beta_{j-1}u_{\text{old}})/\beta_j$
  if $m - u_{\text{new}}^2 < 0$ then set $\tilde{n} = j$ and goto label1: endif:
  $m = m - u_{\text{new}}^2$
endfor j:
$u_{\text{old}} = u\ ;\ u = u_{\text{new}}$
label1:
$u_{\text{new}} = \sqrt{m}$
$\beta_{\tilde{n}} = ((x - \alpha_{\tilde{n}})u - \beta_{\tilde{n}-1}u_{\text{old}})/u_{\text{new}}$ #new α and β
$\alpha_{\tilde{n}+1} = x - \beta_{\tilde{n}}u/u_{\text{new}}$
$c = u_{\text{new}}/\sqrt{u_{\text{new}}^2 + u^2}$ #get rotation for implicit step
$s = -u/\sqrt{u_{\text{new}}^2 + u^2}$
  for $i = 1, 2, \cdots, \tilde{n}$
    $q_i = 0$ #$Q\mathbf{e}_{\tilde{n}+1}$ of (5.3)
  endfor i:
$\alpha_{tmp} = \alpha_{\tilde{n}+1}$
$p = -s\beta_{\tilde{n}-1}$
$r = cs(\alpha_{tmp} - \alpha_{\tilde{n}}) + (c^2 - s^2)\beta_{\tilde{n}}$
$\alpha_{\tilde{n}+1} = c^2\alpha_{tmp} + s^2\alpha_{\tilde{n}} - 2cs\beta_{\tilde{n}}$
$\alpha_{\tilde{n}} = c^2\alpha_{\tilde{n}} + 2cs\beta_{\tilde{n}} + s^2\alpha_{tmp}$
$\beta_{\tilde{n}-1} = c\beta_{\tilde{n}-1}$
$\beta_{\tilde{n}} = 0$
  $d_{\tilde{n}+1} = -sd_{\tilde{n}}$
  $d_{\tilde{n}} = cd_{\tilde{n}}$
  $q_{\tilde{n}} = s$
  $q_{\tilde{n}+1} = c$
for $j = \tilde{n} - 1, \tilde{n} - 2, \cdots, 1$ #chase non-zeros from bottom to top
  $s = r/\sqrt{\beta_j^2 + r^2}$
  $c = \beta_j/\sqrt{\beta_j^2 + r^2}$
  $\alpha_{tmp} = \alpha_{\tilde{n}+1}$
  $\alpha_{\tilde{n}+1} = c^2\alpha_{tmp} + s^2\alpha_j - 2csp$
  $\beta_j = c\beta_j + sr$
  $r = cs(\alpha_{tmp} - \alpha_j) + (c^2 - s^2)p$
  $\alpha_j = c^2\alpha_j + 2csp + s^2\alpha_{tmp}$
  if $j > 1$ then
    $p = -s\beta_{j-1}$
    $\beta_{j-1} = \beta_{j-1}c$

endif:
$d_{tmp} = cd_j + sd_{\tilde{n}+1}$
$d_{\tilde{n}+1} = -sd_j + cd_{\tilde{n}+1}$
$d_j = d_{tmp}$
$q_{tmp} = cq_j + sq_{\tilde{n}+1}$
$q_{\tilde{n}+1} = -sq_j + cq_{\tilde{n}+1}$
$q_j = q_{tmp}$
endfor j:
for $i = 1, 2, \cdots, \tilde{n}$ #add the correction term of (5.3)
$d_i = d_i + (wy - d_{\tilde{n}+1})q_i/q_{\tilde{n}+1}$
endfor i:

6.3. The Lanczos-type algorithms for up- and downdating. All of the algorithms in this section are variations on the LTL method referred to in § 4.3 and are based on recurrences which come from (4.17) and (4.18). For clarity we restate these here in the way that they are used below.

$\alpha_1 = \alpha_1 + \beta_1(\tau_1/\rho_1)$
for $j = 2, 3, \cdots, N$
$\alpha_j = \alpha_j + \beta_j(\tau_j/\rho_j) - \beta_{j-1}(\tau_{j-1}/\rho_{j-1})$
$\beta_{j-1} = \beta_{j-1}\rho_j/\rho_{j-1}$
endfor j:
$\alpha_{N+1} = x - \beta_N(\tau_N/\rho_N)$
$\beta_N = \beta_N\rho_{N+1}/\rho_N$

It was pointed out in § 4.4 that these recurrences are applicable when just the diagonal and subdiagonal, ρ_j and τ_j, elements of the L matrix are available. Note that for this calculation the term β_N is needed at start to advance the algorithm. However, $\tilde{J}_{N+1}$ is independent of the input value of β_N as can be seen by inspection of the complete algorithms in this section. We therefore always set $\beta_N = 1$ on input rather than normalizing the root polynomial by setting $\beta_{N+1} = 1/\rho_{N+1}$ at output.

Another feature common to the algorithms in this section is that they all require the calculation of the elements $\psi_j = wp_{j-1}(x)$, $j = 1, 2, \cdots$ of an eigenvector $\mathbf{u}$ by the forward substitution

$$\beta_j\psi_{j+1} = (x - \alpha_j)\psi_j - \beta_{j-1}\psi_{j-1}.$$

When x lies inside the interval containing the eigenvalues of J_n, the values of the ψ_{j-1} are bounded by a reasonably small constant and may even vanish. For x outside this interval however, the ψ_{j-1} will never vanish and indeed will grow rapidly with j, possibly causing overflow. To avoid this possibility we present a second version of one of the algorithms which uses the ratios ψ_j/ψ_{j-1}. This *scaled* version should not be used when x is inside the interval referred to because some ratio may become infinite.

6.3.1. TLD—Triangular Lanczos method using determinants: Unscaled version. This is an implementation of the update method of § 4.4 and its downdating variant based on § 5.2. In the case of the updating version of this algorithm, the first two tests are trivially false and the third will be false whenever x is an eigenvalue of J_n.

$\tilde{n} = n + 1$
$q = w/\sigma$ # ψ_0
$p = (x - \alpha_1)q/\beta_1$ # ψ_1
if $\sigma^2 \pm w^2 \leqq 0$ then set $\tilde{n} = 0$ and exit. endif:

$\tilde{\sigma} = \sqrt{\sigma^2 \pm w^2}$
$t = 1 \pm q^2$
$\mathbf{r} = \sqrt{t}\mathbf{e}_1$
$\mathbf{q} = \alpha_1 \mathbf{r}$
$d_1 = (d_1 \pm wyq)/\sqrt{t}$
$\theta = pq/t$ # τ_1/ρ_1
$\alpha_1 = \alpha_1 \pm \theta\beta_1$
$\mathbf{q} = (\alpha_1 \mathbf{r} - \mathbf{q})/\beta_1$
$\beta_n = 1$ # needed for last polynomial
for $i = 2, 3, \cdots, \tilde{n} - 1$
$t_{\text{new}} = t \pm p^2$
if $t_{\text{new}}(t \mp q^2) \leqq 0$ then set $\tilde{n} = i - 1$ and exit. endif:
$\beta_{\text{new}} = \beta_{i-1}\sqrt{t_{\text{new}}(t \mp q^2)}/t$ # $\tilde{\beta}_{i-1}$
$\rho = \sqrt{t_{\text{new}}/t}$
$t = t_{\text{new}}$
$r = q$ # ψ_{i-1}
$q = p$ # ψ_i
$p = ((x - \alpha_i)q - \beta_{i-1}r)/\beta_i$ # ψ_{i+1}
$d_i = (d_i \pm wyq - \mathbf{d}^T\mathbf{q})/\rho$
$\mathbf{e}_i^T\mathbf{q} = \rho$
$\mathbf{v} = \mathbf{q}$
$\mathbf{q} = \alpha_i\mathbf{q} + \beta_{i-1}\mathbf{r}$
$\mathbf{r} = \mathbf{v}$
$\alpha_i = \alpha_i \mp \theta\beta_{i-1}$ # part of $\tilde{\alpha}_i$
$\theta = pq/t$ # τ_i/ρ_i
$\alpha_i = \alpha_i \pm \theta\beta_i$ # $\tilde{\alpha}_i$ completed
$\beta_{i-1} = \beta_{\text{new}}$ # $\tilde{\beta}_{i-1}$
$\mathbf{q} = (\tilde{J}_i\mathbf{r} - \mathbf{q})/\beta_i$ # $\tilde{J}_i$ is the updated part of $\tilde{J}_n$
endfor i:
if downdate or $|p|(t - q^2) \leqq 0$ then
set $\tilde{n} = n$ and exit.
else
$\alpha_{\tilde{n}} = x - \theta\beta_n$ # adjustment for larger J
$\beta_n = \beta_n |p| \sqrt{t - q^2}/t$
$d_{\tilde{n}} = (ywp - \mathbf{d}^T\mathbf{q})\sqrt{t}/|p|$
endif:

6.3.2. TLDS—Triangular Lanczos method using determinants: Scaled version. This is an implementation of the method of the previous section scaled to avoid overflow. The operations for up/downdating the vector $\mathbf{d}_n$ have been omitted. In the case of the updating version of this algorithm, the first two tests are trivially false and the third will be false whenever x is an eigenvalue of J_n.

$\tilde{n} = n + 1$
$q = w/\sigma$
$p = (x - \alpha_1)/\beta_1$
if $\sigma^2 \pm w^2 \leqq 0$ then set $\tilde{n} = 0$ and exit. endif:
$\tilde{\sigma} = \sqrt{\sigma^2 \pm w^2}$
$t = 1 \pm 1/q^2$
$\theta = p/t$
$\alpha_1 = \alpha_1 \pm \theta\beta_1$

$\beta_n = 1$
for $i = 2, 3, \cdots, \tilde{n} - 1$
    $t_{\text{new}} = 1 \pm t/p^2$
    if $t_{\text{new}}(t \mp 1) \leqq 0$ then set $\tilde{n} = i - 1$ and exit. endif:
    $\beta_{\text{new}} = \beta_{i-1}\theta\sqrt{t_{\text{new}}(t \mp 1)}$
    $t = t_{\text{new}}$
    $q = p$
    $p = ((x - \alpha_i) - \beta_{i-1}/q)/\beta_i$
    $\alpha_i = \alpha_i \mp \theta\beta_{i-1}$
    $\theta = p/t$
    $\alpha_i = \alpha_i \pm \theta\beta_i$
    $\beta_{i-1} = \beta_{\text{new}}$
endfor i:
if downdate or $|p|(t - 1) \leqq 0$ then
    set $\tilde{n} = n$ and exit.
else
    $\alpha_{\tilde{n}} = x - \theta\beta_n$
    $\beta_n = \beta_n |p| \sqrt{t - 1}/t$
endif:

6.3.3. TLS—Triangular Lanczos method using the special form of L: Unscaled version. This is an implementation of the update method described in § 4.5 and its downdate variant of § 5.2. In the case of the updating version of this algorithm, the first two tests are trivially false and the third will be false whenever x is an eigenvalue of J_n.

$\tilde{n} = n + 1$
$q = w/\sigma$ # ψ_0
$p = (x - \alpha_1)q/\beta_1$ # ψ_1
if $\sigma^2 \pm w^2 \leqq 0$ then set $\tilde{n} = 0$ and exit. endif:
$\tilde{\sigma} = \sqrt{\sigma^2 \pm w^2}$
$t = 1 \pm q^2$ # ρ_1^2
$z = 1 \mp q^2/t$
    $\hat{d} = d_1 \pm wyq$
    $d_1 = \hat{d}/\sqrt{t}$
$\theta = pq/t$ # τ_1/ρ_1
$\alpha_1 = \alpha_1 \pm \theta\beta_1$
    $s_f = \hat{d}q/t$ # sum for forward substitution
    $\hat{d} = d_2 \pm p(wy - s_f)$
$\beta_n = 1$
for $i = 2, 3, \cdots, \tilde{n} - 1$
    $t_{\text{new}} = 1 \pm p^2 z$
    if $t_{\text{new}} \leqq 0$ then set $\tilde{n} = i - 1$ and exit. endif:
    $\beta_{\text{new}} = \beta_{i-1}\sqrt{t_{\text{new}}/t}$ # $\tilde{\beta}_{i-1}$
    $t = t_{\text{new}}$
    $r = q$ # ψ_{i-1}
    $q = p$ # ψ_i
    $p = ((x - \alpha_i)q - \beta_{i-1}r)/\beta_i$ # ψ_{i+1}
        $d_i = \hat{d}/\sqrt{t}$
        $s_f = s_f + \hat{d}qz/t$
        $\hat{d} = (1 - \delta_{ni})d_{i+1} \pm p(wy - s_f)$
    $\alpha_i = \alpha_i \mp \theta\beta_{i-1}$ # part of $\tilde{\alpha}_i$

$\theta = pqz/t$ # τ_i/ρ_i
$z = z/t$
$\alpha_i = \alpha_i \pm \theta\beta_i$ # $\tilde{\alpha}_i$ completed
$\beta_{i-1} = \beta_{\text{new}}$ # $\tilde{\beta}_{i-1}$
endfor i:
if downdate or $|p|z/t \leqq 0$ then
set $\tilde{n} = n$ and exit.
else
$\alpha_{\tilde{n}} = x - \theta\beta_n$ # adjustment for larger J
$\beta_n = \beta_n|p|\sqrt{z/t}$
$d_{\tilde{n}} = \hat{d}/(|p|\sqrt{z})$
endif:

6.4. Computational complexity of the algorithms. We indicate here the complexity of each algorithm as a number of *flops* (essentially the number of $\times$ or $\div$ operations) per up/downdate showing only the term with the highest power of n:

Complexity

Method	Up/downdate J_n		Up/downdate $\mathbf{d}_n$	
	$\sqrt{\ }$	Flops	$\sqrt{\ }$	Flops
RHR	n	$15n$	—	$4n$
REV	n	$23n$	—	$11n$
TLD	n	$10n$	n	$\frac{7}{2}n^2$
TLDS	n	$10n$	—	—
TLS	n	$11n$	n	$3n$

7. Numerical tests. We wish to demonstrate the methods in this paper by computing a moving least squares polynomial fit to noisy data. The methods tested are:

RHR—rotations update, hyper-rotations downdate,
REV—rotations update, eigenvector/rotations downdate,
TLD—triangular Lanczos method using determinants,
TLS—triangular Lanczos method using the special form of L.

We define

$$x_j = -1 + 2(j-1)/(N-1) + \delta,$$

$$w_j = 1/\sqrt{N} + \delta,$$

$$y_j = 1.5 + \sin(4x_j) + \delta,$$

where δ is a uniformly distributed random variable with $|\delta| < 10^{-4}$. Furthermore, we set $Y_k = \{x_j, w_j, y_j\}_{j=k}^{k+M-1}$, $k = 1, 2, \cdots, N - M + 1$.

In the test we start by computing J_n, $\mathbf{d}_n$ and the least squares polynomial fits q_0, $q_1, \cdots, q_{n-1}$, each q_j of degree j, for the data Y_1 using only the update method of § 6.1. For $k = 2, \cdots, N - M + 1$, we then compute the solution to the data Y_k by first updating with the triple $\{x_{k+M}, w_{k+M}, y_{k+M}\}$ and then downdating with the triple $\{x_{k-1}, w_{k-1}, y_{k-1}\}$. After every n_s steps we compare the solution at this stage with a *reference* solution computed again using the update method of § 6.1. Thus the ith stage corresponds to the solution on Y_{in_s+1}.

The figures show the number of correct decimal digits ($-\log$ | *error*|) in the representation of the polynomials, i.e., we plot the accuracy of the α_j and β_j of J_n and of the

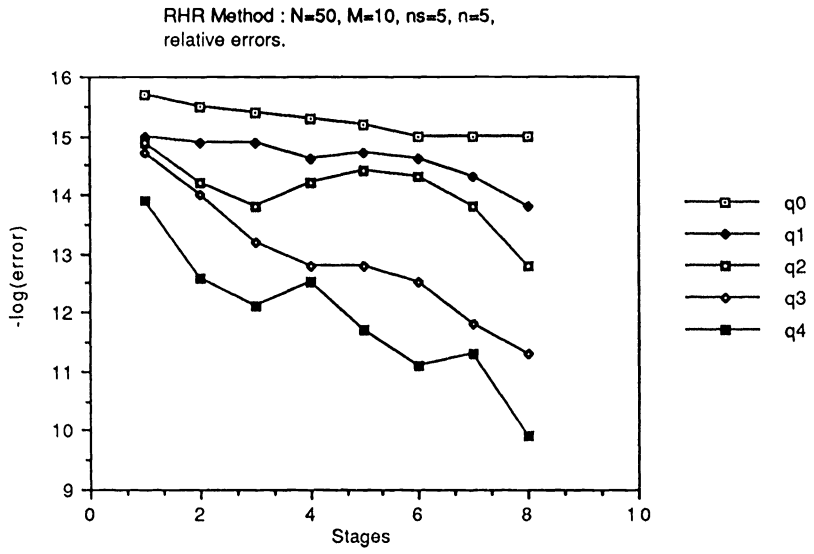


FIG. 1

coefficients d_j against the stage number. We are also interested in how accurately the least squares polynomial q_j is computed: we therefore calculate the relative difference

$$e^2 := \sum_j w_j^2 (q_i^{\text{ref}}(x_j) - q_i(x_j))^2 / y_j^2, \qquad i = 0, 1, \cdots, n-1$$

at each stage.

Figures 1, 2(a), and 2(b) show the errors for method RHR with $N = 50$, a window size of $M = 10$, the order of the polynomial (and therefore the size of the matrix) $n = 5$ staged every $n_s = 5$ steps. We observe that the accuracy decays as we slide the window

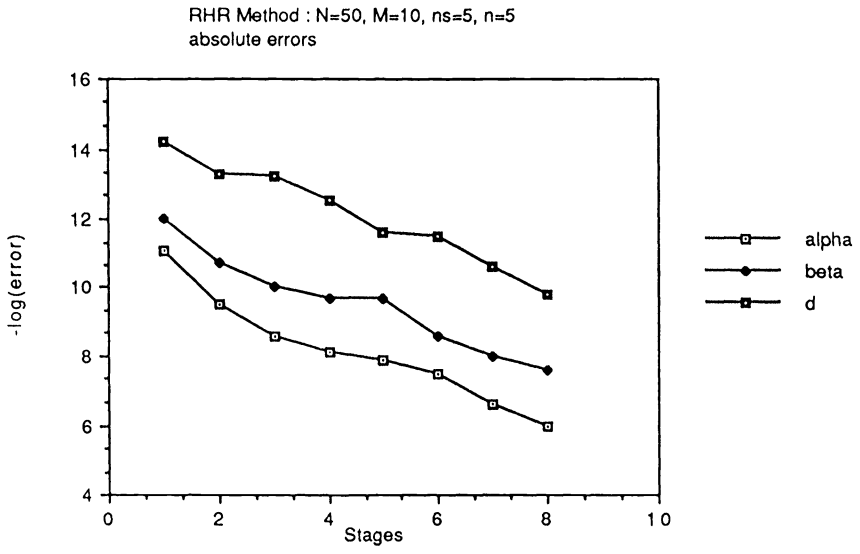


FIG. 2(a)

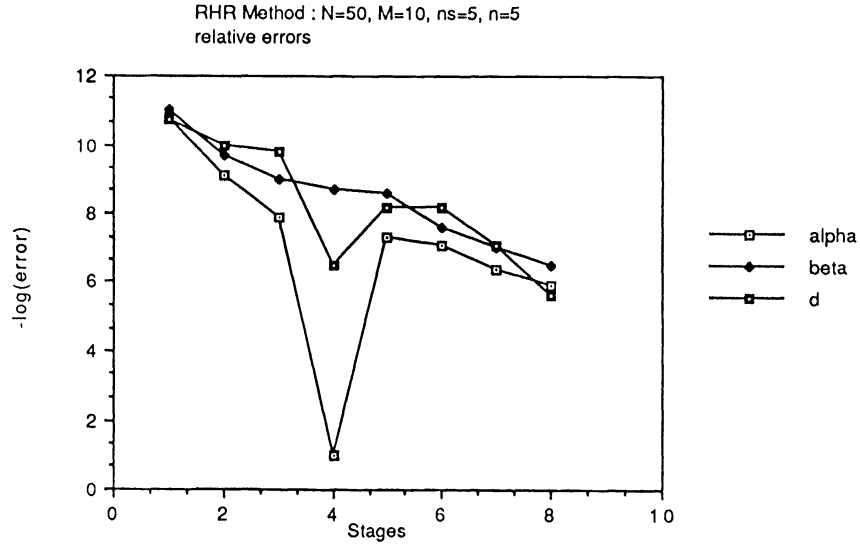


FIG. 2(b)

along the data set and also as we increase the degree of q_j. The relative errors of the diagonal elements of J and the **d** vector, shown in Fig. 2(b), become unduly large at the fourth stage where because of symmetry we are trying to compute zero quantities. Figure 2(a) shows the corresponding absolute errors.

Recall that to produce q_{j-1} for the data set Y_{k+1} requires the application of kj rotations and kj hyper-rotations. While the condition of the rotations is always one it follows from the discussion in § 5.2 that the condition of the product of hyper-rotations

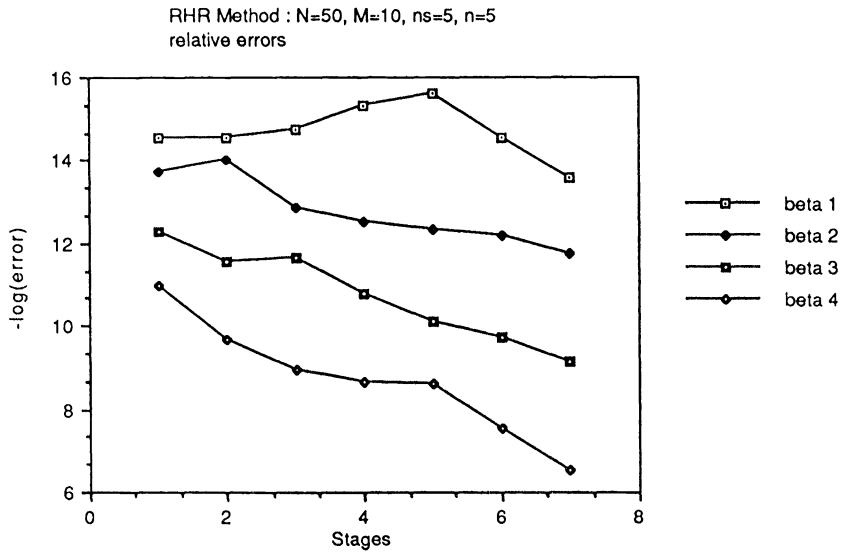


FIG. 3

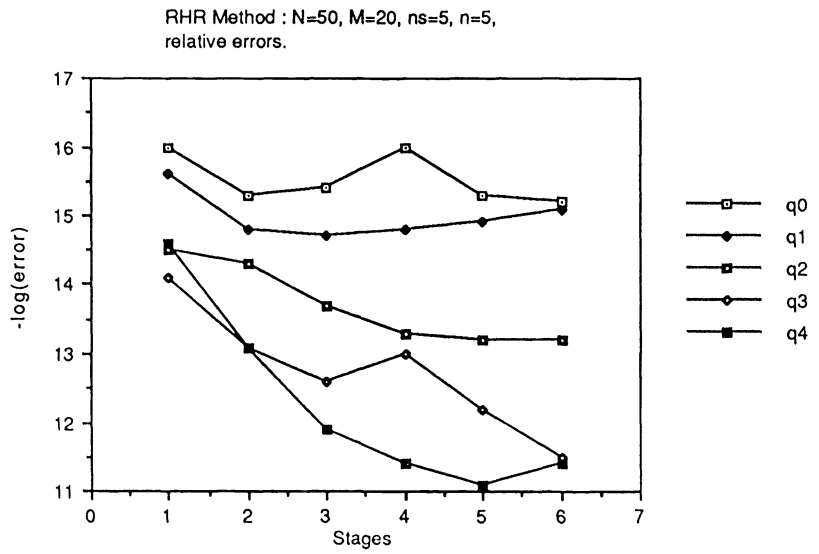


FIG. 4

involved is bounded by

$$\prod_{i=1}^{j}\prod_{l=1}^{k}\frac{|c_{il}|+|s_{il}|}{|c_{il}|-|s_{il}|},$$

$c_{il}^2 - s_{il}^2 = 1$, c_{il} and s_{il} being the c and s of the downdating RHR algorithm in § 6.1. The decay in accuracy observed in Figs. 1 and 2 is consistent with the structure of this bound: for lower degree polynomials the condition of the transformation in each step is quite close to one and the accuracy decreases only slowly as the window slides on.

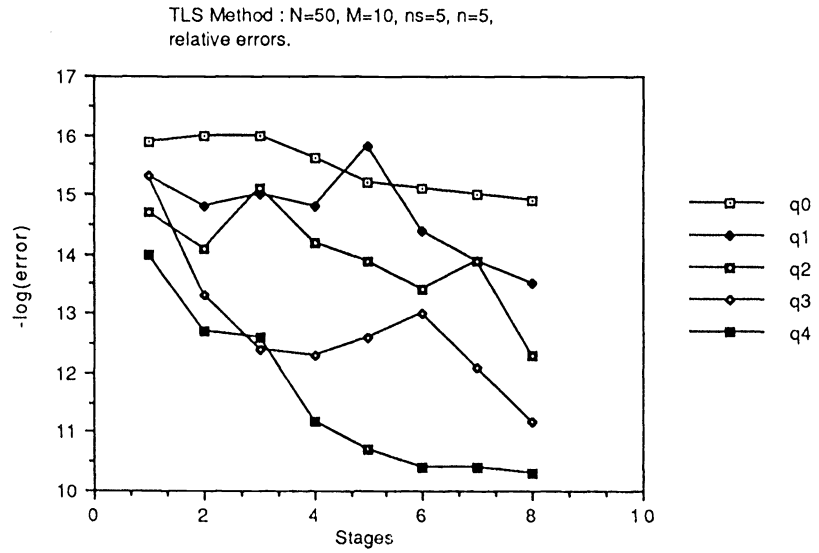


FIG. 5

The propagation of roundoff introduced by the downdating from one stage to the next is similar to the propagation of roundoff in shooting type methods. Thus in this context the test is quite severe.

Figure 3 shows the relative errors of $\beta_1, \beta_2, \cdots, \beta_{n-1}$ at each stage. From Figs. 3 and 2 we infer that the decay in the accuracy of the polynomial fits mirrors that of their representation.

Figure 4, displaying the case of the RHR method as in Fig. 1 but with $M = 20$, shows an insignificant decrease in accuracy for this larger point set. We suggest that this decrease is due to the different location of the points being added and discarded.

Turning now to the other methods, Fig. 5 shows $-\log |e|$ for the TLS method, and Table 1 lists the differences between $-\log |e|$ for these two methods. There is no observable pattern to the differences, the maxima of which occur at a few isolated places. Indeed we observed the same behaviour for the difference in accuracy between all the methods tested. The fact that the maxima are isolated and nowhere exceed 1.8 decimals leads us to not present the figures for the methods REV and TLD.

The numerical behaviour of the methods developed here and their relation to existing methods is not at all clear and warrants further investigation. The test described above reflects one sort of application that occurs in many real situations. In particular we refer to the use of polynomials limited to low degree. In our experience the updating methods not based on rotations and *all* downdating methods may suffer from instability if the degree of the fitted polynomial approaches the degree of the interpolating polynomial. This has also been observed elsewhere [4], [1].

In our other exploratory tests the TLS and RHR methods behave quite similarly. Applied to the updating problem with point distributions that come from some classical Gauss quadratures for finite intervals, the TLS and RHR methods agree to within about 0.5 decimals and lose fewer than about three decimal places accuracy for as many as 1,000 points. On the other hand, applied to updating with equally spaced points, both methods lose about three decimals after about 500 points. But the TLS method on this type of data, even though it is a little more accurate than the RHR, can suffer from extreme exponent growth (even though there is no associated loss of precision) and can fail unless extra care is taken. Of course the RHR method is immune from this problem.

In the case of downdating, it only takes 40 classical Gaussian quadrature points for the TLS and RHR methods to lose three digits of accuracy and both methods lose 16 digits of accuracy on 40 equally spaced points. In both these cases the top part of the resulting Jacobi matrix is still determined to almost full accuracy.

TABLE 1
RHR *and* TLS *methods.*

$M = 50, N = 50, n_s = 5, n = 5$

Stage	q_0	q_1	q_2	q_3	q_4
1	−0.2	−0.3	0.2	−0.6	−0.1
2	−0.5	0.1	0.1	0.7	−0.1
3	−0.6	−0.1	−1.3	0.8	−0.5
4	−0.3	−0.2	0.0	0.5	1.3
5	0.0	−1.1	0.5	0.2	1.0
6	−0.1	0.2	0.9	−0.5	0.7
7	0.0	0.4	−0.1	−0.3	0.9
8	0.1	0.3	0.5	0.1	−0.4

Differences between $-\log(|\mathbf{e}|)$

Thus there remain interesting questions about the effect on the accuracy of the process that the point distribution has and perhaps the order in which the points are added to the inner product.

REFERENCES

[1] S. T. ALEXANDER, C. T. PAN, AND R. J. PLEMMONS, *Analysis of a recursive least squares hyperbolic rotation algorithm for signal processing*, Linear Algebra Appl., 98 (1988), pp. 3–40.

[2] A. W. BOJANCZYK, R. P. BRENT, P. VAN DOOREN, AND F. R. DE HOOG, *A note on downdating the Cholesky factorization*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 210–221.

[3] D. L. BOLEY AND G. H. GOLUB, *A survey of matrix inverse eigenvalue problems*, in Inverse Problems, Physics Trust Publications, Vol. 3, Bristol, England, 1987, pp. 595–622.

[4] W. GAUTSCHI, *On generating orthogonal polynomials*, SIAM J. Sci. Statist. Comput., 3 (1982), pp. 289–317.

[5] P. E. GILL, G. H. GOLUB, W. MURRAY, AND M. A. SAUNDERS, *Methods for modifying matrix factorizations*, Math. Comp., 28 (1974), pp. 505–535.

[6] G. H. GOLUB AND M. A. SAUNDERS, *Linear least squares and quadratic programming*, in Integer and Nonlinear Programming, J. Abadie, ed., North-Holland, Amsterdam, 1970, pp. 229–256.

[7] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[8] G. H. GOLUB AND J. H. WELSCH, *Calculation of Gauss quadrature rules*, Math. Comp., 23 (1969), pp. 221–230.

[9] W. B. GRAGG AND W. J. HARROD, *The numerically stable reconstruction of Jacobi matrices from spectral data*, Numer. Math., 44 (1984), pp. 317–335.

[10] J. KAUTSKY AND G. H. GOLUB, *On the calculation of Jacobi matrices*, Linear Algebra Appl., 52/53 (1983), pp. 439–455.

[11] C. RADER AND A. O. STEINHARDT, *Hyperbolic Householder transformations*, IEEE Trans. Acoust. Speech Signal Process., 1986, pp. 1589–1602.

[12] H. RUTISHAUSER, *Further contributions to the solution of simultaneous linear equations and the determination of eigenvalues*, U.S. Department of Commerce, Applied Mathematics Series 49, National Bureau of Standards, Washington, DC, 1958.

[13] ———, *On Jacobi patterns*, in Experimental Arithmetic, High Speed Computing and Mathematics, Proc. Symposium on Applied Mathematics 15, American Mathematical Society, Providence, RI, 1963, pp. 219–239.

[14] G. W. STEWART, *The effects of rounding error on an algorithm for downdating a Choleski factorization*, J. Inst. Math. Appl., 23 (1979), pp. 203–213.

[15] H. S. WILF, *Mathematics for the Physical Sciences*, John Wiley, New York, 1962, Chap. 2.

[16] D. PARKER, private communication, 1982.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 2, pp. 354–361, April 1991

011

ON GROWTH IN GAUSSIAN ELIMINATION WITH COMPLETE PIVOTING*

NICK GOULD†

Abstract. It has been conjectured that when Gaussian elimination with complete pivoting is applied to a real n-by-n matrix, the maximum possible growth is n. In this note, a 13-by-13 matrix is given, for which the growth is 13.0205. The matrix was constructed by solving a large nonlinear programming problem. Growth larger than n has also been observed for matrices of orders 14, 15, and 16.

Key words. Gaussian elimination, growth, complete pivoting, nonlinear programming methods

AMS(MOS) subject classifications. 65F05, 65G05

1. Introduction. Let A be an n-by-n real matrix, let $A^{(1)} = A$, and let $A^{(k+1)}$, for $k = 1, \cdots, n-1$, be the $n-k$-by-$n-k$ matrix derived from A by elimination operations. That is, if we partition $A^{(k)}$ as

$$A^{(k)} = \begin{pmatrix} \alpha^{(k)} & a_c^{(k)T} \\ a_r^{(k)} & A_B^{(k)} \end{pmatrix}, \tag{1.1}$$

(where the scalar $\alpha^{(k)}$ is known as the *pivot* at the kth stage of the elimination), then

$$A^{(k+1)} = A_B^{(k)} - a_r^{(k)}[\alpha^{(k)}]^{-1} a_c^{(k)T}. \tag{1.2}$$

Alternatively, $A^{(k+1)}$ is the Schur complement of the first k-by-k block of A in the matrix A.

If Gaussian elimination, with complete or partial pivoting, is used to solve the system of linear equations $Ax = b$, Wilkinson [9] showed that the computed solution $\hat{x}$ satisfies the perturbed equations

$$(A+E)\hat{x} = b, \tag{1.3}$$

where the error matrix E satisfies the normwise bound

$$\|E\|_\infty \leqq up(n) g_n(A) \|A\|_\infty. \tag{1.4}$$

Here u is the unit roundoff, $p(n)$ is a cubic polynomial of n, and

$$g_n(A) = \max_{\substack{1 \leqq i,j \leqq n-k+1, \\ 1 \leqq k \leqq n}} |(PAQ)_{ij}^{(k)}| / \max_{1 \leqq i,j \leqq n} |(PAQ)_{ij}|, \tag{1.5}$$

where P and Q represent the pivoting permutations applied to A during the elimination. As the other contributions to the bound (1.4) are beyond our control, it is of interest to know precisely how large the growth factor $g_n(A)$ can be.

We say that A is a *complete elimination matrix* if, at each stage of the elimination, the modulus of each entry in $A^{(k)}$ is no larger than that of the pivot. Such matrices arise when complete pivoting is used to permute the rows and columns of a general matrix during Gaussian elimination (see Wilkinson [11]). The permutation matrices P and Q in (1.5) are both the identity matrix if A is a complete elimination matrix. Moreover,

* Received by the editors September 7, 1990; accepted for publication (in revised form) November 7, 1990.

† Central Computing Department, Rutherford Appleton Laboratory, Chilton, Oxfordshire, OXI1 0QX, England (nimg@ib.rl.ac.uk).

the growth factor is now the ratio of the moduli of the largest pivot to the first. For such matrices, Wilkinson [9] showed that

$$g_n = \sup_A g_n(A) < n^{1/2}(2 3^{1/2} 4^{1/3} \cdots n^{1/n-1})^{1/2} \tag{1.6}$$

and he noted that there were no known examples of matrices for which $g_n(A) > n$ (Wilkinson, [10, p. 97] and [11, p. 213]). Indeed, Cryer [3] hypothesized that $g_n \leqq n$ for all n with equality if and only if there is a Hadamard matrix, that is a matrix with entries ± 1 and orthogonal rows and columns, of dimension n. Higham and Higham [6] give a class of matrices for which $g_n(A) \geqq (n+1)/2$, while simulations by Trefethen and Schreiber [8] on random matrices indicate average growths on random matrices of approximately $n^{1/2}$.

One way of trying to generate large growth factors for complete elimination matrices is to attempt to solve the optimization problem of maximizing the modulus of the nth pivot. (It is always possible to arrange that the maximum growth occurs at this pivot. For, suppose the kth pivot is largest in magnitude. Then the matrix formed by replacing the last k-by-k block of the n-by-n identity matrix, scaled by a_{11}, with the first k-by-k block of A is also a complete elimination matrix with the same growth factor but with the maximum growth now occurring at the nth pivot.) This approach has been considered by Day and Peterson [4] and is also the approach taken in this note. Day and Peterson give lower bounds on the growth for $1 \leqq n \leqq 8$. Here we extend the range to $1 \leqq n \leqq 16$. The major result we obtain is that there are a number of 13-by-13 matrices for which g_{13} is larger than 13, and thus that Cryer's conjecture is false. Examples of growth larger than n have also been observed for matrices of order 14, 15, and 16.

In § 2, we describe the nonlinear programming approach we have taken to this problem. In § 3, the results of our numerical experiments are presented. We give an example where $g_{13} > 13$ in the Appendix.

2. Method. We may formulate the maximum pivot growth problem as a nonlinear optimization problem as follows:

Starting with an n-by-n real matrix $X^{(1)} = A$, we let $X^{(k)}$ be the matrix

$$X^{(k)} = \begin{pmatrix} 0 & 0 \\ 0 & A^{(k)} \end{pmatrix}, \tag{2.1}$$

where $A^{(k)}$ is the kth elimination matrix (1.1). Let $x_{i,j,k}$ be the (i,j)th entry of $X^{(k)}$. We thus wish to maximize $x_{n,n,n}$ subject to the restrictions that the matrices $X^{(k)}$ and $X^{(k+1)}$ are related to each other by elimination operations, that the largest element in $X^{(k)}$ occurs in position (k, k), and that the initial matrix $X^{(1)}$ is scaled so that the largest entry in magnitude is 1. This leads to the problem

$$\text{maximize } x_{n,n,n} \tag{2.2}$$

subject to the elimination constraints:

$$x_{i,j,k+1} - x_{i,j,k} + x_{i,k,k} x_{k,j,k} / x_{k,k,k} = 0, \quad \text{for } k < i, j \leqq n \text{ and } k = 1, \cdots, n-1; \tag{2.3}$$

constraints which make the signs of the pivots unique:

$$x_{k,k,k} \geqq 0 \quad \text{for } k = 1, \cdots, n; \tag{2.4}$$

a normalizing constraint, $x_{1,1,1} = 1$; and complete pivoting constraints:

$$-1 \leqq x_{i,j,1} \leqq 1 \quad \text{for } 1 \leqq i, j \leqq n \tag{2.5}$$

and

$$-x_{k,k,k} \leqq x_{i,j,k} \leqq x_{k,k,k} \quad \text{for} \quad k \leqq i,j \leqq n \quad \text{and} \quad k = 2, \cdots, n-1. \tag{2.6}$$

This formulation involves roughly $n^3/3$ variables, but is a very sparse optimization problem. We chose to solve this problem using our large-scale nonlinear programming package, LANCELOT (Conn, Gould, and Toint [2]), since the package is designed to handle such nonlinear sparsity as appears in problem (2.2)–(2.6) above.

By contrast, Day and Peterson [4] formulate the problem entirely in terms of the n^2 variables $X^{(1)}$, treating all of the remaining variables $X^{(k)}$, $k = 2, \cdots, n$ as implicit functions of $X^{(1)}$. This leads to a problem that is significantly more nonlinear and makes the calculation of analytic derivatives considerably harder. Nonetheless, Day and Peterson report considerable success with the nonlinear programming package NPSOL (Gill, Murray, Saunders, and Wright [5]).

Of course, neither nonlinear programming method is designed to find anything stronger than local solutions to a problem. The problem (2.2)–(2.6) has many local solutions and most of them are highly degenerate. The problem is thus challenging for a nonlinear programming algorithm and the values given in the next section are the result of many runs from different starting points in an attempt to find the global solution to the problem.

3. Results. In Table 3.1, we give the results obtained by running LANCELOT on the problem posed in § 2. LANCELOT is written in standard Fortran 77, compiled in double precision with the SUN Fortran 1.3 compiler; the problems were solved on a SUN SPARCstation 1. Each problem has many local solutions; we cannot, of course, guarantee that the values reported are the largest growths that can be obtained, merely that they are the largest values we encountered.

Of particular interest are the values obtained for $n = 13$, 14, 15, and 16, for here we see growth of more than n. We also observe that for $n = 16$, where a complete elimination Hadamard matrix exists and gives rise to growth of 16, other complete elimination matrices give larger growth. Thus Cryer's [3] conjecture is false.

TABLE 3.1
Maximum growth factors encountered.

n	Growth size	Comments
1	1.0	trivial
2	2.0	trivial
3	2.25	optimal (see Cohen, [1])
4	4.0	Hadamard matrix, optimal (see Cryer [3])
5	4.1325	agrees with Day and Peterson [4]
6	5.0	agrees with Day and Peterson [4]
7	6.0	agrees with Day and Peterson [4]
8	8.0	Hadamard matrix
9	8.4305	
10	9.5294	
11	10.4627	
12	12.0	Hadamard matrix
13	13.0205	
14	14.5949	
15	16.1078	
16	18.0596	not a Hadamard matrix

The matrices that give rise to the growth factors reported in Table 3.1 are often extremely sensitive to small perturbations in their entries in that tiny perturbations to a complete elimination matrix rarely results in another such matrix. This phenomenon was observed by Day and Peterson [4] and may explain why examples of large growth have proved elusive in previous attempts to find them. It also makes it rather difficult to specify matrices which give rise to large growth. Indeed, we had to solve the optimization problem of § 2 to very high accuracy, requiring the residuals of the nonlinear constraints (2.3) to be of the order of the unit roundoff. In some cases, this meant that we had to take the best solution that we obtained on the SUN as a starting point for a further run in extended precision on the CRAY X-MP/416 at Rutherford to reduce the residuals to the desired level. Even then, the mere fact of rounding the CRAY values to 16 decimal places frequently prevented the computed matrix from being a complete elimination matrix when the operations (1.2) were performed in double precision on the SUN. The values obtained had to be adjusted by eye to obtain a suitable floating-point complete elimination matrix.

We specify a 13-by-13 matrix that gives rise to growth of slightly more than 13.0205, in IEEE double precision arithmetic on a SUN SPARCstation 1, in the Appendix to this paper. The values must be read in Fortran 1P,D24.16 format. It is not known whether there are matrices with simple fractional entries that give rise to such large growth. Other 13-by-13 matrices that give rise to growth of larger than 13 were encountered.

The results of applying the elimination operations (1.2) to this matrix are given in Table 3.2. The size of the pivot and the largest nonpivot in absolute value at each stage of the elimination are shown. Note that the pivots are far from monotonic and that there is a "surge" of growth in the last few stages. Such a surge has been observed for Hadamard matrices by other authors [3], [4]. Indeed, values for the last six pivots for such matrices are known [4, Prop. 5.5]. Also observe how close the largest nonpivot at each stage is to the pivot and thus how tiny perturbations to the matrix elements may completely alter the pivot sequence.

4. Conclusions. We have shown that growth of larger than n is possible when Gaussian elimination with complete pivoting is performed on real n-by-n matrices by exhibiting a 13-by-13 matrix for which this is true. If A is an n-by-n complete elimination

TABLE 3.2
Details of the elimination.

Pivot	Pivot size	Largest modulus of nonpivot entry
1	1.0000000000000000	1.0000000000000000
2	2.0000000000000000	2.0000000000000000
3	2.0000000000000000	2.0000000000000000
4	2.5964300000000002	2.5964300000000002
5	2.3776999999999999	2.3776999543751263
6	2.3038700000000003	2.3038700000000003
7	2.9587400000000001	2.9587398634283884
8	3.5890399999999998	3.5890399999999998
9	4.1163800000000004	4.1163800000000004
10	3.3550400000000007	3.3550399999999998
11	6.5102699999999984	6.5102698773166514
12	6.5102700000000011	6.5102699999999567
13	13.0205000013724188	—

matrix with growth $g(A)$, and P is the matrix which permutes the first $2n$ integers to $\{1, n+1, 2, n+2, \cdots, n, 2n\}$, then (see [7] and [4, Prop. 5.12])

$$P^T \begin{pmatrix} A & A \\ A & -A \end{pmatrix} P \tag{4.1}$$

is also a complete elimination matrix with growth $2g(A)$. Thus there are an infinite number of matrices, of dimensions $13 \cdot 2^k$ for nonnegative k, which give rise to growth larger than their dimension. We suspect that there are examples of large growth for many other dimensions—we have encountered such examples for $n = 14$, 15, and 16—and that

$$\limsup_{n \to \infty} g_n / n$$

is unbounded. It is not known if there are matrices of dimension smaller than 13 for which growth larger than n is possible, nor is it known quite how close the growth factors given in this paper are to g_n.

We have observed that examples of large growth in complete elimination are very unstable in that very small perturbations to the matrix entries give rise to radically different pivot sequences. We suspect that this is why such examples have not been observed in practice. We also realize that the examples given here are extremely unlikely to—nor indeed should they—discourage people from using Gaussian elimination with pivoting. The potentially less stable partial and threshold pivoting strategies are used with impunity, and considerable success, throughout the scientific world.

Appendix. Here we give a 13-by-13 complete elimination matrix for which the growth is slightly over 13.0205 when the elimination operations (1.2) are performed in IEEE double precision arithmetic on a SUN SPARCstation 1. The values should be read in Fortran 1P,D24.16 format.

```
row  1

 1.0000000000000000D+00 -1.0000000000000000D+00 -1.0000000000000000D+00
 6.6084891857885364D-01  3.5076867724029653D-01  1.3913093634808771D-01
 1.0000000000000000D+00 -1.0000000000000000D+00  9.4546309508853699D-01
-6.4358761317393848D-02 -4.7259056539260776D-02  9.8144752878695718D-01
 1.0000000000000000D+00

row  2

 1.0000000000000000D+00  1.0000000000000000D+00 -1.0000000000000000D+00
-1.0000000000000000D+00 -8.8262544148845457D-01 -7.9349789219584022D-01
-1.0000000000000000D+00 -7.0049633754068708D-01  1.0000000000000000D+00
 1.0000000000000000D+00 -1.0000000000000000D+00  1.0000000000000000D+00
-6.5149858941930272D-01

row  3

 1.0000000000000000D+00  4.9321847997082674D-01  1.0000000000000000D+00
 5.2321986889464023D-01  1.0000000000000000D+00  9.3147802581501915D-01
-1.0000000000000000D+00 -1.0000000000000000D+00 -1.0000000000000000D+00
 9.0634017140409751D-01  1.0000000000000000D+00  1.9635994245021532D-01
 5.2020043801610605D-01
```

```
row  4

-8.5237723616654504D-01  1.0000000000000000D+00 -7.9959593728640932D-01
 1.0000000000000000D+00 -6.1395029873598805D-01 -1.0000000000000000D+00
-1.0000000000000000D+00  1.0000000000000000D+00  1.0000000000000000D+00
 1.0000000000000000D+00  1.0000000000000000D+00 -1.0000000000000000D+00
 1.0000000000000000D+00

row  5

-6.4197976615948327D-01  1.0000000000000000D+00 -8.2347773920951672D-01
-1.0000000000000000D+00  1.0000000000000000D+00 -1.0000000000000000D+00
-1.0000000000000000D+00  1.0000000000000000D+00 -1.0000000000000000D+00
-1.0000000000000000D+00 -9.8047514562210913D-01  1.0000000000000000D+00
 1.0000000000000000D+00

row  6

-7.5746114421052313D-01  8.7625388681860783D-01 -1.0000000000000000D+00
-1.0000000000000000D+00 -8.1410469390205387D-01  1.0000000000000000D+00
 1.0000000000000000D+00 -1.0000000000000000D+00 -1.0000000000000000D+00
 1.0000000000000000D+00 -1.0000000000000000D+00 -1.0000000000000000D+00
 1.0000000000000000D+00

row  7

 1.0000000000000000D+00  1.0000000000000000D+00  1.0000000000000000D+00
 5.8822529846976079D-01  1.0000000000000000D+00 -1.0000000000000000D+00
 1.1780693451504934D-01 -1.0000000000000000D+00  1.0000000000000000D+00
-1.0000000000000000D+00 -1.0000000000000000D+00 -1.0000000000000000D+00
 1.0000000000000000D+00

row  8

 1.0000000000000000D+00  1.0000000000000000D+00 -1.2365439895441106D-01
-1.0000000000000000D+00 -1.0000000000000000D+00  1.0000000000000000D+00
 1.0000000000000000D+00  1.0000000000000000D+00  1.0000000000000000D+00
-1.0000000000000000D+00  1.0000000000000000D+00  1.0000000000000000D+00
 1.0000000000000000D+00

row  9

-1.0000000000000000D+00  1.6728019890561854D-01 -1.0000000000000000D+00
-1.0000000000000000D+00  1.0000000000000000D+00  6.7037707945403946D-01
-1.0000000000000000D+00 -1.0000000000000000D+00  1.0000000000000000D+00
-1.0000000000000000D+00  1.0000000000000000D+00 -1.0000000000000000D+00
-1.0000000000000000D+00
```

```
row 10

-1.0000000000000000D+00  1.0000000000000000D+00  7.3451234413636224D-01
 7.7420992278979484D-01  1.0000000000000000D+00  1.0000000000000000D+00
 1.0000000000000000D+00  1.0000000000000000D+00  1.0000000000000000D+00
 1.0000000000000000D+00 -1.0000000000000000D+00  1.0000000000000000D+00
-1.0000000000000000D+00

row 11

-1.0000000000000000D+00 -3.2294803009723511D-01  1.0000000000000000D+00
-1.0000000000000000D+00  5.9471427088948606D-02 -1.0000000000000000D+00
 1.0000000000000000D+00 -7.7305121515367092D-01  1.0000000000000000D+00
 1.0000000000000000D+00  1.0000000000000000D+00  1.0000000000000000D+00
 1.0000000000000000D+00

row 12

-1.0000000000000000D+00 -1.7007857952327707D-01  1.0000000000000000D+00
 1.0000000000000000D+00 -1.0000000000000000D+00  1.0000000000000000D+00
-1.0000000000000000D+00 -1.0000000000000000D+00  9.1898031012251935D-01
-1.0000000000000000D+00 -1.0000000000000000D+00  2.5049340232649964D-01
 1.0000000000000000D+00

row 13

 9.6143110935926346D-01 -1.0000000000000000D+00  7.2409299018425932D-01
-1.0000000000000000D+00  1.0000000000000000D+00  1.0000000000000000D+00
-1.0000000000000000D+00  1.0000000000000000D+00  1.0000000000000000D+00
 1.0000000000000000D+00 -1.0000000000000000D+00 -1.0000000000000000D+00
 1.0000000000000000D+00
```

Acknowledgment. The author would like to thank two other Nicks (Higham and Trefethen) for stimulating discussions on this topic over the years and Iain Duff for his helpful comments on this paper. He is also grateful to Philippe Toint and Andy Conn for their interest in the problem and collaboration on the LANCELOT project. Finally, thanks are due to the associate editor and referees for the prompt and efficient processing of the paper.

REFERENCES

[1] A. M. COHEN, *A note on the pivot size in Gaussian elimination*, Linear Algebra Appl., 8 (1974), pp. 361–368.

[2] A. R. CONN, N. I. M. GOULD, AND PH.L. TOINT, *An introduction to the structure of large scale nonlinear optimization problems and the* LANCELOT *project*, FUNDP Report 89/19, Facultés Universitaires Notre-Dame de la Paix, Namur, Belgium, 1989; Also in Proc. 9th Internat. Conference on Computing Methods in Applied Sciences and Engineering, Paris, 1990.

[3] C. W. CRYER, *Pivot growth in Gaussian elimination*, Numer. Math., 12 (1968), pp. 335–345.

[4] J. DAY AND B. PETERSON, *Growth in Gaussian elimination*, Amer. Math. Monthly, June, 1988, pp. 489–513.

[5] P. E. GILL, W. MURRAY, M. A. SAUNDERS, AND M. H. WRIGHT, *User's guide for* SOL/NPSOL: *A Fortran package for nonlinear programming*, Tech. Report SOL 83-12, Systems Optimization Laboratory, Stanford University, Stanford, CA, 1983.

[6] N. J. HIGHAM AND D. J. HIGHAM, *Large growth factors in Gaussian elimination with pivoting*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 155–164.

[7] L. TORNHEIM, *Maximum pivot size in Gaussian elimination with complete pivoting*, Tech. Report, Chevron Research Company, Richmond, CA, 1970.

[8] L. N. TREFETHEN AND R. S. SCHREIBER, *Average-case stability of Gaussian elimination*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 335–360.

[9] J. H. WILKINSON, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8 (1961), pp. 281–330.

[10] ———, *Rounding errors in algebraic processes*, Notes on Applied Science No. 32, 1963, Her Majesty's Stationery Office, London.

[11] ———, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 2, pp. 362–373, April 1991

012

REDUCTION OF A GENERAL MATRIX TO TRIDIAGONAL FORM*

GEORGE A. GEIST†

Abstract. An algorithm for reducing a nonsymmetric matrix to tridiagonal form as a first step toward finding its eigenvalues is described. The algorithm uses a variation of threshold pivoting, where at each step, the pivot is chosen to minimize the maximum entry in the transformation matrix that reduces the next column and row of the matrix. Situations are given where the tridiagonalization process breaks down, and two recovery methods are presented for these situations. Although no existing tridiagonalization algorithm is guaranteed to succeed, this algorithm is found to be very robust and fast in practice. A gradual loss of similarity is also observed as the order of the matrix increases.

Key words. tridiagonalization, nonsymmetric, eigenvalues

AMS(MOS) subject classification. 15

1. Introduction.

The standard method for computing all of the eigenvalues of a dense matrix is based on the QR iteration scheme [5]. In this scheme, orthogonal similarity transformations are successively applied to the matrix to reduce it to quasi-triangular form, so that the eigenvalues appear on the diagonal. Repeated application of these transformations to a general matrix is prohibitively expensive, however, so that in practice the original matrix is first reduced to a simpler form that can be preserved during the subsequent iterative phase. For a general matrix, the initial reduction is usually to upper Hessenberg form (upper triangular except for one additional subdiagonal) by elementary or orthogonal similarity transformations. The initial reduction to Hessenberg form requires $O(n^3)$ operations, where n is the order of the matrix. Computation of the eigenvalues of the reduced matrix usually requires only a few QR iterations per eigenvalue, totaling another $O(n^3)$ operations. Both the initial and iterative phases are costly, but less costly than iterating directly with the original matrix. This two-phase approach is implemented in the standard EISPACK software for the general eigenvalue problem [16].

If the original matrix is symmetric, then that symmetry can be preserved by using orthogonal transformations in the initial reduction, so that the result is in fact tridiagonal. Although the reduction to tridiagonal form costs $O(n^3)$ operations, the subsequent iterations preserve the tridiagonal form and are much less expensive, so that the total cost of the iterative phase is reduced to $O(n^2)$ operations. Again, standard software is available in EISPACK implementing this two-phase approach for the symmetric case [16].

The attractively low operation count of iterating with a tridiagonal matrix suggests that the tridiagonal form would be extremely beneficial in the nonsymmetric case as well. There are two difficulties with such an approach: First, QR iteration does not preserve the structure of a nonsymmetric tridiagonal matrix. This problem can be overcome by using LR iteration [15] instead, which preserves the tridiagonal form. Second, it is difficult to reduce a nonsymmetric matrix to tridiagonal form by similarity transformations in a numerically stable manner. This second problem is the primary focus of this paper.

* Received by the editors April 17, 1989; accepted for publication (in revised form) May 3, 1990. This research was supported by the Applied Mathematical Sciences Research Program, Office of Energy Research, U.S. Department of Energy, under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

† Mathematical Sciences Section, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, Tennessee 37831-8083.

The following notational conventions will be used throughout this paper. Lower case Greek letters will denote scalars; lower case Latin letters will denote vectors. Components of vectors are denoted by subscripts. Upper case Latin letters will denote square matrices and a single subscript, when present, denotes the matrix dimension. Throughout this paper, N is used to represent a matrix that applies a rank one change to another matrix. Special cases of N include N_c and N_r, which zero out the next column and row of a matrix, respectively.

In the early 1960's there was a great amount of interest and research devoted to finding a stable way to reduce a general matrix via similarity transformations to tridiagonal form [12], [14], [17]. The problem is addressed in some detail by Wilkinson [21], and several algorithms are given, but the overall conclusion was that no general purpose algorithm existed. Because of the success and numerical stability of the QR iteration scheme, little research was directed at the problem of reduction to tridiagonal form for nearly 15 years.

One reason for renewed interest in tridiagonalization is the relatively poor performance of the QR iteration on advanced computers. Algorithms for vector supercomputers [10] and parallel architectures [7] have been developed for reducing nonsymmetric matrices to tridiagonal form.

In 1981 Dax and Kaniel published a paper [2] that inspired most of the recent interest in the problem. They describe experiments with reduction from upper Hessenberg form to tridiagonal form using elementary similarity transformations. During the reduction, they monitor the size of the multipliers as follows. They define a *control parameter* for the reduction of row k as $m_k = \max_{i>k+1}\{|H_{k,i}/H_{k,k+1}|\}$. If m_k is greater than a specified value, μ, then *breakdown* is said to have occurred, and their algorithm aborts. They observe that for 100 random test matrices of order 50×50 the number of breakdowns as a function of the specified value μ is:

$\mu = 2^r$	r =	16	12	10	8	7
breakdowns		0	1	5	20	41

Dax and Kaniel refer to Wilkinson's detailed error analysis in [21] and conclude that with judicious use of double precision there is a low probability of having large errors in eigenvalues computed with the tridiagonal matrix, even when using control parameters as large as 2^{16}.

Wachspress [18] and Watkins [19] focus on the fact that in [2] Dax and Kaniel did not address possible ways to recover from breakdown during the reduction to tridiagonal form. Wilkinson states [21, p. 404] that

> If breakdown occurs in the rth step of the reduction of a Hessenberg matrix to tridiagonal form we must return to the beginning and compute NAN^{-1} for some N in the hope that failure will be avoided in this matrix.

This recovery method is actually too restrictive. Wachspress and Watkins both describe efficient methods for finding matrices similar to A without returning to the beginning and wasting work already performed on the matrix. Hare and Tang [11] describe a combination of recovery methods and also investigate the effects of interleaving orthogonal and elementary similarity transformations during the tridiagonalization to reduce the number of multipliers that are greater than one.

In the next section we describe the inherent problems of tridiagonalizing nonsymmetric matrices. In §3 we present a reduction algorithm that incorporates a pivoting scheme designed to produce better conditioned transformation matrices than previous algorithms. We describe two recovery algorithms in §4 that significantly improve the

robustness of the reduction algorithm. Section 5 presents empirical results showing the accuracy and performance of the new algorithm. While no finite stable tridiagonalization algorithm is known [13], the new algorithm significantly broadens the class of matrices that can be successfully reduced.

2. Tridiagonalization. The direct reduction of a general matrix to tridiagonal form is difficult because the elementary similarity transformations, which must be used at some point in the reduction, may have large multipliers. This phenomenon is illustrated by the following example. First note that computations of the form

$$\begin{pmatrix} I_{k-1} & & \\ & 1 & \\ & & G_{n-k} \end{pmatrix} \begin{pmatrix} F_{k-1} & & \\ & \alpha & w^T \\ & v & B_{n-k} \end{pmatrix} \begin{pmatrix} I_{k-1} & & \\ & 1 & \\ & & G^{-1} \end{pmatrix} = \begin{pmatrix} F_{k-1} & & \\ & \alpha & w^T G^{-1} \\ & Gv & GBG^{-1} \end{pmatrix}$$

preserve the inner product of the kth row and column, since $w^T G^{-1} G v = w^T v$. The tridiagonalization algorithms in [2], [6], [10], [11], [12], [17], [19] are all affected by this property.

Let the partially reduced matrix have the form shown in Fig. 1. Let $w^T v = 0$

$$\left(\begin{array}{cc|cc} T_{k-1} & & \times & \\ & & & \\ \hline & \times & \alpha & w^T \\ & & v & B_{n-k} \end{array}\right)$$

FIG. 1. *Partially reduced matrix.*

and $\bar{v} = Gv$, where G is designed to eliminate all but the first element of v. Let $\bar{w}^T = w^T G^{-1}$ and partition $\bar{w}^T = (\bar{w}_1 \tilde{w}^T)$. Since $w^T v = \bar{v}_1 \bar{w}_1$, $\bar{w}_1 = 0$. After all but the first entry of v have been eliminated, the matrix has the form

$$\left(\begin{array}{cc|ccc} T_{k-1} & & \times & & \\ & & & & \\ \hline & \times & \alpha & 0 & \tilde{w}^T \\ & & \bar{v}_1 & & \\ & & 0 & & \bar{B}_{n-k} \end{array}\right).$$

Any attempt to avoid the use of the zero as the pivot now destroys the existing tridiagonal form. This zero pivot will occur regardless of the pivot selection in v or whether orthogonal transformations are used to eliminate v.

Algorithms that include a stable reduction to upper Hessenberg form as an initial step to tridiagonal form will likely encounter small pivots during the reduction of the rows. Stable reduction of the columns tend to make $\bar{v}_1$ large. For example, stable elementary transformations choose $\bar{v}_1 = \max(v_i)$, and orthogonal transformations make $\bar{v}_1 = ||v||_2$. Let $\bar{w}_1$ be the first entry in $w^T G^{-1}$. Since the product of $\bar{v}_1$ and $\bar{w}_1$, the eventual pivot for the row, is fixed, $\bar{w}_1$ tends to be small, which can lead to breakdown when reducing the rows.

If $w^T v = 0$, then a breakdown condition will occur no matter what transformation is used. In this case, the algorithm must abort or apply some recovery method.

3. A tridiagonalization algorithm. In this section we present an algorithm that reduces the matrix directly to tridiagonal form by eliminating columns and rows

using elementary similarity transformations so that the matrix always has the form shown in Fig. 1. This matrix structure allows us the freedom to pivot at each step to improve the overall stability of the algorithm. For example, the pivot could be chosen to minimize the maximum multiplier in the column and row reduction, or the pivot could be chosen to minimize the condition number of the transformation matrices. While these pivoting heuristics work well, the heuristic we found that works at least as well and sometimes better is to choose the pivot that minimizes the norm of the transformation matrix that reduces both the column and row. If C denotes this transformation matrix, then the norm used is $n\{\max|C_{ij}| : i, j = 1, 2, \cdots, n\}$ because it can be computed in constant time for each possible permutation.

```
A2TRI( a, n, tol )

    maxtol = tol
    cnt = 0
    m = 1
    for k = 1 to n − 2
label:
        Check number of recovery attempts
        if(cnt > 2) then
            maxtol = 10 * maxtol, print warning of increase.
            cnt = 0
            if(maxtol > 10 * tol) return and execute NEWSTART
        end if
        Find suitable pivot
        PIVOT(a, n, k, piv, maxmult, err)
        Check for deflation
        if( err = 1 ) m = k + 1, next k
        Interchange row(piv) and row(k)
        Interchange column(piv) and column(k)
        Check maximum multiplier against tolerance
        if( err = 2 or maxmult > maxtol ) then
            FIXUP( a, k, m, n )
            cnt = cnt + 1, print warning
            go to label:
        endif
        Zero out column k
        for i = k + 2 to n
            for j = k + 1 to n
                a_ij = a_ij − a_k+1j * a_ik/a_k+1k
            for j = k to n
                a_jk+1 = a_jk+1 + a_ij * a_ik/a_k+1k
        Zero out row k
        for i = k + 2 to n
            for j = k + 1 to n
                a_k+1j = a_k+1j − a_ij * a_ki/a_kk+1
            for j = k + 1 to n
                a_ji = a_ji + a_jk+1 * a_ki/a_kk+1
    end for
```

FIG. 2. *Algorithm for reducing an $n \times n$ matrix A to tridiagonal form while trying to bound all multipliers below tol.*

At step k of the algorithm shown in Fig. 2, the matrix has the form shown in Fig. 1. If v or $w = 0$, then the matrix has been deflated, and step k can be skipped.

Otherwise the algorithm finds the permutation that minimizes the maximum element in $N_r^{-1}N_c$, where N_r and N_c are elementary matrices such that $N_cAN_c^{-1}$ reduces column k and $N_r^{-1}(N_cAN_c^{-1})N_r$ reduces row k.

This minimization can be done efficiently because of the special structure of $N_r^{-1}N_c$, which is

$$\left(\begin{array}{c|cc} I_k & & \\ \hline & \gamma & u^T \\ & x & I_{n-k-1} \end{array}\right).$$

The vector x contains the multipliers used in reducing column k, and u contains the negatives of the multipliers used in reducing row k. The pivoting algorithm shown in Fig. 3 finds the permutation at step k that minimizes the maximum multiplier used in the column and row reduction and γ. The term γ equals $1 - u^Tx$, which can be simplified to w_1v_1/w^Tv.

```
PIVOT(a, n, k, piv, maxmult, err )

    err = 0
    maxmult = ∞
    Find maximum and next-to-maximum entries in row k and column k
    maxcol = max( |a_ik| |i = k + 1 to n)
    pivc = index of maxcol
    nmxcol = next-to-max( |a_ik| |i = k + 1 to n)
    maxrow = max( |a_ki| |i = k + 1 to n)
    pivr = index of maxrow
    nmxrow = next-to-max( |a_ki| |i = k + 1 to n)
    inprod = Σ_{i=k+1}^{n} a_ki * a_ik

    Check if maximum element in row or column is zero
    if( maxcol = 0 or maxrow = 0 ) err = 1, return
    Check if inner product is zero
    if( inprod = 0 ) then
        piv = index of max(maxcol, maxrow)
        err = 2
        return
    endif
    Calculate maximum entry of (N_r N_c)_i over all permutations i
    for i = k + 1 to n
        if( i = pivc ) maxnc = |nmxcol/a_ik|
        else maxnc = |maxcol/a_ik|
        if( i = pivr ) maxnr = |a_ik * nmxrow/inprod|
        else maxnr = |a_ik * maxrow/inprod|
        maxdiag = |a_ik * a_ki/inprod|
        temp = max(maxnr, maxnc, maxdiag)
        if( temp < maxmult ) then
            maxmult = temp
            piv = i
        endif
    end for
```

FIG. 3. *Algorithm for finding the pivot that minimizes the maximum element in* $N_r^{-1}N_c$, *where* $N_r^{-1}N_cAN_c^{-1}N_r$ *reduces column* k *and then row* k *of the* $n \times n$ *matrix* A.

If $w^Tv = 0$, then the minimization problem has no solution. In this case

$\max(|v_i|, |w_i|)$ is permuted into the pivot location before calling the recovery routine, FIXUP (see Fig. 4). The recovery routine is also called when the maximum

```
FIXUP( a, k, m, n )

    Apply a random shift
    r = random()
    a_mm = a_mm + r * a_m+1m
    a_mm+1 = a_mm+1 + r * (a_m+1m+1 - a_mm)
    a_mm+2 = r * a_m+1m+1

    Chase bulge down to row k - 1
    for i = m + 1 to k - 1
        m = a_i-1i+1 / a_i-1i
        a_i-1i+1 = 0
        a_ii = a_ii + m * a_i+1i
        a_ii+1 = a_ii+1 + m * (a_i+1i+1 - a_ii)
        a_ii+2 = m * a_i+1i+2
        a_i+1i+1 = a_i+1i+1 - m * a_i+1i
    end for
    Fill in row k - 1
    if (k = m + 1) m = r
    for i = k + 2 to n - 1
        a_k-1i = m * a_ki
    end for
    Eliminate row k - 1
    for i = k + 1 to n - 1
        m = a_k-1i / a_k-1k
        a_k-1i = 0
        for j = k to n - 1
            a_kj = a_kj + m * a_ij
        for j = k to n - 1
            a_ji = a_ji - m * a_jk
    end for
```

FIG. 4. *Recovery algorithm to apply an implicit single-shift* LR *iteration to rows m through k of the partially reduced matrix.*

element in $N_r^{-1}N_c$ exceeds a bound set by the user. If the maximum element is less than the bound, then the algorithm simply reduces column k followed by row k.

CLAIM. *The minimization problem can be solved in* $O(n-k)$ *time by observing that for a given permutation, the maximum multipliers in column k and row k, respectively, are:*

$$m_c = \frac{\max_{i>1}|v_i|}{|v_1|}$$

$$m_r = \frac{|v_1|\max_{i>1}|w_i|}{|w^T v|}.$$

Proof. Using Fig. 1 as a reference, given that column k is reduced first by an elementary similarity transformation $N_c A N_c^{-1}$, the expression for m_c is obvious. The form of N_c is

$$\left(\begin{array}{c|c} I_k & \\ \hline & G_{n-k} \end{array}\right),$$

and after the transformation is applied, $\bar{v} = Gv$ and $\bar{w}^T = w^T G^{-1}$. Thus, $\bar{w}^T \bar{v} = w^T G^{-1} G v = w^T v$. Since $\bar{v}_i = 0$ for $i > 1$, $\bar{w}_1 \bar{v}_1 = w^T v$. Since N_c is elementary, $\bar{v}_1 = v_1$ so $v_1 \bar{w}_1 = w^T v$ or $\bar{w}_1 = w^T v / v_1$. Therefore,

$$m_r = \frac{\max|w_i|}{|\bar{w}_1|} = \frac{|v_1| \max|w_i|}{|w^T v|}; \qquad i > 1.$$

For each possible choice of permutation only three terms must be evaluated: m_c, m_r, and γ. At step k there are only $n-k-1$ possible permutations. Thus the permutation that minimizes the maximum element can be found in $O(n-k)$ time, which totals $O(n^2)$ for the entire reduction.

The complexity of the overall tridiagonalization algorithm given in Fig. 2 is $(4/3)n^3 + O(n^2)$ flops, where a flop is defined as a floating point operation of the form $a + b*c$. This complexity is based on the assumption that the recovery routines, which we discuss in the next section, are called only a constant number of times.

An error analysis of tridiagonalization methods is given in [21]. Dax and Kaniel [2] also give an error analysis that shows that the bound on the eigenvalue errors depends on the spectral condition number of the tridiagonal matrix. The potential for $w^T v$ to vanish means that no finite tridiagonalization algorithm can be guaranteed to succeed. Even if the multipliers are all bounded below some modest value, say 10, there is the potential for catastrophic roundoff error.

On the other hand, this large growth has not been observed in practice using our algorithm. Instead, a gradual loss of similarity is observed as the matrix size increases. This degradation is conjectured to be caused by accumulated roundoff from using multipliers larger than one. Research continues into bounding the expected growth, and a future report will describe the results.

4. Recovery methods. In this section we describe the two recovery algorithms used in conjunction with the threshold pivoting algorithm. In most tridiagonalization schemes breakdown is defined as the situation where a multiplier (in our case an element in $N_r^{-1} N_c$) has exceeded some tolerance. When breakdown occurs, a number of options are available to circumvent the problem. Sometimes a local transformation can decrease the size of the multiplier so that the reduction can continue [11], but local methods cannot be robust because the tridiagonal form MAM^{-1} is unique once the first column and row of the transformation matrix M are fixed [13]. Thus, if this unique form has a small pivot, breakdown cannot be avoided without changing the first row or column. In [21], Wilkinson states that if a breakdown occurs, one can go back to the beginning and apply the transformation NAN^{-1} in the hope that breakdown will not occur again. No method of choosing N has been found that guarantees that the breakdown condition found in A will not exist in NAN^{-1}. For this reason, all proposed recovery methods choose another N and repeat the process if the previous choice of N fails to eliminate the breakdown condition.

The two recovery methods we propose differ in their choice of N, the amount of work they perform, and the matrix to which they are applied. In the first method, which is a variant of recovery methods proposed by Wachspress [8] and Watkins [19], a single random implicit single-shift LR iteration is applied to the matrix from the point of the last deflation down to row k. Since the partially reduced matrix is tridiagonal down to row k, one can start the iteration with either of the following forms of N:

$$\left(\begin{array}{cc|c} 1 & r & \\ 0 & 1 & O \\ \hline & O & I_{n-2} \end{array}\right) \qquad \left(\begin{array}{cc|c} 1 & 0 & \\ r & 1 & O \\ \hline & O & I_{n-2} \end{array}\right).$$

Our first recovery method applies these two starting matrices alternately and uses a random value uniformly distributed on [0.1,1] for r. Figure 4 shows the FIXUP algorithm, which uses the left starting matrix above. Assuming no deflations have occurred, the first operations of the FIXUP algorithm introduce a nonzero in the a_{13} position. This "bulge" is then chased down the matrix with elementary similarity transformations to the point of the previous breakdown. Given that the breakdown occurs at row k, this chasing procedure fills in row $k-1$, which then must be annihilated to return the matrix to its prerecovery structure.

If breakdown occurs during the recovery or if the original breakdown condition persists after the recovery step, the recovery method is repeated with the alternate form of N. After three consecutive unsuccessful recovery attempts, the multiplier tolerance is temporarily increased by a factor of 10. After three additional unsuccessful attempts at this higher tolerance, the recovery attempts on the partially reduced matrix stop, and our second recovery method, NEWSTART, is initiated.

A small number of consecutive failures of the first recovery method is usually indicative of a matrix with a large number of small inner products. When this occurs, a random orthogonal matrix Q is applied to the original matrix, and the reduction is restarted with the modified matrix QAQ^T. The purpose of this operation is to reduce the probability of small inner products occurring in the modified matrix. The algorithm is simple and efficient to apply, requiring only $O(n^2)$ flops to execute, because Q is chosen to be a Householder transformation $Q = (I - 2ww^T)$. This routine, which we call NEWSTART, is initiated only as a last resort because it requires restarting the reduction.

5. Results. We report on empirical studies of three aspects of the new algorithm: its speed, robustness, and accuracy. All of the studies are based on finding the eigenvalues of nonsymmetric matrices, which is the primary use of the tridiagonalization algorithm. All computations were performed in double precision on a Sun 3/280.

To perform these studies we developed an algorithm, which we will refer to as TLR, for finding the eigenvalues of nonsymmetric tridiagonal matrices. TLR initially applies a diagonal similarity transformation to the tridiagonal matrix to scale the superdiagonal to contain all ones. Wilkinson [21] suggests this transformation because the superdiagonal is invariant under implicit LR iterations. Thus, the transformation saves space and floating point operations. In fact, the storage and flops per iteration are the same as for the symmetric tridiagonal case when using LR iteration. TLR applies implicit double shift LR iterations to the scaled tridiagonal matrix until all the eigenvalues are found. If the LR iteration breaks down due to encountering a small pivot element (which can occur because pivoting is not performed) or it fails to converge to an eigenvalue after 30 iterations, then an arbitrary shift is applied to the matrix.

The potential dangers of using LR iteration are well documented [21], although some recent research [20] has attempted to put the algorithm on firmer theoretical ground. We chose LR iteration because it preserves nonsymmetric tridiagonal form. Our experience with TLR has been positive, as it has never failed to converge. On the other hand, we have seen tridiagonal matrices where the eigenvalues computed with TLR are not as accurate as the results from the standard EISPACK routines. For the interested reader, Dax and Kaniel [2] present more elaborate methods to improve the stability of the LR iteration and to refine the eigenvalues of the tridiagonal matrix iteratively.

A second alternative, used in [10], is to transform the real nonsymmetric tridiagonal matrix into a complex symmetric tridiagonal matrix. This is done by scaling the ith subdiagonal and superdiagonal entries to $\sqrt{b_i c_i}$, where b_i and c_i are opposing subdiagonal and superdiagonal entries, respectively. Then a complex arithmetic version of the QL iteration can be applied to finding all the eigenvalues. Details of the algorithm can be found in [1] along with a discussion of complex symmetric tridiagonal matrices and potential problems with finding their eigenvalues.

Table 1 compares the execution times in seconds of our algorithm with the EISPACK routines ELMHES, ORTHES, and HQR, for a series of test matrices ranging in size from 50 to 300. The matrices were random with entries distributed uniformly over the interval $[-1, 1]$. ELMHES reduces A to Hessenberg form using stabilized elementary similarity transformations while ORTHES reduces A to Hessenberg form H using Householder transformations. HQR finds the eigenvalues of H using an implicit double shift QR iteration. Our algorithms are presented in the table as A2TRI and TLR. A2TRI reduces A directly to tridiagonal form T as described in §3. TLR finds the eigenvalues of T. The time for either ELMHES or ORTHES should be added to the time for HQR and compared with the sum of the times for A2TRI and TLR.

TABLE 1
Execution times in seconds on a Sun 3/280 for our new routines and the standard EISPACK *routines.*

	EISPACK			NEW	
n	ELMHES	ORTHES	HQR	A2TRI	TLR
50	2.70	4.66	12.92	3.80	0.70
100	21.58	37.60	92.08	32.90	2.70
128	44.92	80.06	208.62	65.01	4.46
150	72.28	136.88	334.56	110.62	6.72
200	173.20	314.78	667.78	240.02	10.86
250	338.54	631.42	1388.36	509.94	16.94
300	582.64	1136.34	2305.14	893.48	23.84

It is clear from the table that our method can find the eigenvalues of a dense nonsymmetric matrix much faster than the EISPACK routines. A complexity analysis, where low order terms are ignored, shows that TLR requires $5n$ flops per iteration versus $4n^2$ flops for HQR. While the number of iterations varies between TLR and HQR, they both require only a few, usually fewer than 5, iterations per eigenvalue. Further, the arithmetic complexities of ELMHES and ORTHES are $(5/6)n^3$ and $(5/3)n^3$, respectively, while the complexity of A2TRI is $(4/3)n^3$ flops, assuming A2TRI needs to apply FIXUP only a constant number of times. The results in Table 1 reflect speedups greater than three for A2TRI/TLR over ELMHES/HQR, which are consistent with the relative complexities of the routines.

Random matrices are not necessarily good choices for testing the robustness of the tridiagonalization. Therefore, we input most of the nonsymmetric eigenvalue test matrices contained in the book by Gregory and Karney [9] as well as the EISPACK test suite of real general matrices into our algorithm. The test set included ill-conditioned, defective, and derogatory matrices in sizes up to 20×20. All were reduced successfully, although several required calls to the routine FIXUP. One matrix, Wilkinson's notoriously ill-conditioned matrix, required a call to the routine NEWSTART before it could be reduced. The eigenvalues calculated from the test matrices in Gregory and Karney were accurate to the expected number of digits given the condition numbers of the problems except for Wilkinson's matrix where only three digits of accuracy were obtained, instead of the expected eight digits. Table 2 gives the number of fixups

executed by A2TRI and the largest relative error in any eigenvalue for each problem in the EISPACK test suite. Whenever the error is greater than 10^{-13}, the condition number of the corresponding eigenvalue is also given. In four cases, no error is made because A2TRI is able to permute the matrix into triangular form. In the two cases in which the relative error is greater than 1, the tridiagonal matrix returned by A2TRI is similar to the original matrix and all the error occurs in TLR. (This is not true in general.) There are also four cases where a significant amount of work is avoided because the problem deflates during the reduction to tridiagonal form. The apparent bad behavior in problem 2 is deceptive, because this problem has eigenvalues (λ_i) spread over six orders of magnitude. Let $\bar{\lambda}_i$ be the exact eigenvalues of A. Wilkinson [21] gives a bound on the absolute error of the eigenvalues as

$$|\lambda_i - \bar{\lambda}_i| \leq \frac{\| A \| \epsilon}{s_i},$$

where ϵ is machine precision and s_i is the inner product of the normalized left and right eigenvectors of λ_i. For problem 2 $\| A \| \approx 10^{10}$ and $s_i \approx 1$. Assuming $\epsilon = 10^{-16}$, the absolute error for the λ_i in problem 2 should be better than 10^{-6} and in fact we see an absolute error of 10^{-8}.

To determine the relative accuracy of the new algorithms, two comparisons were performed on a range of problem sizes. In the first comparison, the eigenvalues of the tridiagonal matrix were computed with HQR and compared with the corresponding eigenvalues of the original matrix as computed with ORTHES/HQR. This measures the loss of similarity caused by the reduction to tridiagonal form. The second comparison was between the eigenvalues computed by A2TRI/TLR and the eigenvalues computed by ORTHES/HQR. Figure 5 presents the accuracy seen during these comparisons.

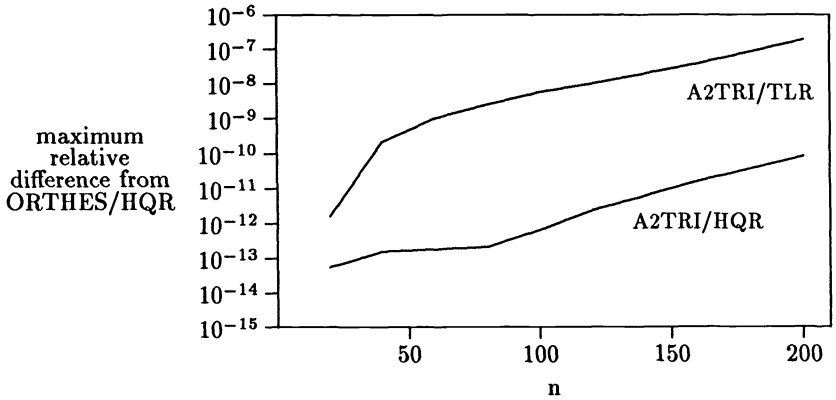


FIG. 5. *Typical degradation of eigenvalue accuracy seen with* A2TRI *and* TLR *for random orthogonal matrices.*

The graph clearly shows that the accuracy decreases as the matrix size increases. Similar results have been observed with other tridiagonalization methods [2], [11]. All calculations were performed in double precision. To avoid variations due to ill-conditioned eigenvalues, random orthogonal matrices were used in these tests. The overall error of the new algorithms on random nonorthogonal matrices was comparable but showed a larger variation between problems.

The main advantage of a faster algorithm is the ability to solve larger problems, but the results of our study indicate that the accuracy of the larger problems may be

TABLE 2

Number of fixups, maximum relative error in any eigenvalue, and comments ($1/s_i$ *is the condition number of the corresponding eigenvalue*) *for* A2TRI/TLR *on the* EISPACK *test suite.*

EISPACK Test Suite of Real General Matrices			
Problem number	Number of fixups	Relative error in λ_i	Comments ($1/s_i$ =condition number of λ_i)
1	0	10^{-13}	
2	0	10^{-11}	$1/s_i = 1$
3	3	10^{-6}	$1/s_i = 10^7$
4	0	10^{-15}	
5	0	10^{-15}	
6	0	10^{-13}	
7	9	10^{-9}	$1/s_i = 10^{15}$
8	0	0	zero matrix
9	0	10^{-6}	$1/s_i = 10^{16}$
10	1	10^{-14}	
11	0	10^{-15}	
12	0	2.5	$1/s_i = 10^{15}$, Tridiagonal OK
13	2	10^{-14}	
14	2	10^{-15}	
15	0	10^{-15}	
16	0	0	permuted to triangular form
17	0	0	permuted to triangular form
18	0	0	permuted to triangular form
19	0	10^{-7}	$1/s_i = 10^{12}$
20	1	10^{-16}	
21	1	10^{-16}	
22	3	10^{-14}	
23	4	10^{-13}	
24	4	10^{-9}	$1/s_i = 10^{23}$
25	2	10^{-16}	
26	0	10^{-16}	deflated during reduction
27	0	10^7	$1/s_i = \infty$, Tridiagonal OK
28	0	10^{-16}	
29	1	10^{-16}	deflated during reduction
30	6	10^{-13}	
31	0	10^{-12}	$1/s_i = 10^{13}$
32	0	10^{-3}	$1/s_i = 10^{11}$
33	0	10^{-16}	deflated during reduction
34	0	10^{-16}	deflated during reduction
35	0	10^{-1}	$1/s_i = 10^{14}$

poor. Methods exist for iteratively refining the accuracy of eigenvalues [4]. Presently, we are investigating an algorithm that improves the accuracy of the eigenvalues determined by TLR and avoids factorization of the original matrix by exploiting the already reduced tridiagonal form T. The algorithm differs from the iterative refinement in [2] in that the eigenvalues converge to the eigenvalues of the original matrix rather than those of T. Details of this work can be found in [3].

We have presented an algorithm for reducing a general matrix directly to tridiagonal form. Pivots are chosen that minimize the maximum element in the transformation matrices. We have described situations where the condition number of the transformation matrices can be large, and we have presented two recovery methods, which work well in practice when such situations arise. The new algorithm is fast and significantly broadens the class of matrices that can be successfully reduced.

Acknowledgments. The author would like to thank Gene Wachspress, Charles Romine, Beresford Parlett, and Gene Golub for helpful discussions during the course of this work. He would also like to thank the referees for their comments on earlier versions of this paper.

REFERENCES

[1] J. K. CULLUM AND R. A. WILLOUGHBY, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations*, Birkhäuser-Verlag, Boston, 1985.

[2] A. DAX AND S. KANIEL, *The* ELR *method for computing the eigenvalues of a general matrix*, SIAM J. Numer. Anal., 18 (1981), pp. 597–605.

[3] J. J. DONGARRA, G. A. GEIST, AND C. H. ROMINE, *Computing the eigenvalues and eigenvectors of a general matrix by reduction to general tridiagonal form*, Tech. Report ORNL/TM-11669, Oak Ridge National Laboratory, Oak Ridge, TN, October 1990.

[4] J. J. DONGARRA, C. B. MOLER, AND J. H. WILKINSON, *Improving the accuracy of computed eigenvalues and eigenvectors*, SIAM J. Numer. Anal., 20 (1983), pp. 23–45.

[5] J. G. F. FRANCIS, *The QR transformation—Part* 2, Comput. J., 4 (1961), pp. 332–345.

[6] G. A. GEIST, *Reduction of a general matrix to tridiagonal form*, Tech. Report ORNL/TM-10991, Oak Ridge National Laboratory, Oak Ridge, TN, February 1989.

[7] ———, *Reduction of a general matrix to tridiagonal form using a hypercube multiprocessor*, in Hypercube Concurrent Computers and Applications 1989, J. L. Gustafson, ed., Golden Gate Enterprises, Los Altos, CA, 1990, pp. 665–670.

[8] G. A. GEIST, A. LU, AND E. L. WACHSPRESS, *Stabilized Gaussian reduction of an arbitrary matrix to tridiagonal form*, Tech. Report ORNL/TM-11089, Oak Ridge National Laboratory, Oak Ridge, TN, February 1989.

[9] R. T. GREGORY AND D. L. KARNEY, *A Collection of Matrices for Testing Computational Algorithms*, Robert E. Krieger Publishing Company, Huntington, NY, 1978.

[10] R. G. GRIMES AND H. D. SIMON, *A new tridiagonalization algorithm for unsymmetric matrices*, Tech. Report SCA-TR-118, Boeing Computer Services, Seattle, WA, 1987.

[11] D. E. HARE AND W. P. TANG, *Toward a stable tridiagonalization algorithm for general matrices*, Tech. Report CS-89-03, Computer Science Dept., University of Waterloo, Waterloo, Ontario, Canada, January 1989.

[12] C. D. LABUDDE, *The reduction of an arbitrary real square matrix to tridiagonal form using similarity transformations*, Math. Comp., 17 (1963), pp. 443–447.

[13] B. N. PARLETT, *Reduction to tridiagonal form and minimal realizations*, SIAM J. Matrix Anal. Appl., submitted.

[14] ———, *A note on LaBudde's algorithm*, Math. Comp., 19 (1964), pp. 505–506.

[15] H. RUTISHAUSER, *Solution of eigenvalue problems with the LR transformation*, Nat. Bur. Standards Appl. Math. Ser., 49 (1958), pp. 47–81.

[16] B. T. SMITH, J. M. BOYLE, J. J. DONGARRA, B. S. GARABOW, Y. IKEBE, V. C. KLEMA, AND C. B. MOLER, *Matrix Eigensystem Routines*—EISPACK *Guide*, Springer-Verlag, Heidelberg, 1974.

[17] C. STRACHEY AND J. G. F. FRANCIS, *The reduction of a matrix to codiagonal form by elimination*, Comput. J., 4 (1961), p. 168.

[18] E. L. WACHSPRESS, *ADI solution of Lyapunov equations,* talk at Minnesota Supercomputer Institute Workshop on Practical Iterative Methods for Large-scale Computations, Minneapolis, MN, October 1988.

[19] D. WATKINS, *Use of the LR algorithm to tridiagonalize a general matrix,* talk at Society for Industrial and Applied Mathematics Annual Meeting, Minneapolis, MN, July 1988.

[20] D. WATKINS AND L. ELSNER, *Self-equivalent flows associated with the generalized eigenvalue problem*, Linear Algebra Appl., 118 (1989), pp. 107–127.

[21] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, U.K., 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 2, pp. 374–384, April 1991

013

CHASING ALGORITHMS FOR THE EIGENVALUE PROBLEM*

D. S. WATKINS† AND L. ELSNER‡

Abstract. A generic chasing algorithm for the matrix eigenvalue problem is introduced and studied. This algorithm includes, as special cases, the implicit, multiple-step QR and LR algorithms and similar bulge-chasing algorithms for the standard eigenvalue problem. The scope of the generic chasing algorithm is quite broad; it encompasses a number of chasing algorithms that cannot be analyzed by the traditional (e.g., implicit Q theorem) approach. These include the LR algorithm with partial pivoting and other chasing algorithms that employ pivoting for stability, as well as hybrid algorithms that combine elements of the LR and QR algorithms. The main result is that each step of the generic chasing algorithm amounts to one step of the generic GR algorithm. Therefore the convergence theorems for GR algorithms that were proven in a previous work [D. S. Watkins and L. Elsner, *Linear Algebra Appl.*, 143 (1991), pp. 19–47] also apply to the generic chasing algorithm.

Key words. eigenvalue, QR algorithm, GR algorithm, chasing the bulge, subspace iteration

AMS(MOS) subject classifications. 65F15, 15A18

1. Introduction.

Two of the best known algorithms for calculating eigenvalues and eigenvectors of matrices are the QR and LR algorithms [15], [12]. There are other, not so well known, algorithms of the same type, e.g., the SR algorithm [8], [9], [6] and the HR algorithm [5], [7], [6], which can be useful in special situations. In [14] we developed a general convergence theory of GR algorithms that includes the QR, LR, SR, HR, and similar algorithms as special cases. In this paper we consider in general terms the question of how such algorithms can be implemented.

Algorithms in this class are usually implemented implicitly, as chasing algorithms: The matrix whose eigenvalues we would like to know is first reduced to upper Hessenberg form. Then the chasing algorithm is set in motion by a similarity transformation that introduces a bulge in the Hessenberg form near the upper left-hand corner of the matrix. A sequence of similarity transformations then chases the bulge downward and to the right, until the Hessenberg form is restored. At this point the first chasing step is complete. Chasing steps are repeated until (hopefully) the matrix converges to upper triangular or block triangular form, exposing the eigenvalues. There are a number of types of similarity transformations that can be used to chase the bulge. For example, certain unitary transformations can be used, in which case each step of the chasing algorithm amounts to a step of the QR algorithm. If, on the other hand, lower triangular transformation matrices are used, each step of the chasing algorithm amounts to a step of the LR algorithm.

In this paper we introduce and study a generic chasing algorithm. After describing the algorithm at the beginning of §2, we state and prove the main result, which is that no matter what kind of transformations are used to chase the bulge, each chasing step amounts to one step of the generic GR algorithm [14]. Consequently all of the observations that we made in [14] concerning the generic GR algorithm apply to the chasing algorithm as well. To wit, each step of the chasing algorithm amounts to a step of nested subspace iteration combined with a change of coordinate system. All of the convergence theorems of [14] apply. Roughly speaking, the chasing

* Received by the editors March 6, 1989; accepted for publication (in revised form) May 24, 1990.
† Department of Pure and Applied Mathematics, Washington State University, Pullman, Washington 99164-3113 (watkins@wsumath.bitnet). The research of this author was supported by National Science Foundation grant DMS-8800437.
‡ Fakultät für Mathematik, Universität Bielefeld, Postfach 8640, D-4800 Bielefeld 1, Federal Republic of Germany (umatfl05@dbuni11.bitnet).

algorithm will converge, provided that (i) reasonable choices of shifts are made, and (ii) the condition numbers of the transforming matrices are kept under control. If the generalized Rayleigh quotient shifting strategy is used, quadratic, and in some cases cubic, convergence can be achieved. We close §2 with a brief discussion of some of the types of transformation that can be used to implement the chasing algorithm.

Our approach to chasing algorithms differs from the traditional one. For purposes of illustration, let us consider the standard way of justifying the implicit QR algorithm. A QR step consists of a similarity transformation $B = Q^{-1}AQ$, where the transforming matrix Q is unitary. One can show that the unitary Q is more or less uniquely determined by its first column. (This fact is known as the *implicit Q theorem*; see, for example, [12, Thm. 7.4.2].) The implicit QR (chasing) algorithm performs a different similarity transformation $\tilde{B} = \tilde{Q}^{-1}A\tilde{Q}$, but $\tilde{Q}$ is constructed in such a way that its first column is proportional to the first column of Q. It follows from the implicit Q theorem that Q and $\tilde{Q}$ are essentially the same, and consequently B and $\tilde{B}$ are essentially the same.

By contrast, our generic chasing algorithm performs repeated similarity transformations $B = G^{-1}AG$, where the nature of G is left unspecified, except that it is nonsingular and its first column is given. In this more general context, we cannot assert that G is more or less uniquely determined. All we can say is that no matter how the chasing step is carried out, it effects one step of the generic GR algorithm. But this is all we need!

Our approach has the following advantages: (i) It covers implicit variants of the QR, LR, SR, and HR algorithms all at once. (ii) It covers the implicit LR algorithm with partial pivoting and other chasing algorithms that employ pivoting for stability, none of which are covered by the traditional approach. (iii) It covers hybrid chasing algorithms as well. For example, an algorithm that uses a mixture of unitary and lower triangular transformations to chase the bulge is a QR-LR hybrid that cannot be analyzed by the traditional approach. Thus our approach encompasses a much broader class of chasing algorithms.

The theorems associated with the traditional approach (e.g., implicit Q theorem) can be derived via our approach by considering the effect of restricting the types of transformations that can be used to do the chasing. This is the main business of §3. Here we restrict our attention, for clarity, to the nonsingular case. This is the generic case, in which none of the shifts (defined in §2) are eigenvalues of A.

In §4 we consider what happens during a singular chasing step. We show that if ν of the shifts are eigenvalues, a $\nu \times \nu$ block can, in principle, be deflated from the matrix after the chasing step. As far as the rest of the matrix is concerned, normal progress is made during the chasing step. That is, a step of nested subspace iteration, combined with a change of coordinate system, takes place. Also considered in §4 is the connection between our approach and the traditional approach in the singular case. Actually the results of §4 include those of §3 as a special case. We have chosen to present the nonsingular case separately because (i) it is generic, and (ii) it is much simpler.

Finally, we wish to emphasize that the theorems in this paper are not at all difficult, nor are the tools used to prove them by any means novel. This paper's main contribution is a new, more flexible, way of looking at chasing algorithms that allows for greater generality than has been attained previously.

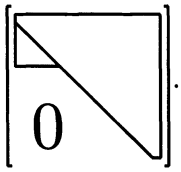
2. The generic GR and chasing algorithms. We described the generic GR algorithm in [14]. For completeness, we briefly repeat the description here. Each GR

algorithm is based on a GR decomposition, which is a rule that specifies a unique way of decomposing any matrix C in some large class of matrices $\mathcal{C}$ into a product $C = GR$, where G is nonsingular, and R is upper triangular. Well-known examples of GR decompositions are the QR and LR decompositions. Corresponding to each GR decomposition is a GR algorithm, an iterative algorithm for finding the eigenvalues of matrices. Given a matrix $A \in \mathbb{C}^{n\times n}$, whose eigenvalues we would like to know, the GR algorithm produces a sequence of similar matrices that, hopefully, converges to upper triangular or block triangular form. A GR step on A is performed as follows. Choose a positive integer m, the *multiplicity* of the step. Choose m *shifts* $\sigma_1, \cdots, \sigma_m$, complex numbers that approximate eigenvalues of A. Let $p(A) = (A-\sigma_1)\cdots(A-\sigma_m)$. Find the GR decomposition of $p(A)$: $p(A) = GR$. Finally, replace A by the similar matrix $B = G^{-1}AG$.

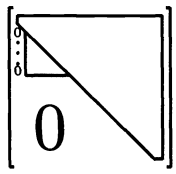
The generic chasing algorithm. A matrix $A = (a_{ij}) \in \mathbb{C}^{n\times n}$ is said to be in *upper Hessenberg* form if $a_{ij} = 0$ for all $i > j+1$. It is in *irreducible* upper Hessenberg form if it satisfies the additional condition $a_{ij} \neq 0$ for all $i = j+1$.

Let $A \in \mathbb{C}^{n\times n}$ be in irreducible upper Hessenberg form, let m be a positive integer, let $\sigma_1, \cdots, \sigma_m$ be approximations to eigenvalues of A, and let $p(A) = (A-\sigma_1)\cdots(A-\sigma_m)$. A step of the generic chasing algorithm of multiplicity m replaces A by $B = G^{-1}AG$, where B is upper Hessenberg, and the first column of G is proportional to the first column of $p(A)$.

The algorithm begins by computing the first column of $p(A)$. This is given by $x = p(A)e_1$, where $e_1 = [1, 0, \cdots, 0]^T$. To avoid potential problems with overflow or underflow, we actually compute $z = x/\|x\|$ using any convenient norm. This can be done efficiently, without actually forming $p(A)$, by the recursion $z \leftarrow (A-\sigma_i)z$, $z \leftarrow z/\|z\|$, $i = 1, \cdots, m$, starting from $z \leftarrow e_1$. Because A is irreducible upper Hessenberg and p has degree m, z satisfies $z_{m+1} \neq 0$, and $z_i = 0$ for $i > m+1$. The next step is to determine a nonsingular matrix $G_0 = \mathrm{diag}\{\tilde{G}_0, I_{n-m-1}\}$, with $\tilde{G}_0 \in \mathbb{C}^{(m+1)\times(m+1)}$, whose first column is proportional to z. Obviously $\tilde{G}_0$ is not uniquely determined. Some common ways of constructing $\tilde{G}_0$ will be discussed below. Once we have G_0, we use it to transform A to $A_0 = G_0^{-1}AG_0$. It is a simple matter to check that A_0 has the form



That is, it is almost in upper Hessenberg form, except that it has a triangular *bulge* with height m rows, base m columns, and vertex at position $(m+2, 1)$. The rest of the algorithm consists of returning A_0 to upper Hessenberg form. The first step in this direction is to transform A_0 to $A_1 = G_1^{-1}A_0G_1$, where G_1 has the following form: $G_1 = \mathrm{diag}\{1, \tilde{G}_1, I_{n-m-2}\}$, $\tilde{G}_1 \in \mathbb{C}^{(m+1)\times(m+1)}$, and the first column of $\tilde{G}_1$ is proportional to the vector $y_1 \in \mathbb{C}^{m+1}$ consisting of the $(2,1)$ through $(m+2,1)$ entries of A_0. It is clear that $\tilde{G}_1$ is not uniquely determined and that any of the techniques for constructing $\tilde{G}_0$ can also be used to construct $\tilde{G}_1$. Since $\tilde{G}_1^{-1}y_1 = ce_1$ for some nonzero constant c, left multiplication of A_0 by G_1^{-1} will create zeros in the first column of the product, below the $(2,1)$ entry. The subsequent right multiplication of $G_1^{-1}A_0$ by G_1 leaves the first column unaltered, but it does create new nonzero entries in row $m+3$, in positions $(m+3, 2)$ through $(m+3, m+1)$. Thus A_1 has the form



The bulge has been "chased" one position down and to the right, so that its vertex now lies at position $(m+3,2)$. The next step produces $A_2 = G_2^{-1}A_1G_2$, where $G_2 = \text{diag}\{1,1,\tilde{G}_2,I_{n-m-3}\}$, $\tilde{G}_2$ being defined analogously to $\tilde{G}_1$. A_2 has a bulge that is down and to the right one position from that of A_1. After $n-m-2$ such steps, the bulge will have been chased to the lower right-hand corner of the matrix. An additional m steps shrink the bulge until it disappears completely. Another viewpoint is that the bulge is pushed off of the edge of the matrix. The similarity transformations used in this phase have the form $G_k = \text{diag}\{I_k,\tilde{G}_k\}$, where $\tilde{G}_k \in \mathbb{C}^{(n-k)\times(n-k)}$, $k = n-(m+1),\cdots,n-2$. After a total of $n-2$ steps we are done; we let $B = A_{n-2} = G^{-1}AG$, where $G = G_0G_1G_2\cdots G_{n-2}$. Each of the matrices $G_1,\cdots,G_{n-2}$ has e_1 as its first column, so the first column of G is the same as the first column of G_0, which is proportional to the first column of $p(A)$.

In the past it has been customary to take m to be a small number, say one or two. The advantage of taking larger values of m is that it improves the vectorizability of the code. The main operations can be expressed as matrix-vector products, and level 2 BLAS [11] can be used. It is also possible to organize the algorithm so that several columns are chased at a time, using tools such as the WY representation of reflectors [4]. This allows the main operations to be expressed as matrix-matrix products, and the algorithm can be coded using level 3 BLAS [10]. This increases the scope for parallelization and efficient use of hierarchical memory. Bai and Demmel [3] have implemented such a version of the QR algorithm and have experimented with values of m as high as 20.

The main point of this paper is that the chasing algorithm effectively performs a step of the generic GR algorithm. The proof follows from three lemmas, whose proofs are easy exercises. We begin with some terminology. Given $x \in \mathbb{C}^n$ and $A \in \mathbb{C}^{n\times n}$, the *Krylov matrix* $K(A,x) \in \mathbb{C}^{n\times n}$ is defined by $K(A,x) = [x, Ax, A^2x, \cdots, A^{n-1}x]$. Clearly $K(A,\beta x) = \beta K(A,x)$ for every scalar β.

LEMMA 2.1. *If $x = p(A)e_1$, then $K(A,x) = p(A)K(A,e_1)$.*

LEMMA 2.2. *For every nonsingular $G \in \mathbb{C}^{n\times n}$, $G^{-1}K(A,x) = K(G^{-1}AG, G^{-1}x)$.*

Lemmas 2.1 and 2.2 are closely related. In fact, Lemma 2.1 for nonsingular $p(A)$ is a special case of Lemma 2.2. However, Lemma 2.1 is valid regardless of whether or not $p(A)$ has an inverse, the key point being that $p(A)$ commutes with A.

LEMMA 2.3. *If A is upper Hessenberg, then $K(A,e_1)$ is upper triangular. Furthermore A is an irreducible upper Hessenberg matrix if and only if $K(A,e_1)$ is upper triangular and nonsingular.*

We are now set to prove the main result.

THEOREM 2.4. *Let $A \in \mathbb{C}^{n\times n}$ be an irreducible upper Hessenberg matrix, and let p be a polynomial. Let G be a nonsingular matrix whose first column is proportional to $x = p(A)e_1$, such that $B = G^{-1}AG$ is upper Hessenberg. Then there exists an upper triangular matrix R such that $p(A) = GR$.*

Proof. By hypothesis, $Ge_1 = \alpha x$ for some nonzero α. Applying Lemmas 2.1 and 2.2 we find that $G^{-1}p(A)K(A,e_1) = G^{-1}K(A,x) = \alpha^{-1}K(B,e_1)$. By Lemma 2.3, $K(A,e_1)$ is both nonsingular and upper triangular, and $K(B,e_1)$ is upper triangular. Therefore $p(A) = G\alpha^{-1}K(B,e_1)K(A,e_1)^{-1} = GR$, where R is the upper triangular

matrix $\alpha^{-1}K(B,e_1)K(A,e_1)^{-1}$. □

Theorem 2.4 shows that the generic chasing algorithm performs a generic GR step implicitly. In this case the rule for calculating the GR decomposition (i.e., constructing G from $p(A)$) is given by the chasing algorithm itself.

There are numerous ways of constructing the $\tilde{G}_i$ in the chasing algorithm. Each $\tilde{G}$ satisfies $\tilde{G}e_1 = \beta y$, or equivalently $\tilde{G}^{-1}y = \beta^{-1}e_1$, for some y. One way to do this, which works if $y_1 \neq 0$, is to define $\tilde{G}$ to be a Gauss transformation:

$$\tilde{G} = \begin{bmatrix} 1 & & & \\ \ell_2 & 1 & & \\ \vdots & & \ddots & \\ \ell_{m+1} & & & 1 \end{bmatrix}, \tag{1}$$

where $\ell_i = y_i/y_1$. If this choice is used, the chasing algorithm amounts to the LR algorithm without pivoting (cf. Example L below). We can eliminate the requirement that $y_1 \neq 0$ by defining $\tilde{G} = PL$, where P is either the identity matrix or a permutation matrix, and L has the form (1). If $|y_1| = \max\{|y_1|, \cdots, |y_{m+1}|\}$, we define P to be the identity matrix. Otherwise we take P to be the transposition matrix whose action on a column vector is to interchange its first and ith entries, where i is the first index for which $|y_i| = \max\{|y_1|, \cdots, |y_{m+1}|\}$. Defining a new vector z by $Py = z$, we then take L to have the form (1), where $\ell_i = z_i/z_1$. This choice yields the LR algorithm with partial pivoting. It is also possible to take the $\tilde{G}$ to be unitary matrices, for example, reflectors (Householder transformations) [12]. In this case the chasing algorithm amounts to the QR algorithm (cf. Example Q below).

For some problems the choice of transformation type is dictated by the structure of the matrix. For example, if A_0 is normal, and we wish to preserve that property, we should use only unitary transformations. If A_0 is Hamiltonian, and we wish to preserve that property, we should use symplectic transformations [6]. This gives the SR algorithm (cf. Example S below).

Other problems have no special structure to exploit. For these problems the transformation type is chosen on the basis of efficiency and stability. From [14] we know that no matter how we choose the G_i, we are doing subspace iteration. The theory developed in [14] suggests that our only consideration is to make the transforming matrices as well conditioned as possible. We might then conclude that we should use only unitary transformations, which are optimally conditioned; that is, we should use the QR algorithm. This conclusion ignores the question of cost. A chasing step using reflectors has about double the flop count of a chasing step using Gauss transformations. Thus LR steps are about half as expensive as QR steps. Of course, the use of Gauss transformations without pivoting is risky; matrices of the form (1) can be made arbitrarily ill conditioned by making the multipliers ℓ_i large. On the other hand, Gauss transformations with partial pivoting, $\tilde{G} = PL$, tend to be well conditioned, as the multipliers never exceed one in modulus. Furthermore, as the iterates approach triangular or block triangular form, the multipliers in the transformations that are generated tend toward zero. As the multipliers approach zero, the condition numbers approach one. Of course, this does not guarantee that the condition numbers of products of many such transformations will remain small. Another possibility, which we already mentioned in the Introduction, is to mix Gauss transformations with unitary transformations. Our theory allows us to do this. All that matters is that the condition numbers of the transforming matrices be kept under control. A hybrid algorithm of this type might well possess a superior combination of speed and robustness.

3. Connection with the traditional approach. In this section we show how our approach can be used to establish the implicit

Q theorem and other results associated with the traditional approach to chasing algorithms. For clarity we restrict our attention to the nonsingular case. That is, we assume that $p(A)$ is nonsingular, which is the same as to say that none of the shifts $\sigma_1, \cdots, \sigma_m$ are eigenvalues of A. We begin by noting that in this case, the matrix B produced by the chasing algorithm is in irreducible upper Hessenberg form.

THEOREM 3.1. *Let $A \in \mathbb{C}^{n\times n}$ be an irreducible upper Hessenberg matrix, and let p be a polynomial for which $p(A)$ is nonsingular. Let G be a nonsingular matrix whose first column is proportional to $x = p(A)e_1$, such that $B = G^{-1}AG$ is upper Hessenberg. Then B has irreducible upper Hessenberg form.*

Proof. The hypotheses are the same as in Theorem 2.4, except that now we are assuming that $p(A)$ is nonsingular. As in the proof of Theorem 2.4, we have $K(B, e_1) = \alpha G^{-1}p(A)K(A, e_1)$. Since G^{-1}, $p(A)$ and $K(A, e_1)$ are all nonsingular, $K(B, e_1)$ must also be nonsingular, in addition to being upper triangular. Therefore, by Lemma 2.3, B has irreducible upper Hessenberg form. □

As we have already mentioned in the Introduction, the transforming matrix G is not uniquely determined by its first column. However, G does have some structure that is specified uniquely, namely, its flag. This useful concept comes from geometry [1], [2]. A *flag* in $\mathbb{C}^n$ is just a nested sequence of subspaces of dimensions $1, 2, \cdots, n$. Given a nonsingular matrix $S \in \mathbb{C}^{n\times n}$ with columns $s_1, \cdots, s_n$, we define the *flag* of S, denoted flag(S), to be the sequence $\{\langle s_1\rangle, \langle s_1, s_2\rangle, \langle s_1, s_2, s_3\rangle, \cdots, \langle s_1, s_2, \cdots, s_n\rangle\}$ determined by the columns of S. It is a simple matter to prove the following lemma.

LEMMA 3.2. *Two nonsingular matrices $S, G \in \mathbb{C}^{n\times n}$ have the same flag if and only if there is a nonsingular upper triangular matrix R such that $S = GR$.*

Theorem 2.4 shows that $p(A) = GR$ for some upper triangular R. Since we are now assuming that $p(A)$ is nonsingular, R must also be nonsingular. Therefore flag(G) = flag($p(A)$). This holds regardless of the type of transformations that are used to build G. We did not use the term "flag" in [14]. However, the nested subspace iterations that are effected by GR steps can be seen to be a consequence of this equality of flags.

From Lemma 2.3 we know that $K(A, e_1)$ is upper triangular and nonsingular. Therefore, by Lemma 2.1, flag($p(A)$) = flag($K(A, x)$). Consequently flag(G) = flag($K(A, x)$). This is actually a special case of a known result that characterizes transformations that reduce a matrix to upper Hessenberg form: Let $A \in \mathbb{C}^{n\times n}$, and suppose there is a vector $x \in \mathbb{C}^n$ for which $K(A, x)$ is nonsingular. Let G be a nonsingular matrix whose first column is proportional to x. Then $B = G^{-1}AG$ is upper Hessenberg if and only if flag(G) = flag($K(A, x)$). If B is upper Hessenberg, then it is irreducible. See especially [6, Satz 4.4.1], but also [12, Thm. 7.4.3]. The case of singular $K(A, x)$ is also covered in [6], but we will give a more general formulation of that case in §4. For now we will state a portion of this result as a lemma for immediate use.

LEMMA 3.3. *Let $A \in \mathbb{C}^{n\times n}$ and let G be a nonsingular matrix such that $B = G^{-1}AG$ is in irreducible upper Hessenberg form. Let $x \in \mathbb{C}^n$ be a vector proportional to the first column of G. Then $K(A, x)$ is nonsingular, and* flag(G) = flag($K(A, x)$).

Proof. By Lemma 2.2, $K(A, x) = G\alpha^{-1}K(B, e_1)$. Since B is irreducible upper Hessenberg, $K(B, e_1)$ is upper triangular and nonsingular. Thus flag(G) = flag($K(A, x)$). □

The transforming matrices G utilized by GR algorithms always lie in $GL_n(\mathbb{C})$,

the group of nonsingular matrices in $\mathbb{C}^{n\times n}$. Nothing more than that is said, in general. However, certain GR algorithms (e.g., the QR algorithm) use only transforming matrices that lie in some proper subgroup $\mathcal{G}$ (e.g., the unitary group). Similarly, one may be able to implement the chasing algorithm in such a way that the transforming matrices all lie in $\mathcal{G}$ (e.g., implicit QR algorithm). The next theorem shows that in such cases the transforming matrices produced by the two algorithms are the same up to right multiplication by a matrix in a certain subgroup $\mathcal{T}$, which we call the *trivial* group. If $\mathcal{G}$ is not too large, then $\mathcal{T}$ really is trivial, and we can conclude that the chasing algorithm and the GR algorithm produce essentially the same result. Examples are given below.

THEOREM 3.4. *Let $\mathcal{G}$ be a subgroup of $GL_n(\mathbb{C})$. Define the* trivial *group $\mathcal{T}$ associated with $\mathcal{G}$ by $\mathcal{T} = \mathcal{G} \cap \mathcal{U}$, where $\mathcal{U}$ denotes the subgroup of $GL_n(\mathbb{C})$ consisting of upper triangular matrices. Let $A \in \mathbb{C}^{n\times n}$, let $G, \tilde{G} \in \mathcal{G}$ have proportional first columns, let $B = G^{-1}AG$ and $\tilde{B} = \tilde{G}^{-1}A\tilde{G}$, and suppose B and $\tilde{B}$ are both irreducible upper Hessenberg. Then there exists $T \in \mathcal{T}$ such that $\tilde{G} = GT$ and $\tilde{B} = T^{-1}BT$.*

Proof. By Lemma 3.3, $\mathrm{flag}(\tilde{G}) = \mathrm{flag}(K(A, x)) = \mathrm{flag}(G)$, where x is a vector proportional to the first columns of G and $\tilde{G}$. Therefore there exists $T \in \mathcal{U}$ such that $\tilde{G} = GT$. But $T = G^{-1}\tilde{G}$, so $T \in \mathcal{G}$. Thus $T \in \mathcal{G} \cap \mathcal{U} = \mathcal{T}$. □

Example Q. The m-step QR algorithm performs a similarity transformation $\tilde{B} = \tilde{Q}^{-1}A\tilde{Q}$, where $\tilde{Q}$ is unitary and $p(A) = \tilde{Q}\tilde{R}$. Similarly, if the chasing algorithm is carried out using unitary transformations exclusively, it performs a similarity transformation $B = Q^{-1}AQ$, where Q is unitary and $p(A) = QR$. Since $\tilde{Q}$ and Q must have proportional first columns, Theorem 3.4 can be applied, with the role of $\mathcal{G}$ played by the unitary group. The group $\mathcal{T}$ associated with this choice of $\mathcal{G}$ is the group of diagonal matrices whose main diagonal entries have unit modulus. Thus $\tilde{Q} = QD$, where D is diagonal with $|d_{ii}| = 1$ for $i = 1, \cdots, n$. This is the complex version of the implicit Q theorem. Thus the chasing algorithm using unitary transformations produces essentially the same result as the QR algorithm.

Example L. The m-step LR algorithm without pivoting performs a similarity transformation $\tilde{B} = \tilde{L}^{-1}A\tilde{L}$, where $\tilde{L}$ is unit lower triangular, and $p(A) = \tilde{L}\tilde{R}$. Similarly, if the chasing algorithm is carried out using Gauss transformations (1) exclusively, it performs a similarity transformation $B = L^{-1}AL$, where L is unit lower triangular, and $p(A) = LR$. Thus Theorem 3.4 applies with $\mathcal{G}$ taken to be the group of unit lower triangular matrices. Then $\mathcal{T} = \{I\}$, so $\tilde{G} = G$ and $\tilde{B} = B$. We conclude that the chasing algorithm using Gauss transformations without pivoting produces exactly the same result as the LR algorithm without pivoting.

Example S. Let n and m be even. The m-step SR algorithm performs a similarity transformation by a symplectic matrix. Taking $\mathcal{G}$ to be the symplectic group (in shuffled form (cf. [14])), we find that $\mathcal{T}$ is the group of all block diagonal matrices $T = \mathrm{diag}\{T_1, \cdots, T_k\}$ (where $k = n/2$), for which each block has the form

$$T_i = \begin{bmatrix} a_i & b_i \\ 0 & a_i^{-1} \end{bmatrix}.$$

Thus the chasing algorithm using symplectic transformations produces a result that differs from the output of the SR algorithm only by a similarity transformation of this simple form.

Remark. In the case $\mathcal{G} = GL_n(\mathbb{C})$, Theorem 3.4 reduces to part (iii) of [6, Satz 4.4.1].

4. The singular case. Before considering the singular case, we introduce some new terminology and present some preliminary results. Given $B \in \mathbb{C}^{n\times n}$ and $j \in \{1,\cdots,n-1\}$, we will say that B is *j-Hessenberg* if its first j columns are in upper Hessenberg form; that is, if B has the form

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

where $B_{11} \in \mathbb{C}^{j\times j}$ is upper Hessenberg, and $B_{21} \in \mathbb{C}^{(n-j)\times j}$ consists entirely of zeros, with the possible exception of the single entry $b_{j+1,j}$ in the upper right-hand corner. We will call B *j-reducible* j-Hessenberg if it is j-Hessenberg, B_{11} is irreducible upper Hessenberg, and $B_{21} = 0$. For the sake of completeness we also include the case $j = n$; the term n-reducible n-Hessenberg will be a synonym for irreducible upper Hessenberg. The first lemma generalizes Lemma 2.3. Again, we leave the proof as an exercise.

LEMMA 4.1. *Let $B \in \mathbb{C}^{n\times n}$ and suppose $K(B,e_1)$ has rank j. Then the following four conditions are equivalent:*

(i) *$K(B,e_1)$ is upper triangular.*

(ii) *$K(B,e_1)$ has the form*

$$K(B,e_1) = \begin{bmatrix} S_{11} & S_{12} \\ 0 & 0 \end{bmatrix}, \tag{2}$$

where $S_{11} \in \mathbb{C}^{j\times j}$ is upper triangular and nonsingular.

(iii) *B is j-Hessenberg.*

(iv) *B is j-reducible j-Hessenberg.*

Extending the definition of the Krylov matrix $K(A,x)$, we define the $n \times j$ *Krylov matrix* $K(A,x,j)$ by $K(A,x,j) = [x, Ax, A^2x, \cdots, A^{j-1}x]$. We also need to extend the definition of the flag of a matrix. Let $S \in \mathbb{C}^{n\times j}$ have linearly independent columns $s_1,\cdots,s_j$. We define the *flag* of S to be the nested seqence of j subspaces $\{\langle s_1\rangle, \langle s_1,s_2\rangle, \langle s_1,s_2,s_3\rangle, \cdots, \langle s_1,s_2,\cdots,s_j\rangle\}$. Generalizing Lemma 3.2, we have Lemma 4.2.

LEMMA 4.2. *Two full-rank matrices $S,G \in \mathbb{C}^{n\times j}$ have the same flag if and only if there is a nonsingular upper triangular matrix $R \in \mathbb{C}^{j\times j}$ such that $S = GR$.*

The next theorem extends parts (i) and (ii) of [6, Satz 4.4.1].

THEOREM 4.3. *Let $A \in \mathbb{C}^{n\times n}$ and $x \in \mathbb{C}^n$, $x \neq 0$, with* $\operatorname{rank}(K(A,x)) = j$. *Let $G \in \mathbb{C}^{n\times n}$ be a nonsingular matrix whose first column is proportional to x, and let $B = G^{-1}AG$. Define submatrices $G_1 \in \mathbb{C}^{n\times j}$ and $G_2 \in \mathbb{C}^{n\times(n-j)}$ by $G = [G_1, G_2]$. Then the following three conditions are equivalent:*

(i) *B is j-Hessenberg.*

(ii) *B is j-reducible j-Hessenberg.*

(iii) $\operatorname{flag}(K(A,x,j)) = \operatorname{flag}(G_1)$.

Proof. Since $x = \alpha G e_1$ for some nonzero α, we have

$$K(A,x) = \alpha G K(B,e_1) \tag{3}$$

by Lemma 2.2. Thus the hypothesis of Lemma 4.1, $\operatorname{rank}(K(B,e_1)) = j$, holds. Therefore (i) and (ii) are equivalent. We now show that (iii) is equivalent to (i) and (ii). Suppose B is j-Hessenberg. Then $K(B,e_1)$ is upper triangular and has the special form (2). Writing (3) in block form, we find that it implies $K(A,x,j) = G_1(\alpha S_{11})$, where αS_{11} is upper triangular and nonsingular. Thus $\operatorname{flag}(K(A,x,j)) = \operatorname{flag}(G_1)$.

Conversely, suppose $\text{flag}(K(A,x,j)) = \text{flag}(G_1)$. Then $K(A,x,j) = G_1S$, where S is upper triangular and nonsingular. We find by inspection that $AK(A,x,j) = K(A,x,j)C$, where $C \in \mathbb{C}^{j\times j}$ is a companion matrix:

$$C = \begin{bmatrix} 0 & \cdots & 0 & * \\ 1 & & 0 & * \\ & \ddots & & \vdots \\ & & 1 & * \end{bmatrix}.$$

In particular, C is irreducible upper Hessenberg. Combining the equations $AK(A,x,j) = K(A,x,j)C$ and $K(A,x,j) = G_1S$, we find that $AG_1S = G_1SC$, or $AG_1 = G_1H$, where $H = SCS^{-1}$ is irreducible upper Hessenberg. Define $F \in \mathbb{C}^{n\times n}$ by $F^* = G^{-1}$, and make the partition $F = [F_1, F_2]$, where $F_1 \in \mathbb{C}^{n\times j}$. Then $F_1^*G_1 = I \in \mathbb{C}^{j\times j}$ and $F_2^*G_1 = 0 \in \mathbb{C}^{(n-j)\times j}$. Also

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} F_1^*AG_1 & F_1^*AG_2 \\ F_2^*AG_1 & F_2^*AG_2 \end{bmatrix}.$$

Thus $B_{11} = F_1^*AG_1 = F_1^*G_1H = H$ and $B_{21} = F_2^*AG_1 = F_2^*G_1H = 0$. Therefore B has j-reducible j-Hessenberg form. □

We are now ready to consider a step of the chasing algorithm for which $p(A)$ is singular. Write $p(A)$ in the factored form

$$p(A) = (A - \sigma_1)(A - \sigma_2)\cdots(A - \sigma_m).$$

$p(A)$ is singular if and only if at least one of the shifts σ_i is an eigenvalue of A. Let ν denote the number of shifts that are equal to eigenvalues of A. Here we count a repeated shift according to its multiplicity as a zero of p, except that the number of times we count it must not exceed its multiplicity as a zero of the characteristic polynomial of A (algebraic multiplicity).

LEMMA 4.4. *The rank of $p(A)$ is $n - \nu$.*

Proof. Since A has irreducible upper Hessenberg form, its eigenspaces are one-dimensional. Thus A has just one Jordan block [13] associated with each eigenvalue; that is, A is nonderogatory. Let $J = \text{diag}\{J_1, \cdots, J_k\}$ be the Jordan canonical form of A. Since the Jordan blocks correspond to distinct eigenvalues, each shift can be an eigenvalue of at most one block. For $i = 1, \cdots, k$, let λ_i be the eigenvalue associated with the block J_i, and let n_i be the dimension of the block. Let $\tilde{\nu}_i$ be the number of shifts that are equal to λ_i, and let $\nu_i = \min\{\tilde{\nu}_i, n_i\}$. Then $\nu = \sum_{i=1}^{k} \nu_i$. For each i, consider the factored form $p(J_i) = (J_i - \sigma_1)\cdots(J_i - \sigma_m)$. The factor $J_i - \sigma_l$ is nonsingular if and only if $\sigma_l \neq \lambda_i$. If $\sigma_l = \lambda_i$, then $J_i - \sigma_l = N$, where N is the nilpotent matrix with ones on the superdiagonal and zeros elsewhere. Since $\tilde{\nu}_i$ of the factors are equal to N, $p(J_i)$ has the form $p(J_i) = MN^{\tilde{\nu}_i} = MN^{\nu_i}$, where M is nonsingular. The nullity of N^{ν_i}, hence also of $p(J_i)$, is ν_i. The nullity of $p(J)$ is the sum of the nullities of the blocks, which is ν. Thus $\text{rank}(p(A)) = \text{rank}(p(J)) = n - \nu$. □

THEOREM 4.5. *Let $B = G^{-1}AG$ be the outcome of one step of the generic chasing algorithm in which $\text{rank}(p(A)) = n - \nu = j$. Then $p(A) = GR$, where R is an upper triangular matrix of the form*

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix},$$

with $R_{11} \in \mathbb{C}^{j\times j}$ upper triangular and nonsingular. Furthermore B has j-reducible j-Hessenberg form:

$$B = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}.$$

The eigenvalues of $B_{22} \in \mathbb{C}^{\nu\times\nu}$ are just the ν shifts σ_i that are eigenvalues of A.

Proof. From the proof of Theorem 2.4 we know that $p(A) = GR$, where $R = \alpha^{-1}K(B,e_1)K(A,e_1)^{-1}$. Since R must have rank j, so must $K(B,e_1)$. Therefore, by Lemma 4.1, $K(B,e_1)$ must have the form (2). It follows immediately that R has the same form. The form of B also follows from Lemma 4.1.

It would appear to be an easy result that the eigenvalues of B_{22} are exactly the shifts that are eigenvalues of A. Consider first the case in which the eigenvalues of A are distinct. The eigenvalues of B_{11} are just those of $A|_{R(p(A))}$. The range of $p(A)$ is exactly the invariant subspace of A associated with the eigenvalues that are not among the shifts. The eigenvalues of B_{22} are the remaining eigenvalues of A, namely, those that are shifts. If A has multiple eigenvalues, this argument is clouded by the multiplicity question: it is possible that B_{11} and B_{22} have common eigenvalues. However, a careful inspection of $p(J)$, where J is the (nonderogatory) Jordan form of A, reveals that the argument can be extended to the general situation: If λ_i is an eigenvalue of J of multiplicity n_i and is used as a shift of multiplicity ν_i, with $\nu_i < n_i$, then λ_i is an eigenvalue of $J|_{R(p(J))}$ (hence of B_{11}) with multiplicity $n_i - \nu_i$. Therefore λ_i must be an eigenvalue of B_{22} of multiplicity ν_i. □

Theorem 4.5 shows that singular $p(A)$ are desirable, as they allow the problem to be deflated after one step. Of course this result ignores the effect of roundoff errors, which will cause $b_{j+1,j}$ to be nonzero in practice. Experience suggests that the computed $b_{j+1,j}$ will usually be large enough to prevent deflation.

Our final task is to extend Theorem 3.4 and its corollaries (Examples Q, R, and S). Let $\mathcal{G}$ be a subgroup of $GL_n(\mathbb{C})$, and for $j = 1,\cdots,n$, let $\mathcal{G}_j$ denote the subset of $GL_j(\mathbb{C})$ consisting of all $G \in \mathbb{C}^{j\times j}$ for which there exist $X \in \mathbb{C}^{j\times(n-j)}$ and $Y \in \mathbb{C}^{(n-j)\times(n-j)}$ such that $\left[\begin{smallmatrix} G & X \\ 0 & Y \end{smallmatrix}\right] \in \mathcal{G}$. It is easy to show that $\mathcal{G}_j$ is a subgroup of $GL_j(\mathbb{C})$. Let $\mathcal{U}_j$ denote the upper triangular subgroup of $GL_j(\mathbb{C})$, and let $\mathcal{T}_j = \mathcal{G}_j \cap \mathcal{U}_j$.

Example Q′. If $\mathcal{G}$ is the unitary group, then $\mathcal{G}_j$ is the unitary group in $GL_j(\mathbb{C})$, so $\mathcal{T}_j$ is the group of $j \times j$ diagonal matrices with main diagonal elements of unit modulus.

Example L′. If $\mathcal{G}$ is the group of unit lower triangular matrices in $GL_n(\mathbb{C})$, then $\mathcal{G}_j$ is the group of unit lower triangular matrices in $GL_j(\mathbb{C})$, so $\mathcal{T}_j$ is the subgroup of $GL_j(\mathbb{C})$ consisting of the single element I.

Example S′. If $\mathcal{G}$ is the symplectic group, and j is even, then $\mathcal{G}_j$ is the symplectic group in $GL_j(\mathbb{C})$, so $\mathcal{T}_j$ is the group of block diagonal matrices in $GL_j(\mathbb{C})$ with 2×2 blocks of the form given in Example S.

THEOREM 4.6. *Let $x \in \mathbb{C}^n$ and $A \in \mathbb{C}^{n\times n}$, with* $\operatorname{rank}(K(A,x)) = j$. *Let $\mathcal{G}$ be a subgroup of $GL_n(\mathbb{C})$, and let $G, \tilde{G} \in \mathcal{G}$ be matrices whose first columns are proportional to x. Suppose $B = G^{-1}AG$ and $\tilde{B} = \tilde{G}^{-1}A\tilde{G}$ both have j-Hessenberg form. Then both are j-reducible, and there exists $T \in \mathcal{T}_j$ such that $\tilde{B}_{11} = T^{-1}B_{11}T$.*

Proof. By Theorem 4.3 we know that B and $\tilde{B}$ are both j-reducible. Furthermore, $\operatorname{flag}(G_1) = \operatorname{flag}(K(A,x,j)) = \operatorname{flag}(\tilde{G}_1)$, where G_1 is defined as in Theorem 4.3, and $\tilde{G}_1$ is defined analogously. Thus there is a $T \in \mathcal{U}_j$ such that $\tilde{G}_1 = G_1T$. We will show that $T \in \mathcal{G}_j$ also, so that in fact $T \in \mathcal{T}_j$. Obviously $G^{-1}\tilde{G} \in \mathcal{G}$. Defining

$F = [F_1, F_2]$ by $F^* = G^{-1}$, as in Theorem 4.3, we have $F_1^* \tilde{G}_1 = F_1^* G_1 T = T$ and $F_2^* \tilde{G}_1 = F_2^* G_1 T = 0$, so

$$G^{-1}\tilde{G} = \begin{bmatrix} F_1^*\tilde{G}_1 & F_1^*\tilde{G}_2 \\ F_2^*\tilde{G}_1 & F_2^*\tilde{G}_2 \end{bmatrix} = \begin{bmatrix} T & F_1^*\tilde{G}_2 \\ 0 & F_2^*\tilde{G}_2 \end{bmatrix}.$$

This proves that $T \in \mathcal{G}_j$, whence $T \in \mathcal{T}_j$. The equation $B = G^{-1}AG$ implies $AG = GB$. Since B is j-reducible, this implies in turn that $AG_1 = G_1 B_{11}$. Similarly $A\tilde{G}_1 = \tilde{G}_1 \tilde{B}_{11}$. Thus $\tilde{G}_1 \tilde{B}_{11} = A\tilde{G}_1 = AG_1 T = G_1 B_{11} T = \tilde{G}_1 (T^{-1} B_{11} T)$. Since $\tilde{G}_1$ has full rank, we can conclude that $\tilde{B}_{11} = T^{-1} B_{11} T$. □

REFERENCES

[1] G. AMMAR, *Geometric aspects of Hessenberg matrices,* Contemp. Math., 68 (1987), pp. 1–21.

[2] G. AMMAR AND C. MARTIN, *The geometry of matrix eigenvalue methods,* Acta Appl. Math., 5 (1986), pp. 239–278.

[3] Z. BAI AND J. DEMMEL, *On a block implementation of Hessenberg multishift iteration,* Internat. J. High Speed Comput., 1 (1989), pp. 97–121.

[4] C. BISCHOF AND C. VAN LOAN, *The WY representation for products of Householder matrices,* SIAM J. Sci. Statist. Comput., 8 (1987), pp. s2–s13.

[5] M. A. BREBNER AND J. GRAD, *Eigenvalues of $Ax = \lambda Bx$ for real symmetric matrices A and B computed by reduction to a pseudosymmetric form and the HR process,* Linear Algebra Appl., 43 (1982), pp. 99–118.

[6] W. BUNSE AND A. BUNSE-GERSTNER, *Numerische lineare Algebra,* Teubner, Stuttgart, 1985.

[7] A. BUNSE-GERSTNER, *An analysis of the HR algorithm for computing the eigenvalues of a matrix,* Linear Algebra Appl., 35 (1981), pp. 155–178.

[8] A. BUNSE-GERSTNER AND V. MEHRMANN, *A symplectic QR-like algorithm for the solution of the real algebraic Riccati equation,* IEEE Trans. Automat. Control, 31 (1986), pp. 1104–1113.

[9] A. BUNSE-GERSTNER, V. MEHRMANN, AND D. WATKINS, *An SR algorithm for Hamiltonian matrices based on Gaussian elimination,* Methods Oper. Res., 58 (1989), pp. 339–358.

[10] J. DONGARRA, J. DU CROZ, S. HAMMARLING, AND I. DUFF, *A set of level 3 basic linear algebra subprograms,* ACM Trans. Math. Software, 16 (1990), pp. 1–17.

[11] J. DONGARRA, J. DU CROZ, S. HAMMARLING, AND R. HANSON, *An extended set of Fortran basic linear algebra subprograms,* ACM Trans. Math. Software, 14 (1988), pp. 1–17.

[12] G. GOLUB AND C. VAN LOAN, *Matrix Computations,* Second Edition, The Johns Hopkins University Press, Baltimore, MD, 1989.

[13] P. LANCASTER AND M. TISMINETSKY, *The Theory of Matrices,* Second Edition, Academic Press, Orlando, FL, 1985.

[14] D. S. WATKINS AND L. ELSNER, *Convergence of algorithms of decomposition type for the eigenvalue problem,* Linear Algebra Appl., 143 (1991), pp. 19–47.

[15] J. H. WILKINSON, *The Algebraic Eigenvalue Problem,* Oxford University Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 2, pp. 385–400, April 1991

014

IMPLICIT SHIFTING IN THE QR AND RELATED ALGORITHMS*

G. S. MIMINIS† AND C. C. PAIGE‡

Abstract. A new approach is suggested for deriving the theory of implicit shifting in the QR algorithm applied to a Hessenberg matrix. This is less concise than Francis' original approach ([Comput. J., 4(1961), pp. 265–271], [Comput. J., 4(1962), pp. 332–345]) but is more instructive, and extends easily to more general cases. For example, it enables us to design implicitly shifted QR algorithms for band and block Hessenberg matrices. It can also be applied to related algorithms such as the LR algorithm, and to algorithms which do not produce triangular matrices in the factorization step. The approach provides details that can be useful in designing numerically effective algorithms in various areas.

In addition to the above, the standard theory describing the result of the QR algorithm with k shifts on a Hessenberg matrix A is extended to the case where some of the shifts can be eigenvalues. This has a practical value in special cases such as eigenvalue allocation. The extension is given for both the explicitly and implicitly shifted QR algorithms, and shows to what extent the latter mimics the former. The new approach to the theory again handles the implicit case simply and clearly.

Key words. matrix eigenproblem, implicit shift, QR algorithm, pole placement

AMS(MOS) subject classifications. 65F15, 15A18, 15A21, 93B55

1. Introduction.

Francis' implicitly shifted QR algorithm [5], [6] is one of the most effective computational tools in numerical linear algebra. It is based on the QR factorization of a matrix, and was originally used for finding eigenvalues of matrices. Its uses are wide, and continue to increase. The theory of implicit shifting given in [5], [6] holds for Hessenberg matrices (upper Hessenberg matrices have all zeros below the first subdiagonal) and has been sufficient for most applications to date, but now algorithms for more general problems involving, for example, block Hessenberg matrices appear to need a new approach.

Here we suggest an approach to implicit shift theory that is different from Francis' approach and seems to be more general in that it can be applied to QR algorithms for block Hessenberg matrices of any form, as well as to many algorithms which differ from, but are in some ways like, the QR algorithm. In fact, the approach here does not depend on orthogonality or the production of triangular structure, and could contribute to many areas where implicit shifting is used, for example the LR and Cholesky LL^T algorithms [14], the QZ [13] and LZ [8] algorithms for the generalized eigenvalue problem, the HR algorithm (see, for example, [2]), and algorithms taking advantage of special structure (see, for example, [4], [3]), to mention just two of the many recent algorithms in this area.

We present this new approach by only considering implicit shifting for the QR algorithm and some variants of it. This should be sufficient for others to extend the ideas to other related algorithms. We first apply the approach to show how to derive the usual implicitly shifted QR algorithm for an upper Hessesnberg matrix A, and later indicate its generality by applying it to the novel case of block upper Hessenberg A. The case of upper Hessenberg A allows a close comparison with the approach

* Received by the editors November 28, 1988; accepted for publication (in revised form) June 25, 1990. This research was supported by Natural Sciences and Engineering Research Council of Canada grants A0944 and A9236.

† Department of Computer Science, Memorial University, St. John's, Newfoundland, Canada A1C 5S7 (george@garfield.cs.mun.edu).

‡ School of Computer Science, McGill University, Montreal, Quebec, Canada H3A 2A7 (chris@cs.mcgill.ca).

in [5], [6]. It will be seen that for this case the new approach is less concise, but more instructive in that it gives a clearer derivation of the implicit shift algorithm, and reveals useful details that are hidden by the approach in [5], [6]. To this end, §2 summarizes the necessary theory for the *explicitly* shifted QR algorithm, then summarizes the standard theory for the implicitly shifted QR algorithm applied to upper Hessenberg A, and discusses some of the advantages and limitations of the standard approach. The new approach is outlined in §3, as well as a comparison with the standard approach for the special case of upper Hessenberg A.

The theory for the shifted QR algorithm for finding eigenvalues is usually given assuming that none of the shifts are eigenvalues. However, recently, a class of algorithms based on some of the QR algorithm theory has provided *direct* (rather than iterative) algorithms for allocating eigenvalues [11], and here the shifts *are* eigenvalues, and we need to extend the usual results to show just what must be done to allocate the required eigenvalues in different cases. For this reason we give in §4 the theory for the QR algorithm with k shifts for general upper Hessenberg A with no restrictions at all on the shifts. This parallels §2 in first giving the results for the explicitly shifted algorithm, then using our new approach to show that the implicitly shifted algorithm parallels the explicit one to the required extent.

Finally, in §5, it is shown how the approach to implicit shifting outlined in §3 can be applied with shifts $\mu_1, \cdots, \mu_k$ to matrices A that are not upper Hessenberg. Two examples are given: A with several nonzero subdiagonals, and block upper Hessenberg A. It is shown that unlike the upper Hessenberg case, more than one column of $N = (A - \mu_1 I) \cdots (A - \mu_k I)$ must effectively be formed in order to carry through the implicit shift algorithm, and it becomes clear just what parts of N are required. Block Hessenberg matrices can arise when we want to introduce a significant amount of parallelism in the computation of eigenvalues using the QR algorithm (see, for example, [1], [15]). Also there are problems where we are unable to reduce the matrix to Hessenberg form, for example in the multi-input pole placement problem. A method suggested in [11] and [12] for this problem uses a QR-like approach (called the QS algorithm) on a block Hessenberg form, which is not reduced to triangular, or even Hessenberg, form at any stage. The approach to implicit shifting to be described here was very successful for this problem; in fact it was the reason we developed this approach—we could find no other way.

2. The QR algorithm with shifts. For a given $n \times n$ matrix $A_1 = A$, k steps of Francis' QR algorithm [5], [6] with explicit shifts $\mu_1, \cdots, \mu_k$ produce the following relationships (see, for example, [16, p. 524]). For $i = 1, \cdots, k$, unitary Q_i is chosen to give upper triangular R_i in

$$Q_i^H (A_i - \mu_i I) = R_i,$$

$$A_{i+1} = R_i Q_i + \mu_i I = Q_i^H A_i Q_i = Q_i^H \cdots Q_1^H A Q_1 \cdots Q_i. \tag{2.1}$$

Such unitary similarity transformations preserve eigenvalues, and in Francis' algorithm for finding eigenvalues, the shifts μ_i are chosen so that A_{i+1} should swiftly converge to upper, or near upper, triangular form (see, for example, [7, Chap. 7]).

Three matrices that will appear regularly in this paper are

$$N = (A - \mu_k I) \cdots (A - \mu_1 I), \qquad Q = Q_1 \cdots Q_k, \qquad R = R_k \cdots R_1, \tag{2.2}$$

so that Q is unitary and R is upper triangular. The two main relations on which implicit shifting techniques are based are then

$$A_{k+1} = Q^H A Q, \tag{2.3}$$

(2.4)

$$\begin{aligned} Q^H N &= Q_k^H \cdots Q_1^H (A - \mu_k I) Q_1 \cdots Q_{k-1} Q_{k-1}^H \cdots Q_1^H (A - \mu_{k-1} I) \cdots \\ &\quad \cdots (A - \mu_2 I) Q_1 Q_1^H (A - \mu_1 I) \\ &= Q_k^H (A_k - \mu_k I) Q_{k-1}^H (A_{k-1} - \mu_{k-1} I) \cdots Q_2^H (A_2 - \mu_2 I) Q_1^H (A_1 - \mu_1 I) \\ &= R_k R_{k-1} \cdots R_2 R_1 \\ &= R. \end{aligned}$$

This shows that Q in (2.1) and (2.2) always gives a QR factorization of N.

If none of the shifts is an eigenvalue of A, then N is nonsingular, and Q and R in (2.4) are uniquely defined by the QR factorization of N if we insist the diagonal elements of R are real and positive (see, for example, [16, p. 241]). Thus for nonsingular N, the Q in the QR factorization of N is identical to Q in (2.2). This, as we shall see, allows us to work directly with $Q = Q_1 \cdots Q_k$ without forming or using the individual Q_i matrices.

The aim of implicit shifting is to produce A_{k+1} by applying unitary transformations directly to A, rather than to shifted matrices as in (2.1). Francis considered the case where A is upper Hessenberg, the obvious computation in (2.1) ensuring that A_{k+1} is also upper Hessenberg. The implicit shift computation for an unreduced upper Hessenberg matrix $A_1 = A$ when none of the shifts is an eigenvalue will now be described. Unreduced means that there are no zero elements on the first subdiagonal. This, with no shift being an eigenvalue, is the condition used in Francis' proof that implicit shifting works [6, Thm. 11]. This also ensures A_{k+1} is unreduced, which is the condition for the proof in [16, p. 529; pp. 352–353]. Note that the single shift method is obtained with $k = 1$, and the double with $k = 2$.

Let P_1 be a unitary matrix chosen to zero all but the first element of the first column of N (just as in the QR factorization (2.4))

$$P_1^H N e_1 = e_1 \rho_1. \tag{2.5}$$

Form $P_1^H A P_1$, and choose unitary similarity transformations based, for example, on Householder transformations $P_2, \cdots, P_{n-1}$, each having (1,1) element unity, to transform $P_1^H A P_1$ back to upper Hessenberg form H, giving

$$AP = PH, \qquad P = P_1 P_2 \cdots P_{n-1}. \tag{2.6}$$

Paralleling this, we have from (2.3)

$$AQ = Q A_{k+1}. \tag{2.7}$$

It can be shown that $P = QD$ and $H = D^H A_{k+1} D$, where D is a diagonal matrix with elements of modulus unity, which we denote $|D| = I$. So effectively A_{k+1} has been found by applying unitary similarity transformations directly to A. The usual approach to this proof can be outlined briefly as follows (see, for example,

[16, pp. 528–537] for a more complete proof for the double step QR algorithm of Francis). From (2.5) and the form of P in (2.6) we see $Ne_1 = P_1 e_1 \rho_1 = P e_1 \rho_1$. But from (2.4) $Ne_1 = QRe_1$, so the first columns of P and Q are identical up to a scalar multiplier of modulus unity. This approach then uses the result that since A_{k+1} is unreduced, Q and A_{k+1} in (2.7) are effectively uniquely determined by the first column of Q. It follows from (2.6) that P corresponds to Q, and H corresponds to A_{k+1}. Computationally, the shifts are used to produce the first column of N from which P_1 is obtained, then $H = D^H A_{k+1} D$ is obtained from the unitary similarity transformation of $P_1^H A P_1$ back to upper Hessenberg form.

Note that in (2.1) we have k explicit steps, and in (2.5) and (2.6) we have $n-1$ implicit steps. The ith explicit step corresponds to the design and application of Q_i in (2.1), while the ith implicit step corresponds to the design and application of P_i in (2.5) or (2.6).

The above approach to implicit shifting is concise and reasonable to follow for those well versed in the area; however it has two drawbacks which the approach we will present in §3 does not have. First it is restricted to A of Hessenberg form. A more general theory was required in [11] and [12] where it was not possible to reduce A to Hessenberg form in the algorithm used to *allocate* eigenvalues of $A = A_0 - BF$, A_0 and B given, by choosing F when the rank of B was greater than one. Second, by appealing to the uniqueness result for (2.7), this approach hides some important relationships that can be useful in designing numerical algorithms. As a simple example, it obscures the fact that the $P_1, \cdots, P_{n-1}$ in (2.6) are exactly those transformations required for the QR factorization $Q^H N = R$.

Relationships like these help in designing algorithms such as eigenvalue allocation algorithms [12], where it is crucial that the shift not be significantly degraded by the presence of rounding errors. These are *direct* algorithms, and each shift (which corresponds to an eigenvalue being allocated) has only one chance—if it is significantly degraded, then an incorrect eigenvalue is allocated. In iterative algorithms such degradation is not so important, as the worst an incorrect shift can do is upset convergence.

These relationships could also be useful in designing some iterative algorithms: for example, applying the QR algorithm to an A which is only defined implicitly, or in other cases where it could be difficult to detect or make use of any splitting of the matrix.

Another possible advantage of the approach we present is that it is straightforward and motivates the implicit shift computation. It may be easier to follow for beginners, and could be useful for teaching.

Following a presentation of this work [10], some ideas here have been used in a block version of the QR algorithm for parallel computation [1].

3. A new approach to implicit shift theory. The usual description of implicit shifting presents the algorithm (2.5) to (2.6), and then proves that it works. We could do this with our approach too, but we can also derive the algorithm step by step, and as this provides greater motivation and understanding, it is the course we take here. The comments regarding nonsingular N are there to help understanding, but it is in no way essential for N to be nonsingular for this theory.

We have seen that with shifts $\mu_1, \cdots, \mu_k$, k steps of the explicitly shifted QR algorithm give $A_{k+1} = Q^H A Q$, where with $N = (A - \mu_k I) \cdots (A - \mu_1 I)$ this unitary Q gives upper triangular R in

$$Q^H N = R. \tag{3.1}$$

Thus consider the usual QR factorization of N giving upper triangular N_n

$$P_{n-1}^H \cdots P_2^H P_1^H N = N_n. \tag{3.2}$$

Here P_1 is the same as in (2.5). If N is nonsingular, then from uniqueness of the QR factorization, $P_1 \cdots P_{n-1} = QD$, $|D| = I$, $N_n = D^H R$, and we will prove that $P_1, \cdots, P_{n-1}$ are essentially those in (2.6), which is why we use the same symbols.

To develop the theory for implicit shifting we write $N_1 = N, H_1 = A$, and define for $i = 1, 2, \cdots, n-1$

$$N_{i+1} = P_i^H N_i = P_i^H \cdots P_1^H N \tag{3.3}$$

to describe the intermediate steps of the QR factorization, and

$$H_{i+1} = P_i^H H_i P_i = P_i^H \cdots P_1^H A P_1 \cdots P_i \tag{3.4}$$

to describe the corresponding unitary similarity transformations of A.

Since N is a polynominal in A, we have the key relationship between H_i and N_i:

$$\begin{aligned} N_i A &= P_{i-1}^H \cdots P_1^H N A \\ &= P_{i-1}^H \cdots P_1^H A N \\ &= P_{i-1}^H \cdots P_1^H A P_1 \cdots P_{i-1} P_{i-1}^H \cdots P_1^H N \\ &= H_i N_i, \qquad i = 1, 2, \cdots, n. \end{aligned} \tag{3.5}$$

Of course this also holds for any polynomial N in A and any sequence of unitary $P_1, P_2, \cdots$, and requires no restrictions on the form of $n \times n$ A or the shifts μ_i.

This relationship between H_i and N_i will allow us to handle the most general case of implicit shifting, but as an introduction we will parallel the usual approach by assuming N is nonsingular and A is an unreduced upper Hessenberg matrix. The case of singular N will be treated in §4. We already have for nonsingular N that $H_n = P_{n-1}^H \cdots P_1^H A P_1 \cdots P_{n-1} = D^H Q^H A Q D = D^H A_{k+1} D$, $|D| = I$, so H_n is the desired matrix. But since it is not computationally acceptable to form all of N and compute the P_i from the QR factorization of N, we will use (3.5) to show that only P_1 need be computed by forming the first column of N, then for $i = 2, \cdots, n-1$, P_i can be computed directly from H_i, and applied to give H_{i+1}. We illustrate this for $n = 6$ and a double shift, $k = 2$, which gives N two subdiagonals. The structure of (3.5) when the first two columns of N have been reduced is

$$\underbrace{\begin{bmatrix} \rho_1 & \times & \times & \times & \times & \times \\ & \rho_2 & \times & \times & \times & \times \\ & & \tau_3 & \times & \times & \times \\ & & \sigma_3 & \times & \times & \times \\ & & \nu_3 & \times & \times & \times \\ & & & \times & \times & \times \end{bmatrix}}_{N_3} \underbrace{\begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \alpha_1 & \times & \times & \times & \times & \times \\ & \alpha_2 & \times & \times & \times & \times \\ & & \times & \times & \times & \times \\ & & & \times & \times & \times \\ & & & & \times & \times \end{bmatrix}}_{A}$$

(3.6)

$$= \underbrace{\begin{bmatrix} \times & \times & \times & \times & \times & \times \\ \eta_1 & \times & \times & \times & \times & \times \\ & \beta_2 & \times & \times & \times & \times \\ & \gamma_2 & \times & \times & \times & \times \\ & \delta_2 & \times & \times & \times & \times \\ & & & & \times & \times \end{bmatrix}}_{H_3} \underbrace{\begin{bmatrix} \rho_1 & \times & \times & \times & \times & \times \\ & \rho_2 & \times & \times & \times & \times \\ & & \tau_3 & \times & \times & \times \\ & & \sigma_3 & \times & \times & \times \\ & & \nu_3 & \times & \times & \times \\ & & & \times & \times & \times \end{bmatrix}}_{N_3}.$$

Since N is nonsingular, ρ_1 and ρ_2 will be nonzero, and the first column above with A unreduced ensures that the first column of H_3 is zero except for the first two elements, and that η_1 is nonzero. From the second column on each side we see

$$\alpha_2[\ \tau_3 \quad \sigma_3 \quad \nu_3\ ] = \rho_2[\ \beta_2 \quad \gamma_2 \quad \delta_2\ ], \qquad \alpha_2\rho_2 \neq 0,$$

so any transformation that reduces $[\ \tau_3 \quad \sigma_3 \quad \nu_3\ ]$ to $[\ \rho_3 \quad 0 \quad 0\ ]$ will reduce $[\ \beta_2 \quad \gamma_2 \quad \delta_2\ ]$ to $[\ \eta_2 \quad 0 \quad 0\ ]$, and vice versa. It follows that P_3 can be defined either from N_3 or H_3. Designing P_3 from H_3 gives the usual implicit shift computation, which is what we wanted to show. Thus after P_1 is designed to give $P_1^H N e_1 = e_1\rho_1$, for $i = 2, \cdots, n-1$ each P_i is designed so $H_{i+1} = P_i^H H_i P_i$ has its first $i-1$ columns in upper Hessenberg form, giving the required $H_n = D^H A_{k+1} D$, $|D| = I$. This emphasizes that the P_i used to produce H_{i+1} from H_i is exactly the P_i used in the ith step of the QR transformation of N, a result which is not immediately obvious from the usual derivation of implicit shift techniques.

This last observation is more than a passing comment. The P_i can be designed from either H_i or N_i, so, for example, if P_i was not well defined numerically by H_i, as might be the case with very small $[\ \beta_2 \quad \gamma_2 \quad \delta_2\ ]$ above, then $[\ \tau_3 \quad \sigma_3 \quad \nu_3\ ]$ could perhaps be computed, and P_i computed from this. Such care can be important in allocating eigenvalues.

The way P_1 and P_2 are chosen in (3.6) ensures that the (6,3) element of N_3 will be zero, and the (6,2), (6,3), and (6,4) elements of H_3 will also be zero. However, all that is required here is that N be reduced in $n-1$ steps to upper triangular form, so P_1 and P_2 could be more general. As a result our illustration covers the general k shift case with unreduced upper Hessenberg A and nonsingular N. It is an easy exercise to go through the above with $k = 1$, and $N = A - \mu I$ an unreduced upper Hessenberg matrix.

Later in this paper we will indicate how the approach taken here can be used to produce other results, however it may help to summarize this comparison with the standard approach for unreduced upper Hessenberg A when no shifts are eigenvalues, since this is the setting the standard approach was designed for. The standard approach requires us first to prove that the A_{k+1} arising from explicit shifts in (2.1) is an unreduced upper Hessenberg matrix, and then to use this to prove by uniqueness that (2.6) and (2.7) essentially describe the same transformation of A, so H is effectively A_{k+1} (see the material immediately following (2.7)). This use of uniqueness allows the standard approach to omit most of the details of obtaining H from A, and in so doing holds for any number k of shifts. The present approach uses (3.5) to show the connection between the QR factorization (3.3) and the corresponding unitary similarity transformation of A in (3.4), thus showing how P_i can be derived directly from H_i instead of from N_i. However, (3.6) requires different details for different k (as does the algorithm of course). These details lengthen the description, but add to our understanding, and allow us to *derive* the implicit shift method constructively, rather than just present it as the standard approach does.

Note that this new approach is not an unravelling of the theory of uniqueness of (2.7). The standard approach proves the uniqueness of the *columns* of Q in (2.7) (see, for example, [16, pp. 352–353]), whereas the key relation $N_i A = H_i N_i$ used in this new approach shows directly how to *design* the next P_i, and this is more germane to the implicit shift computation than the *columns* of Q. In fact, this new approach still requires a uniqueness result, the uniqueness of Q in the QR factorization, whereas the standard approach just uses uniqueness of the first column of Q, and uniqueness

of the rest of Q and A_{k+1} from (2.7). The new approach provides an alternative which is both different from and more detailed than the standard approach. Both have their uses.

4. Theory for the general upper Hessenberg case. This section includes a rigorous treatment of a case not treated in §§2 and 3, that of singular N. Readers not interested in this theory can omit this section, as the remainder of the paper can be understood without it.

In the previous sections we concentrated on the case where none of the QR algorithm shifts $\mu_1, \cdots, \mu_k$ were eigenvalues of A. This ensured that $N = (A - \mu_1 I) \cdots (A - \mu_k I)$ was nonsingular, and considerably simplified the analysis. However, when we use variants of the QR algorithm to allocate eigenvalues, our shifts *are* eigenvalues, and we need the theory for this case in order to design correct algorithms. The case of general singular N has not been given in the literature for explicit or implicit shifting (for example, [5], [6] assume nonsingular N, while [16, p. 36] and [7, Thm. 7.5.1] only consider a single step of the QR algorithm with an explicit shift of one eigenvalue). We now consider what happens *mathematically* when some of the k QR algorithm shifts are eigenvalues of unreduced upper Hessenberg A.

This section will parallel §2 in first giving in Theorem 4.1, and its corollaries, the general result for the explicitly shifted QR algorithm. Then Theorem 4.4 will show how the implicitly shifted QR algorithm mimics the explicit case. The theorems will be given for unreduced upper Hessenberg A, and a comment at the end of the section will show how the results apply to upper Hessenberg matrices with possible zeros on the subdiagonal.

For completeness, the results here will be fully general for unreduced upper Hessenberg A in that there will be no restrictions at all on the shifts $\mu_1, \cdots, \mu_k$. All statements about eigenvalues will take algebraic multiplicities into account. For example, "s eigenvalues of A" means a collection of s values taken from the n eigenvalues of $n \times n$ A, and so can include repeats up to the multiplicity of the corresponding eigenvalue. Thus if 1, 1, 2, 2, 3 are the eigenvalues of 5×5 A, then the collection of values $\{1, 2, 2, 2\}$ contains exactly three eigenvalues of A, and these are 1, 2, 2. Clearly any eigenvalue of an unreduced upper Hessenberg matrix has geometric multiplicity one [7, Thm. 7.4.4]. To simplify the wording, the term "unreduced upper Hessenberg matrix" will be assumed to apply to a 1×1 matrix.

We first give the results corresponding to equations (2.1)–(2.4) for the explicitly shifted QR algorithm. Since $A = A_1$ is an unreduced upper Hessenberg matrix here, each A_i in (2.1) will be upper Hessenberg, but need not be unreduced. This may occasionally allow an arbitrary rotation in the reduction of $A_i - \mu_i I$ to R_i. To define the algorithm fully, assume the trivial rotation (the unit matrix) is used in all such cases.

The full theorem is somewhat long, as we are trying to cover all the possibly useful details. Basically it shows that if exactly s of the k shifts are eigenvalues of A then the explicitly shifted QR algorithm gives A_{k+1} with its last $s \times s$ block upper triangular with these eigenvalues on the diagonal, and its leading $n - s$ square block unreduced upper Hessenberg. Also, $Q^H N$ is upper triangular with its last s rows zero.

THEOREM 4.1. *Let $A = A_1$ be an $n \times n$ unreduced upper Hessenberg matrix, and let $\mu_1, \cdots, \mu_k$ be given complex scalars. For $i = 1, \cdots, k$ let $Q_i = P_{1,2}^{(i)} \cdots P_{n-1,n}^{(i)}$, with each $P_{j,j+1}^{(i)}$ a unitary rotation (trivial where possible) in the $(j, j+1)$ plane chosen so*

that

$$Q_i^H(A_i - \mu_i I) = R_i \tag{4.1}$$

is upper triangular. Define

$$A_{i+1} = R_i Q_i + \mu_i I, \tag{4.2}$$

which is necessarily upper Hessenberg by construction, and let

$$N = (A - \mu_1 I)\cdots(A - \mu_k I), \qquad Q = Q_1 \cdots Q_k, \qquad R = R_k \cdots R_1. \tag{4.3}$$

Assume $\{\mu_1, \cdots, \mu_k\}$ *contains exactly* s *eigenvalues of* A *and denote these by* $\nu_1, \cdots, \nu_s$ *with the ordering they had in* $\mu_1, \cdots, \mu_k$. *Then* (s *can be zero*)

$$Q^H A Q = A_{k+1} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & \underbrace{A_{22}}_{s} \end{bmatrix}, \qquad Q = \begin{bmatrix} \underbrace{Q^{(1)}}_{n-s}, & \underbrace{Q^{(2)}}_{s} \end{bmatrix}, \tag{4.4}$$

with $(n-s)\times(n-s)$ A_{11} *unreduced upper Hessenberg and* $s \times s$ A_{22} *upper triangular with* $\nu_s, \cdots, \nu_1$ *on the diagonal. If* $\nu_1, \cdots, \nu_k$ *is also a reordering of* $\mu_1, \cdots, \mu_k$ *then none of* $\nu_{s+1}, \cdots, \nu_k$ *are eigenvalues of* A_{11}. *The rows of* $Q^{(2)H}$ *span the left invariant subspace of* A *corresponding to the* s *left principal vectors of lowest grades associated with* $\nu_1, \cdots, \nu_s$. *That is, if the value* ν_i *appears with multiplicity* r_i *in* $\nu_1, \cdots, \nu_s$, *then its left principal vectors of grades* $1, 2, \cdots, r_i$ *lie in this space, but none of higher grades do.* (See, for example, [16, pp. 42–43] for a description and properties of principal vectors and their grades.)

For N, Q, *and* R *in* (4.3) *we also have*

$$Q^H N = R = \begin{bmatrix} R_{11} & R_{12} \\ \underbrace{0}_{n-s} & \underbrace{0}_{s} \end{bmatrix} \tag{4.5}$$

with $(n-s)\times(n-s)$ R_{11} *nonsingular and upper triangular, so that* N *has rank* $n-s$ *and its first* $n-s$ *columns are linearly independent.*

Proof. If $n \times n$ A_i is an unreduced upper Hessenberg matrix then in the QR factorization $Q_i^H(A_i - \mu_i I) = R_i$ no rotations are trivial and the first $n-1$ diagonal elements of R_i are nonzero. If μ_i is an eigenvalue of A_i, then the (n,n) element of R_i must be zero, so that $A_{i+1} = R_i Q_i + \mu_i I$ is unreduced upper Hessenberg in its leading principal $(n-1)\times(n-1)$ block, and has last row $(0, \cdots, 0, \nu_1)$, where $\nu_1 = \mu_i$. If μ_i is not an eigenvalue of A_i, then R_i has no zero diagonal element and the leading unreduced upper Hessenberg block of A_{i+1} (A_{i+1} itself) does not have μ_i as an eigenvalue. This gives the initial result for the following induction proof.

Now suppose $\{\mu_1, \cdots, \mu_{i-1}\}$ contains exactly s_i eigenvalues $\nu_1, \cdots, \nu_{s_i}$ of A, and

$$A_i = \begin{bmatrix} A_{11}^{(i)} & A_{12}^{(i)} \\ 0 & \underbrace{A_{22}^{(i)}}_{s_i} \end{bmatrix}, \qquad R_{i-1}\cdots R_1 = \begin{bmatrix} R_{11}^{(i)} & R_{12}^{(i)} \\ 0 & \underbrace{0}_{s_i} \end{bmatrix} \tag{4.6}$$

with $(n-s_i)\times(n-s_i)$ $A_{11}^{(i)}$ unreduced upper Hessenberg, $s_i \times s_i$ $A_{22}^{(i)}$ upper triangular with diagonal elements $\nu_{s_i}, \cdots, \nu_1$, none of the remaining $i-1-s_i$ values in $\{\mu_1, \cdots, \mu_{i-1}\}$ are eigenvalues of $A_{11}^{(i)}$, and $(n-s_i)\times(n-s_i)$ $R_{11}^{(i)}$ is nonsingular and upper triangular. This is clearly true for $s_i = 0$ or $s_i = 1$ from the previous paragraph.

Since $A_{22}^{(i)} - \mu_i I$ is already upper triangular it will not be altered in (4.1), and it follows from the previous paragraph that A_{i+1} and $R_i \cdots R_1$ will also have their forms described by (4.6), with $s_{i+1} = s_i$ if μ_i is not an eigenvalue of $A_{11}^{(i)}$, or $s_{i+1} = s_i + 1$ if it is. But μ_i can be an eigenvalue of $A_{11}^{(i)}$ only if it is an eigenvalue of A of multiplicity r_i say, and appears in $\nu_1, \cdots, \nu_{s_i}$ fewer than r_i times, in which case $\{\mu_1, \cdots, \mu_i\}$ contains exactly $s_i + 1$ eigenvalues of A and $\nu_{s_i+1} = \mu_i$. In either case, none of the $i - s_{i+1}$ elements of $\{\mu_1, \cdots, \mu_i\}$ less $\{\nu_1, \cdots, \nu_{s_{i+1}}\}$ is an eigenvalue of $A_{11}^{(i+1)}$. It follows that (4.4) holds if and only if $\{\mu_1, \cdots, \mu_k\}$ contains exactly s eigenvalues $\nu_s, \cdots, \nu_1$ of A, that these are the diagonal elements of A_{22}, and that none of the remaining $k - s$ elements of $\{\mu_1, \cdots, \mu_k\}$ is an eigenvalue of A_{11}. Also, $R = R_k \cdots R_1$ has the form in (4.5) with $(n-s) \times (n-s)$ R_{11} nonsingular and upper triangular. We have shown the form of R, but we have not shown that $Q^H N$ gives this R. This follows from (2.2) and (2.4), so (4.5) holds. It follows that N has rank $n - s$ and its first $n - s$ columns are linearly independent.

From (4.4), $Q^{(2)H} A = A_{22} Q^{(2)H}$, so let $X^{-1} A_{22} X = J_{22}$ be the Jordan canonical form of A_{22}. Define $Y^H = X^{-1} Q^{(2)H}$, so $Y^H A = J_{22} Y^H$. The rows of Y^H are then seen to be the s left principal vectors of A of lowest grades associated with $\nu_1, \cdots, \nu_s$. Of course the rows of $Q^{(2)H}$ span the same subspace as those of Y^H, which completes the proof of Theorem 4.1. □

For simplicity in presenting the proof, Theorem 4.1 assumed $\{\mu_1, \cdots, \mu_k\}$ contained exactly s eigenvalues of A and showed what followed. But since s must take just one of the values $0, 1, \cdots, n$, it follows that these are "if and only if" results, and this is stated in the following two corollaries.

COROLLARY 4.2. *Suppose A is an unreduced upper Hessenberg matrix, and* (4.1)–(4.3) *are as in Theorem* 4.1. *If the form* (4.4) *results with A_{11} an unreduced upper Hessenberg matrix, then $\{\mu_1, \cdots, \mu_k\}$ contains exactly s eigenvalues of A, and all the other results in Theorem* 4.1 *follow.* □

COROLLARY 4.3. *Suppose A is an unreduced upper Hessenberg matrix, and N is defined as in* (4.3). *If N has rank r, its first r columns are linearly independent (so $Ne_1 = 0$ if and only if $N = 0$) and exactly $s = n - r$ of the $\mu_1, \cdots, \mu_k$ are eigenvalues of A.* □

The results in Theorem 4.1 and Corollary 4.2 are for Q obtained from the explicitly shifted QR algorithm. We now want to see what happens in the implicit algorithm. Our new approach reveals this with little effort. We see from Corollary 4.3 that the leading columns of N (and so N_i in (3.3), (3.5), and (3.6)), are linearly independent. This is not only a generally valuable result, it is the key to the theoretical behaviour of the implicit algorithm.

THEOREM 4.4. *Let A be an $n \times n$ unreduced upper Hessenberg matrix, let $\mu_1, \cdots, \mu_k$ be given complex scalars, and $N_1 \equiv N = (A - \mu_1 I) \cdots (A - \mu_k I)$. Define the implicit shift algorithm with k shifts by*

$$P_1^H N_1 e_1 = e_1 \rho_1, \qquad P_1 \text{ unitary;} \tag{4.7}$$

take $H_1 \equiv A$, form $H_2 = P_1^H H_1 P_1$, and design unitary P_i to give

$$H_{i+1} = P_i^H H_i P_i, \qquad P_i^H e_1 = e_1, \qquad i = 2, 3, \cdots, n-1 \tag{4.8}$$

with the leading $i - 1$ columns of H_{i+1} having upper Hessenberg form.

Then this implicit shift algorithm mimics the result of k steps of the explicitly shifted QR algorithm in Theorem 4.1 *in the following sense:*

If $\{\mu_1, \cdots, \mu_k\}$ contains exactly s eigenvalues of A, after $r = n - s$ implicit steps, we obtain

$$P \equiv P_1 \cdots P_r = \left[\underbrace{P^{(1)}}_{r}, \underbrace{P^{(2)}}_{s}\right], \qquad P^H A P = H_{r+1} = \begin{bmatrix} H_{11} & H_{12} \\ 0 & \underbrace{H_{22}}_{s} \end{bmatrix}, \tag{4.9}$$

with $r \times r$ H_{11} being unreduced upper Hessenberg, and

$$N_{r+1} \equiv P^H N = P^H [N^{(1)}, N^{(2)}] = \begin{bmatrix} N_{11} & N_{12} \\ \underbrace{0}_{r} & \underbrace{0}_{s} \end{bmatrix} \tag{4.10}$$

with $r \times r$ N_{11} nonsingular and upper triangular. The algorithm in Theorem 4.1 *splits A_{k+1} in a similar way to H_{r+1}, and* (4.9) *and* (4.10) *parallel* (4.4) *and* (4.5) *in that*

$$P^{(1)} = Q^{(1)} D, \qquad H_{11} = D^H A_{11} D, \qquad [N_{11}, N_{12}] = D^H [R_{11}, R_{12}] \tag{4.11}$$

with $|D| = I$. The columns of $P^{(2)}$ span the same space as those of $Q^{(2)}$, the eigenvalues of H_{22} are the s shifts which are eigenvalues of A, and none of the remaining elements of $\{\mu_1, \cdots, \mu_k\}$ are eigenvalues of H_{11}.

Proof. If $s = n$, then $r = 0$, $N = 0$, and the theorem is true trivially. Now assume $s < n$, so N has rank $r = n - s > 0$, and its leading r columns are linearly independent. For $i = 1, \cdots, r$ define $N_{i+1} = P_i^H N_i$. We will first show that N_{i+1} is upper triangular in its first i columns. Our choice of P_i ensures $N_{i+1} e_1 = e_1 \rho_1$, and (3.5) holds for this N_{i+1} and H_{i+1} in (4.8), so we can use (3.5) to show that the first two columns of N_{i+1} are upper triangular, and so on. The idea is illustrated by (3.6), which assumes $n = 6$ and $k = 2$. When P_2 has transformed H_2 so the first column of H_3 has upper Hessenberg form, the first column of (3.6) shows the first two columns of N_3 are in upper triangular form, and so on. This proof that the *usage* is correct is just the counterpart of our earlier *derivation* of the implicit shift algorithm. Thus if $r > 2$ (in fact, $r \geq 4$ if $k = 2$ and $n = 6$), the second column of (3.6) gives $\rho_2 \neq 0$ and

$$\alpha_2 [\ \tau_3 \quad \sigma_3 \quad \nu_3\ ] = \rho_2 [\ \beta_2 \quad \gamma_2 \quad \delta_2\ ]$$

is also nonzero since α_2 is, and $[\ \tau_3 \quad \sigma_3 \quad \nu_3\ ]$ is since the first three columns of N_3 are linearly independent. This shows $[\ \beta_2 \quad \gamma_2 \quad \delta_2\ ]^T$ is nonzero, and P_3 is effectively uniquely designed to transform this to a multiple of e_1. However, if $r = 2$ (imagine (3.6) with $k \geq 4$), then $[\ \tau_3 \quad \sigma_3 \quad \nu_3\ ] = 0$, so $[\ \beta_2 \quad \gamma_2 \quad \delta_2\ ] = 0$. In either case, we see we that eventually obtain the form in (4.9) and (4.10), and that these are true in general.

We see in (4.10) that $N^{(1)}$ has linearly independent columns, and $P^H N^{(1)}$ has upper triangular form. But $Q^H N^{(1)}$ has the same form in (4.5), so by uniqueness of the QR factorization, $P^{(1)} = Q^{(1)} D$, $|D| = I$, and the rest of the theorem follows from (4.9), (4.10), and Theorem 4.1. □

Because the (2,1) block in H_{r+1} in (4.9) is zero, P_{r+1}, which is meant to make the first r columns of H_{r+2} upper Hessenberg, is clearly arbitrary to an obvious extent. At this point we say the standard implicit shift algorithm breaks down in that the relationship with the explicit shift algorithm is lost. However, if we are computing eigenvalues, it has done its job in separating H_{22} with known eigenvalues from H_{11} with unknown eigenvalues, and we would usually stop here. Note that N_{r+1} has its last s rows zero, and H_{r+1} has $\eta_r = 0$ in (3.6), so (3.5) and (3.6) tell us nothing about H_{22}, which is understandable because H_{22} is not completely defined by this implicit shift algorithm.

We have shown that when the set of shifts $\{\mu_1, \cdots, \mu_k\}$ contains exactly s eigenvalues of A, the first $n - s$ steps of the implicit shift algorithm are uniquely defined, and the relations in (3.5) play a key role to this point, both in justifying and understanding the algorithm. These $n - s$ implicit steps mimic the result of the full k explicit steps to the extent shown in the theorem, as would be hoped. For algorithms which find eigenvalues it is useful to know this rigorously and to have a completely general result. That N of rank r has its first r columns linearly independent leading to (4.5) was unexpected, and the effect of *repeated* shifts corresponding to eigenvalues of A was not initially clear to the authors. We expect the resulting structure to be extremely useful in the design of algorithms.

A general upper Hessenberg matrix A may have some zero $(j + 1, j)$ elements, so that A may be partitioned to be block upper triangular with each submatrix on the diagonal being unreduced upper Hessenberg. Applying the *explicitly* shifted QR algorithm to A results in applying the algorithm with the same shifts to each such submatrix. But the implicitly shifted algorithm "breaks down" when the first zero $\alpha_{j+1,j}$ is encountered. For example, if $\alpha_2 = 0$ in (3.6), the second column tells us nothing about the third column of N_3, and in fact we cannot design P_3 successfully from H_3 since β_2, γ_2, and δ_2 will all be zero. In a sense this is acceptable, since only the first two rows of N_3 and H_3 will differ from N and A, respectively, so H_3 will already be upper Hessenberg, and the algorithm could stop. But the only way to mimic the explicit step would be to form the necessary part of the third column of N_3 (i.e., of N) and design P_3 on this. Of course, in the QR algorithm for finding eigenvalues, we welcome such splitting, and proceed to treat each submatrix on the diagonal separately.

For a matrix which splits, the first theorem and its corollaries consider the explicitly shifted QR algorithm, and so these apply to *each* unreduced upper Hessenberg matrix on the diagonal. Theorem 4.4 considers the implicit algorithm using the first column of N only, so it only applies to the first unreduced upper Hessenberg block.

5. Block and band QR algorithms. The explicitly shifted QR algorithm and its theory in (2.1)–(2.4) hold for $n \times n$ A of any structure, but there is no generally available extension of the standard *implicit* shift theory to other than Hessenberg matrices. Fortunately the new theory for implicitly shifted algorithms in (3.1)–(3.5) extends easily to other matrices. Here we illustrate how to use this theory to design implicit shift algorithms for two classes of matrices. First we consider t-unreduced upper Hessenberg (t-uuh) A, meaning A has $\alpha_{t+j,j}$ nonzero, $j = 1, \cdots, n - t$, with all subdiagonals below that being zero. Then we consider block upper Hessenberg matrices. This class is quite general and includes t-uuh matrices, but the t-uuh case is important in illustrating what we must do to start all these more general cases, that is, it shows us the equivalent of the $P_1^H N e_1 = e_1 \rho_1$ step in the ordinary upper Hessenberg case (1-uuh).

Variants of the implicitly shifted QR algorithm for block upper Hessenberg A can be used to solve eigenvalue allocation problems, and may be useful for parallel implementations of block QR algorithms. Examples are all that are required to show how effective the approach of §3 is, but in any particular problem the relevant version of (3.6) must be analyzed. For simplicity, in these examples we will assume none of the shifts are eigenvalues, so that N is nonsingular, and only consider single and double shift algorithms.

The general approach is to realize that Q from the explicit shift algorithm gives the QR factorization of N in (2.4), so Q is effectively unique when no shifts are

eigenvalues. Next consider obtaining this unique QR factorization from N as in (3.2), and try to show for $i > m$, for some $m \geq 1$, how each P_i in (3.2) may also be developed from the H_i in (3.4) (rather than the N_i in (3.3)) by looking at a detailed version of (3.5). These P_i will be designed to bring H_n back to the original form of A, and the detailed version of (3.5) will show that such P_i also produce the required reduction in (3.3), and so give the unique $P_1 \cdots P_{n-1} = QD, |D| = I$, and $H_n = D^H A_{k+1} D$. Of course the P_i will often be designed to make certain *blocks* zero in the n-block by n-block case. See [15] for the use of block reflectors to do this, or [9] for the possible use of unsymmetric generalizations of Householder transformations.

Consider a 5×5, 2-uuh matrix A with a single shift μ_1, so $k = 1$ and $N = N_1 = A - \mu_1 I$. After designing P_1 via $P_1^H N e_1 = e_1 \rho_1$ and applying it, (3.5) becomes

$$\underbrace{\begin{bmatrix} \rho_1 & \times & \times & \times & \times \\ & \tau_2 & \times & \times & \times \\ & \sigma_2 & \times & \times & \times \\ & \alpha_2 & \times & \times & \times \\ & & \alpha_3 & \times & \times \end{bmatrix}}_{N_2} \underbrace{\begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \alpha_1 & \times & \times & \times & \times \\ & \alpha_2 & \times & \times & \times \\ & & \alpha_3 & \times & \times \end{bmatrix}}_{A}$$

$$= \underbrace{\begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \square & \times & \times & \times & \times \\ \square & \square & \times & \times & \times \end{bmatrix}}_{H_2} \underbrace{\begin{bmatrix} \rho_1 & \times & \times & \times & \times \\ & \tau_2 & \times & \times & \times \\ & \sigma_2 & \times & \times & \times \\ & \alpha_2 & \times & \times & \times \\ & & \alpha_3 & \times & \times \end{bmatrix}}_{N_2},$$

where the $\square$ denote elements introduced into $H_2 = P_1^H A P_1$. Note that the first column does *not* show us how to design P_2 from H_2, because α_1 in A picks up the third, not second, column of N_2. Thus we still have to design P_2 on N_2, but after applying this we have

$$\underbrace{\begin{bmatrix} \rho_1 & \times & \times & \times & \times \\ & \rho_2 & \times & \times & \times \\ & & \tau_3 & \times & \times \\ & & \nu_3 & \times & \times \\ & & \alpha_3 & \times & \times \end{bmatrix}}_{N_3} \underbrace{\begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \alpha_1 & \times & \times & \times & \times \\ & \alpha_2 & \times & \times & \times \\ & & \alpha_3 & \times & \times \end{bmatrix}}_{A}$$

$$= \underbrace{\begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \beta_1 & \times & \times & \times & \times \\ \gamma_1 & \times & \times & \times & \times \\ \delta_1 & \times & \times & \times & \times \end{bmatrix}}_{H_3} \underbrace{\begin{bmatrix} \rho_1 & \times & \times & \times & \times \\ & \rho_2 & \times & \times & \times \\ & & \tau_3 & \times & \times \\ & & \nu_3 & \times & \times \\ & & \alpha_3 & \times & \times \end{bmatrix}}_{N_3}.$$

The first column on each side shows us we can design P_3 on either N_3 or H_3 to give the required implicit computation. Clearly, from this point we can carry out the computation on the H_i without reference to the N_i.

The point here is that for t-uuh A, the transformation matrices $P_1, \cdots, P_t$ must be designed from N, whether by forming the first t columns of N and upper triangularizing this part, or by some more subtle procedure designed to improve speed or accuracy. The remaining P_i, $i = t+1, \cdots, n-1$ may then be designed from and applied to the H_i, giving the implicit shift algorithm. Note that this statement is

independent of k, the number of shifts used. We see that increasing k can increase the complexity of each P_i, but does not alter the structure of A, and the position of the last nonzero element in the first column of A is what determines the number of transformations t that must be designed on N.

Finally we consider the general case of *block* upper Hessenberg A, with initially no restriction on the block structure other than that the blocks on the diagonal be square. Here we use $H_{i,j}^{(s)}$ to represent the (i,j) block of H_s. For example, a 6-block by 6-block A which is 9×9 may have block and scalar representations

$$\begin{bmatrix} \times & \times & \times & \times & \times & \times \\ A_{2,1} & \times & \times & \times & \times & \times \\ & A_{3,2} & \times & \times & \times & \times \\ & & A_{4,3} & \times & \times & \times \\ & & & A_{5,4} & \times & \times \\ & & & & A_{6,5} & \times \end{bmatrix}$$

$$(5.1) \qquad = \left[\begin{array}{cccc|cc|ccc} \times & \times & \times & \times & \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times & \times & \times & \times & \times \\ \hline & \times & \times & \times & \times & \times & \times & \times & \times \\ & \times & \times & \times & \times & \times & \times & \times & \times \\ \hline & & & & \times & \times & \times & \times & \times \\ \hline & & & & \times & \times & \times & \times & \times \\ \hline & & & & & & \times & \times & \times \\ & & & & & \times & \times & \times & \times \\ \hline & & & & & & \times & & \times \end{array}\right],$$

where so far we have made no restrictions on the ranks of the $A_{j+1,j}$, in fact, $A_{3,2}$ is zero above. Here N, and each N_i, will have 6-block by 6-block structure, and each H_i will have the same block structure as A. In general we let each of $A_{j,j}$, $N_{j,j}^{(i)}$, $H_{j,j}^{(i)}$ be $m_j \times m_j$, $j = 1, \cdots, n$, so A is an $m \times m$ matrix with $m = m_1 + \cdots + m_n$.

Once again we assume that no shift is an eigenvalue, so that each N_i is nonsingular, and we give a small example which represents a computation for any number k of shifts. We consider a 4-block by 4-block case, where some of the indicated subdiagonal blocks of N_i and H_i may be zero if k is small enough. If $m \times m_j$ $N_j^{(i)}$ represents the jth block of columns of N_i, then the initial step is to design a unitary matrix P_1 so that

$$(5.2) \quad P_1^H N_1^{(1)} = \begin{bmatrix} R_1 \\ 0 \end{bmatrix}, \qquad R_1 \text{ nonsingular, and usually upper triangular.}$$

This P_1, applied to give $H_2 = P_1^H A_1 P_1$, and $N_2 = P_1^H N$ in theory, results in (3.5) of the form

$$\underbrace{\begin{bmatrix} R_1 & \times & \times & \times \\ & N_{2,2}^{(2)} & \times & \times \\ & N_{3,2}^{(2)} & \times & \times \\ & N_{4,2}^{(2)} & \times & \times \end{bmatrix}}_{N_2} \underbrace{\begin{bmatrix} \times & \times & \times & \times \\ A_{2,1} & \times & \times & \times \\ & A_{3,2} & \times & \times \\ & & A_{4,3} & \times \end{bmatrix}}_{A}$$

$$= \underbrace{\begin{bmatrix} \times & \times & \times & \times \\ H_{2,1}^{(2)} & \times & \times & \times \\ H_{3,1}^{(2)} & \times & \times & \times \\ H_{4,1}^{(2)} & \times & \times & \times \end{bmatrix}}_{H_2} \underbrace{\begin{bmatrix} R_1 & \times & \times & \times \\ & \times & \times & \times \\ & \times & \times & \times \\ & \times & \times & \times \end{bmatrix}}_{N_2},$$

and from the first column, by defining $\tilde{N}_2^{(2)}$ and $\tilde{H}_1^{(2)}$ we have

$$\tilde{N}_2^{(2)} A_{2,1} = \begin{bmatrix} N_{2,2}^{(2)} \\ N_{3,2}^{(2)} \\ N_{4,2}^{(2)} \end{bmatrix} A_{2,1} = \begin{bmatrix} H_{2,1}^{(2)} \\ H_{3,1}^{(2)} \\ H_{4,1}^{(2)} \end{bmatrix} R_1 = \tilde{H}_1^{(2)} R_1. \tag{5.3}$$

If $A_{2,1}$ has full row rank m_2, then since N_2 is nonsingular we see $\tilde{N}_2^{(2)}$ has rank m_2 and R_1 is nonsingular, so that $\tilde{H}_1^{(2)}$ has rank m_2. But then

$$\tilde{N}_2^{(2)} = \tilde{H}_1^{(2)} R_1 A_{2,1}^T (A_{2,1} A_{2,1}^T)^{-1},$$

and a unitary matrix $\tilde{P}_2$ (the relevant part of P_2) that makes all but the first m_2 rows of $\tilde{P}_2^H \tilde{H}_1^{(2)}$ zero, will necessarily do the same for $\tilde{N}_2^{(2)}$. But if R_1 is upper triangular, and we also require R_2 to be upper triangular, then $A_{2,1}$ must come into the computation. The easiest approach is to insist

$$A_{2,1} = [0, \tilde{A}_{2,1}]$$

with $\tilde{A}_{2,1}$ nonsingular and upper triangular. The first $m_1 - m_2$ columns of $\tilde{H}_1^{(2)}$ will then be zero, and $\tilde{P}_2^H$, which transforms the last m_2 columns of $\tilde{P}_2^H \tilde{H}_1^{(2)}$ to an upper triangle in the first m_2 rows and zero elsewhere, will do the same to $\tilde{N}_2^{(2)}$, which is the required form.

Note that a block variant of the QR algorithm, which only gives block upper triangular R_i in (2.1) and N_n in (3.2), is also possible, but then $P_1 \cdots P_{n-1} = QD$ would only be unique up to a block diagonal unitary matrix D.

On the other hand, if $A_{2,1}$ has rank $r_2 < m_2$, then it is impossible to define the desired P_2 by working with $\tilde{H}_1^{(2)}$ alone. For example, suppose R_1 is upper triangular and

$$A_{2,1} = \begin{bmatrix} 0 & \tilde{A}_{2,1} \\ 0 & 0 \end{bmatrix}, \qquad r_2 \times r_2 \qquad \tilde{A}_{2,1} \text{ upper triangular.}$$

From (5.3) the first $m_1 - r_2$ columns of $\tilde{H}_1^{(2)}$ will be zero, and the matrix that transforms the last r_2 columns of $\tilde{H}_1^{(2)}$ in the usual way will also give the correct result for the first r_2 columns of $\tilde{N}_2^{(2)}$, but not for the remaining $m_2 - r_2$ columns. For this particular block structure these last $m_2 - r_2$ columns of $\tilde{N}_2^{(2)}$ will have to be formed and used to complete the design of P_2 (or some equivalent computation).

When P_2 has been designed, it can be applied to give $H_3 = P_2^H H_2 P_2$, etc. But then we will have a similar equation to (5.3) to use in the design of P_3, and a similar argument will hold here and in each succeeding step. It follows that for the implicit computation to be carried out on H_i without recourse to N_i, $i > 1$, we need $A_{i,i-1}$ to have full row rank. This was assured by each $A_{i,i-1}$ being $t \times t$ and nonsingular in our earlier t-uuh example. If $m_i \times m_{i-1}$ $A_{i,i-1}$ has rank $r_i < m_i$, then we need to consider the equivalent of $m_i - r_i$ columns of N_i, as well as the relevant columns of H_i in order to design P_i correctly to give H_{i+1} and, in theory, N_{i+1}. This can always be done, and is simple and cheap when $k = 1$, but becomes more complicated and expensive as k increases. A familiar example is upper Hessenberg A with some $\alpha_{i,i-1} = 0$ so that the matrix splits. Here $m_i = 1$, $r_i = 0$, and an extra $m_i - r_i = 1$ column of N_i is needed to continue the implicit algorithm.

6. Conclusions and suggestions. This paper has presented two new pieces of work. Section 3 described a new approach to the theory and development of implicit shifting, and illustrated it with the QR algorithm applied to an unreduced upper Hessenberg matrix A. The power of this approach was exhibited in §4, by showing how it handled the fully general Hessenberg case, and in §5, by showing how it could be used to derive implicit shift algorithms for more general matrices A, in particular block Hessenberg A of any form having square diagonal blocks. This work could be continued to show how this approach can be used to develop implicitly shifted algorithms for related methods applied to more general matrices than Hessenberg. Another direction is to show how this approach can also be used to develop implicitly shifted algorithms where the intermediate factorizations do not necessarily give upper triangular matrices, such as was done for the QS algorithm in [11] and [12] to allocate eigenvalues in a multi-input linear constant coefficient control system with state feedback.

The second piece of work was given in §4, where the theory for the shifted QR algorithm applied to upper Hessenberg A was extended to cover the case where any of the k shifts could be eigenvalues. This was done for both the explicit and implicit shift algorithms. It is needed, for example, to complete the theory for eigenvalue allocation algorithms working with Hessenberg A, where shifts *are* eigenvalues of the matrix being designed—unlike the usual case of finding eigenvalues by the QR algorithm. An obvious continuation of this work would be to give the equivalent theory for shifted QR algorithms (with no restrictions on the shifts) applied to the more general matrices A (t-uuh and block upper Hessenberg) dealt with in §5. The equivalent theory could also be developed for other than the QR algorithm, wherever such results are found to be useful.

Acknowledgments. We would like to thank Linda Kaufman and the referees for their helpful comments.

REFERENCES

[1] Z. Bai and J. Demmel, *On a block implementation of Hessenberg multishift* QR *iteration*, Internat. J. High Speed Comput., 1 (1989), pp. 97–121.

[2] A. Bunse-Gerstner, *An analysis of the* HR *algorithm for computing the eigenvalues of a matrix*, Linear Algebra Appl., 35 (1981), pp. 155–173.

[3] A. Bunse-Gerstner, R. Byers, and V. Mehrmann, *The quaternion* QR *algorithm*, Numer. Math., 55 (1989), pp. 83–95.

[4] R. Byers, *A Hamiltonian* QR *algorithm*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 212–229.

[5] J. G. F. Francis, *The* QR *transformation. A unitary analogue to the* LR *transformation, Part* 1, Comput. J., 4 (1961), pp. 265–271,

[6] ———, *The* QR *transformation. A unitary analogue to the* LR *transformation, Part* 2, Comput. J., 4 (1962), pp. 332–345.

[7] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Second Edition, The Johns Hopkins University Press, Baltimore, MD, 1989.

[8] L. Kaufman, *The* LZ*-algorithm to solve the generalized eigenvalue problem*, SIAM J. Numer. Anal., 11 (1973), pp. 997–1024.

[9] ———*The generalized Householder transformation and sparse matrices*, Linear Algebra Appl., 90 (1987), pp. 221–234.

[10] G. S. Miminis and C. C. Paige, *Implicit shifting in the* QR *and related algorithms,* talk given at the Xth Gatlinburg Meeting on Numerical Linear Algebra, Fairfield Glade, TN, October 19–23, 1987.

[11] G. S. Miminis, *Numerical algorithms for the pole placement problem*, Ph.D. dissertation, Computer Science Department, McGill University, Montreal, Canada, 1985.

[12] G. S. Miminis and C. C. Paige, *A double step algorithm for pole assignment of time invariant multi-input linear systems using state feedback*, in preparation.

[13] C. B. MOLER AND G. W. STEWART, *An algorithm for generalized matrix eigenvalue problems*, SIAM J. Numer. Anal., 10 (1973), pp. 241–256.
[14] H. RUTISHAUSER, *Solution of eigenvalue problems with the* LR*-transformation*, Appl. Math. Ser. Nat. Bur. Stand., 49 (1958), pp. 47–81.
[15] R. SCHREIBER AND B. N. PARLETT, *Block reflectors: Theory and computation*, SIAM J. Numer. Anal., 25 (1987), pp. 189–205.
[16] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, U.K., 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 401–425, July 1991

THE RESTRICTED SINGULAR VALUE DECOMPOSITION: PROPERTIES AND APPLICATIONS*

BART L. R. DE MOOR† AND GENE H. GOLUB‡

Abstract. The *restricted singular value decomposition* (RSVD) is the factorization of a given matrix, relative to two other given matrices. It can be interpreted as the *ordinary singular value decomposition* with different inner products in row and column spaces. Its properties and structure, as well as its connection to generalized eigenvalue problems, canonical correlation analysis, and other generalizations of the singular value decomposition, are investigated in detail.

Applications that are discussed include the analysis of the extended shorted operator, unitarily invariant norm minimization with rank constraints, rank minimization in matrix balls, the analysis and solution of linear matrix equations, rank minimization of a partitioned matrix, and the connection with generalized Schur complements, constrained linear and total linear least squares problems with mixed exact and noisy data, including a generalized Gauss–Markov estimation scheme.

Key words. generalized SVD, generalized matrix inverses, (total) linear least squares, (generalized) Schur complements, matrix balls, shorted operator

AMS(MOS) subject classifications. 15A09, 15A18, 15A21, 15A24, 65F20

1. Introduction.

The *ordinary singular value decomposition* (OSVD) has a long history with original contributions by Beltrami (1873) [2], Sylvester (1889) [26], Autonne (1902) [1], Eckart and Young (1936) [12] and many others (see, e.g., the references in [15], [21], [27]). It has become an important tool in the analysis and numerical solution of numerous problems arising in such diverse applications as psychometrics, statistics, signal processing, and system theory. Not only does it allow for an elegant problem formulation, but at the same time it provides geometrical and algebraic insight together with an immediate numerically robust implementation [15].

Recently, several generalizations to the OSVD have been proposed and their properties analysed. The one that is best known is the *generalized* SVD as introduced by Paige and Saunders in 1981 [22], which we propose to rename as the *Quotient* SVD (QSVD) [8]. Another example is the *Product* SVD (PSVD) as proposed by Fernando and Hammarling in 1987 [14] and further analysed in [10]. The third one is the *Restricted* SVD (RSVD), introduced in its explicit form by Zha in [32] and further developed and discussed in this paper. In [8] we have proposed a standardized nomenclature for the singular value decomposition and its generalizations. This set of names has the advantage of being *alphabetic* and *mnemonic*, **O-P-Q-R-SVD**. For the structure and properties of the OSVD, PSVD, and QSVD, we also refer to [8].

The RSVD, which is the main subject of this paper, applies for a given triplet of matrices A, B, C of compatible dimensions (Theorem 1). In essence, the RSVD provides a factorization of the matrix A, relative to matrices B and C. It could be

* Received by the editors June 8, 1989; accepted for publication (in revised form) September 12, 1990. Part of this work was supported by the United States Army under contract DAAL03-87-K-0095.

† Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven, Kardinaal Mercierlaan 94 B-3001, Leuven, Belgium (demoor@esat.kuleuven.ac.be). This author was a visiting research associate at the Computer Science Department and the Department of Electrical Engineering (Information Systems Laboratory) of Stanford University, where he was supported by an Advanced Research Fellowship in Science and Technology of the North Atlantic Treaty Organization (NATO) Science Fellowships Program and by a grant from IBM. He is now a research associate of the Belgian National Fund for Scientific Research (NFWO).

‡ Department of Computer Science, Stanford University, Stanford, California 94305 (na.golub@na-net.stanford.edu).

considered as the OSVD of the matrix A, but with different (possibly nonnegative-definite) inner products in its column and in its row space. It will be shown that the RSVD not only allows for an elegant treatment of algebraic and geometric problems in a wide variety of applications, but that its structure provides a powerful tool in simplifying proofs and derivations that are algebraically rather complicated.

Soon after the present paper was completed, Zha and de Moor discovered that the RSVD is only one of the three possible SVD-like factorizations for three matrices. Similar generalizations of the OSVD are not only limited to two or three matrices, but can be derived for 4, 5, $\cdots$, i.e., any number of matrices of compatible dimensions. The PSVD and the QSVD serve as basic building blocks in this infinite tree of generalizations of the OSVD. For instance, the RSVD which is analysed in this paper can also be considered as a double QSVD. This is the reason why we have called it the **QQ-SVD** in [11], where the complete structure of this tree of generalizations is also developed in detail.

This paper is organised as follows. In §2, the main structure of the RSVD is analysed in terms of the ranks of the concatenation of certain matrices. The factorization is related to a generalized eigenvalue problem (§2.2.1). A variational characterization is provided in §2.2.2. A generalized dyadic decomposition is explored in §2.2.3 together with a geometrical interpretation. It is shown how the RSVD contains other generalizations of the OSVD, such as the PSVD and the QSVD, as special cases in §2.2.4. In §3, several applications are discussed:

- *Rank minimization* and the *extended shorted operator* are the subject of §3.1, as well as *unitarily invariant norm minimization with rank constraints* and the relation with *matrix balls*. We also investigate a certain linear matrix equation which is directly related to the Moore–Penrose pseudo-inverse of a matrix.
- The *low rank approximation of a partitioned matrix* when only one of its blocks can be modified is explored in §3.2, together with *total least squares with mixed exact and noisy data and linear constraints*. While the role of the Schur complement and its close connection to least squares estimation is well understood, it will be shown in this section that there exists a similar relation between constrained total linear least squares solutions and a *generalized Schur complement*.
- *Generalized Gauss–Markov models*, possibly with constraints, are discussed in §3.3 and it is shown how the RSVD simplifies the solution of linear least squares problems with constraints.

In §4 the main conclusions are presented together with some perspectives. Let us conclude this Introduction by referring to the reports mentioned in [9] for a detailed constructive proof of the main theorem of this paper.

Notation, conventions, and abbreviations. Throughout the paper, capitals denote matrices. The lower case letters $i, j, k, l, m, n, p, q, r$ are nonnegative integers. Other lower case letters denote vectors. The set of real numbers is denoted by $\Re$. Scalars (possibly complex) are denoted by Greek letters. The matrices A $(m \times n)$, B $(m \times p)$, C $(q \times n)$ are given matrices. Their ranks will be denoted by r_a, r_b, r_c. D is a $p \times q$ matrix. M is the matrix with A, B, C, D^* as its blocks: $M = \left(\begin{smallmatrix} A & B \\ C & D^* \end{smallmatrix}\right)$. We shall also frequently use the following ranks: $r_{ac} = \operatorname{rank}\left(\begin{smallmatrix} A \\ C \end{smallmatrix}\right), r_{abc} = \operatorname{rank}\left(\begin{smallmatrix} A & B \\ C & 0 \end{smallmatrix}\right)$, $r_{ab} = \operatorname{rank}(\begin{smallmatrix} A & B \end{smallmatrix})$. A^t is the transpose of a (possibly complex) matrix A and $\overline{A}$ is the complex conjugate of A. A^* denotes the complex conjugate transpose of a (complex) matrix: $A^* = \overline{A}^t$. The matrix A^{-*} represents the inverse of A^*. I_k is the $k \times k$

identity matrix. The subscript is omitted when the dimensions are clear from the context. Identity vectors with the ith component equal to 1 and all others zero, are denoted by e_i $(m \times 1)$. A matrix X is called an $A(i, j, \cdots)$-inverse of the matrix A if it satisfies equation $i, j, \cdots$ of the following:

1. $AXA = A$,
2. $XAX = X$,
3. $(AX)^* = AX$,
4. $(XA)^* = XA$.

An $A(1)$ inverse is also called an inner inverse and denoted by A^-. The $A(1,2,3,4)$ inverse is the Moore–Penrose pseudo-inverse denoted by A^+ and it is unique. We shall also need the following lemmas.

LEMMA 1 (inner inverse of a factored matrix). *Let the matrix A be factored as*

$$A = P^{-*} \begin{pmatrix} D_a & 0 \\ 0 & 0 \end{pmatrix} Q^{-1}$$

where D_a is square $r_a \times r_a$ nonsingular. Then, every inner inverse A^- can be written as

$$A^- = Q \begin{pmatrix} D_a^{-1} & Z_{12} \\ Z_{21} & Z_{22} \end{pmatrix} P^* \tag{1}$$

where Z_{12}, Z_{21}, Z_{22} are arbitrary matrices. Conversely, every matrix A^- of this form is an inner inverse of A.

For a detailed discussion of generalized inverses, we refer to [21]. The matrices U_a $(m \times m)$, V_a $(n \times n)$, V_b $(p \times p)$, U_c $(q \times q)$ are unitary, i.e., $U_a U_a^* = I_m = U_a^* U_a, V_a V_a^* = I_n = V_a^* V_a$, $V_b V_b^* = I_p = V_b^* V_b$, $U_c U_c^* = I_q = U_c^* U_c$. The matrices P $(m \times m)$ and Q $(n \times n)$ are square nonsingular. The nonzero elements of the diagonal matrices S_1, S_2, and S_3, which appear in the theorems, are denoted by α_i, β_i, and γ_i. The vector a_i denotes the ith column of the matrix A. The range (column space) of the matrix A is denoted by $\mathbf{R}(A) = \{y | y = Ax\}$. The row space of A is denoted by $\mathbf{R}(A^*)$. The null space of the matrix A is represented as $\mathbf{N}(A) = \{x | Ax = 0\}$. The symbol $\bigcap$ denotes the intersection of two vector spaces. We shall use the following well-known result.

LEMMA 2 (the dimension of the intersection of subspaces).

$$\dim(\mathbf{R}(A) \bigcap \mathbf{R}(B)) = r_a + r_b - r_{ab}$$

$$\dim(\mathbf{R}(A^*) \bigcap \mathbf{R}(C^*)) = r_a + r_c - r_{ac}.$$

$\|A\|$ is any unitarily invariant matrix norm while $\|A\|_F$ is the Frobenius norm: $\|A\|_F^2 = \text{trace}(AA^*)$. The norm of the vector a is denoted by $\|a\|_2$ where $\|a\|_2^2 = a^* a$. Moreover, we will adopt the following convention for block matrices: Any (possibly rectangular) block of zeros is denoted by 0, the precise dimensions being obvious from the block dimensions. The symbol I represents a matrix block corresponding to the square identity matrix of appropriate dimensions. Whenever a dimension indicated by an integer in a block matrix is zero, the corresponding block row or block column should be omitted and all expressions and equations in which a block matrix of that block row or block column appears, can be disregarded. An equivalent formulation would be that we allow $0 \times n$ or $n \times 0$ $(n \neq 0)$ blocks to appear in matrices. This permits an elegant treatment of several cases at once. Finally, we would like to introduce the term *quasi-diagonal* matrix for a matrix, the block rows and block columns of which are a permutation of a diagonal matrix.

2. The restricted singular value decomposition (RSVD). The idea of a generalization of the OSVD for three matrices is implicit in the S,T-singular value decomposition of Van Loan [30] via its relation to a generalized eigenvalue problem. Zha [32] introduced an explicit formulation of the RSVD constructing it through the use of several OSVDs and QSVDs (see also [9]). For the sake of brevity, we have omitted our constructive proof based on a sequence of OSVDs and PSVDs. It can be found in [9]. In this section, we first state the main theorem (§2.1), which describes the structure of the RSVD, followed by a discussion of the main properties in §2.2, including the connection to generalized eigenvalue problems, a generalized dyadic decomposition, geometrical insights, and the demonstration that the RSVD contains the OSVD, the PSVD, and the QSVD as special cases.

2.1. The RSVD theorem. With the notation and conventions of §1, we have the following theorem.

THEOREM 1 (the restricted singular value decomposition). *Every triplet of matrices A $(m \times n)$, B $(m \times p)$, and C $(q \times n)$ can be factorized as*

$$A = P^{-*} S_a Q^{-1},$$
$$B = P^{-*} S_b V_b^{*},$$
$$C = U_c S_c Q^{-1},$$

where P $(m \times m)$ and Q $(n \times n)$ are square nonsingular, and V_b $(p \times p)$ and U_c $(q \times q)$ are unitary. S_a $(m \times n)$, S_b $(m \times p)$, and S_c $(q \times n)$ are real quasi-diagonal matrices with nonnegative elements and the following block structure:

$$\begin{pmatrix} S_a & S_b \\ S_c & \end{pmatrix} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} S_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \\ \\ \begin{pmatrix} I & 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & S_3 & 0 \end{pmatrix} \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & S_2 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ \\ \end{array}$$

The block dimensions of the matrices S_a, S_b, S_c are the following.

	Block columns of S_a and S_c	*Block columns of S_b*
1.	$r_{abc} + r_a - r_{ac} - r_{ab}$	$r_{abc} + r_a - r_{ac} - r_{ab}$
2.	$r_{ab} + r_c - r_{abc}$	$r_{ac} + r_b - r_{abc}$
3.	$r_{ac} + r_b - r_{abc}$	$p - r_b$
4.	$r_{abc} - r_b - r_c$	$r_{ab} - r_a$
5.	$r_{ac} - r_a$	
6.	$n - r_{ac}$	
	Block rows of S_a and S_b	*Block rows of S_c*
1.	$r_{abc} + r_a - r_{ab} - r_{ac}$	$r_{abc} + r_a - r_{ab} - r_{ac}$
2.	$r_{ab} + r_c - r_{abc}$	$r_{ab} + r_c - r_{abc}$
3.	$r_{ac} + r_b - r_{abc}$	$q - r_c$
4.	$r_{abc} - r_b - r_c$	$r_{ac} - r_a$
5.	$r_{ab} - r_a$	
6.	$m - r_{ab}$	

The matrices S_1, S_2, S_3 are square nonsingular diagonal with positive diagonal elements.

Let $\alpha_i, \beta_j, \gamma_k$ be the diagonal elements of the matrices S_1, S_2, S_3. We propose to call the following triplets of numbers the *restricted singular value triplets*:

- $r_{abc} + r_a - r_{ab} - r_{ac}$ triplets of the form $(\alpha_i, 1, 1)$ with $\alpha_i > 0$. By convention, they will be ordered as

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_{r_{abc}+r_a-r_{ab}-r_{ac}} > 0.$$

- $r_{ab} + r_c - r_{abc}$ triplets of the form $(1, 0, 1)$.
- $r_{ac} + r_b - r_{abc}$ triplets of the form $(1, 1, 0)$.
- $r_{abc} - r_b - r_c$ triplets of the form $(1, 0, 0)$.
- $r_{ab} - r_a$ triplets of the form $(0, \beta_j, 0)$, $\beta_j > 0$ (elements of S_2).
- $r_{ac} - r_a$ triplets of the form $(0, 0, \gamma_k)$, $\gamma_k > 0$ (elements of S_3).
- $\min(m - r_{ab}, n - r_{ac})$ trivial triplets $(0, 0, 0)$.

We propose to call the factorization of a matrix triplet, as described in Theorem 1, the *restricted* singular value decomposition because the RSVD allows us to analyse matrix problems that can be stated in terms of the matrices $A + BDC$ and

$$M = \begin{pmatrix} A & B \\ C & D^* \end{pmatrix},$$

in which the matrices B and C represent certain *restrictions* on the type of operations that are allowed. Typically, we are interested in the ranks of these matrices as the matrix D is modified. The rank of the matrix $A + BDC$ can only be reduced by modifications that belong to the column space of B and the row space of C. It will be shown how the rank of M can be analysed via a generalized Schur complement, which is of the form $D^* - CA^-B$, where again, C and B represent certain restrictions and A^- is an inner inverse of A. Moreover, the RSVD yields the restriction of the linear operator represented by the matrix A to the column space of B and the row space of C. Finally, the RSVD can be interpreted as an OSVD but with certain restrictions on the inner products to be used in the column and row space of the matrix A (see §2.2.1).

Some algorithmic issues related to the RSVD are discussed in [11], [13], [29], [28], and [33], though a full portable and documented algorithm for the RSVD is still to be developed.

2.2. Properties of the RSVD. The OSVD, as well as the PSVD and the QSVD, can all be related to a certain (generalized) eigenvalue problem. It comes as no surprise that this is also the case for the RSVD. First, the generalized eigenvalue problem for the RSVD will be analysed in §2.2.1 and we shall point out an interesting connection with canonical correlation analysis. A variational characterization of the RSVD is provided in §2.2.2. A generalized dyadic decomposition and some geometrical properties are investigated in §2.2.3. In §2.2.4, it is shown how the OSVD, PSVD, and QSVD are special cases of the RSVD.

2.2.1. Relation to a generalized eigenvalue problem. Consider the generalized eigenvalue problem

$$\begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} BB^* & 0 \\ 0 & C^*C \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix} \lambda. \tag{2}$$

Let p_i be the ith column of P and q_i the ith column of Q. Obviously, the column vector $(p_i^* \quad q_i^*)^*$ is a generalized eigenvector of the pencil (2). There are four types of generalized eigenvalues (finite nonzero, zero, infinite, and arbitrary), which can be related to the restricted singular value triplets of Theorem 1.

Note that if $BB^* = I_m$ and $C^*C = I_n$, the eigenvalues λ are $\pm$ the singular values of the matrix A. In the case that the matrices BB^* and C^*C are nonsingular, it can be shown that the generalized eigenvalue problem (2) is equivalent to a singular value decomposition. It follows from (2) that

$$\begin{aligned} Aq_i &= BB^*p_i\lambda_i, \\ A^*p_i &= C^*Cq_i\lambda_i. \end{aligned}$$

If BB^* and C^*C are both nonsingular, then there exist square nonsingular matrices W_b and W_c (for example, the Cholesky decomposition) such that $BB^* = W_b^*W_b$ and $C^*C = W_c^*W_c$. Then, we have that

$$\begin{aligned} (W_b^{-*}AW_c^{-1})(W_cq_i) &= (W_bp_i)\lambda_i, \\ (W_c^{-*}A^*W_b^{-1})(W_bp_i) &= (W_cq_i)\lambda_i. \end{aligned}$$

From Theorem 1, it follows that $P^*(BB^*)P = S_bS_b^t$ and $Q^*(C^*C)Q = S_c^tS_c$. Hence, if BB^* is nonsingular, the column vectors of P are orthogonal with respect to the inner product provided by the positive-definite matrix BB^*. A similar observation applies for the column vectors of Q with respect to C^*C. The BB^*-orthogonality of the vectors p_i and the C^*C-orthogonality of the vectors q_i implies that the vectors W_bp_i and W_cq_i are (multiples of) the left and right singular vectors of the matrix $W_b^{-*}AW_c^{-1}$.

Consider the RSVD of the matrix triplet (A^*B, A^*, B) and its related generalized eigenvalue problem:

$$\begin{pmatrix} 0 & A^*B \\ B^*A & 0 \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} A^*A & 0 \\ 0 & B^*B \end{pmatrix} \begin{pmatrix} p \\ q \end{pmatrix} \sigma.$$

This is nothing more than the eigenvalue problem that arises in *canonical correlation analysis* (principal angles and vectors between subspaces; see, e.g., [3], [15]). There exist applications where the matrices BB^* and C^*C are (almost) singular (see, e.g., [13], [18]). The matrices BB^* and C^*C can be (sample) covariance matrices that are (almost) singular. This is, for instance, the case in [18], where a generalized type of canonical correlation analysis is required, allowing singular covariance matrices. Another example is generalized Gauss–Markov estimation as described in §3.3. It is in these situations that the RSVD may provide essential insight into the geometry of the singularities and at the same time yield a numerically robust and elegant implementation of the solution by avoiding the explicit solution (with its "implicit squaring") of the generalized eigenvalue problem.

2.2.2. A variational characterization. Let $\phi(x, y) = x^*Ay$ be a bilinear form of 2 vectors x and y. We wish to maximize $\phi(x, y)$ over all vectors x, y subject to $x^*BB^*x = 1$ and $y^*C^*Cy = 1$. It follows directly from the RSVD that a solution exists only if one of the following situations occurs:

- $r_{abc} + r_a - r_{ab} - r_{ac} \neq 0$. In this case, the maximum is equal to the largest diagonal element of S_1 and the optimizing vectors are $x = p_1$ (first column vector of P) and $y = q_1$ (first column vector of Q) so that $\phi(p_1, q_1) = \alpha_1$.

- $r_{abc} + r_a - r_{ab} - r_{ac} = 0$. The norm constraints on x and y can only be satisfied if

$$r_{ac} + r_b - r_{abc} > 0 \quad \text{or} \quad r_{ab} - r_a > 0$$

and

$$r_{ab} + r_c - r_{abc} > 0 \quad \text{or} \quad r_{ac} - r_a > 0.$$

In either case, the maximum is 0.

If none of these conditions is satisfied, there is *no* solution.

Assume that the maximum is achieved for the vectors $x_1 = p_1$ and $y_1 = q_1$. Then, other extrema of the objective function $\phi(x,y) = x^*Ay$, constrained to lie in subspaces that are BB^*-orthogonal to p_1 and C^*C-orthogonal to q_1, can be found in an obvious recursive manner. All of these extrema are then generated by the columns of the matrices P and Q.

2.2.3. A generalized dyadic decomposition and geometrical properties. Denote $P' = P^{-*}$ and $Q^{-1} = Q'^*$. Then, with an appropriate partitioning of the matrices P', Q', U_c, and V_b, corresponding to the diagonal structure of the matrices S_a, S_b, S_c of Theorem 1, it is straightforward to obtain the following sums:

$$\begin{aligned} A &= P'_1 S_1 Q'^*_1 + P'_2 Q'^*_2 + P'_3 Q'^*_3 + P'_4 Q'^*_4, \\ B &= P'_1 V^*_{b1} + P'_3 V^*_{b2} + P'_5 S_2 V^*_{b4}, \\ C &= U_{c1} Q'^*_1 + U_{c2} Q'^*_2 + U_{c4} S_3 Q'^*_5. \end{aligned}$$

Hence,

$$\begin{aligned} \mathbf{R}(P'_1) + \mathbf{R}(P'_3) &= \mathbf{R}(A) \bigcap \mathbf{R}(B), \\ \mathbf{R}(Q'^*_1) + \mathbf{R}(Q'^*_2) &= \mathbf{R}(A^*) \bigcap \mathbf{R}(C^*). \end{aligned}$$

The decomposition of A can be interpreted as a decomposition relative to $\mathbf{R}(B)$ and $\mathbf{R}(C^*)$: The four terms of this decomposition can be classified geometrically as follows:

	in $\mathbf{R}(B)$	not in $\mathbf{R}(B)$
in $\mathbf{R}(C^*)$	$P'_1 S_1 Q'^*_1$	$P'_2 Q'^*_2$
not in $\mathbf{R}(C^*)$	$P'_3 Q'^*_3$	$P'_4 Q'^*_4$

Obviously, the term $P'_1 S_1 Q'^*_1$ represents the *restriction* of the linear operator represented by the matrix A to the column space of the matrix B **and** the row space of the matrix C, while the term $P'_4 Q'^*_4$ is the restriction of A to the orthogonal complements of $\mathbf{R}(B)$ and $\mathbf{R}(C^*)$.

Also, we find that

$$\begin{aligned} \mathbf{R}(B^*) &= \mathbf{R}(V^*_{b1}) + \mathbf{R}(V^*_{b2}) + \mathbf{R}(V^*_{b4}), \\ \mathbf{R}(C) &= \mathbf{R}(U_{c1}) + \mathbf{R}(U_{c2}) + \mathbf{R}(U_{c4}), \end{aligned}$$

and

$$\begin{aligned} BV_{b3} = 0 \quad &\Longrightarrow \quad \mathbf{N}(B) = \mathbf{R}(V_{b3}), \\ U^*_{c3} C = 0 \quad &\Longrightarrow \quad \mathbf{N}(C^*) = \mathbf{R}(U_{c3}). \end{aligned}$$

Finally, some of the block dimensions in the RSVD of the matrix triplet (A, B, C) can be related to geometrical interpretations by repeated application of Lemma 2.

$$\dim\left[\mathbf{R}\begin{pmatrix}A\\C\end{pmatrix}\bigcap\mathbf{R}\begin{pmatrix}B\\0\end{pmatrix}\right] = r_{ac} + r_b - r_{abc},$$

$$\dim[\,\mathbf{R}(A\ \ B)^*\bigcap\mathbf{R}(C\ \ 0)^*\,] = r_{ab} + r_c - r_{abc},$$

$$\dim[\,\mathbf{R}(A)\bigcap\mathbf{R}(B)\,] = r_a + r_b - r_{ab},$$

$$\dim[\,\mathbf{R}(A^*)\bigcap\mathbf{R}(C^*)\,] = r_a + r_c - r_{ac}.$$

It is easy to show that

$$\mathbf{R}(Q_6') = \mathbf{N}(A)\bigcap\mathbf{N}(C),$$

$$\mathbf{R}(P_6') = \mathbf{N}(A^*)\bigcap\mathbf{N}(B^*).$$

Hence Q_6' provides a basis for the common null space of A and C, which is of dimension $n - r_{ac}$, while P_6' provides a basis for the common null space of A^* and B^*, which is of dimension $m - r_{ab}$.

2.2.4. Relation to (generalized) SVDs. The RSVD reduces to the OSVD, the PSVD, or the QSVD for special choices of the matrices A, B, and/or C. For the precise structure of the PSVD and the QSVD, we refer to [8].

THEOREM 2 (special cases of the RSVD).

1. RSVD *of* (A, I_m, I_n) *is an* OSVD *of* A.
2. RSVD *of* (I_m, B, C) *is a* PSVD *of* (B^*, C).
3. RSVD *of* (A, B, I_n) *is a* QSVD *of* (A, B).
4. RSVD *of* (A, I_m, C) *is a* QSVD *of* (A, C).

Proof. 1. $B = I_m, C = I_n$. Consider the RSVD of (A, I_m, I_n). By definition, $I_m = P^{-*}S_bV_b^*$ and $I_n = U_cS_cQ^{-1}$. This implies $P^{-*} = V_bS_b^{-1}$ and $Q^{-1} = S_c^{-1}U_c^*$. Hence, we find that $A = V_b(S_b^{-1}S_aS_c^{-1})U_c^*$, which is an OSVD of A.

2. $A = I_m$. Consider the RSVD of (I_m, B, C). Then $I_m = P^{-*}S_aQ^{-1}$, which implies $Q^{-1} = S_a^{-1}P^*$. Hence, $B^* = V_bS_b^tP^{-1}$, $C = U_c(S_cS_1^{-1})P^*$, which is nothing else than a PSVD of (B^*, C).

3. $C = I_n$. Consider the RSVD of (A, B, I_n). Then $I_n = U_cS_cQ^{-1}$, which implies $Q^{-1} = S_c^{-1}U_c^*$. Then, $A = P^{-*}(S_aS_c^{-1})U_c^*$, $B = P^{-*}S_bV_b^*$, which is (up to a diagonal scaling) a QSVD of the matrix pair (A, B).

4. $B = I_m$. The proof is similar to part 3. □

3. Applications. In this section, we shall first explore the use of the RSVD in the analysis of problems related to expressions of the form $A+BDC$ where A, B, C are given matrices. The connection with Mitra's concept of the extended shorted operator [20] and with matrix balls will be discussed, as will the solution of the matrix equation $BDC = A$, which led Penrose to rediscover the pseudo-inverse of a matrix [24], [25]. In §3.2, it is shown how the RSVD can be used to solve constrained total linear least squares problems with exact, noiseless rows and columns and the close connection to Carlson's generalized Schur complement [4] is emphasized. In §3.3, we discuss the application of the RSVD in the analysis and solution of generalized Gauss–Markov models, with and without constraints.

Throughout this section, we shall use a matrix E, defined as

$$E = V_b^*DU_c \tag{3}$$

with a block partitioning derived from the block structure of S_b and S_c as follows:

(4)

$$\begin{array}{c} \\ r_{abc}+r_a-r_{ab}-r_{ac} \\ r_{ac}+r_b-r_{abc} \\ p-r_b \\ r_{ab}-r_a \end{array} \begin{array}{c} \begin{array}{cccc} r_{abc}+r_a-r_{ab}-r_{ac} & r_{ab}+r_c-r_{abc} & q-r_c & r_{ac}-r_a \end{array} \\ \begin{pmatrix} E_{11} & E_{12} & E_{13} & E_{14} \\ E_{21} & E_{22} & E_{23} & E_{24} \\ E_{31} & E_{32} & E_{33} & E_{34} \\ E_{41} & E_{42} & E_{43} & E_{44} \end{pmatrix} \end{array}.$$

3.1. On the structure of A + BDC. The RSVD provides geometrical insight into the structure of a matrix A relative to the column space of a matrix B and the row space of a matrix C. As will now be shown, it is an appropriate tool to analyse expressions of the form $A + BDC$ where D is an arbitrary $p \times q$ matrix. The RSVD allows us to analyse and solve the following questions:

1. What is the range of ranks of $A + BDC$ over all possible $p \times q$ matrices D (§3.1.1)?
2. When is the matrix D that minimizes the rank of $A + BDC$ unique (§3.1.2)?
3. When is the term BDC that minimizes $\operatorname{rank}(A + BDC)$ unique? It will be shown how this corresponds to Mitra's extension of the shorted operator [20] in §3.1.3.
4. In the case of nonuniqueness, what is the minimum norm solution (for unitarily invariant norms) D that minimizes $\operatorname{rank}(A + BDC)$ (§3.1.4)?
5. The reverse question is the following: Assume that $\|D\| \leq \delta$ where δ is a given positive real scalar. What is the minimum rank of $A + BDC$? This can be linked to rank minimization problems in so-called matrix balls (§3.1.5).
6. An extreme case occurs if we look for the (minimum norm) solution D to the linear matrix equation $BDC = A$. The RSVD provides the necessary and sufficient conditions for consistency and allows us to parameterize all solutions (§3.1.6).

3.1.1. The range of ranks of A+BDC. The range of ranks of $A + BDC$ for all possible matrices D is described in the following theorem.

THEOREM 3 (on the rank of $A + BDC$).

$$r_{ab} + r_{ac} - r_{abc} \leq \operatorname{rank}(A + BDC) \leq \min(r_{ab}, r_{ac}).$$

For every number r in between these bounds, there exists a matrix D such that $\operatorname{rank}(A + BDC) = r$.

Proof. The proof uses the RSVD structure of Theorem 1:

$$\begin{aligned} A + BDC &= P^{-*} S_a Q^{-1} + P^{-*} S_b V_b^* D U_c S_c Q^{-1} \\ &= P^{-*}(S_a + S_b E S_c) Q^{-1}, \end{aligned}$$

where $E = V_b^* D U_c$. Because of the nonsingularity of P, Q, U_c, V_b, we have that $\operatorname{rank}(A + BDC) = \operatorname{rank}(S_a + S_b E S_c)$. Using elementary row and column operations and the block partitioning of E as in (4), it is easy to show that

$$\operatorname{rank}(A + BDC) = \operatorname{rank} \begin{pmatrix} S_1 + E_{11} & 0 & 0 & 0 & E_{14} S_3 & 0 \\ 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & I & 0 & 0 \\ S_2 E_{41} & 0 & 0 & 0 & S_2 E_{44} S_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{5}$$

the block dimensions of which are the same as those of S_a in Theorem 1. Obviously, a lower bound is achieved for $E_{11} = -S_1$, $E_{14} = 0$, $E_{41} = 0$, $E_{44} = 0$. The upper bound is achieved for almost every ("random") choice of $E_{11}, E_{14}, E_{41}, E_{44}$. □

Observe that, if $r_a = r_{ab} + r_{ac} - r_{abc}$, then there is no S_1 block in S_a and the minimum rank of $A + BDC$ will be r_a. Also observe that the minimum achievable rank, $r_{ab} + r_{ac} - r_{abc}$, is precisely the number of restricted singular values triplets of the form $(1,0,1)$, $(1,1,0)$, and $(1,0,0)$.

3.1.2. The unique rank minimizing matrix D. When is the matrix D that minimizes the rank of $A+BDC$ unique? The answer is given in the following theorem.

THEOREM 4. *Let D be such that* $\operatorname{rank}(A + BDC) = r_{ab} + r_{ac} - r_{abc}$ *and assume that $r_a > r_{ab} + r_{ac} - r_{abc}$. Then the matrix D that minimizes the rank of $A + BDC$ is unique if and only if:*

1. $r_c = q$,
2. $r_b = p$,
3. $r_{abc} = r_{ab} + r_c = r_{ac} + r_b$.

In the case where these conditions are satisfied, the matrix D is given as

$$D = V_b \begin{pmatrix} -S_1 & 0 \\ 0 & 0 \end{pmatrix} U_c^*.$$

Observe that the expression for the matrix D is nothing more than an OSVD!

Proof. It can be verified from the matrix in (5) that the rank of $A + BDC$ is independent of the block matrices $E_{12}, E_{13}, E_{21}, E_{22}, E_{23}, E_{24}, E_{31}, E_{32}, E_{33}, E_{34}, E_{42}, E_{43}$. Hence, the rank minimizing matrix D will not be unique, whenever one of the corresponding block dimensions is not zero, in which case it is parameterized by the blocks E_{ij} in

$$D = V_b \begin{pmatrix} -S_1 & E_{12} & E_{13} & 0 \\ E_{21} & E_{22} & E_{23} & E_{24} \\ E_{31} & E_{32} & E_{33} & E_{34} \\ 0 & E_{42} & E_{43} & 0 \end{pmatrix} U_c^*. \tag{6}$$

Setting the expressions for these block dimensions equal to zero results in the necessary conditions. The unique optimal matrix D is then given by $D = V_b E U_c^*$, where

$$E = \begin{matrix} \\ p + r_a - r_{ab} \\ r_{ab} - r_a \end{matrix} \begin{matrix} \begin{matrix} q + r_a - r_{ac} & r_{ac} - r_a \end{matrix} \\ \begin{pmatrix} E_{11} & E_{14} \\ E_{41} & E_{44} \end{pmatrix} \end{matrix} = \begin{pmatrix} -S_1 & 0 \\ 0 & 0 \end{pmatrix}. \qquad \square$$

3.1.3. On the uniqueness of BDC: The extended shorted operator. A question related to the one of §3.1.2 concerns the uniqueness of the product term BDC that minimizes the rank of $A + BDC$. As a matter of fact, this problem has received much attention in the literature where the term BDC is called the *extended shorted operator* and was introduced in [20]. It is an extension to rectangular matrices, of the shorting of an operator considered by Krein, Anderson, and Trapp only for positive operators (see [20] for references).

DEFINITION 1 (the extended shorted operator[1]). Let A $(m \times n)$, B $(m \times p)$, and C $(q \times n)$ be given matrices. A shorted matrix $\mathcal{S}(A|B,C)$ is any $m \times n$ matrix that satisfies the following conditions:

[1] We have slightly changed the notation that is used in [20].

1.

$$\mathbf{R}(\mathcal{S}(A|B,C)) \subseteq \mathbf{R}(B), \qquad \mathbf{R}(\mathcal{S}(A|B,C)^*) \subseteq \mathbf{R}(C^*).$$

2. If F is an $m \times n$ matrix satisfying $\mathbf{R}(F) \subseteq \mathbf{R}(B)$ and $\mathbf{R}(F^*) \subseteq \mathbf{R}(C^*)$, then

$$\operatorname{rank}(A-F) \geq \operatorname{rank}(A - \mathcal{S}(A|B,C)).$$

Hence, the shorted operator is a matrix whose column space belongs to the column space of B, whose row space belongs to the row space of C, and which minimizes the rank of $A - F$ over all matrices F, satisfying these conditions. From this, it follows that the shorted operator can be written as

$$\mathcal{S}(A|B,C) = BDC$$

for a certain $p \times q$ matrix D. This establishes the direct connection of the concept of extended shorted operator with the RSVD.

The shorted operator is not always unique, as can be seen from the following example. Let

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Then, all matrices of the form

$$S = \begin{pmatrix} 1 & 0 & 0 \\ \alpha & \beta & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

minimize the rank of $A - S$, which equals 2, for arbitrary α and β.

Necessary conditions for uniqueness of the shorted operator can be found in a straightforward way from the RSVD.

THEOREM 5 (on the uniqueness of the extended shorted operator). *Let the* RSVD *of the matrix triplet* (A,B,C) *be given as in Theorem* 1. *Then*

$$\mathcal{S}(A|B,C) = P^{-*}\mathcal{S}(S_a|S_b,S_c)Q^{-1}.$$

The extended shorted operator $\mathcal{S}(A|B,C)$ *is unique if and only*

1. $r_{abc} = r_c + r_{ab}$,
2. $r_{abc} = r_b + r_{ac}$,

and is given by

$$\mathcal{S}(A|B,C) = P^{-*}\begin{pmatrix} -S_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} Q^{-1}.$$

Proof. It follows from Theorem 3 that the minimum rank of $A + BDC$ is $r_{ab} + r_{ac} - r_{abc}$, and that in this case $E_{11} = -S_1, E_{14} = 0, E_{41} = 0, E_{44} = 0$. A short computation shows that

$$BDC = P^{-*}\begin{pmatrix} -S_1 & E_{12} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ E_{21} & E_{22} & 0 & 0 & E_{24}S_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & S_2E_{42} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} Q^{-1}.$$

Hence, the matrix BDC is unique if and only if the blocks E_{12}, E_{22}, E_{42}, E_{21}, E_{22}, and E_{24} do not appear in this decomposition. Setting the corresponding block dimensions equal to zero proves the theorem. □

Observe that the conditions for uniqueness of the extended shorted operator BDC are less restrictive than the uniqueness conditions for the matrix D (Theorem 4). As a consequence of Theorem 5, we also obtain a parameterization of all shorted operators in the case where the uniqueness conditions are not satisfied. All possible shorted operators are then parameterized by the matrices E_{12}, E_{21}, E_{22}, E_{24}, E_{42}. Observe that the shorted operator is independent of the matrices E_{13}, E_{23}, E_{31}, E_{32}, E_{33}, E_{34}, E_{43}. The result of Theorem 5, derived via the RSVD, corresponds to Theorem 4.1 and Lemma 5.1 of [20]. Some connections with the generalized Schur complement and statistical applications of the shorted operator can also be found in [20].

3.1.4. The minimum norm solutions D that reduce the rank of A+BDC. Consider the problem of finding the matrix D of minimal (unitarily invariant) norm $\|D\|$ such that:

$$\operatorname{rank}(A + BDC) = r < r_a,$$

where r is a prescribed nonnegative integer.

It follows from Theorem 3 that a necessary condition for a solution to exist is that $r_a > r \geq r_{ab} + r_{ac} - r_{abc}$. Observe that if $r_a = r_{ab} + r_{ac} - r_{abc}$, no solution exists. In this case, there is no diagonal matrix S_1 in S_a of Theorem 1. Assume that the required rank r equals the minimal achievable: $r = r_{ab} + r_{ac} - r_{abc}$. Then, if the conditions of Theorem 4 are satisfied, the optimal D is unique and follows directly from the RSVD. The interesting case occurs whenever the rank minimizing D is not unique. Before examining matrices D that *minimize* the rank of $A+BDC$, note that, whenever $\min(r_{ab}, r_{ac}) - r_a > 0$, there exist many matrices that will *increase* the rank of $A + BDC$. In this case,

$$\inf_{\epsilon}\{\, \epsilon = \|D\| \mid \operatorname{rank}(A + BDC) > r_a\} = 0, \tag{7}$$

which implies that there exist arbitrarily "small" matrices D that will increase the rank.

THEOREM 6. *Consider all matrices D satisfying*

$$r_{ab} + r_{ac} - r_{abc} \leq r = \operatorname{rank}(A + BDC) < r_a$$

where r is a given integer and let $\|.\|$ be any unitarily invariant norm. A matrix D of minimal norm $\|D\|$ is given by

$$D = -V_b \begin{pmatrix} S_1^r & 0 \\ 0 & 0 \end{pmatrix} U_c^*$$

where S_1^r is a singular diagonal matrix

$$S_1^r = \begin{matrix} r + r_{abc} - r_{ab} - r_{ac} \\ r_a - r \end{matrix} \begin{pmatrix} \overset{r + r_{abc} - r_{ac} - r_{ab}}{0} & \overset{r_a - r}{0} \\ 0 & S_d \end{pmatrix}.$$

S_d contains the $r_a - r$ smallest diagonal elements of S_1.

Proof. From the RSVD of the matrix triplet A, B, C it follows that

$$\begin{aligned} A + BDC &= P^{-*}(S_a + S_b(V_b^* D U_c)S_c)Q^{-1} \\ &= P^{-*}(S_a + S_b E S_c)Q^{-1} \end{aligned}$$

with $\|E\| = \|V_b^* D U_c\| = \|D\|$. The result follows immediately from the partitioning of E as in (4) and from equation (5). □

We could use Theorem 6 to define the *restricted singular values* σ_k as

$$\sigma_k = \inf_{\epsilon} \{\, \epsilon = \sigma_{\max}(D) \mid \operatorname{rank}(A + BDC) = k - 1 \,\}$$

where $\sigma_{\max}(.)$ denotes the maximum ordinary singular value. Because the rank of $A + BDC$ cannot be reduced below $r_{ab} + r_{ac} - r_{abc}$, there will be $r_{ab} + r_{ac} - r_{abc}$ infinite restricted singular values. There are $r_a + r_{abc} - r_{ab} - r_{ac}$ finite restricted singular values, corresponding to the diagonal elements of S_1. From (5), it can be seen that the diagonal elements of S_2 and S_3 can be used to increase the rank of $A + BDC$ to $\min(r_{ab}, r_{ac})$. However, from (7) it is obvious that $\min(r_{ac} - r_a, r_{ab} - r_a)$ restricted singular values will be zero.

3.1.5. The reverse problem: Given ||D||, what is the minimal rank of A+BDC? The results of §§3.1.3 and 3.1.4 allow us to obtain in a simple fashion the answer to the reverse question: Assuming we are given a positive real number δ such that $\|D\| \leq \delta$, what is the minimum rank $r_{\min}$ of $A + BDC$?

The answer is an immediate consequence of Theorem 6. Note that the optimal matrix D is given as the product of three matrices, which form its OSVD! Hence, $\|D\| = \|S_1^r\|$ and the integer $r_{\min}$ can be determined as follows. Let S_i be the $i \times i$ diagonal matrix that contains the i smallest elements of S_1. Then,

$$r_{\min} = r_a - (\max_i \{\text{size}(S_i) \text{ such that } \|S_i\| \leq \delta\}). \tag{8}$$

It is interesting to note that expressions of the form $A + BDC$ with restrictions on the norm of D can be related to the notion of *matrix balls*, which show up in the analysis of so-called completion problems [6].

DEFINITION 2 (matrix ball). For given matrices A $(m \times n)$, B $(m \times p)$, and C $(q \times n)$, the closed matrix ball $\mathcal{R}(A|B,C)$ with center A, left semiradius B, and right semiradius C is defined by

$$\mathcal{R}(A|B,C) = \{\, X \mid X = A + BDC \text{ where } \|D\|_2 \leq 1\}.$$

Using Theorem 6 and (8), we can find all matrices of least rank within a certain given matrix ball by simply requiring that $\sigma_{\max}(D) \leq 1$. The solution is obtained from the appropriate truncation of S_1^r in Theorem 6. Since the solution of the completion problems investigated in [6] are described in terms of matrix balls, it follows that we can find the minimal rank solution in the matrix ball of all solutions of the completion problems, using the RSVD.

3.1.6. The matrix equation BDC = A. Consider the problem of investigating the consistency of, and, if consistent, finding a (minimum norm) solution to, the linear equation in the unknown matrix D:

$$BDC = A.$$

This equation has an historical significance because it led Penrose to rediscover what is now called the Moore–Penrose pseudo-inverse [21], [24]. Of course, this problem can be viewed as an extreme case of Theorems 3 and 6, with the prescribed integer $r = 0$.

THEOREM 7. *The matrix equation $BDC = A$ in the unknown matrix D is consistent if and only if*

$$r_{ab} = r_b, \qquad r_{ac} = r_c, \qquad r_{abc} = r_b + r_c.$$

All solutions are then given by

$$D = V_b \begin{pmatrix} S_1 & E_{13} & 0 \\ E_{31} & E_{33} & E_{34} \\ 0 & E_{43} & 0 \end{pmatrix} U_c^*$$

and a minimum norm solution corresponds to $E_{13} = 0$, $E_{31} = 0$, $E_{33} = 0$, $E_{34} = 0$, $E_{43} = 0$.

Proof. Let $E = V_b^* D U_c$ and partition E as in (4). The consistency of $BDC = A$ depends on whether the following is satisfied with equality

$$\begin{pmatrix} E_{11} & E_{12} & 0 & 0 & E_{14}S_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ E_{21} & E_{22} & 0 & 0 & E_{24}S_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ S_2E_{41} & S_2E_{42} & 0 & 0 & S_2E_{44}S_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} S_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Comparing the diagonal blocks, the conditions for consistency follow immediately as $r_{abc} = r_{ab} + r_c = r_{ac} + r_b = r_b + r_c$, which implies $r_{ab} = r_b$ and $r_{ac} = r_c$. These conditions express the fact that the column space of A should be contained in the column space of B and that the row space of A should be contained in the row space of C. If these conditions are satisfied, the matrix equation $BDC = A$ is consistent and the matrix $E = V_b^* D U_c$ is given by

$$E = \begin{array}{c} \\ r_a \\ p - r_b \\ r_b - r_a \end{array} \begin{array}{c} \begin{array}{ccc} r_a & q - r_c & r_c - r_a \end{array} \\ \begin{pmatrix} E_{11} & E_{13} & E_{14} \\ E_{31} & E_{33} & E_{34} \\ E_{41} & E_{43} & E_{44} \end{pmatrix} \end{array}.$$

The equation $BDC = A$ is equivalent to

$$\begin{pmatrix} E_{11} & E_{14}S_3 & 0 \\ S_2E_{41} & S_2E_{44}S_3 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} S_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

This is solved for $E_{11} = S_1$, $E_{14} = 0$, $E_{41} = 0$, $E_{44} = 0$. Observe that the solution is independent of the blocks E_{13}, E_{31}, E_{33}, E_{34}, E_{43}. Hence, all solutions can be parameterized as

$$D = \begin{pmatrix} V_{b1} & V_{b3} & V_{b4} \end{pmatrix} \begin{pmatrix} S_1 & E_{13} & 0 \\ E_{31} & E_{33} & E_{34} \\ 0 & E_{43} & 0 \end{pmatrix} \begin{pmatrix} U_{c1}^* \\ U_{c3}^* \\ U_{c4}^* \end{pmatrix}.$$

The minimum norm solution follows immediately. □

Penrose originally proved [21], [24] that a necessary and sufficient condition for $BDC = A$ to have a solution is:

$$BB^{-}AC^{-}C = A, \tag{9}$$

where B^{-} and C^{-} are inner inverses of B and C. All solutions D can then be written as

$$D = B^{-}AC^{-} + Z - BB^{-}ZC^{-}C, \tag{10}$$

where Z is an arbitrary $p \times q$ matrix. It requires a tedious though straightforward calculation to verify that our solution of Theorem 7 coincides with (10). In order to verify this, consider the RSVD of A, B, C and use Lemma 1 to obtain an expression for the inner inverses of B and C, which will contain arbitrary matrices. Using the block dimensions of S_a, S_b, S_c as in Theorem 1, it can be shown that the consistency conditions of Theorem 7 coincide with the consistency condition (9).

Before concluding this section, it is worth mentioning that all results of this section can be specialized for the case where either B or C equals the identity matrix. In this case, the RSVD specializes to the QSVD (Theorem 2) and mutatis mutandis, the same type of questions, now related to two matrices, can be formulated and solved using the QSVD such as shorted operators, minimum norm rank minimization, solution of the matrix equation $DC = A$, etc.

3.2. On low rank approximations of a partitioned matrix. In this section, the RSVD will be used to analyse and solve problems that can be stated in terms of the matrix[2] $M = \left(\begin{smallmatrix} A & B \\ D & D^* \end{smallmatrix}\right)$ where A, B, C, D are given matrices. The main results include the analysis of the (generalized) Schur complement [4] in terms of the RSVD (§3.2.1), the range of ranks of the matrix M as D is modified, and the analysis of the (non)unique matrix D that minimizes the rank of M (§3.2.2), and finally the solution of the constrained total least squares problem with exact and noisy data by imposing additional norm constraints on D (§3.2.3).

3.2.1. (Generalized) Schur complements and the RSVD. The notion of a Schur complement S of the matrix A in M (which is $S = D^* - CA^{-1}B$ when A is square nonsingular), can be generalized to the case where the matrix A is rectangular and/or rank deficient [4] as follows.

DEFINITION 3 ((Generalized) Schur complement). A generalized Schur complement of A in $M = \left(\begin{smallmatrix} A & B \\ C & D^* \end{smallmatrix}\right)$ is any matrix $S = D^* - CA^{-}B$ where A^{-} is an inner inverse of A.

In general, there are many generalized Schur complements, because from Lemma 1 we know that there are many inner inverses. However, the RSVD allows us to investigate the dependency of S on the choice of the inner inverse.

THEOREM 8. *The Schur complement $S = D^* - CA^{-}B$ is independent of A^{-} if and only if $r_a = r_{ab} = r_{ac}$. In this case, S is given by*

$$S = U_c \begin{pmatrix} E_{11}^* - S_1^{-1} & E_{21}^* & E_{31}^* \\ E_{12}^* & E_{22}^* & E_{32}^* \\ E_{13}^* & E_{23}^* & E_{33}^* \end{pmatrix} V_b^*.$$

[2] In order to keep the notation consistent with that of §3.1, we use the matrix D^*, which is the complex conjugate transpose of D in §3.1, as the lower right block of M. This allows us, for instance, to use the same matrix E as defined in (3) and (4).

Proof. Consider the factorization of A as in the RSVD. From Lemma 1, every inner inverse of A can be written as

$$A^- = Q \begin{pmatrix} S_1^{-1} & 0 & 0 & 0 & X_{15} & X_{16} \\ 0 & I & 0 & 0 & X_{25} & X_{26} \\ 0 & 0 & I & 0 & X_{35} & X_{36} \\ 0 & 0 & 0 & I & X_{45} & X_{46} \\ X_{51} & X_{52} & X_{53} & X_{54} & X_{55} & X_{56} \\ X_{61} & X_{62} & X_{63} & X_{64} & X_{65} & X_{66} \end{pmatrix} P^*$$

for certain block matrices X_{ij}, where the block dimensions correspond to the block dimensions of the matrix S_a^* of Theorem 1. It is straightforward to show that

$$CA^-B = U_c \begin{pmatrix} S_1^{-1} & 0 & 0 & X_{15}S_2 \\ 0 & 0 & 0 & X_{25}S_2 \\ 0 & 0 & 0 & 0 \\ S_3X_{51} & S_3X_{53} & 0 & S_3X_{55}S_2 \end{pmatrix} V_b^*.$$

Hence, this product is dependent on the blocks X_{15}, X_{25}, X_{51}, X_{53}, X_{55}. The corresponding block dimensions are 0 if and only if $r_a = r_{ab} = r_{ac}$. □

Observe that the theorem is equivalent with the statement that the (generalized) Schur complement $S = D^* - CA^-B$ is independent of the precise choice of A^- if and only if $\mathbf{R}(B) \subset \mathbf{R}(A)$ and $\mathbf{R}(C^*) \subset \mathbf{R}(A^*)$. This corresponds to Carlson's statement of the same result (Proposition 1 of [4]). In the case that these conditions are not satisfied, all possible generalized Schur complements are parameterized by the blocks X_{51}, X_{53}, X_{15}, X_{25}, and X_{55} as

$$S = U_c \begin{pmatrix} E_{11}^* - S_1^{-1} & E_{21}^* & E_{31}^* & E_{41}^* - X_{15}S_2 \\ E_{12}^* & E_{22}^* & E_{32}^* & E_{42}^* - X_{25}S_2 \\ E_{13}^* & E_{23}^* & E_{33}^* & E_{43}^* \\ E_{14}^* - S_3X_{51} & E_{24}^* - S_3X_{53} & E_{34}^* & E_{44}^* - S_3X_{55}S_2 \end{pmatrix} V_b^*. \tag{11}$$

3.2.2. How does the rank of M change with changing D? Define the matrix $M(\tilde{D}) = \left(\begin{smallmatrix} A & B \\ C & D^* - \tilde{D}^* \end{smallmatrix}\right)$. We shall also use $\tilde{D} = D - \tilde{D}$. How can we modify the rank of $M(\tilde{D})$ by changing the matrix $\tilde{D}$? Before answering this question, we need to state the following (well-known) lemma.

LEMMA 3 (rank of a partitioned matrix and the Schur complement). *If A is square and nonsingular, then*

$$\operatorname{rank} \begin{pmatrix} A & B \\ C & D^* \end{pmatrix} = \operatorname{rank}(A) + \operatorname{rank}(D^* - CA^{-1}B).$$

Proof. Observe that:

$$\begin{pmatrix} A & B \\ C & D^* \end{pmatrix} = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & D^* - CA^{-1}B \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix}. \qquad \square$$

Thus we have Theorem 9.

THEOREM 9.

$$\operatorname{rank} \begin{pmatrix} A & B \\ C & D^* \end{pmatrix} = r_{ab} + r_{ac} - r_a + \operatorname{rank} \begin{pmatrix} E_{11}^* - S_1^{-1} & E_{21}^* & E_{31}^* \\ E_{12}^* & E_{22}^* & E_{32}^* \\ E_{13}^* & E_{23}^* & E_{33}^* \end{pmatrix}.$$

Proof. From the RSVD, it follows immediately that the required rank is equal to the rank of the matrix

$$\begin{pmatrix} S_a & S_b \\ S_c & E^* \end{pmatrix} = \begin{pmatrix} S_1 & 0 & 0 & 0 & 0 & 0 & I & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 & 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & S_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ I & 0 & 0 & 0 & 0 & 0 & E_{11}^* & E_{21}^* & E_{31}^* & E_{41}^* \\ 0 & I & 0 & 0 & 0 & 0 & E_{12}^* & E_{22}^* & E_{32}^* & E_{42}^* \\ 0 & 0 & 0 & 0 & 0 & 0 & E_{13}^* & E_{23}^* & E_{33}^* & E_{43}^* \\ 0 & 0 & 0 & 0 & S_3 & 0 & E_{14}^* & E_{24}^* & E_{34}^* & E_{44}^* \end{pmatrix}.$$

From the nonsingularity of S_2 and S_3, it follows that the rank is independent of $E_{41}, E_{42}, E_{43}, E_{14}, E_{24}, E_{34}, E_{44}$. The result then follows immediately from Lemma 3, taking into account the block dimensions of the matrices. □

A consequence of Theorem 9 is the following result.

COROLLARY 1. *The range of ranks r of M attainable by an appropriate choice of $\tilde{D}$ in $M = \left(\begin{smallmatrix} A & B \\ C & D^* - \tilde{D}^* \end{smallmatrix}\right)$ is*

$$r_{ab} + r_{ac} - r_a \leq r \leq \min(p + r_{ac}, q + r_{ab}).$$

The minimum is attained for

$$\tilde{D}^* = U_c \begin{pmatrix} E_{11}^* - S_1^{-1} & E_{21}^* & E_{31}^* & \tilde{E}_{41}^* \\ E_{12}^* & E_{22}^* & E_{32}^* & \tilde{E}_{42}^* \\ E_{13}^* & E_{23}^* & E_{33}^* & \tilde{E}_{43}^* \\ \tilde{E}_{14}^* & \tilde{E}_{24}^* & \tilde{E}_{34}^* & \tilde{E}_{44}^* \end{pmatrix} V_b^* \tag{12}$$

where the matrices $\tilde{E}_{14}$, $\tilde{E}_{24}$, $\tilde{E}_{34}$, $\tilde{E}_{41}$, $\tilde{E}_{42}$, $\tilde{E}_{43}$, and $\tilde{E}_{44}$ are arbitrary matrices.

Compare the expression of $\tilde{D}$ of Corollary 1 with the expression for the generalized Schur complement of A in M, as given by (11). Obviously, the set of matrices $\tilde{D}$ contains all generalized Schur complements, which are those matrices $\tilde{D}$ for which $\tilde{E}_{34} = E_{34}$ and $\tilde{E}_{43} = E_{43}$. If these blocks are not present in E, there are no matrices $\tilde{D}$, other than generalized Schur complements, that minimize the rank of M. Hence, we have proved the following theorem.

THEOREM 10. *The rank of $M(\tilde{D})$ is minimized for $\tilde{D}$ equal to a generalized Schur complement of A in M. The rank of $M(\tilde{D})$ is minimized* **only** *for $\tilde{D} = D^* - CA^-B$ where A^- is an inner inverse of A, if and only if $r_{ab} = r_a$ or $r_c = q$* **and** *$r_{ac} = r_c$ or $r_b = p$. If $r_a = r_{ab} = r_{ac}$, then the minimizing $\tilde{D}$ is unique.*

Proof. The fact that each generalized Schur complement minimizes the rank of $M(\tilde{D})$ follows directly from the comparison of $\tilde{D}$ in Corollary 2 with the expression for the generalized Schur complement in (11). The rank conditions follow simply from setting the block dimensions of E_{34} and E_{43} in (4) equal to 0. The condition for uniqueness of $\tilde{D}$ follows from Theorem 8. □

This theorem can also be found as Theorem 3 of [4], where it is proved via a different approach. Related results can be found in [7] and [31].

3.2.3. Total linear least squares with exact rows and columns. The nomenclature *total linear least squares* was introduced in [16]. The technique is an extension of least squares fitting in the case where there are errors in both the observation vector b and the data matrix A for overdetermined sets of linear equations

$Ax \approx b$. The analysis and solution is given completely in terms of the OSVD of the concatenated matrix $(A \ b)$. In the case where some of the columns of A are noise-free while the others contain errors, a mixed least squares–total least squares strategy was developed in [17]. The problem where some rows are also error-free was analysed via a Schur complement-based approach in [7]. One of the key canonical decompositions (Lemma 2 of [7]) and related results concerning rank minimization were described earlier in [4]. Another useful reference is [31]. We shall now show how the RSVD allows us to treat the general situation in an elegant way. Again, let the data matrix be given as $M = \left(\begin{smallmatrix} A & B \\ C & D^* \end{smallmatrix}\right)$ where A, B, C are free of error and only D is contaminated by noise. It is assumed that the data matrix is of full row rank.

The *constrained total linear least squares problem* is the following.

Find the matrix $\hat{D}$ and the nonzero vector x such that

$$\begin{pmatrix} A & B \\ C & \hat{D}^* \end{pmatrix} x = 0,$$

and $\|D - \hat{D}\|_F$ is minimized.

A slightly more general problem is the following.

Find the matrix $\hat{D}$ such that $\|D - \hat{D}\|_F$ is minimal and

$$\text{rank} \begin{pmatrix} A & B \\ C & \hat{D}^* \end{pmatrix} \leq r. \tag{13}$$

The error matrix $D - \hat{D}$ will be denoted by $\tilde{D}$. Assume that a solution x is found. By partitioning x conformally to the dimensions of A and B, we find that the vector x satisfies

$$\begin{aligned} Ax_1 + Bx_2 &= 0, \\ Cx_1 + \hat{D}^* x_2 &= 0. \end{aligned}$$

Hence, the total least squares problem can be interpreted as follows. The rows of A and B correspond to linear constraints on the solution vector x. The columns of the matrix C contain error-free (noiseless) data while those of the matrix D are corrupted by noise. In order to find a solution, we must modify the matrix D with minimum effort, as measured by the Frobenius norm of the "error matrix" $\tilde{D}$, into the matrix $\hat{D}$. Without the constraints imposed by matrices A and B, the problem reduces to a mixed linear–total linear least squares problem, as is analysed and solved in [17].

From the results in §3.2.2, we already know that a necessary condition for a solution to exist is $r \geq r_{ab} + r_{ac} - r_a$ (Corollary 1). The class of rank minimizing matrices $\tilde{D}$ is described by Corollary 1 when $r = r_{ab} + r_{ac} - r_a$. Theorem 9 shows how the generalized Schur complements of A in M form a subset of this set. From Corollary 1, it is straightforward to find the minimum norm matrix $\tilde{D}$ that reduces the rank of $M(\tilde{D})$ to $r = r_{ab} + r_{ac} - r_a$. It is given by

$$\tilde{D}^* = U_c \begin{pmatrix} E_{11}^* - S_1^{-1} & E_{21}^* & E_{31}^* & 0 \\ E_{12}^* & E_{22}^* & E_{32}^* & 0 \\ E_{13}^* & E_{32}^* & E_{33}^* & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} V_b^*.$$

The *minimum norm generalized Schur complement* that minimizes the rank of M is

given by

$$S = U_c \begin{pmatrix} E_{11}^* - S_1^{-1} & E_{21}^* & E_{31}^* & 0 \\ E_{12}^* & E_{22}^* & E_{32}^* & 0 \\ E_{13}^* & E_{23}^* & E_{33}^* & E_{43}^* \\ 0 & 0 & E_{34}^* & 0 \end{pmatrix} V_b^* .$$

This corresponds to a choice of inner inverse in (11) given by $X_{15} = E_{41}^* S_2^{-1}$, $X_{25} = E_{42}^* S_2^{-1}$, $X_{51} = S_3^{-1} E_{14}^*$, $X_{53} = S_3^{-1} E_{24}^*$, $X_{55} = S_3^{-1} E_{44}^* S_2^{-1}$.

We shall now investigate two solution strategies, both of which are based on the RSVD. The first one is an immediate consequence of Theorems 6, but, while elegant and extremely simple, might be considered as suffering from some "overkill." It is a direct application of the insights obtained in analysing the sum $A + BDC$. The second one is less elegant but is more in the line of results reported in [4] and [7]. It exploits the insights obtained from analysing the partitioned matrix $M = \left(\begin{smallmatrix} A & B \\ C & D^* \end{smallmatrix}\right)$.

3.2.3.1. Constrained total linear least squares directly via the RSVD. It is straightforward to show that the constrained total least squares problem can be recast as a minimum norm problem as discussed in Theorem 6 as follows.

Find the matrix $\tilde{D}$ of minimum norm $\|\tilde{D}\|$ such that

$$\operatorname{rank}\left(\begin{pmatrix} A & B \\ C & D^* \end{pmatrix} + \begin{pmatrix} 0_{m\times q} \\ I_q \end{pmatrix} \tilde{D}^* \begin{pmatrix} 0_{p\times n} & I_p \end{pmatrix}\right) \leq r.$$

The solution is an immediate consequence of Theorem 6.

COROLLARY 2. *The solution of the constrained total linear least squares problem follows from the application of Theorem* 6 *to the matrix triplet* A', B', C' *where*

$$A' = \begin{pmatrix} A & B \\ C & D^* \end{pmatrix}, \qquad B' = \begin{pmatrix} 0_{m\times q} \\ I_q \end{pmatrix}, \qquad C' = \begin{pmatrix} 0_{p\times n} & I_p \end{pmatrix}.$$

Hence, all that we need is the RSVD of the matrix triplet (A', B', C') and the truncation of the matrix S_1 as described in Theorem 6. It is interesting to also apply Theorem 3 to the matrix triplet (A', B', C'):

$$r_{a'b'} = \operatorname{rank}\begin{pmatrix} A & B & 0 \\ C & D^* & I_q \end{pmatrix} = r_{ab} + q,$$

$$r_{a'c'} = \operatorname{rank}\begin{pmatrix} A & B \\ C & D^* \\ 0 & I_p \end{pmatrix} = r_{ac} + p,$$

$$r_{a'b'c'} = \operatorname{rank}\begin{pmatrix} A & B & 0 \\ C & D^* & I_q \\ 0 & I_p & 0 \end{pmatrix} = r_a + p + q.$$

Hence, from Theorem 3, the minimum achievable rank is $r_{a'b'} + r_{a'c'} - r_{a'b'c'} = r_{ab} + r_{ac} - r_a$, which corresponds precisely to the result from Corollary 1.

As a special case, consider the Golub–Hoffman–Stewart result [17] for the total linear least squares solution of $(A \;\; B)x \approx 0$, where A is noise-free and B is contaminated with errors. Instead of applying the QR-SVD-least squares solution as discussed in [17], we could as well achieve the mixed linear–total linear least squares solution from the following.

Minimize $\|\tilde{B}\|$ such that

$$\operatorname{rank}((A \;\; B) - \tilde{B}(0_{p\times n} \;\; I_p)) \leq r,$$

where r is a prespecified integer. This can be done directly via the QSVD of the matrix pair $((A \ B), (0_{p\times n} \ I_p))$ and it is not too difficult to provide another proof of the Golub–Hoffman–Stewart result derived in [17], now in terms of the properties of the QSVD.

As a matter of fact, the RSVD of the matrix triplet of Corollary 2 allows us to provide a geometrical proof of constrained total linear least squares, in the line of the Golub–Hoffman–Stewart result, taking into account the structure of the matrices B' and C'. We shall not, however, consider this any further in this paper.

3.2.3.2. Solution via RSVD–OSVD. While the solution to the constrained total least squares problem as presented in Corollary 2 is extremely simple, we might object to it because of the apparent "overkill" in computing the RSVD of the matrix triplet (A', B', C'), where B' and C' have an extremely simple structure (zeros and the identity matrix). It will now be shown that the RSVD, combined with the OSVD, may lead to a computationally simpler solution, which more closely follows the lines of the solution as presented in [7].

Using the RSVD, we find that

$$\begin{pmatrix} A & B \\ C & D^* \end{pmatrix} = \begin{pmatrix} P^{-*} & 0 \\ 0 & U_c \end{pmatrix} \begin{pmatrix} S_a & S_b \\ S_c & U_c^* D^* V_b \end{pmatrix} \begin{pmatrix} Q^{-1} & 0 \\ 0 & V_b^* \end{pmatrix}.$$

Let $E^* = U_c^* D^* V_b$. Since U_c and V_b are unitary matrices, the problem can be restated as follows.

Find $\hat{E}$ such that $\|E - \hat{E}\|_F$ is minimal and

$$\operatorname{rank} \begin{pmatrix} S_a & S_b \\ S_c & \hat{E}^* \end{pmatrix} \leq r.$$

The constrained total least squares problem can now be solved as follows.

THEOREM 11 (RSVD–OSVD solution of constrained total least squares). *Consider the* OSVD

$$\begin{pmatrix} E_{11} - S_1^{-1} & E_{12} & E_{13} \\ E_{21} & E_{22} & E_{23} \\ E_{31} & E_{32} & E_{33} \end{pmatrix} = \sum_{i=1}^{r_e} u_i^e \sigma_i^e (v_i^e)^*,$$

where r_e is the rank of this matrix. The modification of minimal Frobenius norm follows immediately from the OSVD *of this matrix by truncating its dyadic decomposition after $r - r_{ab} - r_{ac} + r_a$ terms. Let*

$$\hat{E} = \sum_{i=1}^{r-r_{ab}-r_{ac}+r_a} u_i^e \sigma_i^e (v_i^e)^*.$$

Then the optimal $\hat{D}$ is given by

$$\hat{D} = V_b \begin{pmatrix} \hat{E} & 0 \\ 0 & 0 \end{pmatrix} U_c^*.$$

Proof. From Theorem 9, it follows that the rank of $\left(\begin{smallmatrix} A & B \\ C & D^* \end{smallmatrix}\right)$ can be reduced by reducing the rank of the matrix

$$\begin{pmatrix} E_{11} - S_1^{-1} & E_{12} & E_{13} \\ E_{21} & E_{22} & E_{23} \\ E_{31} & E_{32} & E_{33} \end{pmatrix}.$$

The matrix $\tilde{D}$ is then obtained from (12) by setting the blocks $\tilde{E}_{14}$, $\tilde{E}_{24}$, $\tilde{E}_{34}$, $\tilde{E}_{41}$, $\tilde{E}_{42}$, $\tilde{E}_{43}$, $\tilde{E}_{43}$ to 0 in order to minimize the Frobenius norm and then truncating the OSVD of the matrix above. □

We conclude this section by pointing out that more results as well as algorithms to solve total least squares problems with and without constraints and given covariance matrices, can be found in [7], [28], [29], and [31].

3.3. Generalized Gauss–Markov models with constraints. Consider the problem of minimizing $\|y\|^2 + \|z\|^2 = y^*y + z^*z$ over all vectors x, y, z satisfying

$$b = Ax + By, \qquad z = Cx$$

where A, B, C, b are given.

This formulation is a generalization of the conventional least squares problem where $B = I_m$ and $C = 0$. The formulation above admits singular or ill-conditioned matrices B and C. The problem formulation as presented here could be considered as a "square root" version of the problem as follows.

Find x such that

$$\|b - Ax\|_{W_b} + \|x\|_{W_c}$$

is minimized, where $\|u\|_{W_b} = u^*W_bu$ and W_b and W_c are nonnegative-definite symmetric matrices.

In the case that BB^* is nonsingular, we can put $W_b = (BB^*)^{-1}$ and $W_c = C^*C$. The solution can then be obtained as follows.

Minimize $\|y\|^2 + \|z\|^2$ where

$$y^*y = (b - Ax)^*W_b(b - Ax),$$
$$z^*z = x^*C^*Cx.$$

Setting the derivative with respect to x equal to 0, results in

$$x = (A^*W_bA + W_c)^{-1}A^*W_bb. \tag{14}$$

In the case where $W_b = I_m$ and $C = 0$, (14) reduces to the classical least squares expression. For the more general case, we can see a connection with so-called regularization problems. Consider the case where $C \neq 0$ and $B = I_m$. If the matrix A is ill conditioned (because of so-called collinearities, which are (almost) linear dependencies among the columns of A), the addition of the term C^*C may possibly make the sum better suited for numerical inversion than the original product A^*A, hence stabilizing the solution x.

The matrix B acts as a "static" noise filter: Typically, it is assumed that the vector y is normally distributed with the covariance matrix $E(yy^*)$ being a multiple of the identity. The error vector By for the first equation can only be in a direction which is present in the column space of B. If the observation vector b has some component in a certain direction not present in the column space of B, this component should be considered as error-free. The matrix C represents a weighting on the components of x. It reflects possible a priori information concerning the unknown components of x or may reflect the fact that certain components of x (or linear combinations thereof) are more "likely" or less costly than others. The fact that we try to minimize $y^*y + z^*z$ reflects the intention to explain as much as possible (i.e., min y^*y) in terms of the data (columns of the matrix A), taking into account a priori knowledge of the

geometrical distribution of the noise (the weighting W_b). The matrix C reflects the cost per component, expressing the preference (or prejudice?) of the modeller to use more of one variable in explaining the phenomenon than of another. In applications, however, typically the matrix A contains many more rows than columns, which corresponds to the fact that better results are to be expected if there are more equations (measurements) than unknowns. However, the condition that BB^* is nonsingular requires a priori knowledge concerning the statistics of the noise. Because typically this knowledge is rather limited, B will have fewer columns than rows, implying that BB^* is singular and (14) does not apply. In this case, however, the RSVD can be applied. It provides important geometrical information on the sensitivity of the solution. Inserting the RSVD of the matrix triplet (A, B, C), the problem can be rewritten as

$$\begin{aligned}(P^*b) &= S_a(Q^{-1}x) + S_b(V_b^*y),\\ (U_c^*z) &= S_c(Q^{-1}x).\end{aligned}$$

Define $b' = P^*b, x' = Q^{-1}x, y' = V_b^*y, z' = U_c^*z$. Then, with obvious partitionings of b', x', y', z', it follows that

$$\begin{array}{ll} b'_1 = S_1x'_1 + y'_1, & z'_1 = x'_1,\\ b'_2 = x'_2, & z'_2 = x'_2,\\ b'_3 = x'_3 + y'_2, & z'_3 = 0,\\ b'_4 = x'_4, & z'_4 = S_3x'_5,\\ b'_5 = S_2y'_4, & \\ b'_6 = 0. & \end{array}$$

Observe that $b'_6 = 0$ is a *consistency* condition. It reflects the fact that b is not allowed to have a component in a direction that is not present in the column space of $(A\ B)$. The components of x'_2 and x'_4 can be estimated without error while the fact that $b'_5 = S_2y'_4$ could be exploited to estimate the variance of the noise.

Most terms in the object function $y^*y + z^*z$ can now be expressed with the subvectors $x'_i, (i = 1, \cdots, 6)$,

$$\begin{aligned}y^*y + z^*z = {} & b'^*_1b'_1 + x'^*_1S_1^2x'_1 - 2b'^*_1S_1x'_1 + b'^*_3b'_3 + x'^*_3x'_3 - 2b'^*_3x'_3\\ & + y'^*_3y'_3 + b'^*_5S_2^{-2}b'_5 + x'^*_1x'_1 + x'^*_5S_3^2x'_5 + b'^*_2b'_2.\end{aligned}$$

The minimum solution follows from differentiation with respect to these vectors and results in

$$\begin{array}{lll} x'_1 = (I + S_1^2)^{-1}S_1b'_1, & y'_1 = (I + S_1^2)^{-1}b'_1, & z'_1 = (I + S_1^2)^{-1}S_1b'_1,\\ x'_2 = b'_2, & y'_2 = 0, & z'_2 = b'_2,\\ x'_3 = b'_3, & y'_3 = 0, & z'_3 = 0,\\ x'_4 = b'_4, & y'_4 = S_2^{-1}b'_5, & z'_4 = 0,\\ x'_5 = 0, & & \\ x'_6 = \text{arbitrary}. & & \end{array}$$

Statistical properties, such as (un)biasedness and consistency, can be analysed in the same spirit as in [23], where Paige has related the Gauss–Markov model without the z-equation, to the QSVD. Similarly, the RSVD also allows us to analyse the sensitivity of the solution. If, for instance, S_2 is ill conditioned, then the minimum of the object function will tend to be high, whenever b'_5 has strong components among the "weak" singular vectors of S_2, because of the term $b'^*_5S_2^{-2}b'_5$.

A related problem is the following.

Minimize y^*y subject to $b = Ax + By$ and $Cx = c$ where A, B, C, b, c are given.

This is also a Gauss–Markov linear estimation problem as in [23], but now with constraints. The solution is again straightforward from the RSVD. With $b' = P^*b$, $x' = Q^{-1}x$, $y' = V_b^*y$, $c' = U_c^*c$, and an appropriate partitioning, we find

$$\begin{array}{ll} x'_1 = c'_1, & y'_1 = b'_1 - S_1 c'_1, \\ x'_2 = c'_2 = b'_2, & y'_2 = 0, \\ x'_3 = b'_3, & y'_3 = 0, \\ x'_4 = b'_4, & y'_4 = S_2^{-1} b'_5, \\ x'_5 = S_3^{-1} c'_4, & \\ x'_6 = \text{arbitrary}. & \end{array}$$

Observe that $c'_2 = b'_2$ and $c'_3 = 0$ are two *consistency* conditions.

4. Conclusions and perspectives. In this paper, we have derived a generalization of the OSVD, the *restricted singular value decomposition* (RSVD), which has the OSVD, PSVD, and QSVD as special cases. A constructive proof, based upon a sequence of OSVDs and PSVDs can be found in [9]. We have also analysed in detail its structural and geometrical properties and its relations to generalized eigenvalue problems and canonical correlation analysis. It was shown how the RSVD is a valuable tool in the analysis and solution of rank minimization problems with restrictions. First, we have shown how to study expressions of the form $A+BDC$ and find matrices D of minimum norm that minimize the rank. It was demonstrated how this problem is connected to the concept of shorted operators and matrix balls. Second, we have analysed in detail low rank approximations of a partitioned matrix, when only one of its blocks can be modified. The close relation with generalized Schur complements was discussed and it was shown how the RSVD permits us to solve constrained total linear least squares problems with mixed exact and noisy data. Third, it was demonstrated how the RSVD provides an elegant solution to Gauss–Markov models with constraints. The fact that the RSVD is only the tip of an iceberg of generalizations of the OSVD for 2, 3, 4, $\cdots$ matrices, is fully explored in [11].

Acknowledgments. We would like to thank Hongyuan Zha, Sabine Van Huffel and the referees for their detailed comments and suggestions, which allowed us to improve considerably an earlier version of this paper. We would also like to thank Dan Boley for providing us with an English translation of Beltrami's original paper.

REFERENCES

[1] L. AUTONNE, *Sur les groupes linéaires, réels et orthogonaux*, Bull. Sci. Math., France, 30 (1902), pp. 121–133.

[2] E. BELTRAMI, *Sulle funzioni bilineari*, in Giornale di Mathematiche, G. Battagline and E. Fergola, eds., 11 (1873), pp. 98–106.

[3] Å. BJÖRCK AND G. H. GOLUB, *Numerical methods for computing angles between linear subspaces*, Math. Comp., 27 (1973) pp. 579–594.

[4] D. CARLSON, *What are Schur complements, anyway?*, Linear Algebra Appl., 74 (1986), pp. 257–275.

[5] J. S. CHIPMAN, *Estimation and aggregation in econometrics: An application of the theory of generalized inverses*, in Generalized Inverses and Applications, Academic Press, New York, 1976, pp. 549–769.

[6] C. DAVIS, W. M. KAHAN, AND H. F. WEINBERGER, *Norm-preserving dilations and their application to optimal error bounds*, SIAM J. Numer. Anal., 19 (1982), pp. 445–469.

[7] J. W. DEMMEL, *The smallest perturbation of a submatrix which lowers the rank and constrained total least squares problems*, SIAM J. Numer. Anal., 24 (1987), pp. 199–206.

[8] B. De Moor and G. H. Golub, *Generalized singular value decompositions: A proposal for a standardized nomenclature*, Numerical Analysis Project Manuscript NA-89-03, Department of Computer Science, Stanford University, Stanford, CA, April 1989; also in ESAT-SISTA Report 1989-10, Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium, April 1989.

[9] ———, *The restricted singular value decomposition: Properties and applications*, Numerical Analysis Project Manuscript NA-89-04, Department of Computer Science, Stanford University, Stanford, CA, April 1989; also in ESAT-SISTA Report 1989-09, Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium, April 1989.

[10] B. De Moor, *On the structure and geometry of the product singular value decomposition*, Numerical Analysis Project Manuscript NA-89-05, Department of Computer Science, Stanford University, Stanford, CA, May 1989; also in ESAT-SISTA Report 1989-12, Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium, May 1989.

[11] B. De Moor and H. Zha, *A tree of generalizations of the ordinary singular value decomposition*, ESAT-SISTA Report 1989-21 and revised version ESAT-SISTA Report 1990-11, Department of Electrical Engineering, Katholieke Universiteit Leuven, Belgium; also in the special issue on Matrix Canonical Forms, Linear Algebra Appl., to appear.

[12] C. Eckart and G. Young, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218.

[13] L. M. Ewerbring and F. T. Luk, *Canonical correlations and generalized* SVD: *Applications and new algorithms*, Proc. SPIE, Vol. 977, Real Time Signal Processing XI, paper 23, 1988.

[14] K. V. Fernando and S. J. Hammarling, *A product induced singular value decomposition for two matrices and balanced realisation*, in Linear Algebra in Signal Systems and Control, B. N. Datta, C. R. Johnson, M. A. Kaashoek, R. Plemmons, and E. Sontag, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1988, pp. 128–140.

[15] G. H. Golub and C. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[16] ———, *An analysis of the total least squares problem*, SIAM J. Numer. Anal., 17 (1980), pp. 883–893.

[17] G. H. Golub, A. Hoffman, and G. W. Stewart, *A generalization of the Eckart–Young–Mirsky matrix approximation theorem*, Linear Algebra Appl., 88/89 (1987), pp. 317–327.

[18] W. E. Larimore, *Identification of nonlinear systems using canonical variate analysis*, in Proc. 26th Conference on Decision and Control, Los Angeles, CA, December 1987.

[19] C. Lawson and R. Hanson, *Solving Least Squares Problems*, Prentice–Hall, Englewood Cliffs, NJ, 1974.

[20] S. K. Mitra and M. L. Puri, *Shorted matrices—An extended concept and some applications*, Linear Algebra Appl., 42 (1982), pp. 57–79.

[21] M. Z. Nashed, ed., *Generalized Inverses and Applications*, Academic Press, New York, 1976.

[22] C. C. Paige and M. A. Saunders, *Towards a generalized singular value decomposition*, SIAM J. Numer. Anal., 18 (1981), pp. 398–405.

[23] C. C. Paige, *The general linear model and the generalized singular value decomposition,* Linear Algebra Appl., 70 (1985), pp. 269–284.

[24] R. Penrose, *A generalized inverse for matrices*, Proc. Cambridge Philos. Soc., 51 (1955), pp. 406–413.

[25] ———, *On best approximate solutions of linear matrix equations*, Proc. Cambridge Philos. Soc., 52 (1956), pp. 17–19.

[26] J. J. Sylvester *Sur la réduction biorthogonale d'une forme linéo-linéaire à sa forme canonique*, Comptes Rendus, CVIII, 1889, pp. 651–653.

[27] J. Vandewalle and B. De Moor, *A variety of applications of the singular value decomposition,* in SVD and Signal Processing: Algorithms, Applications and Architectures, E. Deprettere, ed., North-Holland, Amsterdam, 1988, pp. 43–91.

[28] S. Van Huffel and J. Vandewalle, *Analysis and properties of the generalized total least squares problem $Ax = B$ when somme or all columns in A are subject to error*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 294–315.

[29] S. Van Huffel and H. Zha, *Restricted total least squares: A unified approach for solving (generalized) (total) least squares problems with(out) equality constraints*, ESAT-SISTA Report 1989-05, Department of Electrical Engineering, Katholieke Universiteit Leuven, Leuven, Belgium, March 1989.

[30] C. F. Van Loan, *Generalizing the singular value decomposition*, SIAM J. Numer. Anal., 13 (1976), pp. 76–83.

[31] G. A. WATSON, *The smallest perturbation of a submatrix which lowers the rank of the matrix*, IMA J. Numer. Anal., 8 (1988), pp. 295–303.

[32] H. ZHA, *Restricted* SVD *for matrix triplets and rank determination of matrices*, Scientific Report 89-2, Konrad-Zuse-Zentrum für Informationstechnik, Berlin, Germany, 1989.

[33] ———, *A numerical algorithm for computing the* RSVD *for matrix triplets*, Scientific Report 89-1, Konrad-Zuse-Zentrum für Informationstechnik, Berlin, Germany, 1989.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 426–431, July 1991

$O(n^2)$ REDUCTION ALGORITHMS FOR THE CONSTRUCTION OF A BAND MATRIX FROM SPECTRAL DATA*

GREGORY S. AMMAR† AND WILLIAM B. GRAGG‡

Abstract. Efficient rotation patterns are presented that provide stable $O(n^2)$ algorithms for the construction of a real symmetric band matrix having specified eigenvalues and first p components of its normalized eigenvectors. These methods can also be used in the second phase of the construction of a band matrix from the interlacing eigenvalues as described in [*Linear Algebra Appl.*, 40 (1981), pp. 79–87]. Previously presented algorithms for these reductions that use elementary orthogonal similarity transformations require $O(n^3)$ arithmetic operations.

Key words. band matrix, inverse eigenvalue problem, Givens rotations

AMS(MOS) subject classification. 65F30

1. Introduction. Let A be a real symmetric $(2p+1)$-band matrix of order n, and let A_k denote the trailing principal submatrix of $A = A_n$ of order k. It is well known that the eigenvalues of A_k interlace those of A_{k+1} for each $k < n$, and moreover, given real numbers $\lambda_j^{(k)}$ $(1 \leqq j \leqq k,\ n - p \leqq k \leqq n)$ satisfying

$$\lambda_j^{(k+1)} \leqq \lambda_j^{(k)} \leqq \lambda_{j+1}^{(k+1)}, \tag{1}$$

there is a $(2p+1)$-band matrix $A = A_n$ such that the eigenvalues of A_k are $\{\lambda_j^{(k)}\}_{j=1}^{k}$ for each k. In general, this band matrix is not uniquely determined.

The problem of constructing a band matrix from the interlacing eigenvalues (1) is considered in [2] and [1]. A survey of this problem and some related inverse eigenvalue problems is given in [3]. In [2] the interlacing eigenvalues are used to determine the first p components of the normalized eigenvectors of A, and the remaining components of the eigenvectors (and hence A) are constructed using a block Lanczos process. In [1] a matrix of bordered structure (where the trailing principal submatrix of order p is diagonal) is constructed that satisfies the required spectral conditions. Householder transformations that preserve the eigenvalues of the trailing submatrices are then applied to reduce this bordered matrix to band form. This reduction procedure uses $O(n^3)$ arithmetic operations.

In this note we present efficient rotation patterns that provide stable $O(n^2)$ procedures that can be used in the second step (the reduction step) of either of the above methods. These algorithms provide solutions to the open problem posed in [3, p. 615]. The first rotation pattern we present can be considered as the generalization to band matrices of Rutishauser's procedure for the construction of Jacobi matrices from spectral data presented in [4].

2. Efficient reduction algorithms. The reduction step in [2] can be described as follows. Given $\{\lambda_j\}_{j=1}^{n}$ and an $n \times p$ matrix Q_1 with orthonormal columns, construct a

* Received by the editors October 24, 1988; accepted for publication (in revised form) January 31, 1990. This research was supported in part by National Science Foundation grant DMS-8704196.

† Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115 (ammar@math.niu.edu).

‡ Department of Mathematics, Naval Postgraduate School, Monterey, California 93943. The research of this author was supported in part by the Foundation Research Program of the Naval Postgraduate School (na.gragg@na-net.ornl.gov).

$(2p+1)$-band matrix A having eigenvalues λ_j such that Q_1^T forms the first p rows of the (orthogonal) eigenvector matrix for A. This reduction can be performed using a sequence of orthogonal similarity transformations whose composition results in an orthogonal transformation Q such that

$$\begin{bmatrix} I_p & 0 \\ 0 & Q^T \end{bmatrix}\begin{bmatrix} X & Q_1^T \\ Q_1 & \Lambda \end{bmatrix}\begin{bmatrix} I_p & 0 \\ 0 & Q \end{bmatrix}=\begin{bmatrix} X & I_p & 0 \\ I_p & & \\ 0 & & A \end{bmatrix} \tag{2}$$

is a $(2p+1)$-band matrix of order $n+p$. The trailing principal submatrix $A = A_n$ then satisfies the required spectral conditions, and Q_1 comprises the first p columns of Q. (The matrix X is arbitrary and remains unchanged.)

In the algorithm given in [1], an $n \times n$ matrix of the bordered form

$$B=\begin{bmatrix} B_0 & B_1^T \\ B_1 & D \end{bmatrix}, \tag{3}$$

where D is a diagonal matrix of order $n-p$, is constructed such that the trailing principal submatrices of orders $n-p$ through n of B have prescribed eigenvalues. Householder transformations that do not involve the first p coordinate axes are then used to transform B to a $(2p+1)$-band matrix A while preserving the eigenvalues of the trailing principal submatrices. In particular, the composition of these Householder transformations yields an orthogonal matrix U of order $n-p$ such that

$$A=\begin{bmatrix} I_p & 0 \\ 0 & U^T \end{bmatrix}\begin{bmatrix} B_0 & B_1^T \\ B_1 & D \end{bmatrix}\begin{bmatrix} I_p & 0 \\ 0 & U \end{bmatrix} \tag{4}$$

is a $(2p+1)$-band matrix of order n. Thus, the reduction of the matrices in (2) and (4) is essentially the same problem. (Observe that the identity matrix in (2) arises because the columns of Q_1 are orthonormal.) We will describe our efficient rotation patterns in terms of the reduction of a matrix in the bordered form (3).

The efficient reduction to band form that generalizes the algorithm of [4] is obtained by performing plane rotations to introduce appropriate zeros in B row-by-row beginning at row $p+2$, in such a way that *the intermediate matrices remain sparse*. In contrast, a Householder transformation to introduce zeros in the first column of the matrix will result in a full matrix, and the subsequent Householder transformations must be performed on full matrices.

Let $R(A, j, k, l) = GAG^T$, where G is the elementary Givens rotation in the (j, k)-plane that annihilates a_{kl}. Thus, G is the identity matrix if $a_{kl} = 0$. If $a_{kl} \neq 0$, then G is the identity matrix apart from the 2×2 submatrix formed from rows and columns j and k, which is given by

$$G\begin{bmatrix} j, & k \\ j, & k \end{bmatrix}=\begin{bmatrix} c & s \\ -s & c \end{bmatrix},$$

where $c := a_{jl}/\sqrt{a_{jl}^2 + a_{kl}^2}$ and $s := a_{kl}/\sqrt{a_{jl}^2 + a_{kl}^2}$. Our algorithm for reducing the bordered matrix to band form is then given as follows.

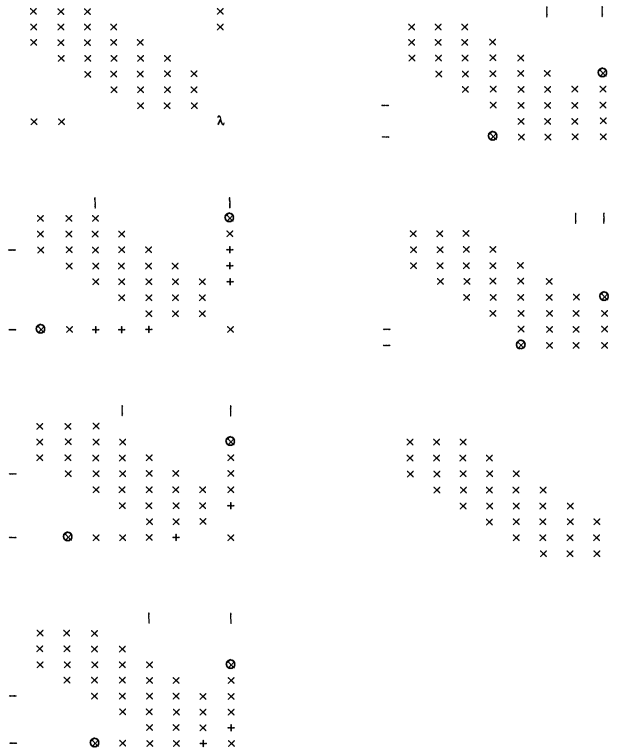


FIG. 1. *Rotations are performed in coordinate planes* (3, 8), (4, 8), (5, 8), (6, 8), *and* (7, 8) *to introduce the appropriate zeros in the eighth row.*

ALGORITHM 1.

for $k = p+2, \cdots, n$
  for $j = p+1, \cdots, k-1$
    $A := R(A, j, k, j-p)$

To see how the sparsity is preserved, consider the example in Fig. 1. There $n = 8$, $p = 2$, and the necessary zeros have already been introduced in rows 4 through 7. Nonzero entries are represented by $\times$, a Givens rotation is performed in the indicated planes to annihilate the circled entry, and the symbol $+$ indicates the "fillin" (i.e., the additional nonzero entries) introduced by the rotation. The first rotation, in the (3, 8) plane, annihilates $a_{8,1}$ and creates $p + 1 = 3$ additional nonzero entries. (We count a_{ij} and a_{ji} as one element.) The successive rotations introduce at most one additional nonzero element each, so there are at most $2p + 1 = 5$ nonzero entries on the eighth row at any time. We can therefore perform each elementary similarity transformation on A in $O(p)$ arithmetic work. Thus the amount of computation required by the reduction is $O(pn^2)$.

Below is an explicit description of Algorithm 1 that involves only the lower-triangular part of the symmetric matrix A.

ALGORITHM 1.

Input: a symmetric matrix $A = [a_{j,k}]_{j,k=1}^{n}$ whose trailing principal submatrix of order $n - p$ is diagonal.

Output: a symmetric $(2p+1)$-band matrix A whose trailing principal submatrices of orders $n-p$ through n are orthogonally similar with those of the input matrix.

for $k=p+2, \cdots, n$
  for $j=p+1, \cdots, k-1$
    if $a_{k,j-p} \neq 0$ **then**
      $\rho := \sqrt{a_{j,j-p}^2 + a_{k,j-p}^2}$;
      $c := a_{j,j-p}/\rho$; $s := a_{k,j-p}/\rho$;
      $a_{j,j-p} := \rho$; $a_{k,j-p} := 0$;
      for $i=p-1, p-2, \cdots, 1$

$$\begin{bmatrix} a_{j,j-i} \\ a_{k,j-i} \end{bmatrix} := \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} a_{j,j-i} \\ a_{k,j-i} \end{bmatrix}$$

      for $i=j+1, j+2, \cdots, \min\{j+p, k-1\}$.

$$\begin{bmatrix} a_{i,j} \\ a_{k,i} \end{bmatrix} := \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} a_{i,j} \\ a_{k,i} \end{bmatrix}$$

      $u := a_{j,j}$; $v := a_{k,k}$; $w := a_{k,j}$;
      $a_{j,j} := c^2u + s^2v + 2csw$; $a_{k,k} := c^2v + s^2u - 2csw$;
      $a_{k,j} := cs(v-u) + (c^2-s^2)w$.

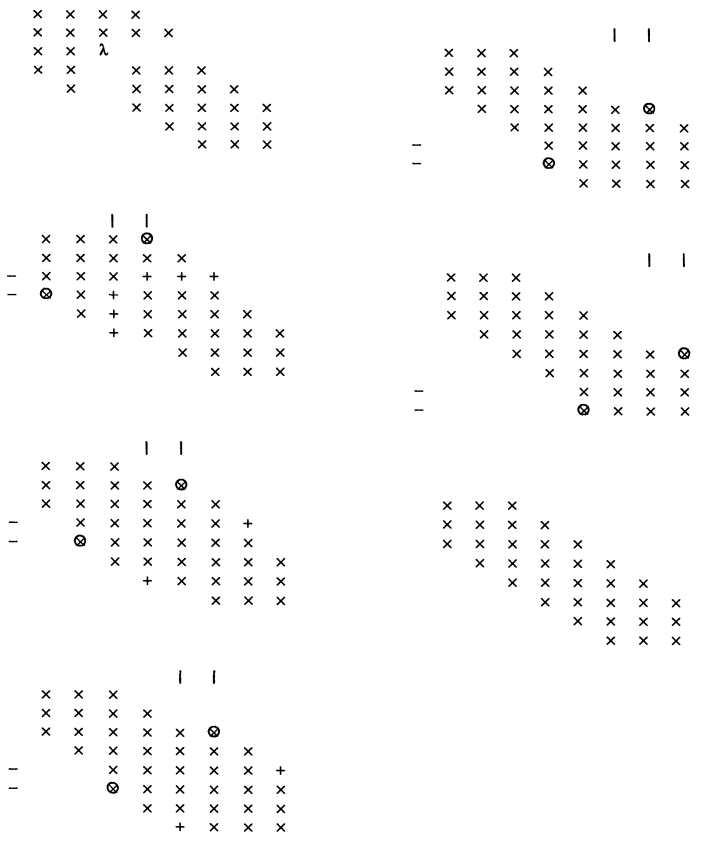


FIG. 2. *Rotations are performed in coordinate planes* (3, 4), (4, 5), (5, 6), (6, 7), *and* (7, 8) *to introduce the appropriate zeros.*

Another rotation pattern can be obtained by introducing the zeros in (3) from the bottom up along downwardly sloping diagonals.

ALGORITHM 2.

for $k = n-1, n-2, \cdots, p+1$

  for $j = k+1, \cdots, n$

    $A := R(A, k-1, k, j-k)$

One step of this procedure is illustrated in Fig. 2. Observe that Algorithm 2 creates the same amount of fillin as Algorithm 1.

In fact, several patterns of rotations exist that preserve the sparsity of the intermediate matrices. For example, Algorithms 1 and 2 can be combined to build the band matrix according to any ordering for which the intermediate band (sub)matrices occupy contiguous rows and columns of the work array.

3. Numerical results. Numerical experiments verify that our efficient rotation pattern produces accurate results in lower-order work than the Householder reduction technique. These experiments were performed on the VAX 11/750 at Northern Illinois University. We will not attempt to analyze the numerical sensitivity of the inverse eigenproblem. Our only aim is to show that an efficient rotation pattern produces errors comparable with the Householder reduction technique in lower-order work.

The following experiment was performed. The method of [1] was used to create a bordered matrix whose trailing principal matrices of order $n - p$ through n have specified eigenvalues. This matrix was then reduced to $(2p + 1)$-band form using

I. The Householder reduction procedure of [1];

II. The efficient rotation pattern of Algorithm 1.

We calculated the average and maximum error among the assigned eigenvalues of the trailing principal submatrices of orders $n - p$ through n relative to the Frobenius norm of the band matrix. The results displayed in Table 1 were obtained by assigning the eigenvalues of A_k, $n - p \leqq k \leqq n$, to be the integers $2j + (n - k - 1)$, $1 \leqq j \leqq k$. Experiments were carried out on a variety of other problems with similar results.

Tables 2(a) and 2(b) show average CPU times used by each reduction scheme for various values of n and p. Table 2(c) shows the corresponding ratios of the time used by the Householder reduction to that of our rotation pattern. These ratios represent the speedup factors of Procedure II relative to Procedure I. Note that for fixed n, the amount

TABLE 1
Relative errors in eigenvalues.

		Average error		Maximum error	
n	p	I	II	I	II
10	2	3.39e−08	1.58e−08	2.61e−07	5.23e−08
20	2	3.58e−08	2.16e−08	2.21e−07	9.42e−08
50	2	2.40e−08	2.96e−08	1.31e−07	1.12e−07
10	4	3.52e−08	1.50e−08	1.57e−07	5.23e−08
20	4	2.00e−08	2.86e−08	1.10e−07	1.11e−07
50	4	2.71e−08	2.94e−08	1.12e−07	1.68e−07
10	6	2.08e−08	1.29e−08	7.84e−08	5.23e−08
20	6	2.73e−08	3.18e−08	7.39e−08	1.11e−07
50	6	2.91e−08	5.04e−08	1.68e−07	1.50e−07

TABLE 2(a)
Average timings for Procedure I (CPU *seconds*).

n	10	20	30	40	50	100	200
p							
1	0.029	0.182	0.550	1.231	2.342	17.858	140.070
2	0.023	0.163	0.534	1.199	2.286	17.632	139.693
5	0.013	0.131	0.456	1.081	2.119	17.127	137.837
10		0.072	0.327	0.868	1.796	15.852	133.120
20			0.117	0.476	1.178	13.503	123.227

TABLE 2(b)
Average timings for Procedure II (CPU *seconds*).

n	10	20	30	40	50	100	200
p							
1	0.022	0.087	0.207	0.381	0.596	2.493	10.273
2	0.018	0.099	0.244	0.453	0.734	3.112	13.037
5	0.009	0.103	0.302	0.618	1.044	4.807	20.757
10		0.063	0.287	0.692	1.275	6.937	32.130
20			0.104	0.451	1.110	9.250	50.007

TABLE 2(c)
Ratios of CPU *times.*

n	10	20	30	40	50	100	200
p							
1	1.346	2.096	2.661	3.232	3.931	7.162	13.634
2	1.333	1.639	2.188	2.645	3.114	5.666	10.715
5	1.364	1.266	1.511	1.748	2.030	3.563	6.641
10		1.147	1.136	1.253	1.408	2.285	4.143
20			1.128	1.055	1.062	1.460	2.464

of computation required by Procedure I decreases as p increases, while that of Procedure II is often increasing as a function of p when p is small. These results show that our rotation pattern is consistently more efficient than the Householder reduction technique. The relative efficiency of the rotation pattern generally increases as n increases and decreases as p increases.

REFERENCES

[1] F. W. BIEGLER-KÖNIG, *Construction of band matrices from spectral data*, Linear Algebra Appl., 40 (1981), pp. 79–87.

[2] D. BOLEY AND G. H. GOLUB, *Inverse eigenvalue problems for band matrices*, in Proc. Biennial Conference Held at Dundee, 1977, Lecture Notes in Mathematics 630, Numerical Analysis, G. A. Watson, ed., Springer-Verlag, New York, 1978, pp. 23–31.

[3] ———, *A survey of matrix inverse eigenvalue problems*, Inverse Problems, 3 (1987), pp. 595–622.

[4] W. B. GRAGG AND W. J. HARROD, *The numerically stable reconstruction of Jacobi matrices from spectral data*, Numer. Math., 44 (1984), pp. 317–355.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 432–448, July 1991

003

TRIDIAGONAL APPROACH TO THE ALGEBRAIC ENVIRONMENT OF TOEPLITZ MATRICES, PART II: ZERO AND EIGENVALUE PROBLEMS*

P. DELSARTE† AND Y. GENIN†

Abstract. The central subject of this paper is the three-term recurrence formula satisfied by the symmetric (first-kind) and antisymmetric (second-kind) polynomials relative to a given sequence of reflection coefficients, the last element of which has unit modulus. The theory of these polynomials is shown to have interesting analogies with the classical theory of orthogonal polynomials on the real line. In the case of real data, the former is equivalent to a special case of the latter (by a change of variable). The main application considered here is the problem of computing the zeros of the highest-degree symmetric polynomial, which is identified as a predictor polynomial. This problem occurs not only in some modelling techniques for digital signal processing, but can also be interpreted as the eigenvalue problem for a unitary Hessenberg matrix. Attractive solution methods are derived from the "tridiagonal approach," based on the three-term recurrence relation.

Key words. three-term recurrence, positive-definite tridiagonal matrix, unitary Hessenberg matrix, rational lossless functions, orthogonal polynomials

AMS(MOS) subject classifications. 42C05, 30C15, 15A18

1. Introduction.

In a companion paper [9], we have seen how to associate a family of *symmetric polynomials* $p_0(z), \cdots, p_n(z)$ with a sequence of reflection coefficients $\rho_1, \cdots, \rho_n$ satisfying $|\rho_k| < 1$ for $k = 1, \cdots, n-1$ and $|\rho_n| = 1$. This family is specified by a fixed complex number of unit modulus, denoted by ζ_0 and called the circle parameter. The symmetric polynomials $p_k(z)$ obey a *three-term recurrence relation* of the simple form $p_{k+1}(z) = (\alpha_k + \bar{\alpha}_k z)p_k(z) - zp_{k-1}(z)$, where the coefficients α_k can be determined explicitly from the data $\zeta_0, \rho_1, \cdots, \rho_n$. By construction, the final polynomial $p_n(z)$ is proportional to the classical ρ_n-symmetric predictor polynomial $a_n(z)$ produced by the Szegö–Levinson formula $a_k(z) = a_{k-1}(z) + \rho_k z\hat{a}_{k-1}(z)$ for $k = 1, \cdots, n$.

The present contribution provides further results on these polynomials $p_k(z)$, based essentially on the three-term recurrence relation. As far as applications are concerned, the main objective is to derive some methods for computing the zeros of the predictor polynomial $a_n(z)$. In this framework it is assumed that the relevant data are the reflection coefficients, although the proposed methods can be modified so as to deal with the case where the data are the entries of the corresponding Toeplitz matrix C_n. (Recall that the coefficient vector $\mathbf{a}_n$ of $a_n(z)$ is the solution of the homogeneous linear system $C_n\mathbf{a}_n = \mathbf{0}$.) The zeros $\zeta_1, \cdots, \zeta_n$ of $a_n(z)$ are known to be distinct and to have unit modulus [19]. In digital signal processing applications, the problem of computing these zeros occurs in the *Pisarenko modelling technique* [22], and in the *composite sinusoidal modelling technique* [24]. Furthermore, the same problem is essentially equivalent to that of *computing the eigenvalues of a unitary Hessenberg matrix* [2], [17], [18], which is a significant subject since any unitary matrix can be reduced to Hessenberg form by simple transformations.

Section 2 contains some preliminary material concerning the symmetric polynomials $p_k(z)$. Without real loss of generality, we restrict our attention to the regular case $a_n(\zeta_0) \neq 0$, implying $p_k(\zeta_0) \neq 0$ for all k. Our results include a duality relation induced by reversing the sequence of the recurrence coefficients α_k, and some important *Chris-*

* Received by the editors March 24, 1988; accepted for publication (in revised form) December 13, 1989.

† Philips Research Laboratory, Av. Albert Einstein, 4, 1348 Louvain-la-Neuve, Belgium (phd@prlb.philips.be and yg@prlb.philips.be).

toffel–Darboux-type formulas. When written in matrix form, these formulas involve an accretive bidiagonal matrix or, after symmetrization, a *positive-definite tridiagonal matrix* (see [9]) built from the coefficients α_k. This explains our "tridiagonal approach" terminology.

Section 3 is concerned with two *lossless rational functions*, dual of each other (in the sense above), having $p_n(z)$ as their denominators. In fact, it deals with *Gaussian quadrature on the unit circle*. (The intimate connection between this subject and unitary Hessenberg matrices is discussed in [16] and [18].) The poles of these lossless functions, which are the zeros $\zeta_1, \cdots, \zeta_n$ of $a_n(z)$, yield the *mass points* of the discrete positive measure (defined on the unit circle) with respect to which the polynomials $\hat{a}_k(z)$ are pairwise orthogonal, and the corresponding residues yield the appropriate *weights*. We derive Christoffel-type formulas, involving the values $p_k(\zeta_j)$, for the weights associated with the mass points.

A general technique for computing the zeros of $a_n(z)$ is developed in § 4. These zeros are shown to be obtainable as the generalized eigenvalues of a well-defined lower and upper bidiagonal accretive-type pencil constructed directly from the recurrence coefficients α_k. By use of a simple change of variable, this pencil can be transformed into a *tridiagonal Hermitian-positive pencil*. As an application, we consider the problem of computing the eigenvalues of a unitary Hessenberg matrix. The exact connection with our zero problem and the applicability of the proposed technique is explained from the parameterization of such a matrix in terms of a sequence of reflection coefficients [17], [21]. This provides a clear interpretation of the "duality operation" $\rho_k \rightarrow \rho_n \bar{\rho}_{n-k}$ investigated in [9].

The case where the reflection coefficients ρ_k are *real*, and the circle parameter ζ_0 is chosen to be 1 or -1, deserves a special study and is treated in detail in § 5. Here, instead of a change of variable expressed by a cotangent function, as in the preceding section, we use a cosine function (roughly speaking). This transforms the family of symmetric polynomials $p_k(z)$ into a family of even/odd real polynomials $\phi_k(x)$ which are orthogonal on the interval $[-1, 1]$ with respect to an appropriate positive measure [5], [12]; in fact, they satisfy the standard recurrence relation $\phi_{k+1}(x) = 2|\alpha_k|\phi_k(x) - \phi_{k-1}(x)$. As a result, we can reduce the zero problem for the predictor polynomial $a_n(z)$ to the *eigenvalue problem for any of four well-defined tridiagonal matrices* constructed from the reflection coefficients (see [2], [4], and [5] in that respect).

2. Preliminary results. To start, let us consider a family of polynomials $p_0(z)$, $p_1(z), \cdots, p_n(z)$, with complex coefficients, satisfying the Frobenius-type *three-term recurrence relation*

$$p_{k+1}(z) = (\alpha_k + \bar{\alpha}_k z)p_k(z) - zp_{k-1}(z), \tag{2.1}$$

for $k = 0, \cdots, n-1$, with $p_{-1}(z) = 0$ and $p_0(z) = p_0$, a nonzero real number. Given a sequence of nonzero complex numbers $\alpha_0, \alpha_1, \cdots, \alpha_{n-1}$, the formula (2.1) produces a polynomial $p_k(z)$ of degree k that enjoys the *symmetry* property $\hat{p}_k(z) = p_k(z)$, for $0 \leqq k \leqq n$. Without loss of generality, to agree with [9], we make the normalization assumption

$$p_0^2 |\alpha_0|^2 = 1. \tag{2.2}$$

Such a family of symmetric polynomials $p_k(z)$ has been constructed in the companion paper [9] from a *nonnegative-definite Toeplitz matrix* $C_n = [c_{i-j} : 0 \leqq i, j \leqq n]$ *of order* $n+1$ *and rank* n or, equivalently, from a *sequence of reflection coefficients* $\rho_1, \cdots, \rho_{n-1}$, ρ_n *satisfying* $|\rho_k| < 1$ *for* $k = 1, \cdots, n-1$ *and* $|\rho_n| = 1$. Let us now briefly recall the explicit connections that are needed in the sequel.

Without loss of generality, we assume that $|\alpha_0| > \frac{1}{2}$ (in order to agree with the setting of [9]). Then there exists a complex number ζ_0 of unit modulus satisfying $\mathrm{Re}\,(\zeta_0^{-1/2}\alpha_0) = \frac{1}{2}$. (Here and in the sequel, $\zeta_0^{1/2}$ denotes either of the square roots of ζ_0.) As in [9], the number ζ_0 will be referred to as the *circle parameter*. From α_0 let us define another unit modulus complex number, that is $\omega_1 = \bar{\alpha}_0/\alpha_0$. Thus we obtain the identity

$$\alpha_0 = (\zeta_0^{-1/2} + \zeta_0^{1/2}\omega_1)^{-1}. \tag{2.3}$$

Note that we have the property $\omega_1\zeta_0 \neq -1$, which means that we consider only the *regular case* of the general theory developed in [9]. Recall that this restriction entails no essential loss of generality (see [9, § 4]).

From the numbers α_k in (2.1) we define the sequence $(\lambda_1, \lambda_2, \cdots, \lambda_n)$ of *Jacobi parameters* λ_k by the recurrence

$$\lambda_{k+1} = 2\,\mathrm{Re}\,(\zeta_0^{-1/2}\alpha_k) - \lambda_k^{-1}, \tag{2.4}$$

with $\lambda_0 = \infty$. Thus we have $\lambda_1 = 1$. As shown in [9], the polynomials $p_k(z)$ correspond to a suitable Toeplitz matrix C_n under the necessary and sufficient condition $\lambda_k > 0$ for $k = 2, \cdots, n$. Using (2.1) and (2.4) we obtain the identity

$$p_k(\zeta_0) = p_0\zeta_0^{k/2}\lambda_1\lambda_2\cdots\lambda_k. \tag{2.5}$$

The recurrence coefficients α_k can be determined from the circle parameter ζ_0 and the reflection coefficients $\rho_1, \cdots, \rho_n$ as follows. First, the sequence of *pseudoreflection coefficients* ω_k is computed by means of the formula

$$\omega_k = (\rho_k + \zeta_0\omega_{k+1})(1 + \zeta_0\omega_{k+1}\bar{\rho}_k)^{-1}, \tag{2.6}$$

for $k = n-1, n-2, \cdots, 1$, with the initial value $\omega_n = \rho_n$. Since $|\rho_n| = 1$, this yields the property $|\omega_k| = 1$, for $n \geq k \geq 1$, by induction. The final value ω_1 allows us to determine α_0 by use of (2.3). Next, the numbers $\alpha_1, \cdots, \alpha_{n-1}$ are obtainable by means of the recurrence formula

$$\alpha_k = \zeta_0\alpha_{k-1}^{-1}(1 + \zeta_0\omega_k\bar{\rho}_{k-1})^{-1}(1 - \bar{\omega}_k\rho_k)^{-1}, \tag{2.7}$$

with $\rho_0 = 1$ (by convention). An interpretation of these results in the Toeplitz environment can be found in [9].

Certain applications involve not only the family of "first-kind" (symmetric) polynomials $p_k(z)$, but also an appropriate family of associated *second-kind polynomials*, denoted by $q_k(z)$. These satisfy exactly the same recurrence relation (2.1) as the first-kind polynomials. On the other hand, they enjoy the *antisymmetry* property $\hat{q}_k(z) = -q_k(z)$. In this paper, we choose the initial conditions

$$q_0(z) = 0, \qquad q_1(z) = 2c_0p_0^{-1}\zeta_0^{-1/2}(\zeta_0 - z). \tag{2.8}$$

They imply $q_k(\zeta_0) = 0$ for all $k \geq 0$, in contrast with the property $p_k(\zeta_0) \neq 0$ for all $k \geq 0$. (Polynomials $q_k(z)$ are called "shifted" second-kind polynomials in [9], and are denoted by $q'_k(z)$.) Using both versions of (2.1), we obtain the remarkable identity

$$p_k(z)q_{k+1}(z) - q_k(z)p_{k+1}(z) = 2c_0\zeta_0^{-1/2}(\zeta_0 - z)z^k, \tag{2.9}$$

for $k = 0, \cdots, n-1$. This makes sense for $k = -1$ if we set $q_{-1}(z) = -z^{-1}q_1(z)$.

Next, let us examine an interesting duality in the theory. (It is different from that investigated in § 4 of [9].) This duality is produced by transforming the coefficient sequence $(\alpha_0, \alpha_1, \cdots, \alpha_{n-1})$ into its *mirror image* $(\check{\alpha}_0, \check{\alpha}_1, \cdots, \check{\alpha}_{n-1})$, where

$$\check{\alpha}_k = \alpha_{n-1-k} \quad \text{for } 0 \leq k \leq n-1. \tag{2.10}$$

Let $\check{p}_k(z)$ and $\check{q}_k(z)$ denote the first-kind and second-kind polynomials of degree k generated by the three-term recurrence (2.1), where α_k is replaced by $\check{\alpha}_k$, while the initial conditions remain the same as in the original case, i.e., $\check{p}_{-1}(z) = 0$, $\check{p}_0(z) = p_0$, and $\check{q}_0(z) = 0$, $\check{q}_1(z) = q_1(z)$. Then we have the simple *duality relations*

$$\begin{aligned} &\check{p}_n(z) = p_n(z), \qquad \check{p}_{n-1}(z) = p_0 q_n(z)/q_1(z), \\ &\check{q}_n(z) = q_1(z)p_{n-1}(z)/p_0, \qquad \check{q}_{n-1}(z) = q_{n-1}(z), \end{aligned} \tag{2.11}$$

between polynomials of degree n and $n-1$.

To prove this, let us introduce the 2×2 matrix $W_k(z)$ associated naturally with (2.1); it is defined by

$$W_k(z) = \begin{bmatrix} \alpha_k + \bar{\alpha}_k z & -1 \\ z & 0 \end{bmatrix}, \tag{2.12}$$

for $k = 0, 1, \cdots, n-1$. It is easily verified that the first-kind and second-kind versions of (2.1) yield the matrix factorization

$$\begin{bmatrix} p_n(z)/p_0 & -p_{n-1}(z)/p_0 \\ zq_n(z)/q_1(z) & -zq_{n-1}(z)/q_1(z) \end{bmatrix} = W_0(z)W_1(z)\cdots W_{n-1}(z). \tag{2.13}$$

There exists a "dual identity," which is the reciprocal version of (2.13); it reads

$$\begin{bmatrix} p_n(z)/p_0 & -q_n(z)/q_1(z) \\ zp_{n-1}(z)/p_0 & -zq_{n-1}(z)/q_1(z) \end{bmatrix} = W_{n-1}(z)\cdots W_1(z)W_0(z). \tag{2.14}$$

In view of (2.10) and (2.12), the desired relations (2.11) follow immediately from a comparison between (2.13) and (2.14).

The remainder of this section is mainly concerned with certain "Christoffel–Darboux-type formulas" that relate the symmetric polynomials $p_k(z)$. From the circle parameter ζ_0 and the recurrence coefficients $\alpha_0, \alpha_1, \cdots, \alpha_{n-1}$ in (2.1), construct the $n \times n$ bidiagonal matrix

$$A = \zeta_0^{1/2} \begin{bmatrix} \bar{\alpha}_0 & & & \\ -1 & \bar{\alpha}_1 & & \\ & \ddots & \ddots & \\ & & -1 & \bar{\alpha}_{n-1} \end{bmatrix}. \tag{2.15}$$

The positivity property $\lambda_k > 0$ of the Jacobi parameters means that A is *strictly accretive*, in the sense that its real part (or Hermitian part), i.e.,

$$\operatorname{Re} A = \tfrac{1}{2}(A + A^*), \tag{2.16}$$

is positive definite. Let us now prove this elementary but important result by exhibiting an explicit factorization of (2.16) that will be useful in the sequel. (An alternative proof can be found at the end of § 3 in [9].)

Set the diagonal matrix $\Delta = \operatorname{diag}(\zeta_0^{k/2} : 0 \leqq k \leqq n-1)$. By (2.15), we have the relation $2 \operatorname{Re} A = \Delta R \bar{\Delta}$, where R is the real symmetric tridiagonal matrix that has the numbers $2 \operatorname{Re}(\zeta_0^{-1/2}\alpha_k)$ on the diagonal, for $k = 0, \cdots, n-1$, and the number -1 above and below the diagonal. Define the $n \times n$ matrices

$$L = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}, \qquad G = \begin{bmatrix} -1 & \lambda_1^{-1} & & \\ & -1 & \ddots & \\ & & \ddots & \lambda_{n-1}^{-1} \\ & & & -1 \end{bmatrix}. \tag{2.17}$$

The set of formulas (2.4) can be written as the matrix identity $R = G^T L G$. Hence we obtain the *Cholesky factorization of the tridiagonal matrix* Re A in the explicit form

$$\text{Re } A = M^* M \quad \text{with } M = \frac{1}{\sqrt{2}} L^{1/2} G \bar{\Delta}. \tag{2.18}$$

Next, we derive an interesting identity that can be viewed as an analogue of the celebrated Christoffel–Darboux formula of orthogonal polynomial theory [25]. Set the n-vector polynomial

$$\mathbf{p}(z) = [p_0(z), p_1(z), \cdots, p_{n-1}(z)]^T. \tag{2.19}$$

Its reciprocal $\hat{\mathbf{p}}(z) = z^{n-1}\mathbf{p}^*(1/\bar{z})$ is $\hat{\mathbf{p}}(z) = [z^{n-1}p_0(z), \cdots, p_{n-1}(z)]$, by the symmetry property of $p_k(z)$. The three-term recurrence (2.1) amounts to the matrix identity

$$(\zeta_0 A^* + zA)\mathbf{p}(z) = \zeta_0^{1/2} p_n(z)\mathbf{e}_n, \tag{2.20}$$

with $\mathbf{e}_n = [0, \cdots, 0, 1]^T$. Replacing z by ζ in (2.20) and taking reciprocals, we deduce an alternative version, that is,

$$\hat{\mathbf{p}}(\zeta)(\zeta_0 A^* + \zeta A) = \zeta_0^{1/2} p_n(\zeta)\mathbf{e}_n^T. \tag{2.21}$$

Premultiplying (2.20) by $\hat{\mathbf{p}}(\zeta)$, postmultiplying (2.21) by $\mathbf{p}(z)$, and subtracting the results, we arrive at the following conclusion.

PROPOSITION 1. *The symmetric polynomials $p_k(z)$ generated by the three-term recurrence* (2.1) *satisfy the Christoffel–Darboux-type formula*

$$(\zeta - z)\hat{\mathbf{p}}(\zeta) A \mathbf{p}(z) = \zeta_0^{1/2}[p_n(\zeta)p_{n-1}(z) - p_{n-1}(\zeta)p_n(z)]. \tag{2.22}$$

It differs markedly from the analogous result of orthogonal polynomial theory by the presence of the *accretive bidiagonal matrix* A (instead of a positive diagonal matrix). Combining (2.22) with its reciprocal version, we obtain a "symmetrized formula," involving the *positive-definite tridiagonal matrix* Re A, that is,

$$(\zeta - z)\hat{\mathbf{p}}(\zeta)\,\text{Re}\, A\,\mathbf{p}(z) = \tfrac{1}{2}\zeta_0^{-1/2}[(\zeta_0 - z)p_n(\zeta)p_{n-1}(z) - (\zeta_0 - \zeta)p_{n-1}(\zeta)p_n(z)]. \tag{2.23}$$

The *confluent version* of (2.23), obtained by letting z tend to ζ, is of special interest in the theory; it reads as follows:

$$\hat{\mathbf{p}}(\zeta)\,\text{Re}\, A\,\mathbf{p}(\zeta) = \tfrac{1}{2}\zeta_0^{-1/2}\{p_{n-1}(\zeta)p_n(\zeta) + (\zeta_0 - z)[p_{n-1}(\zeta)p_n'(\zeta) - p_n(\zeta)p_{n-1}'(\zeta)]\}. \tag{2.24}$$

Furthermore, it is possible to derive a "Green-type formula" that links the first-kind and second-kind families. Let us briefly introduce this question, without going into all the details. Set the n-vector polynomial

$$\mathbf{q}(z) = [q_0(z), q_1(z), \cdots, q_{n-1}(z)]^T. \tag{2.25}$$

Here, we have $\hat{\mathbf{q}}(z) = -[z^{n-1}q_0(z), \cdots, q_{n-1}(z)]$, by antisymmetry. Using the recurrence relation (2.1) for second-kind polynomials, and recalling $q_{-1}(z) = -z^{-1}q_1(z)$, we obtain the matrix identity

$$(\zeta_0 A^* + zA)\mathbf{q}(z) = \zeta_0^{1/2}[q_n(z)\mathbf{e}_n - q_1(z)\mathbf{e}_1], \tag{2.26}$$

where $\mathbf{e}_1 = [1, 0, \cdots, 0]^T$. Then, manipulations similar to those above produce the *bidiagonal Green-type formula*

$$(\zeta - z)\hat{\mathbf{q}}(\zeta) A \mathbf{p}(z) = 2c_0(\zeta_0 - \zeta)\zeta^{n-1} + \zeta_0^{1/2}[q_{n-1}(\zeta)p_n(z) - q_n(\zeta)p_{n-1}(z)]. \tag{2.27}$$

When $\zeta = z$, this reduces to (2.9) where k is replaced by $n - 1$. We shall give neither the symmetrized (tridiagonal) version, nor the confluent versions of the Green-type formula (2.27).

Finally, let us mention that (2.20) yields an interesting expression for the first-kind (symmetric) polynomials, namely,

$$p_n(z) = p_0 \zeta_0^{-n/2} \det(\zeta_0 A^* + zA). \tag{2.28}$$

In view of (2.18), this agrees with (2.5). There exists an analogous expression for the second-kind polynomials, based on (2.26); it will not be given here.

3. Lossless rational functions. Since the Toeplitz matrix C_n is nonnegative-definite and has nullity one, there exists a *unique positive measure* $d\mu$, defined on the interval $[0, 2\pi)$, that admits the entries of C_n as its first $2n + 1$ *trigonometric moments*, in the sense that it satisfies

$$c_k = \int_0^{2\pi} e^{-ik\theta}\, d\mu(\theta), \tag{3.1}$$

for $-n \leqq k \leqq n$; the measure $d\mu$ is discrete and possesses exactly n *mass points* $\theta_1, \theta_2, \cdots, \theta_n$ in the interval $[0, 2\pi)$. This is a classical result [19].

Let $a_n(z)$ denote the comonic polynomial that admits the sequence of reflection coefficients $\rho_1, \cdots, \rho_n$; by definition, it is produced by the *Szegö–Levinson recurrence relation* $a_k(z) = a_{k-1}(z) + \rho_k z \hat{a}_{k-1}(z)$, with $a_0(z) = 1$. Equivalently, $a_n(z)$ is the unique comonic polynomial of degree n whose coefficient vector $\mathbf{a}_n$ satisfies the system of homogeneous linear equations $C_n \mathbf{a}_n = \mathbf{0}$ (see details in [9]). Polynomial $a_k(z)$ will be referred to as the *first-kind predictor* relative to the Toeplitz matrix C_k. Consider the factorization

$$a_n(z) = (1 - \zeta_1^{-1} z)(1 - \zeta_2^{-1} z) \cdots (1 - \zeta_n^{-1} z). \tag{3.2}$$

It is well known that the zeros $\zeta_1, \cdots, \zeta_n$ of the predictor $a_n(z)$ are simple and belong to the unit circle. In fact, they are given by

$$\zeta_j = e^{i\theta_j} \quad \text{for } j = 1, \cdots, n, \tag{3.3}$$

where $\theta_1, \cdots, \theta_n$ are the mass points of the measure $d\mu$. Let h_j denote the *weight* of $d\mu$ at point θ_j. Then (3.1) can be written in the explicit form

$$c_k = \sum_{j=1}^{n} h_j \zeta_j^{-k}. \tag{3.4}$$

Relation (3.4) provides the *Carathéodory representation* of the entries of the Toeplitz matrix C_n (see Grenander and Szegö [19]).

Next, let us examine the question of the weights in some detail. Consider the *Gaussian quadrature formula on the unit circle* relative to $d\mu$, that is,

$$\int_0^{2\pi} v(e^{i\theta})\, d\mu(\theta) = \sum_{j=1}^{n} h_j v(\zeta_j), \tag{3.5}$$

where $v(z)$ is any Laurent polynomial in the variable z; this is a direct consequence of the Carathéodory representation (3.4). The *inner product* $\langle x(z), y(z) \rangle$ of two Laurent polynomials with respect to the measure $d\mu$ is defined as the left-hand side of (3.5) with $v(z) = \bar{x}(1/\bar{z}) y(z)$. Thus we have the identity

$$\langle x(z), y(z) \rangle = \sum_{j=1}^{n} h_j \bar{x}(\zeta_j) y(\zeta_j). \tag{3.6}$$

It is well known that the reciprocals $\hat{a}_k(z)$ of the predictor polynomials $a_k(z)$ constitute the family of *monic polynomials orthogonal on the unit circle* with respect to $d\mu$. Indeed, they satisfy the Szegö orthogonality relations

$$\langle \hat{a}_k(z), \hat{a}_l(z) \rangle = \sigma_k \delta_{k,l}, \tag{3.7}$$

for $0 \leqq k, l \leqq n$. Recall the properties $\sigma_k > 0$ for $k = 0, \cdots, n-1$ and $\sigma_n = 0$. In view of (3.6), a standard manipulation of (3.7) yields the classical expression

$$h_j = \left[\sum_{k=0}^{n-1} \sigma_k^{-1} |a_k(\zeta_j)|^2 \right]^{-1}. \tag{3.8}$$

Alternatively, the weight h_j can be interpreted as explained below. Define $r_k(z)$ as the *second-kind predictor polynomial* relative to C_k, for $k = 0, \cdots, n$; it is obtained by means of the Szegö–Levinson recurrence $r_k(z) = r_{k-1}(z) - \rho_k z \hat{r}_{k-1}(z)$, with the initial condition $r_0(z) = c_0$. As shown in [9], we have the *embedding properties*

$$p_n(z) = p_n(0) a_n(z), \qquad q_n(z) = p_n(0)[r_n(z) + i\gamma a_n(z)], \tag{3.9}$$

with the real number $\gamma = (\alpha_0 - \bar{\alpha}_0 \zeta_0)/ic_0(\alpha_0 + \bar{\alpha}_0 \zeta_0)$. It turns out that the weights h_j can be determined from the polar expansion

$$\frac{q_n(z)}{p_n(z)} = i\gamma + \sum_{j=1}^{n} h_j \frac{\zeta_j + z}{\zeta_j - z} \tag{3.10}$$

of the *lossless function* $h(z) = q_n(z)/p_n(z)$. Indeed, in view of (3.1), (3.4), and (3.5), this expresses the fact that $d\mu$ is the positive measure that occurs in the *Riesz–Herglotz representation* [1] of the lossless function $f(z) = h(z) - i\gamma = r_n(z)/a_n(z)$. Note that $q_k(z)/p_k(z)$ is a lossless function of degree k for each k (with $0 \leqq k \leqq n$). Therefore, *$p_k(z)$ and $q_k(z)$ have simple zeros, located on the unit circle, and the zeros of $p_k(z)$ separate those of $q_k(z)$.*

Let us now derive a useful *Christoffel-type formula* for the weights h_j. In view of (3.10) we have the expression

$$h_j = -q_n(\zeta_j)/2\zeta_j p_n'(\zeta_j). \tag{3.11}$$

Applying the confluent Christoffel–Darboux-type relation (2.24) at a zero ζ_j of $p_n(z)$, we can write the identity

$$\mathbf{p}^*(\zeta_j) \operatorname{Re} A \mathbf{p}(\zeta_j) = \tfrac{1}{2} \zeta_0^{-1/2} (\zeta_0 - \zeta_j) \bar{p}_{n-1}(\zeta_j) p_n'(\zeta_j). \tag{3.12}$$

Besides, we have $2c_0(\zeta_0 - \zeta_j) = \zeta_0^{1/2} \bar{p}_{n-1}(\zeta_j) q_n(\zeta_j)$, by use of (2.9). Combining this with (3.12), we obtain the desired formula; the result can be stated as follows.

PROPOSITION 2. *The weights h_j associated with the mass points θ_j can be expressed in terms of the values $p_k(\zeta_j)$ by means of the Christoffel-type formula*

$$h_j = \frac{c_0 |\zeta_0 - \zeta_j|^2}{2\mathbf{p}^*(\zeta_j) \operatorname{Re} A \mathbf{p}(\zeta_j)}. \tag{3.13}$$

This formula can be used as a substitute for the classical formula (3.8). The positivity property $h_j > 0$ appears here as a consequence of the positive definiteness of the tridiagonal matrix $\operatorname{Re} A$.

Besides the function $h(z)$ in (3.10), there is a second rational lossless function that occurs naturally in the theory (see [6], [10]). To emphasize the role of the duality mentioned in § 2, we denote this function by $\check{h}(z)$; the definition is

$$\check{h}(z) = \frac{2c_0(\zeta_0 - z) p_{n-1}(z)}{p_0^2 \zeta_0^{1/2} p_n(z)}. \tag{3.14}$$

In view of (2.11) and (2.8) we have $\check{h}(z) = \check{q}_n(z)/\check{p}_n(z)$. By duality, this implies that $\check{h}(z)$ is a lossless function of degree n. As a consequence, since a similar result holds when n is replaced by k (with $1 \leqq k \leqq n$), we conclude that *the zeros of* $p_k(z)$ *separate those of* $(\zeta_0 - z)p_{k-1}(z)$ (on the unit circle). Consider the polar expansion

$$\check{h}(z) = i\check{\gamma} + \sum_{j=1}^{n} \check{h}_j \frac{\zeta_j + z}{\zeta_j - z}, \tag{3.15}$$

with $\check{h}_j = c_0(\zeta_j - \zeta_0)p_{n-1}(\zeta_j)/p_0^2 \zeta_0^{1/2} \zeta_j p_n'(\zeta_j)$, in view of (3.14). Making use of (3.11) and (2.9), we obtain a remarkable relation between the weights h_j and $\check{h}_j$, that is,

$$\check{h}_j = h_j |\, p_0^{-1} p_{n-1}(\zeta_j)|^2. \tag{3.16}$$

When $x(z)$ and $y(z)$ are polynomials of degree less than or equal to n in z, the inner product (3.6) can be written in the form $\mathbf{x}^* C_n \mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ denote the coefficient vectors of $x(z)$ and $y(z)$. As an application, let us compute the squared norm $\| p_k(z) \|^2 = \mathbf{p}_k^* C_k \mathbf{p}_k$ of the symmetric polynomial $p_k(z)$, for $k \leqq n-1$. Using the results of [9], we obtain $\| p_k(z) \|^2 = 2c_0 \operatorname{Re}(\zeta_0^{1/2} \alpha_k^{-1})$, so that (3.6) yields the identity

$$\sum_{j=1}^{n} h_j |\, p_k(\zeta_j)|^2 = 2c_0 \operatorname{Re}(\zeta_0^{1/2} \alpha_k^{-1}). \tag{3.17}$$

More generally, we can compute inner products of the form $\langle p_k(z), z^l p_l(z) \rangle$ and thus obtain an explicit congruence relation between the tridiagonal matrix $\operatorname{Re} A$ and the Toeplitz matrix C_{n-1}. Details on this subject can be found in [9]. Note that (3.17) provides an interesting check about (3.16). Indeed, in view of (3.15), the sum of the weights $\check{h}_j$ is the real part of $\check{h}(0)$, which equals $2c_0 p_0^{-2} \operatorname{Re}(\zeta_0^{1/2} \alpha_{n-1}^{-1})$ since $p_n(0) = \alpha_{n-1} p_{n-1}(0)$ by (2.1); according to (3.16), this is nothing but (3.17) with $k = n-1$.

Remark. The *singular case* where the circle parameter ζ_0 coincides with one of the zeros of $a_n(z)$, i.e., the case where (2.3) yields $\alpha_0 = \infty$, can be "regularized" as follows [9]. It suffices to replace the generating sequence $(\alpha_0, \alpha_1, \cdots, \alpha_{n-1})$ of the symmetric polynomials $p_k(z)$ by the *truncated sequence* $(\alpha_1, \alpha_2, \cdots, \alpha_{n-1})$. This sequence generates a family of reduced symmetric polynomials $\tilde{p}_k(z)$, for $k = 0, 1, \cdots, n-1$, with the property that $\tilde{p}_{n-1}(z)$ is proportional to the polynomial $\tilde{a}_{n-1}(z) = a_n(z)/(1 - \zeta_0^{-1} z)$. If the zeros of $\tilde{a}_{n-1}(z)$ are $\zeta_1, \cdots, \zeta_{n-1}$ (i.e., if $\zeta_0 = \zeta_n$), then the weight $\tilde{h}_j$ of the "reduced measure" $d\tilde{\mu}$ that corresponds to the mass point θ_j is given by $\tilde{h}_j = h_j |\, \zeta_0 - \zeta_j|^2$, for $j = 1, \cdots, n-1$. This result is proved implicitly in § 4 of [9].

4. The zeros of $a_n(z)$ as generalized eigenvalues. Let us now examine, in a detailed manner, the question of computing the zeros $\zeta_1, \cdots, \zeta_n$ of the predictor polynomial $a_n(z)$ defined from the singular Toeplitz system $C_n \mathbf{a}_n = \mathbf{0}$. In the sequel it is assumed that *the actual data are the reflection coefficients* $\rho_1, \cdots, \rho_n$ relative to C_n rather than the coefficients $a_{n,i}$ themselves. Note that this information is highly redundant, since $a_n(z)$ is ρ_n-symmetric.

Our approach makes use of the family of symmetric polynomials $p_k(z)$ in which $a_n(z)$ is embedded, in the sense that $a_n(z)$ is proportional to $p_n(z)$. It is explained in § 2 how such a family can be constructed uniquely from the reflection coefficients ρ_k and an "arbitrary" value of the circle parameter ζ_0 (with $|\zeta_0| = 1$). The only constraint we impose is that $a_n(\zeta_0) \neq 0$, i.e., $\zeta_0 \neq \zeta_j$ for $1 \leqq j \leqq n$. (In terms of the data, this amounts to $\omega_1 \zeta_0 \neq -1$.) As explained at the end of § 3, such a restriction implies no essential loss of generality. Indeed, if the singular case $\omega_1 \zeta_0 = -1$ occurs, then a slight modification of the algorithm devised for the regular case produces the zeros $\zeta_1, \cdots, \zeta_{n-1}$ of the reduced

polynomial $\tilde{a}_{n-1}(z) = a_n(z)/(1 - \zeta_0^{-1} z)$ from the given reflection coefficients ρ_k relative to $a_n(z)$.

Remark. In certain applications, such as the Pisarenko modelling problem [22], it is probably more appropriate from an algorithmic viewpoint to consider that the actual data are the entries of the nonnegative-definite Toeplitz matrix C_n. In such a situation, rather than computing the recurrence coefficients α_k (from the reflection coefficients, as explained above), it may be preferable to compute the "distorted recurrence coefficients" α_k^+ from the entries of C_n, by means of the extended split Levinson algorithm, as explained in § 5 of [9]. This produces the desired result since the reduced polynomial $\tilde{p}_n^+(z)$ is proportional to $a_n(z)$. We shall not go into further detail about such a dual computation scheme.

Let $\mathbf{p}(z)$ be the n-vector (2.19), whose entries are the symmetric polynomials $p_k(z)$ with $0 \leqq k \leqq n - 1$. From (2.20) we deduce the identity

$$(\zeta_0 A^* + \zeta_j A)\mathbf{p}(\zeta_j) = \mathbf{0}, \tag{4.1}$$

for $j = 1, \cdots, n$. Thus, we can state the following result.

PROPOSITION 3. *The zeros $\zeta_1, \cdots, \zeta_n$ of the ρ_n-symmetric predictor $a_n(z)$ are the generalized eigenvalues of the pencil $(\zeta_0 A^*, -A)$.*

Next, we examine a simple transformation of the problem (4.1). Consider the change of variable $z = e^{i\theta} \to t = \cot(\theta - \theta_0)/2$, which maps the unit circle to the extended real line. In other terms, we have

$$z = -\zeta_0 \frac{1 - it}{1 + it}, \qquad t = -i\frac{\zeta_0 + \zeta}{\zeta_0 - \zeta}. \tag{4.2}$$

Accordingly, with the zeros $\zeta_j = \exp(i\theta_j)$ of $a_n(z)$ we associate the real numbers

$$t_j = \cot\frac{\theta_j - \theta_0}{2} \quad \text{for } j = 1, \cdots, n. \tag{4.3}$$

Translating (4.1) by means of (4.2), we obtain the linear system

$$(\operatorname{Im} A - t_j \operatorname{Re} A)\mathbf{p}(\zeta_j) = \mathbf{0}, \tag{4.4}$$

where $\operatorname{Im} A = (A - A^*)/2i$ denotes the imaginary part of A. The main conclusion is the following version of the preceding result.

PROPOSITION 4. *The numbers $t_1, \cdots, t_n$ associated with the zeros $\zeta_1, \cdots, \zeta_n$ of $a_n(z)$ are the generalized eigenvalues of the tridiagonal Hermitian-definite pencil* (Im A, Re A).

This result leads to an interesting numerical method for computing the zeros ζ_j of $a_n(z)$ since there exist efficient algorithms to solve that type of generalized eigenvalue problems [15].

Let us make an additional comment concerning the relation between Propositions 3 and 4. In view of (4.1), the numbers ζ_j are the eigenvalues of the matrix $-\zeta_0 A^{-1} A^*$, which is similar to a unitary matrix due to the fact that A is strictly accretive [11]. In the present context, this similarity property can be explained as follows. Define the Hermitian matrix

$$K = (M^{-1})^*(\operatorname{Im} A)M^{-1}, \tag{4.5}$$

where M is the Cholesky factor of Re A given in (2.18). In view of (4.4), the eigenvalues of K are the real numbers t_j. Consider the *Cayley transform* $U = (I + iK)^{-1}(I - iK)$ of (4.5). Since K is Hermitian, U is unitary. Using (4.5), we obtain the factorization

$$U = M(A^{-1}A^*)M^{-1}, \tag{4.6}$$

which shows that $A^{-1}A^*$ is similar to U (as alluded to above). In view of (4.1), the eigenvalues of U are the unit modulus numbers $-\zeta_0^{-1}\zeta_j$ for $j = 1, \cdots, n$.

The weights h_j associated with the zeros ζ_j of $a_n(z)$ can be computed either by use of (3.8) or by use of (3.13). The latter method involves the generalized eigenvectors $\mathbf{p}(\zeta_j)$ of the problems (4.1) and (4.4). Let us now examine the properties of these vectors. Define the vector polynomial

$$\mathbf{g}(z) = M\mathbf{p}(z). \tag{4.7}$$

It is seen from (4.4) that $\mathbf{g}(\zeta_j)$ is the eigenvector of the Hermitian matrix K corresponding to the eigenvalue t_j. Set the squared norm

$$w_j = \|\mathbf{g}(\zeta_j)\|^2 = \mathbf{p}^*(\zeta_j)\,\mathrm{Re}\,A\,\mathbf{p}(\zeta_j). \tag{4.8}$$

The transpose version of the orthogonality relations satisfied by the eigenvectors $\mathbf{g}(\zeta_j)$ can be written in the form

$$\sum_{j=1}^{n} w_j^{-1}\bar{g}_k(\zeta_j)g_l(\zeta_j) = \delta_{k,l}, \tag{4.9}$$

for $0 \leqq k, l \leqq n-1$, with $[g_0(z), \cdots, g_{n-1}(z)]^T = \mathbf{g}(z)$. Now, using (2.17), (2.18), and (4.7), together with the expression of $\hat{a}_k(z)$ in terms of $p_k(z)$ and $p_{k+1}(z)$ given in [9], we obtain

$$g_k(z) = (2\lambda_{k+1}\zeta_0^{k+1})^{-1/2}[\bar{p}_{k+1}(0)(\zeta_0 - z)\hat{a}_k(z) + \delta_{k,n-1}p_n(z)]. \tag{4.10}$$

Since $w_j^{-1} = 2c_0^{-1}|\zeta_0 - \zeta_j|^{-2}h_j$, by (3.13) and (4.8), it can be verified, with the help of (4.10), that (4.9) yields exactly the Szegö orthogonality relations (3.7). Details are omitted.

The methods described above can be applied to several interesting problems where we must compute the zeros of an $a_n(z)$-type polynomial (see [7], [8]). In the remaining part of this section we explain one of the most natural applications, dealing with the spectral decomposition of unitary Hessenberg matrices.

Let $(v_0(z), v_1(z), \cdots, v_n(z))$ be a *graded basis* of the vector space of complex polynomials of degree less than or equal to n in the variable z. Here, "graded" means that $v_k(z)$ has degree k, for $k = 0, 1, \cdots, n$. Consider the expansion of the shifted polynomial $zv_k(z)$ in that basis, for $0 \leqq k \leqq n-1$; it has the form

$$zv_k(z) = \sum_{l=0}^{k+1} d_{k,l}v_l(z), \tag{4.11}$$

for some complex numbers $d_{k,l}$, with $d_{k,k+1} \neq 0$. Let us denote by D the $n \times n$ matrix that represents the shift operator in the given basis. In view of (4.11), it reads

$$D = \begin{bmatrix} d_{0,0} & d_{0,1} & & \\ d_{1,0} & d_{1,1} & d_{1,2} & \\ \vdots & \vdots & \vdots & \ddots \\ d_{n-2,0} & d_{n-2,1} & d_{n-2,2} & \cdots\ d_{n-2,n-1} \\ d_{n-1,0} & d_{n-1,1} & d_{n-1,2} & \cdots\ d_{n-1,n-1} \end{bmatrix}. \tag{4.12}$$

Thus, D is an *irreducible lower Hessenberg matrix*. (Here, "irreducible" means that the entries just above the diagonal are nonzero.) The matrix D is called the *confederate matrix* of $v_n(z)$ with respect to the given graded basis. (See Barnett [3] and the references therein. Note that confederate matrices have been known for a long time; see, for example, Householder [20, pp. 25–26].) The set of relations (4.11) can be written in matrix form as follows:

$$(zI - D)\mathbf{v}(z) = [0, \cdots, 0, d_{n-1,n}v_n(z)]^T, \tag{4.13}$$

with $\mathbf{v}(z) = [v_0(z), \cdots, v_{n-1}(z)]^T$. This shows that the zeros ζ_j of $v_n(z)$ are the eigenvalues of D and that the vectors $\mathbf{v}(\zeta_j)$ are the corresponding eigenvectors (in the case of simple eigenvalues).

Now consider a nonnegative-definite Toeplitz matrix C_n of order $n+1$ and rank n, with the normalization $c_0 = 1$. From the Szegö polynomials $\hat{a}_k(z)$ relative to C_n, define the graded basis $(v_0(z), \cdots, v_n(z))$, called *Szegö basis*, by setting

$$v_n(z) = a_n(z), \qquad v_k(z) = \sigma_k^{-1/2}\hat{a}_k(z) \quad \text{for } 0 \leqq k \leqq n-1, \tag{4.14}$$

with $\sigma_0 = 1$ and $\sigma_k = (1 - |\rho_k|^2)\sigma_{k-1}$. It is easily verified that *the corresponding confederate matrix D is unitary*. Indeed, the orthogonality relations (3.7) yield the identity $VHV^* = I$, with $H = \operatorname{diag}(h_1, \cdots, h_n)$ and

$$V = [\mathbf{v}(\zeta_1), \mathbf{v}(\zeta_2), \cdots, \mathbf{v}(\zeta_n)], \tag{4.15}$$

where $\mathbf{v}(z) = [v_0(z), \cdots, v_{n-1}(z)]^T$ as above. From (4.13) we deduce

$$DV = VZ \quad \text{with } Z = \operatorname{diag}(\zeta_1, \cdots, \zeta_n). \tag{4.16}$$

The identity $D = (VH^{1/2})Z(VH^{1/2})^{-1}$ exhibits D as a product of three unitary matrices, which proves the claim. Note that (4.16) provides the *spectral decomposition of the unitary Hessenberg matrix D*. Indeed, the numbers ζ_j are the eigenvalues of D and the vectors $h_j^{1/2}\mathbf{v}(\zeta_j)$ are the corresponding normalized (pairwise orthogonal) eigenvectors.

For future use, let us give the expression of the entries $d_{k,l}$ in terms of the reflection coefficients. From the Szegö–Levinson formula, used in an inductive manner, we can deduce the expansion of $\hat{v}_k(z)$ in the $v_l(z)$ basis, whence the expansion of $zv_k(z)$ in that basis. Thus, by straightforward computation, we obtain

$$d_{k,l} = \begin{cases} -\rho_l\mu_{l+1}\mu_{l+2}\cdots\mu_k\bar{\rho}_{k+1} & \text{if } l \leqq k, \\ \mu_l & \text{if } l = k+1, \end{cases} \tag{4.17}$$

and $d_{n-1,n} = \bar{\rho}_n\sigma_{n-1}^{-1/2}$, where $\rho_0 = 1; \rho_1, \cdots, \rho_n$ are the reflection coefficients; and $\mu_1, \cdots, \mu_{n-1}$ are positive numbers given by

$$\mu_k = (1 - |\rho_k|^2)^{1/2}. \tag{4.18}$$

We are now in a position to state the following remarkable result [17], [21], which is an exact converse of the results above.

PROPOSITION 5. *Let D be an irreducible lower Hessenberg unitary matrix of order n. Assume that D is normalized in such a way that its $d_{k,k+1}$ entries are positive. Then D is the confederate matrix of a ρ_n-symmetric predictor polynomial $a_n(z)$ with respect to a well-defined Szegö basis. In other words, the entries $d_{k,l}$ of D can be parameterized in the form* (4.17), (4.18), *where $\rho_0, \rho_1, \cdots, \rho_n$ are well-defined complex numbers satisfying $\rho_0 = 1$, $|\rho_k| < 1$ for $k = 1, \cdots, n-1$, and $|\rho_n| = 1$.*

Let us briefly indicate the argument [17], [21]. Consider the well-known *Jacobi–Givens factorization*

$$D = D_n D_{n-1}\cdots D_1, \tag{4.19}$$

with $D_n = I_{n-1} \oplus Q_n$ and $D_k = I_{k-1} \oplus Q_k \oplus I_{n-k-1}$ for $k = 1, \cdots, n-1$; here, the Q's are unitary matrices of the form

$$Q_n = -\bar{\rho}_n, \qquad Q_k = \begin{bmatrix} -\bar{\rho}_k & \mu_k \\ \mu_k & \rho_k \end{bmatrix}, \tag{4.20}$$

for $1 \leqq k \leqq n-1$, with $\mu_k = (1 - |\rho_k|^2)^{1/2}$. This defines the parameters $\rho_1, \cdots, \rho_n$; they enjoy the properties $|\rho_1| < 1, \cdots, |\rho_{n-1}| < 1$, and $|\rho_n| = 1$. It is easily checked,

by use of (4.20), that the factorized form (4.19) produces exactly the desired expressions (4.12) and (4.17).

As a consequence, we can use the methods summarized in Propositions 3 and 4 to *compute the eigenvalues (and the eigenvectors) of a normalized unitary Hessenberg matrix D given in terms of its parameters* $\rho_1, \rho_2, \cdots, \rho_n$. Indeed, we have seen in (4.16) that the eigenvalues of D are the zeros of the predictor $a_n(z)$ generated by the reflection coefficients ρ_k, and the entries of the corresponding eigenvectors are the values $v_k(\zeta_j)$ of the orthonormal Szegö polynomials $v_k(z) = \sigma_k^{-1/2}\hat{a}_k(z)$, for $k = 0, \cdots, n-1$. (Note that these values could be determined from the values $p_k(\zeta_j)$; see [9].) The technique just mentioned is quite different from the QR algorithm [17] and from the divide-and-conquer algorithm [18] which were recently proposed to solve the unitary Hessenberg eigenvalue problem.

The context of unitary Hessenberg matrices provides a transparent interpretation of the *duality relation* $\rho_k \to \rho_k^{\#} = \rho_n\bar{\rho}_{n-k}$ (for $0 \leqq k \leqq n$) between sequences of reflection coefficients, investigated in [9]. Indeed, it follows immediately from (4.17) that the unitary Hessenberg matrix $D^{\#}$ parameterized by the dual sequence $(\rho_k^{\#})_{k=1}^n$ can be written in the form

$$D^{\#} = JD^TJ, \tag{4.21}$$

where J denotes the mirror permutation matrix, i.e., the secondary diagonal matrix. Therefore, D and $D^{\#}$ have the same spectrum. Thus, the eigenvalues of D can be computed, by means of the method suggested above, either from the data sequence $(\rho_k)_{k=1}^n$ or from the dual sequence $(\rho_k^{\#})_{k=1}^n$. Of course, the same conclusion holds for the "general problem" of computing the zeros of the predictor $a_n(z)$, since we have $a_n^{\#}(z) = a_n(z)$.

Not surprisingly, there is a close connection between the unitary matrix U defined in (4.6) from $\alpha_0, \cdots, \alpha_{n-1}$ and the unitary Hessenberg matrix D given by (4.17) in terms of $\rho_1, \cdots, \rho_n$. It is clear that U is lower Hessenberg, since, by use of (2.18), we obtain the expression $U = 2MA^{-1}M^* - I$ where A^{-1} is lower triangular and M is upper bidiagonal. More precisely, it can be shown that the unitary lower Hessenberg matrices D and U are related by the simple identity

$$D = -\zeta_0\Gamma^{-1}U\Gamma, \tag{4.22}$$

where Γ is a well-defined diagonal matrix, the diagonal entries of which have squared modulus $c_0/2$. This produces an explicit interpretation of unitary Hessenberg matrices in the framework of the tridiagonal theory. The proof of (4.22) results from a comparison of (4.1) and (4.13). Indeed, using (4.10) and (4.14) we can write $\mathbf{g}(\zeta_j) = (\zeta_0 - \zeta_j)\Gamma\mathbf{v}(\zeta_j)$, for $j = 1, \cdots, n$, with a suitable diagonal matrix Γ. Hence (4.7) yields

$$\mathbf{p}(\zeta_j) = (\zeta_0 - \zeta_j)M^{-1}\Gamma\mathbf{v}(\zeta_j), \tag{4.23}$$

which implies that (4.1) can be written in the form

$$(\zeta_j I + \zeta_0\Gamma^{-1}U\Gamma)\mathbf{v}(\zeta_j) = \mathbf{0}. \tag{4.24}$$

In view of (4.13), this proves the desired relation (4.22). The successive diagonal entries $\gamma_1, \cdots, \gamma_n$ of Γ are found to be given by

$$\gamma_k = \bar{p}_k(0)[\sigma_{k-1}/2\lambda_k\zeta_0^k]^{1/2}. \tag{4.25}$$

Using formula $\lambda_k = c_0^{-1}\sigma_{k-1}|p_k(0)|^2$ in [9], we deduce the identity $|\gamma_k|^2 = c_0/2$, which completes the proof of the statement.

5. Methods for real data. In this section we consider certain special features exhibited by the theory in the case where the data ρ_k are real. Thus we have $-1 < \rho_k < 1$ for $k = 1, \cdots, n-1$, and $\rho_n = \pm 1$. In this framework it is natural to restrict the circle parameter

ζ_0 to real values, i.e., $\zeta_0 = \pm 1$. More generally (see however the remark at the end of § 2 of [9]), we assume that the reflection coefficients ρ_k satisfy the *reality condition*

$$\text{Im}\,(\rho_k \zeta_0^k) = 0 \quad \text{for } k = 1, \cdots, n, \tag{5.1}$$

with respect to a given complex number ζ_0 of unit modulus. Note that the transformation $\zeta_0 \to -\zeta_0$ preserves this property; therefore, the admissible values of the circle parameter ζ_0 occur in *symmetric pairs* $\{\zeta_0, -\zeta_0\}$.

Equivalently, the assumption (5.1) amounts to the fact that the polynomials $a_k(\zeta_0 z)$ have real coefficients (in the variable z). By induction, it is easily seen that the pseudoreflection coefficients ω_k in (2.6) have the simple form

$$\omega_k = \varepsilon \zeta_0^{-k} \quad \text{with } \varepsilon = \rho_n \zeta_0^n = \pm 1. \tag{5.2}$$

Thus, in contrast with the general complex case, they are independent of the numerical values of $\rho_1, \cdots, \rho_{n-1}$. Besides, our reality condition implies that all coefficients α_k in the recurrence (2.1) have the same phase. More precisely, by use of (5.1) and (5.2), we deduce the property

$$\alpha_k = \nu_k \zeta_0^{1/2} \quad \text{with } \nu_k > 0. \tag{5.3}$$

Hence, all polynomials $\zeta_0^{-k/2} p_k(\zeta_0 z)$ have real coefficients with respect to z.

The cases $\varepsilon = 1$ and $\varepsilon = -1$ in (5.2) correspond to $a_n(\zeta_0) \neq 0$ and to $a_n(\zeta_0) = 0$, respectively. Without loss of generality, we restrict our attention to the first case, that is, $\varepsilon = 1$. The required information concerning the treatment of the second case will be provided later on.

Instead of the "cotangent" transformation (4.2), we consider the "cosine" change of variable $z \to x$ given formally by

$$x = \tfrac{1}{2}(\zeta_0^{-1/2} z^{1/2} + \zeta_0^{1/2} z^{-1/2}), \tag{5.4}$$

i.e., $z = e^{i\theta} \to x = \cos(\theta - \theta_0)/2$. This maps the unit circle to the real interval $[-1, 1]$. Note that the sign ambiguity in (5.4) is harmless; in fact, we are only interested in the correspondence between *pairs* of points, given by

$$z = e^{i(\theta_0 \pm \theta)} \to x = \pm \cos \theta/2. \tag{5.5}$$

Note, however, the exceptional situations $z = \zeta_0 \to x = \pm 1$ and $z = -\zeta_0 \to x = 0$.

With the symmetric polynomials $p_k(z)$ generated by (2.1) let us associate the functions $\phi_k(x)$ defined by

$$\phi_k(x) = z^{-k/2} p_k(z), \tag{5.6}$$

by means of (5.4). It is easily seen that $\phi_k(x)$ is a *polynomial of degree k in the variable x, with real coefficients $\phi_{k,i}$ enjoying the parity property*

$$\phi_k(-x) = (-1)^k \phi_k(x), \tag{5.7}$$

i.e., $\phi_{k,k-1-2i} = 0$ for $i = 0, 1, \cdots, \lfloor (k-1)/2 \rfloor$. It follows from (5.3), (5.4), and (5.6) that the three-term recurrence relation (2.1) can be written in the simple form

$$\phi_{k+1}(x) = 2\nu_k x \phi_k(x) - \phi_{k-1}(x). \tag{5.8}$$

The initial conditions are given by $\phi_{-1}(x) = 0$ and $\phi_0(x) = p_0$. Note that the numbers $\phi_k(1)$ have the same sign, whereas the numbers $\phi_k(-1)$ have alternating signs. Indeed, we have $\phi_k(1) = p_0 \lambda_1 \cdots \lambda_k$ and $\phi_k(-1) = (-1)^k \phi_k(1)$, in view of (2.5) and (5.7).

The properties mentioned above mean exactly that $\phi_0(x), \phi_1(x), \cdots, \phi_n(x)$ constitute a *family of orthogonal polynomials on the interval* $[-1, 1]$, *with respect to an even measure*. (We shall identify the appropriate "minimal discrete measure" in the sequel.)

In particular, (5.8) is the well-known three-term formula of orthogonal polynomial theory [25]. It is useful to "symmetrize" (5.8). To that end, let us introduce the *normalized polynomials*

$$\psi_k(z) = \nu_k^{1/2}\phi_k(x), \tag{5.9}$$

for $k = 0, 1, \cdots, n$, with an arbitrary positive value for ν_n. Thus, we have $\psi_0(x) = |p_0|^{1/2}$. In addition, we set $\psi_{-1}(x) = 0$. Next, let us define the positive numbers $\beta_1, \cdots, \beta_n$ as follows:

$$\beta_k = \tfrac{1}{2}(\nu_{k-1}\nu_k)^{-1/2} = \tfrac{1}{2}(1+\zeta_0^{k-1}\rho_{k-1})^{1/2}(1-\zeta_0^k\rho_k)^{1/2}. \tag{5.10}$$

The rightmost expression in (5.10) is obtained from (2.7), (5.2), and (5.3), with $\varepsilon = 1$. It is valid for $k = 1, \cdots, n-1$ (not for $k = n$). Note that β_k depends only on the reflection coefficients ρ_{k-1} and ρ_k, in a simple manner. By use of (5.9) and (5.10) we can write (5.8) in the standard symmetric form

$$\beta_{k+1}\psi_{k+1}(x) = x\psi_k(x) - \beta_k\psi_{k-1}(x). \tag{5.11}$$

Consider the $n \times n$ *real symmetric tridiagonal matrix* T associated classically with the recurrence (5.11), that is,

$$T = \begin{bmatrix} 0 & \beta_1 & & \\ \beta_1 & 0 & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ & & \beta_{n-1} & 0 \end{bmatrix}. \tag{5.12}$$

Thus, T is the confederate matrix of $\psi_n(x)$ with respect to the graded basis $(\psi_k(x))_{k=0}^n$. More precisely, the set of identities (5.11) with $0 \leqq k \leqq n-1$ can be given the form

$$(xI - T)\psi(x) = [0, \cdots, 0, \beta_n\psi_n(x)]^T, \tag{5.13}$$

with $\psi(x) = [\psi_0(x), \cdots, \psi_{n-1}(x)]^T$. As a consequence, *the zeros* $x_1, \cdots, x_n$ *of* $\psi_n(x)$ *are the eigenvalues of* T, *and the vectors* $\psi(x_1), \cdots, \psi(x_n)$ *are the corresponding eigenvectors*. It is interesting to compare these statements with the analogous statements made in § 4 concerning the role of unitary Hessenberg matrices with respect to orthogonal polynomials on the unit circle.

The result (5.13) leads to an efficient method for computing the zeros $\zeta_j = \exp(i\theta_j)$ of the predictor polynomial $a_n(z)$ from the given reflection coefficients ρ_k. (When it applies, i.e., in the "real case" (5.1), this method is certainly preferable to the general technique described in § 4.) Let us briefly explain the idea. In view of (5.7), the zeros of $\psi_n(x)$ occur in *symmetric pairs* $\{x_j, -x_j\}$, with the exception of the value $x_j = 0$ when n is odd. From (5.5) and (5.6) it follows that these pairs correspond to the conjugate pairs of zeros $\{\zeta_0^{-1}\zeta_j, \zeta_0\zeta_j^{-1}\}$ of the real polynomial $a_n'(z) = a_n(\zeta_0 z)$ via the formula

$$\{x_j, -x_j\} = \{\cos(\theta_j - \theta_0)/2, -\cos(\theta_j - \theta_0)/2\}. \tag{5.14}$$

The main conclusion of this development is the following.

PROPOSITION 6. *The zeros* ζ_j *of* $a_n(z)$ *can be obtained, via* (5.14), *from the eigenvalues* x_j *of the tridiagonal matrix* T, *which is defined by means of* (5.10) *in terms of the reflection coefficients* ρ_k.

Without going into detail, let us mention that the eigenvalue problem for this matrix T is exactly equivalent to the *singular value problem for the* $\lfloor n/2 \rfloor \times \lceil n/2 \rceil$ *bidiagonal matrix* that has the numbers $\beta_1, \beta_3, \cdots$ on the diagonal and the numbers $\beta_2, \beta_4, \cdots$ just below the diagonal [13], [14].

The method outlined above is especially attractive for determining the eigenvalues of a *real orthogonal Hessenberg matrix* D. (In this context, the idea goes back to Rutishauser [23]; an efficient implementation was recently proposed by Ammar, Gragg, and Reichel [2].) Indeed, D has real parameters $\rho_1, \cdots, \rho_n$, via (4.17), and its eigenvalues are the zeros of the corresponding predictor polynomial $a_n(z)$; see § 4. Of course, the appropriate choices for the circle parameter are $\zeta_0 = 1$ and/or $\zeta_0 = -1$ in that case. (The condition $\varepsilon = 1$ amounts to $\zeta_0^n = \rho_n$. Hence, there are zero or two possibilities for ζ_0 when n is even, and there is one possibility when n is odd.)

Next, we examine the question of the *orthogonality relations* satisfied by the polynomials $\psi_k(x)$. The result will provide us with an explicit formula for the weights h_j, which are required in certain applications. We shall prove that the "discrete version" of the orthogonality relations is nothing but

$$\sum_{j=1}^{n} h_j \psi_k(x_j)\psi_l(x_j) = 2c_0\delta_{k,l}, \tag{5.15}$$

for $0 \leqq k, l \leqq n-1$, where $x_1, \cdots, x_n$ are the zeros of $\psi_n(x)$ and where the Christoffel numbers $h_1, \cdots, h_n$ coincide with the weights (3.8) of the measure $d\mu$. According to the classical theory of orthogonal polynomials [25], the only thing we have to show is that the inverse of the weight h_j is given by

$$h_j^{-1} = (2c_0)^{-1}\beta_n \psi_n'(x_j)\psi_{n-1}(x_j). \tag{5.16}$$

From the results of § 3 we deduce

$$h_j^{-1} = c_0^{-1}(\zeta_j - \zeta_0)^{-1}\zeta_0^{1/2}\zeta_j p_n'(\zeta_j)\bar{p}_{n-1}(\zeta_j). \tag{5.17}$$

It is an easy exercise to transform (5.17) into the desired formula (5.16), by use of (5.4), (5.6), and (5.9). Details are omitted. The conclusion can be stated as follows.

PROPOSITION 7. *The weights h_j associated with the mass points θ_j are equal to the Christoffel numbers relative to the zeros x_j of $\psi_n(x)$. More precisely, they are given by the Christoffel formula*

$$h_j = 2c_0\left[\sum_{k=0}^{n-1} \psi_k^2(x_j)\right]^{-1}. \tag{5.18}$$

This should be compared with the analogous expression (3.8) in Szegö's theory. Observe that the zeros x_j and $-x_j$ of $\psi_n(x)$ yield the same weight (or Christoffel number) h_j. Note also that (5.18) can be written in terms of the values $p_k(\zeta_j)$ as follows:

$$h_j = 2c_0\zeta_0^{1/2}\left[\sum_{k=0}^{n-1} \alpha_k |p_k(\zeta_j)|^2\right]^{-1}. \tag{5.19}$$

Therefore, an alternative proof of (5.15) consists of showing that formula (3.13) reduces to (5.19) under the reality assumption (5.1), (5.3).

Recall that the *dual reflection coefficients* $\rho_k^{\#} = \rho_n\bar{\rho}_{n-k}$ generate the same ρ_n-symmetric predictor polynomial $a_n(z)$ as the given reflection coefficients ρ_k (see [9]). Hence, the zeros ζ_j of $a_n(z)$ can be obtained alternatively, via (5.14), from the eigenvalues x_j of the real symmetric tridiagonal matrix $T^{\#}$ whose nontrivial entries $\beta_k^{\#}$ are given by

$$\beta_k^{\#} = \tfrac{1}{2}(1 - \zeta_0^k\rho_k)^{1/2}(1 + \zeta_0^{k+1}\rho_{k+1})^{1/2}, \tag{5.20}$$

for $k = 1, \cdots, n-1$. This expression of $\beta_k^{\#}$ is deduced from (5.10), by use of the assumption $\varepsilon = 1$. (For convenience, we have included the mirror permutation $k \to n-k$ in the definition of $T^{\#}$.)

It is interesting to give a context-free proof of the fact that the matrices T and $T^{\#}$ have the same spectrum. Let us introduce the monic orthogonal polynomials $\pi_k(x) = \psi_k(x)/\psi_{k,k}$. They satisfy the recurrence relation

$$\pi_{k+1}(x) = x\pi_k(x) - \beta_k^2 \pi_{k-1}(x). \tag{5.21}$$

Note that $\pi_k(x)$ is the characteristic polynomial of the top principal submatrix of order k of T. With a similar definition for the dual polynomials $\pi_k^{\#}(z)$ from the matrix $T^{\#}$, we have the remarkable identity

$$\pi_k(x) = \pi_k^{\#}(x) - \tfrac{1}{4}(1 - \zeta_0^{k-1}\rho_{k-1})(1 - \zeta_0^k \rho_k)\pi_{k-2}^{\#}(x), \tag{5.22}$$

for $k = 2, \cdots, n$. This can be proved easily by induction, with the help of (5.10), (5.20), (5.21), and its dual. Details are left to the reader. In particular, (5.3) yields the desired result $\pi_n(x) = \pi_n^{\#}(x)$, since $\zeta_0^n \rho_n = \varepsilon = 1$.

Let us now consider the case $\varepsilon = -1$, which characterizes the fact that ζ_0 is a zero of $a_n(z)$. The final remark in § 3 shows that the results above apply without modification to this case provided α_k is replaced by $\tilde{\alpha}_k = \alpha_{k+1}$. Thus, the remaining zeros $\zeta_1, \cdots, \zeta_{n-1}$ of $a_n(z)$ can be obtained from (5.14) where the numbers $x_1, \cdots, x_{n-1}$ are interpreted as the eigenvalues of the symmetric tridiagonal matrix $\tilde{T}$, of the form (5.10) with $\tilde{n} = n - 1$, whose nontrivial entries $\tilde{\beta}_k$ are given by

$$\tilde{\beta}_k = \tfrac{1}{2}(1 - \zeta_0^k \rho_k)^{1/2}(1 + \zeta_0^{k+1}\rho_{k+1})^{1/2}, \tag{5.23}$$

for $k = 1, \cdots, n - 2$. (Formally, this is identical to (5.20).) Note that changing the sequence $(\rho_k)_{k=1}^n$ into its dual $(\rho_k^{\#})_{k=1}^n$ would yield the same tridiagonal matrix as $\tilde{T}$.

Remark [2]. Recall that we can replace ζ_0 by $-\zeta_0$, i.e., θ_0 by $\theta_0 \pm \pi$ in the method described above. (This preserves the parameter $\varepsilon = \zeta_0^n \rho_n$ if and only if n is even.) Thus, the zeros of $a_n(z)$ can be determined not only, via (5.14), from the eigenvalues x_j of the tridiagonal matrix T or $\tilde{T}$ relative to the point ζ_0, but also, via

$$\{x_j', -x_j'\} = \{\sin(\theta_j - \theta_0)/2, -\sin(\theta_j - \theta_0)/2\}, \tag{5.24}$$

from the eigenvalues x_j' of the tridiagonal matrix T or $\tilde{T}$ relative to the point $-\zeta_0$. It may be interesting to compute both spectra (x_j) and (x_j') in order to determine the zeros of $a_n(z)$ with good accuracy by means of (5.14) *and* (5.24).

Finally, let us examine a rational function $f(x)$ that is of frequent use in the theory of orthogonal polynomials on the real line. The definition is

$$f(x) = \psi_{n-1}(x)/\psi_n(x). \tag{5.25}$$

Formulas (5.6) and (5.9) provide a direct connection between $f(x)$ and the lossless function $\check{h}(z)$ in (3.14), that is,

$$f(x) = c_0^{-1} p_0^2 \beta_n \nu_{n-1}(\zeta_0^{1/2} z^{-1/2} - \zeta_0^{-1/2} z^{1/2})^{-1} \check{h}(z). \tag{5.26}$$

Next, consider the polar expansion

$$f(x) = \sum_{j=1}^{n} f_j (x - x_j)^{-1}, \tag{5.27}$$

with $f_j = \psi_{n-1}(x_j)/\psi_n'(x_j)$. In view of (5.16), we have

$$f_j = (2c_0)^{-1} \beta_n h_j \psi_{n-1}^2(x_j). \tag{5.28}$$

On the other hand, by formal computation based on (5.4), we can express (5.27) in terms of the variable z as follows:

$$f(x) = 2(\zeta_0^{1/2} z^{-1/2} - \zeta_0^{-1/2} z^{1/2})^{-1} \sum_{j=1}^{n} f_j \frac{\zeta_j + z}{\zeta_j - z}. \tag{5.29}$$

Using (5.26) and comparing both expansions (5.28) and (3.15), with $\check{\gamma} = 0$ since $\check{h}(0)$ is real, we deduce the identity

$$f_j = (2c_0)^{-1} p_0^2 \beta_n \nu_{n-1} \check{h}_j. \tag{5.30}$$

By equating the right-hand sides of (5.28) and (5.30), we obtain an alternative proof (in the "real case") of the connection (3.16) between the weights h_j and $\check{h}_j$.

REFERENCES

[1] N. I. AKHIEZER, *The Classical Moment Problem*, Oliver and Boyd, London, 1965.

[2] G. S. AMMAR, W. B. GRAGG, AND L. REICHEL, *On the eigenproblem for orthogonal matrices*, in Proc. 25th Annual IEEE Conference on Decision and Control, Athens, Greece, 1986, pp. 1963–1966.

[3] S. BARNETT, *Congenial matrices*, Linear Algebra Appl., 41 (1981), pp. 277–298.

[4] G. CYBENKO, *Computing Pisarenko frequency estimates*, in Proc. Conference on Information Systems and Sciences, Princeton University, Princeton, NJ, 1984, pp. 587–591.

[5] P. DELSARTE AND Y. GENIN, *The split Levinson algorithm*, IEEE Trans. Acoust. Speech Signal Process., 34 (1986), pp. 470–478.

[6] ———, *The tridiagonal approach to Szegö's orthogonal polynomials, Toeplitz linear systems, and related interpolation problems*, SIAM J. Math. Anal., 19 (1988), pp. 718–735.

[7] ———, *A survey of the split approach based techniques in digital signal processing applications*, Philips J. Res., 43 (1988), pp. 346–374.

[8] ———, *On the split approach based algorithms for* DSP *problems*, in Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, G. H. Golub and P. Van Dooren, eds., Springer NATO ASI Series, Springer-Verlag, Berlin, New York, 1991, pp. 131–148.

[9] ———, *Tridiagonal approach to the algebraic environment of Toeplitz matrices, Part* I: *Basic results*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 220–238.

[10] P. DELSARTE, Y. GENIN, AND Y. KAMP, *Application of the index theory of pseudolossless functions to the Bistritz stability test*, Philips J. Res., 39 (1984), pp. 226–241.

[11] K. FAN, *On strictly dissipative matrices*, Linear Algebra Appl., 9 (1974), pp. 223–241.

[12] Y. GENIN, *On a duality relation in the theory of orthogonal polynomials and its application in signal processing*, in Proc. First Internat. Conference on Industrial and Applied Mathematics, Paris, 1987, J. McKenna and R. Teman, eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1988, pp. 102–113.

[13] G. H. GOLUB, *Least squares, singular values and matrix approximations*, Apl. Mat., 13 (1968), pp. 44–51.

[14] G. H. GOLUB AND W. KAHAN, *Calculating the singular values and pseudoinverse of a matrix*, SIAM J. Numer. Anal. Ser. B, 2 (1965), pp. 205–224.

[15] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, North Oxford Academic, Oxford, 1983.

[16] W. B. GRAGG, *Positive definite Toeplitz matrices, the Arnoldi process for isometric operators, and Gaussian quadrature on the unit circle*, in Numerical Methods in Linear Algebra, E. S. Nikolaev, ed., Moscow University Press, Moscow, 1982, pp. 16–32. (In Russian.)

[17] ———, *The* QR *algorithm for unitary Hessenberg matrices*, J. Comput. Appl. Math., 16 (1986), pp. 1–8.

[18] W. B. GRAGG AND L. REICHEL, *A divide and conquer method for unitary and orthogonal eigenproblems*, Numer. Math., 57 (1990), pp. 695–718.

[19] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and Their Applications*, University of California Press, Berkeley, CA, 1958.

[20] A. S. HOUSEHOLDER, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York, 1964.

[21] H. KIMURA, *Generalized Schwarz form and lattice–ladder realizations of digital filters*, IEEE Trans. Circuits Systems, 32 (1985), pp. 1130–1139.

[22] V. P. PISARENKO, *The retrieval of harmonics from a covariance function*, Geophys. J. R. Astr. Soc., 33 (1973), pp. 347–366.

[23] H. RUTISHAUSER, *Bestimmung der Eigenwerte orthogonaler Matrizen*, Numer. Math., 9 (1966), pp. 104–108.

[24] S. SAGAYAMA AND F. ITAKURA, *Duality theory of composite sinusoidal modeling and linear prediction*, in Proc. IEEE Internat. Conference on Acoustics, Speech, and Signal Processing, Tokyo, 1986, pp. 1261–1264.

[25] G. SZEGÖ, *Orthogonal Polynomials*, American Mathematical Society, New York, 1959.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 449–462, July 1991

004

MONOTONICITY PROPERTIES OF THE TODA FLOW, THE QR-FLOW, AND SUBSPACE ITERATION*

JEFFREY C. LAGARIAS†

Abstract. Let $\mathbf{X}(t)$ denote the Toda flow on the space of $n \times n$ matrices, with $\mathbf{X}(0)$ a symmetric matrix, and let $\mathbf{X}_r(t)$ denote the $r \times r$ upper left corner principal submatrix of $\mathbf{X}(t)$, i.e., $\mathbf{X}_r(t) = \mathbf{E}_r^T \mathbf{X}(t) \mathbf{E}_r$ where $\mathbf{E}_r = \begin{bmatrix} I_r \\ 0 \end{bmatrix}$. Then the r ordered eigenvalues $\lambda_1(\mathbf{X}_r(t)) \geq \lambda_2(\mathbf{X}_r(t)) \geq \cdots \geq \lambda_r(\mathbf{X}_r(t))$ of $\mathbf{X}_r(t)$ are each a nondecreasing function of t, for $1 \leq r \leq n$. A similar result is proved for the QR-flow $\mathbf{Y}(t) = \exp(\mathbf{X}(t))$, for the eigenvalues of $\mathbf{Y}_r(t) = \mathbf{E}_r^T \mathbf{Y}(t) \mathbf{E}_r$. For any generalized Toda flow $f(\mathbf{X}(t))$ with $f(\cdot)$ a nondecreasing function, it is shown that $\operatorname{Tr}(\mathbf{E}_r^T f(\mathbf{X}(t)) \mathbf{E}_r)$ is a nondecreasing function of t. The QR-flow inequalities are used to show that the Ritz values of a symmetric matrix $\mathbf{X}$ on a subspace are nondecreasing under subspace iteration.

Key words. Toda flow, QR-flow, QR-algorithm, subspace iteration, Ritz value

AMS(MOS) subject classifications. 15A42, 65F15, 34A34

1. Introduction. Let $\mathbf{X}$ be an $n \times n$ symmetric matrix and let $\mathcal{S}$ be an r-dimensional subspace of $\mathbb{R}^n$. The *Ritz values* of $\mathbf{X}$ on $\mathcal{S}$ [P, p. 214], denoted by

$$\lambda_1(\mathbf{X}; \mathcal{S}) \geq \cdots \geq \lambda_r(\mathbf{X}; \mathcal{S}),$$

are the eigenvalues of $\mathbf{V}^T \mathbf{X} \mathbf{V}$ where $\mathbf{V}$ is any $n \times r$ partial isometry ($\mathbf{V}^T \mathbf{V} = \mathbf{I}_r$) whose columns span $\mathcal{S}$. The Ritz values depend only on $\mathcal{S}$, for if $\mathbf{V}$, $\tilde{\mathbf{V}}$ are any two partial isometries spanning $\mathcal{S}$, then $\mathbf{V} = \tilde{\mathbf{V}}\mathbf{O}$ for some $r \times r$ orthogonal matrix $\mathbf{O}$.

Subspace iteration [P, Chap. 14] is an extension of the power method for finding the largest eigenvalue of a symmetric matrix. Given a subspace $\mathcal{S}_0$, define subspaces of $\mathcal{S}_i$ recursively by $\mathcal{S}_{i+1} = \mathbf{X}\mathcal{S}_i$. For most subspaces $\mathcal{S}_0$, the Ritz values of $\mathcal{S}_i$ converge to the r largest eigenvalues of $\mathbf{X}$ as $i \to \infty$. This paper shows that the Ritz values are nondecreasing under subspace iteration, i.e., for any symmetric $\mathbf{X}$ and any subspace $\mathcal{S}_0$,

$$\lambda_i(\mathbf{X}; \mathcal{S}_1) \geq \lambda_i(\mathbf{X}; \mathcal{S}_0), \qquad 1 \leq i \leq r. \tag{1.1}$$

This eigenvalue monotonicity property of subspace iteration is well known for the largest eigenvalue, and has apparently been observed in general (see [R, p. 11], [P, p. 295]), but without proof. The inequalities (1.1) are derived from eigenvalue monotonicity properties of a continuous flow interpolating subspace iteration, which itself is a projection of the QR-flow (defined below). More generally, the paper proves monotonicity inequalities for generalized Toda flows, which include the QR-flow as a special case.

The *Toda flow* or *Toda lattice* is a Hamiltonian dynamical system originally proposed to describe the motion of a set of particles moving on a line under the influence of exponentially repulsive potentials between nearest neighbors [T]. Flaschka [F1], [F2] showed that for n particles this system of differential equations can, by a change of variable, be put in the Lax pair form

$$\begin{aligned} \dot{\mathbf{X}}(t) &= [\mathbf{X}(t), \pi_s(\mathbf{X}(t))] \\ &= \mathbf{X}(t)\pi_s(\mathbf{X}(t)) - \pi_s(\mathbf{X}(t))\mathbf{X}(t). \end{aligned} \tag{1.2}$$

* Received by the editors January 18, 1989; accepted for publication (in revised form) February 27, 1990.
† AT&T Bell Laboratories, Murray Hill, New Jersey 07974 (jcl@research.att.com).

Here $\mathbf{X}(t)$ is an $n \times n$ symmetric tridiagonal matrix with entries $x_{ij}(t)$, and $\pi_s(\mathbf{X}(t))$ is the skew-symmetric tridiagonal matrix associated to $\mathbf{X}(t)$ by

$$\pi_s(\mathbf{X}(t)) = \begin{cases} \mathbf{X}_{ij}(t), & i>j, \\ 0, & i=j, \\ -\mathbf{X}_{ji}(t), & i<j. \end{cases}$$

The mapping $\pi_s(\mathbf{X})$ arises from the unique decomposition (for real matrices $\mathbf{X}$)

$$\mathbf{X} = \pi_s(\mathbf{X}) + \pi_u(\mathbf{X}), \tag{1.3}$$

in which $\pi_s(\mathbf{X})$ is skew-symmetric and $\pi_u(\mathbf{X})$ is upper-triangular. The mappings $\pi_s(\cdot)$ and $\pi_u(\cdot)$ are both linear maps. The Lax pair equation (1.2) guarantees that the eigenvalues of $\mathbf{X}(t)$ are independent of t, i.e., the flow $\mathbf{X}(t)$ is *isospectral* (cf. [L]). Flaschka observed that the eigenvalues of $\mathbf{X}(t)$ form a set of commuting Hamiltonians for the flow, which show that it is a completely integrable Hamiltonian dynamical system. In 1975, Moser [M] made a detailed study of this flow, obtaining an explicit solution and, among other things, showing that the flow always converges to a diagonal matrix as $t \to \pm\infty$.

In 1982, Symes [Sy] observed that there is a close connection between the Toda flow and the QR-algorithm for finding the eigenvalues and eigenvectors of a positive-definite symmetric tridiagonal matrix (see [DNT], [Fr], [W1], [W2]). Any invertible real matrix $\mathbf{M}_0$ has a unique QR-decomposition

$$\mathbf{M}_0 = \mathbf{QR},$$

in which $\mathbf{Q}$ is orthogonal and $\mathbf{R}$ is upper triangular with positive elements on the diagonal. This decomposition is computed using the Gram–Schmidt orthogonalization process on the columns of $\mathbf{M}_0$. The QR-algorithm iterate $\mathbf{M}_1$ of $\mathbf{M}_0$ is

$$\mathbf{M}_1 = \mathbf{RQ} = \mathbf{Q}^T\mathbf{M}_0\mathbf{Q}.$$

For any positive-definite symmetric matrix $\mathbf{M}_0$, the QR-algorithm iterates $\{\mathbf{M}_i\}$ of $\mathbf{M}_0$ converge to a diagonal matrix that has the eigenvalues of $\mathbf{M}_0$ on the diagonal. Now associate to the Toda flow $\mathbf{X}(t)$ the flow

$$\mathbf{Y}(t) = \exp(\mathbf{X}(t)), \tag{1.4}$$

which evolves according to the Lax pair equation

$$\dot{\mathbf{Y}}(t) = [\mathbf{Y}(t), \pi_s(\log \mathbf{Y}(t))],$$

where $\log \mathbf{Y}(t)$ denotes the (unique) symmetric logarithm of a positive-definite symmetric matrix $\mathbf{Y}(t)$. Symes [Sy][1] and later Deift, Nanda, and Tomei [DNT] observed that if

$$\mathbf{Y}(t) = \mathbf{Q}(t)\mathbf{R}(t), \tag{1.5}$$

then

$$\mathbf{Y}(t+1) = \mathbf{R}(t)\mathbf{Q}(t) = \mathbf{Q}(t)^T\mathbf{Y}(t)\mathbf{Q}(t).$$

Hence for positive-definite symmetric matrices, the flow $\mathbf{Y}(t)$ gives at integer times the QR-algorithm iterates of $\mathbf{Y}(0)$. For this reason the flow $\mathbf{Y}(t)$ is called the QR-*flow*.

[1] Symes's results actually apply to the QL-flow, but carry over in a straightforward way to the QR-flow.

The QR-decomposition can be used to explicitly exhibit the isospectral nature of the Toda flow (see [Sy]). Consider the QR-decomposition

$$\exp(t\mathbf{X}(0)) = \bar{\mathbf{Q}}(t)\bar{\mathbf{R}}(t). \tag{1.6}$$

Then we have

$$\mathbf{X}(t) = \bar{\mathbf{Q}}(t)^T\mathbf{X}(0)\bar{\mathbf{Q}}(t), \tag{1.7}$$

which exhibits a similarity transformation between $\mathbf{X}(t)$ and $\mathbf{X}(0)$. We also have

$$\mathbf{Y}(t) = \bar{\mathbf{Q}}(t)^T\mathbf{Y}(0)\bar{\mathbf{Q}}(t). \tag{1.8}$$

The Toda flow is defined on the set Σ_n of real symmetric $n \times n$ matrices, and remains a completely integrable Hamiltonian dynamical system on "generic" matrices [DLNT]. For all matrices on Σ_n, it has the property that all off-diagonal elements $\mathbf{X}_{ij}(t) \to 0$ as $t \to \pm\infty$. For a "generic" symmetric starting point $\mathbf{X}(0)$ the limiting values $\mathbf{X}(-\infty)$, as $t \to -\infty$ and $\mathbf{X}(\infty)$ as $t \to \infty$, are given by

$$\mathbf{X}(-\infty) = \begin{pmatrix} \lambda_n & & \\ & \ddots & \\ & & \lambda_1 \end{pmatrix}, \qquad \mathbf{X}(\infty) = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, \tag{1.9}$$

in which $\lambda_1 \geqq \cdots \geqq \lambda_n$ are the eigenvalues of $\mathbf{X}_0$ arranged in *decreasing* order. (This does not hold for all starting points $\mathbf{X}(0)$, e.g., for any diagonal $\mathbf{X}(0)$ the Toda flow is constant, so that $\mathbf{X}(-\infty) = \mathbf{X}(\infty) = \mathbf{X}(0)$ in this case.) Thus the effect of the time evolution of the Toda flow for $t = -\infty$ to $t = +\infty$ on a "generic" symmetric matrix is to reverse the order of eigenvalues on the diagonal, from increasing order at $-\infty$ to decreasing order at $+\infty$.

This paper proves matrix inequalities which show that, as time increases, the Toda flow on symmetric matrices moves eigenvalues up and to the left in a *monotone* fashion on *all* Toda orbits, including "nongeneric" ones. To make precise statements, let $\mathbf{X}^{(r)}(t)$ denote the $r \times r$ principal submatrix of $\mathbf{X}(t)$ and $\lambda_{1,r}(t) \geqq \lambda_{2,r}(t) \geqq \cdots \geqq \lambda_{r,r}(t)$ be the eigenvalues of $\mathbf{X}^{(r)}(t)$ arranged in decreasing order. Note that

$$\mathbf{X}^{(r)}(t) = \mathbf{E}_r^T\mathbf{X}(t)\mathbf{E}_r = \begin{bmatrix} x_{11}(t) \cdots x_{1r}(t) \\ \vdots \qquad\quad \vdots \\ x_{r1}(t) \cdots x_{rr}(t) \end{bmatrix},$$

where $\mathbf{E}_r$ is the $n \times r$ matrix $\begin{bmatrix} I_r \\ 0 \end{bmatrix}$, so that $\mathbf{E}_r^T\mathbf{E}_r = \mathbf{I}_r$, the $r \times r$ identity matrix. The eigenvalues $\lambda_{i,r}(t)$ depend on the initial condition $\mathbf{X}(0) = \mathbf{X}_0$ of the differential equation; this will be indicated as $\lambda_{i,r}(t, \mathbf{X}_0)$ when necessary.

THEOREM 1. *For $1 \leqq r \leqq n$ and any symmetric matrix $\mathbf{X}(0)$, all the ordered eigenvalues of the projected Toda flow orbit $\mathbf{X}^{(r)}(t) = \mathbf{E}_r^T\mathbf{X}(t)\mathbf{E}_r$ are nondecreasing functions of t, i.e., if $t_1 > t_2$, then*

$$\lambda_{j,r}(t_1) \geqq \lambda_{j,r}(t_2), \qquad 1 \leqq j \leqq r. \tag{1.10}$$

This theorem asserts that while $\mathbf{X}(t)$ is an isospectral flow on the space of $n \times n$ matrices, when restricted to $r \times r$ principal submatrices with $r < n$, it is *spectrum-increasing* (actually spectrum-nondecreasing).

Theorem 1 yields a large class of matrix inequalities through rescalings of the eigenvalues of the Toda flow. Any function $f: \mathbb{R} \to \mathbb{R}$ induces an operator-valued mapping, also labeled f, of the space Σ_n of $n \times n$ real symmetric matrices into itself (see [Do], [Lo]) such that if $(\lambda_1, \cdots, \lambda_n)$ are the eigenvalues of $\mathbf{X}$, then $(f(\lambda_1), \cdots, f(\lambda_n))$ are the eigenvalues of $f(\mathbf{X})$. Theorem 1 immediately implies the following result.

THEOREM 2. *Let $f_1(x)$ be any nondecreasing real-valued function. For $1 \leqq r \leqq n$ and any symmetric matrix $\mathbf{X}(0)$, the eigenvalues of $f_1(\mathbf{E}_r^T \mathbf{X}(t)\mathbf{E}_r)$ arranged in decreasing order are $f_1(\lambda_{j,r}(t))$ for $1 \leqq j \leqq r$ and each $f_1(\lambda_{j,r}(t))$ is a nondecreasing function of t. In particular,*

$$\mathrm{T}_{f_1,I}(t) = \mathrm{Tr}\,(f_1(\mathbf{E}_r^T \mathbf{X}(t)\mathbf{E}_r)) \tag{1.11}$$

is a nondecreasing function of t for $-\infty < t < \infty$.

Next we establish similar monotonicity inequality of eigenvalues for all positive powers of the QR-flow. Let $\mathbf{Y}(t; \mathbf{Y}_0)$ denote a QR-flow orbit with $\mathbf{Y}(0) = \mathbf{Y}_0$, and let $\{\lambda_{i,r}(t; \alpha, \mathbf{Y}_0)\colon 1 \leqq i \leqq n\}$ denote the eigenvalues of $\mathbf{E}_r^T \mathbf{Y}(t)^\alpha \mathbf{E}_r$ in decreasing order.

THEOREM 3. *Let $\mathbf{Y}_0$ be a positive-definite symmetric matrix and $\mathbf{Y}(t)$ the* QR*-flow orbit having $\mathbf{Y}(0) = \mathbf{Y}_0$. Then for all $\alpha > 0$ and $1 \leqq r \leqq n$ all the ordered eigenvalues of $\mathbf{Y}_\alpha^{(r)}(t) = \mathbf{E}_r^T \mathbf{Y}(t)^\alpha \mathbf{E}_r$ are nondecreasing functions of t. That is, if $t_1 \geqq t_2$, then*

$$\lambda_{j,r}(t_1; \alpha, \mathbf{Y}_0) \geqq \lambda_{j,r}(t_2; \alpha, \mathbf{Y}_0), \qquad 1 \leqq j \leqq r. \tag{1.12}$$

As an immediate corollary of Theorem 3 we obtain another large class of matrix inequalities through reparameterization of the QR-flow.

THEOREM 4. *Let $f_1(x)$ be any nondecreasing real-valued function for $x > 0$. Let $\mathbf{Y}_0$ be a positive-definite symmetric matrix, and let $\mathbf{Y}(t)$ be the* QR*-flow orbit having $\mathbf{Y}(0) = \mathbf{Y}_0$. Then for $1 \leqq r \leqq n$ and any $\alpha > 0$, the eigenvalues of $f_1(\mathbf{E}_r^T \mathbf{Y}(t)^\alpha \mathbf{E}_r)$ arranged in decreasing order, which are $f_1(\lambda_{j,r}(t; \alpha, \mathbf{Y}_0))$, are each a nondecreasing function of t. In particular,*

$$\mathrm{T}_{f_1,\exp(\alpha)}(t) = \mathrm{Tr}\,(f_1(\mathbf{E}_r^T \mathbf{Y}(t)^\alpha \mathbf{E}_r))$$

is a nondecreasing function of t for $-\infty < t < \infty$.

Section 4 establishes a monotonicity inequality for the trace of a projection of a generalized Toda flow. A *generalized Toda flow* is any reparameterized flow $f(\mathbf{X}(t))$ on the symmetric matrices Σ_n induced from the Toda flow $\mathbf{X}(t)$ by a monotone increasing function $f\colon \mathbb{R} \to \mathbb{R}$, see [C1].

THEOREM 5. *Let $f_2(x)$ be any nondecreasing real-valued function defined on $\mathbb{R}$. For $1 \leqq r \leqq n$ and any symmetric matrix $\mathbf{X}(0)$, the quantity*

$$\mathrm{T}_{I,f_2}(t) = \mathrm{tr}\,(\mathbf{E}_r^T f_2(\mathbf{X}(t))\mathbf{E}_r)$$

is a nondecreasing function of t for $-\infty < t < \infty$.

Section 5 shows that the *simple subspace iteration algorithm* [P, p. 290] for finding the r largest eigenvalues of a symmetric matrix is a projected form of the QR-algorithm, an observation due to Watkins [W1]. Using this result, the Ritz value inequalities (1.1) follow from Theorem 4. Hastie [H] suggested using simple subspace iteration to find a best rank r approximation $\hat{\mathbf{X}}$ to a given symmetric matrix $\mathbf{X}$. He defines $\hat{\mathbf{X}}_j = \mathbf{X}\mathbf{V}_j\mathbf{V}_j^T$ and $\hat{\hat{\mathbf{X}}}_j = \mathbf{V}_j\mathbf{V}_j^T\mathbf{X}\mathbf{V}_j\mathbf{V}_j^T$, where $\mathbf{V}_j$ is the jth simple subspace iteration basis. We show that

$$\|\hat{\mathbf{X}}_{j+1} - \mathbf{X}\|^2 \leqq \|\hat{\mathbf{X}}_j - \mathbf{X}\|^2,$$

$$\|\hat{\hat{\mathbf{X}}}_{j+1} - \mathbf{X}\|^2 \leqq \|\hat{\hat{\mathbf{X}}}_j - \mathbf{X}\|^2,$$

where $\|\mathbf{X}\|^2 = \sum_{i,j} \mathbf{X}_{ij}^2$ is the Frobenius norm. These inequalities answer questions raised by Hastie [H], which were the original motivation for this paper.

It seems likely that eigenvalue monotonicity inequalities hold for all generalized Toda flows.

MONOTONICITY CONJECTURE I. Let f be any nondecreasing real-valued function on $\mathbb{R}$. Let $f(\mathbf{X}(t))$ denote a generalized Toda flow orbit where $\mathbf{X}(0)$ is an $n \times n$ symmetric matrix. Then for $1 \leqq r \leqq n$, the eigenvalues of $\mathbf{E}_r^T f(\mathbf{X}(t))\mathbf{E}_r$ arranged in decreasing order are each nondecreasing functions of t for $-\infty < t < \infty$.

An equivalent form of the Monotonicity Conjecture follows.

MONOTONICITY CONJECTURE II. Let f_1 and f_2 be nondecreasing real-valued functions on $\mathbb{R}$. Let $\mathbf{X}(t)$ denote the Toda flow where $\mathbf{X}(0)$ is a symmetric $n \times n$ matrix. Then for $1 \leqq r \leqq n$, the function

$$T_{f_1,f_2}(t) = \operatorname{Tr}(f_1(\mathbf{E}_r^T f_2(\mathbf{X}(t))\mathbf{E}_r))$$

is a nondecreasing function of t, for $-\infty < t < \infty$.

The equivalence of these conjectures is easily proved. Conjecture II follows from Conjecture I since the trace is a sum of eigenvalues. To show that Conjecture I follows from Conjecture II we establish the contrapositive: If the eigenvalues of $\mathbf{E}_r^T f(\mathbf{X}(t))\mathbf{E}_r$ are somewhere decreasing as a function of t, then $\operatorname{tr}(f_1(\mathbf{E}_r^T f(\mathbf{X}(t))\mathbf{E}_r))$ is somewhere decreasing for a suitable choice of f_1. Suppose that $\lambda_i(t) = \lambda_i(\mathbf{E}_r^T f(\mathbf{X}(t))\mathbf{E}_r)$ is decreasing on the interval $[t_0, t_0 + \delta)$. Choose $f_1(t)$ to be constant outside the interval $[\lambda_i - \varepsilon, \lambda_i]$ and to have slope 1 on this interval, where $\lambda_i = \lambda_i(t_0)$, and ε is small enough to exclude all $\lambda_j(t_0) \neq \lambda_i$, and then $\operatorname{Tr}(f_1(\mathbf{E}_r^T f(\mathbf{X}(t))\mathbf{E}_r))$ will decrease on $[t_0, t_0 + \eta]$ for small enough η.

Each of Theorems 1–5 is a special case of these conjectures.

The results of this paper extend essentially without change to the complex domain, with Hermitian matrices replacing symmetric matrices, and with an appropriate change in the Toda lattice differential equation (1.2). Furthermore, similar eigenvalue monotonicity inequalities hold for the projection onto the lower right corner $r \times r$ principal submatrix of $\mathbf{X}(t)$, i.e., for $\hat{\mathbf{X}}_r(t) = \hat{\mathbf{E}}_r^T \mathbf{X}(t)\hat{\mathbf{E}}_r$ where $\hat{\mathbf{E}}_r = \begin{bmatrix} \mathbf{0} \\ I_r \end{bmatrix}$; in this case the ordered eigenvalues of $\hat{\mathbf{X}}_r(t)$ are *nonincreasing* functions of t, for $1 \leqq r \leqq n$. This result, and the analogous result for the QR-flow, can be proved using the QL-decomposition instead of the QR-decomposition (see [Sy]).

There are various directions for further work. The Toda flow and QR-flows are well defined for nonsymmetric matrices, but have different dynamical properties, e.g., individual orbits $\mathbf{X}(t)$ converge to upper triangular matrices as $t \to \infty$ (see [C2], [C3]). It is possible that some generalization of the results here carries over to that case. One might also study projections of other isospectral flows, the Cholesky and LU-flows (see [W2], [DLT]), and a Hamiltonian QR-flow [By].

2. Eigenvalue monotonicity properties. We recall basic facts about symmetric matrices. Given an $r \times r$ symmetric matrix $\mathbf{S}$, let $\lambda_1(\mathbf{S}) \geqq \cdots \geqq \lambda_r(\mathbf{S})$ denote the eigenvalues of $\mathbf{S}$ in decreasing order. These eigenvalues have the max-min characterization

$$\lambda_j(\mathbf{S}) = \max_{\dim(W) = j} \left(\min_{\substack{\mathbf{x} \in W \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^T \mathbf{S} \mathbf{x}}{\|\mathbf{x}\|^2} \right),$$

for $1 \leqq j \leqq r$ (cf. [G, Thm. 12, p. 321]). An immediate consequence is that if $\mathbf{S}$ is symmetric and $\mathbf{T}$ is positive-semidefinite symmetric, then

$$\lambda_j(\mathbf{S}+\mathbf{T}) \geqq \lambda_j(\mathbf{S}), \qquad 1 \leqq j \leqq r \tag{2.1}$$

(cf. [BE, p. 115]).

Proof of Theorem 1. Let $\mathbf{X}(t)$ evolve according to the Toda flow. A computation shows that

$$\begin{aligned} \dot{\mathbf{X}}^{(r)}(t) &= \mathbf{E}_r^T \dot{\mathbf{X}}(t)\mathbf{E}_r \\ &= \mathbf{E}_r^T [\mathbf{X}(t), \pi_s(\mathbf{X}(t))]\mathbf{E}_r \\ &= [\mathbf{X}^{(r)}(t), \pi_s(\mathbf{X}^{(r)}(t))] + \sum_{j=r+1}^{n} \mathbf{y}_{r,j}(t)\mathbf{y}_{r,j}(t)^T, \end{aligned}$$

where $\mathbf{y}_{r,j}(t)$ is the $r \times 1$ column vector $\mathbf{y}_{r,j}(t) = (x_{1j}(t), \cdots, x_{rj}(t))^T$. Thus

$$\dot{\mathbf{X}}^{(r)}(t) = [\mathbf{X}^{(r)}(t), \pi_s(\mathbf{X}^{(r)}(t))] + \mathbf{Y}_{r,n}(t) \tag{2.2}$$

where $\mathbf{Y}_{r,n}(t)$ is a positive-semidefinite symmetric matrix. The key idea of the proof lies in the form of this differential equation: the commutator term $[\mathbf{X}^{(r)}(t), \pi_s(\mathbf{X}^{(r)}(t))]$ is spectrum-preserving, while the positive-semidefinite term can only increase the spectrum of $\mathbf{X}^{(r)}(t)$.

We remove the commutator term by a suitable orthogonal transformation. Let $\mathbf{Q}(t)$ be the solution of

$$\dot{\mathbf{Q}}(t) = -\mathbf{Q}(t)\pi_s(\mathbf{X}^{(r)}(t)), \tag{2.3}$$

with $\mathbf{Q}(0) = \mathbf{I}$. Then $\mathbf{Q}(t)$ is orthogonal and a calculation using (2.2) yields

$$(\mathbf{Q}(t)\mathbf{X}^{(r)}(t)\mathbf{Q}(t)^T)^{\cdot} = \mathbf{Q}(t)\mathbf{Y}_{r,n}(t)\mathbf{Q}(t)^T. \tag{2.4}$$

The right side of (2.3) is positive semidefinite, so by integration from 0 to a positive t, we obtain

$$\mathbf{Q}(t)\mathbf{X}^{(r)}(t)\mathbf{Q}(t)^T \geqq \mathbf{X}^{(r)}(0).$$

Hence

$$\begin{aligned}\lambda_i(\mathbf{X}^{(r)}(t)) &= \lambda_i(\mathbf{Q}(t)\mathbf{X}^{(r)}(t)\mathbf{Q}(t)^T) \\ &\geqq \lambda_i(\mathbf{X}^{(r)}(0))\end{aligned}$$

for $1 \leqq i \leqq r$. □

As noted in the Introduction, Theorem 2 is an immediate corollary of Theorem 1.

3. Eigenvalue inequalities for the QR-flow. Let $\mathbf{Y}(t) = \mathbf{Y}(t; \mathbf{Y}_0)$ denote a QR-flow orbit with $\mathbf{Y}(0) = \mathbf{Y}_0$. We suppose throughout this section that $\mathbf{Y}_0$ is a positive-definite symmetric matrix. Our object is to prove monotonicity properties of the eigenvalues of $\mathbf{E}_r^T\mathbf{Y}(t)^\alpha\mathbf{E}_r$ (Theorem 3). We shall derive the continuous-time version of the inequalities from the following weaker discrete-time version of the inequalities.

THEOREM 3.1. *Let* $\mathbf{Y}(t)$ *be a* QR-*flow orbit with* $\mathbf{Y}(0)$ *positive-definite symmetric. Denote the eigenvalues of* $\mathbf{E}_r^T\mathbf{Y}(t)^\alpha\mathbf{E}_r$ *in decreasing order by*

$$\lambda_{1,r}(t; \alpha, \mathbf{Y}_0) \geqq \lambda_{2,r}(t; \alpha, \mathbf{Y}_0) \geqq \cdots \geqq \lambda_{r,r}(t; \alpha, \mathbf{Y}(0)).$$

Then for all $\alpha > 0$ *and all* r,

$$\lambda_{i,r}(t+1; \alpha, \mathbf{Y}_0) \geqq \lambda_{i,r}(t; \alpha, \mathbf{Y}_0), \tag{3.1}$$

for $1 \leqq i \leqq r$.

Before proving this result, we show that it implies Theorem 3, using the following lemma that shows that the size of the discrete-time step can be traded off against the exponent α.

LEMMA 3.1. *For positive-definite* $\mathbf{Y}_0$ *and* $\alpha > 0$, $\beta > 0$,

$$\mathbf{Y}(t, \mathbf{Y}_0)^\alpha = \mathbf{Y}\left(\frac{t}{\beta}; \mathbf{Y}_0^\beta\right)^{\alpha/\beta}. \tag{3.2}$$

Proof. Let $\mathbf{Y}_0 = \exp(\mathbf{X}_0)$ with $\mathbf{X}_0$ symmetric. The QR-decomposition

$$\exp(t\mathbf{X}_0) = \mathbf{Q}(t; \mathbf{X}_0)\mathbf{R}(t; \mathbf{X}_0)$$

gives both

$$\exp(\alpha t\mathbf{X}_0) = \mathbf{Q}(\alpha t; \mathbf{X}_0)\mathbf{R}(\alpha t; \mathbf{X}_0),$$

and

$$\exp(\alpha t\mathbf{X}_0) = \mathbf{Q}(t;\alpha\mathbf{X}_0)\mathbf{R}(t;\alpha\mathbf{X}_0).$$

Then the uniqueness of the QR-decomposition for invertible matrices gives

$$\mathbf{Q}(t;\alpha\mathbf{X}_0) = \mathbf{Q}(\alpha t;\mathbf{X}_0). \tag{3.3}$$

The relation between the QR-flow $\mathbf{Y}(t;\mathbf{Y}_0)$ and the associated Toda flow $\mathbf{X}(t;\mathbf{X}_0)$ gives

$$\begin{aligned}\mathbf{Y}(t;\mathbf{Y}_0)^\alpha &= \exp(\alpha\mathbf{X}(t;\mathbf{X}_0)) \\ &= \mathbf{Q}(t;\mathbf{X}_0)^T \exp(\alpha\mathbf{X}_0)\mathbf{Q}(t;\mathbf{X}_0)\end{aligned} \tag{3.4}$$

since

$$\mathbf{X}(t;\mathbf{X}_0) = \mathbf{Q}(t;\mathbf{X}_0)^T\mathbf{X}_0\mathbf{Q}(t;\mathbf{X}_0).$$

Similarly, we have

$$\begin{aligned}\mathbf{Y}\left(\frac{t}{\beta},\mathbf{Y}_0^\beta\right)^{\alpha/\beta} &= \exp\left(\frac{\alpha}{\beta}\mathbf{X}\left(\frac{t}{\beta};\beta\mathbf{X}_0\right)\right) \\ &= \mathbf{Q}\left(\frac{t}{\beta};\beta\mathbf{X}_0\right)^T \exp(\alpha\mathbf{X}_0)\mathbf{Q}\left(\frac{t}{\beta};\beta\mathbf{X}_0\right).\end{aligned} \tag{3.5}$$

Since (3.3) implies that $\mathbf{Q}(t;\mathbf{X}_0) = \mathbf{Q}(t/\beta;\beta\mathbf{X}_0)$, the lemma follows on comparing (3.4) and (3.5). □

This lemma and Theorem 3.1 easily imply Theorem 3.

Proof of Theorem 3. The lemma implies that

$$\mathbf{Y}(t+1;\mathbf{Y}_0)^\alpha = \mathbf{Y}\left(\frac{t}{\beta}+\frac{1}{\beta},\mathbf{Y}_0^\beta\right)^{\alpha/\beta}$$

so that

$$\lambda_{i,r}(t+1;\alpha,\mathbf{Y}_0) = \lambda_{i,r}\left(\frac{t}{\beta}+\frac{1}{\beta};\frac{\alpha}{\beta},\mathbf{Y}_0^\beta\right)$$

and, similarly,

$$\lambda_{i,r}(t;\alpha,\mathbf{Y}_0) = \lambda_{i,r}\left(\frac{t}{\beta};\frac{\alpha}{\beta},\mathbf{Y}_0^\beta\right).$$

Now, for any fixed β, the map $(\alpha, \mathbf{Y}_0) \to (\alpha/\beta, \mathbf{Y}_0^\beta)$ maps the cone $\mathbb{R}^+ \times P_n^+$ onto itself, where P_n^+ is the cone of positive-definite symmetric matrices. Hence the last two equalities show that Theorem 3.1 holds with the timestep $1/\beta$ instead of 1. Since $\beta > 0$ is arbitrary, we obtain for $t_1 \geqq t_2$ that

$$\lambda_{i,r}(t_1;\alpha,\mathbf{Y}_0) \geqq \lambda_{i,r}(t_2;\alpha,\mathbf{Y}_0)$$

holds for all $\alpha > 0$, $1 \leqq r \leqq n$, $1 \leqq i \leqq r$. □

Theorem 4 follows as an immediate corollary of Theorem 3.

It remains to establish Theorem 3.1. The main ideas are a relation between $\mathbf{E}_r\mathbf{Y}(t+1)^\alpha\mathbf{E}_r$ and $\mathbf{E}_r\mathbf{Y}(t)^\alpha\mathbf{E}_r$ arising from the QR-algorithm (Lemma 3.2 below) and matrix inequalities arising from Loewner's theory of operator convexity. General background on operator convexity is available in [D1], [Do], [Lo], and related results in [BS], [D2], [Kr].

LEMMA 3.2. *Let* $\mathbf{Y}(t)$ *denote the* QR-*flow with* $\mathbf{Y}(0)$ *positive-definite symmetric. Consider the factorization* $\mathbf{Y}(t) = \mathbf{Q}^T\mathbf{D}\mathbf{Q}$ *with* $\mathbf{Q}$ *orthogonal and* $\mathbf{D}$ *diagonal with positive entries, and set* $\mathbf{V} = \mathbf{Q}\mathbf{E}_r$. *Then for any* $\alpha > 0$,

$$\mathbf{E}_r^T\mathbf{Y}(t)^\alpha\mathbf{E}_r = \mathbf{V}^T\mathbf{D}^\alpha\mathbf{V}, \tag{3.6}$$

and $\mathbf{E}_r^T\mathbf{Y}(t+1)^\alpha\mathbf{E}_r$ *is similar to the matrix*

$$\hat{\mathbf{Y}} = (\mathbf{V}^T\mathbf{D}^{-2}\mathbf{V})\mathbf{V}^T\mathbf{D}^{\alpha+2}\mathbf{V}. \tag{3.7}$$

Proof. For any positive-definite $\mathbf{M}$ and orthogonal matrix $\mathbf{Q}$, and any $\alpha > 0$,

$$(\mathbf{Q}^T\mathbf{M}\mathbf{Q})^\alpha = \mathbf{Q}^T\mathbf{M}^\alpha\mathbf{Q}, \tag{3.8}$$

by definition of the matrix power operation. Taking $\mathbf{M} = \mathbf{D}$ and multiplying by $\mathbf{E}_r^T$, $\mathbf{E}_r$ on the left and the right gives (3.6).

To prove the second part of the lemma, recall that the QR-algorithm gives

$$\mathbf{Y}(t+1) = \mathbf{Q}(t)^T\mathbf{Y}(t)\mathbf{Q}(t),$$

where $\mathbf{Y}(t) = \mathbf{Q}(t)\mathbf{R}(t)$ is the QR-decomposition of $\mathbf{Y}(t)$. Together with (3.8), this gives

$$\begin{aligned}\mathbf{E}_r^T\mathbf{Y}(t+1)^\alpha\mathbf{E}_r &= (\mathbf{Q}(t)^T\mathbf{Y}(t)\mathbf{Q}(t))^\alpha\mathbf{E}_r \\ &= \mathbf{E}_r\mathbf{Q}(t)^T\mathbf{Y}(t)^\alpha\mathbf{Q}(t)\mathbf{E}_r \\ &= \mathbf{E}_r\mathbf{Q}(t)^T\mathbf{Q}^T\mathbf{D}^\alpha\mathbf{Q}\mathbf{Q}(t)\mathbf{E}_r.\end{aligned} \tag{3.9}$$

Now we have

$$\begin{aligned}\mathbf{Q}(t) &= \mathbf{Y}(t)\mathbf{R}(t)^{-1} \\ &= \mathbf{Q}^T\mathbf{D}\mathbf{Q}\mathbf{R}(t)^{-1}.\end{aligned}$$

Substituting this into (3.9) and using $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}_n$ yields

$$\mathbf{E}_r^T\mathbf{Y}(t+1)^\alpha\mathbf{E}_r = \mathbf{E}_r^T(\mathbf{R}(t)^{-1})^T\mathbf{Q}^T\mathbf{D}^{\alpha+2}\mathbf{Q}\mathbf{R}(t)^{-1}\mathbf{E}_r. \tag{3.10}$$

Now we use a key property of $\mathbf{E}_r$: any upper triangular matrix $\mathbf{R}$ satisfies

$$\mathbf{R}\mathbf{E}_r = \mathbf{E}_r\mathbf{E}_r^T\mathbf{R}\mathbf{E}_r. \tag{3.11}$$

We apply this in (3.10), with $\hat{\mathbf{R}} = \mathbf{E}_r^T\mathbf{R}(t)^{-1}\mathbf{E}_r$, to get

$$\begin{aligned}\mathbf{E}_r^T\mathbf{Y}(t+1)^\alpha\mathbf{E}_r &= \hat{\mathbf{R}}^T\mathbf{E}_r^T\mathbf{Q}^T\mathbf{D}^{\alpha+2}\mathbf{Q}\mathbf{E}_r\hat{\mathbf{R}} \\ &= \hat{\mathbf{R}}^{-1}(\hat{\mathbf{R}}\hat{\mathbf{R}}^T\mathbf{V}^T\mathbf{D}^{\alpha+2}\mathbf{V})\hat{\mathbf{R}}.\end{aligned} \tag{3.12}$$

Note here that since $\mathbf{R}(t)^{-1}$ is upper triangular with nonzero diagonal entries, so is $\hat{\mathbf{R}}$, hence $\hat{\mathbf{R}}^{-1}$ exists.

We claim that

$$\hat{\mathbf{R}}\hat{\mathbf{R}}^T = \mathbf{V}^T\mathbf{D}^{-2}\mathbf{V}. \tag{3.13}$$

Indeed we have

$$\begin{aligned}\hat{\mathbf{R}}\hat{\mathbf{R}}^T &= \mathbf{E}_r^T\mathbf{R}(t)^{-1}\mathbf{E}_r\mathbf{E}_r^T(\mathbf{R}(t)^{-1})^T\mathbf{E}_r \\ &= \mathbf{E}_r^T\mathbf{R}(t)^{-1}(\mathbf{R}(t)^{-1})^T\mathbf{E}_r \qquad \text{(by (3.11))} \\ &= \mathbf{E}_r^T\mathbf{R}(t)^{-1}\mathbf{Q}(t)^T\mathbf{Q}(t)(\mathbf{R}(t)^{-1})^T\mathbf{E}_r \\ &= \mathbf{E}_r^T\mathbf{Y}(t)^{-1}(\mathbf{Y}(t)^{-1})^T\mathbf{E}_r \\ &= \mathbf{E}_r^T\mathbf{Q}^T\mathbf{D}^{-2}\mathbf{Q}\mathbf{E}_r \\ &= \mathbf{V}^T\mathbf{D}^{-2}\mathbf{V},\end{aligned}$$

proving the claim.

The second part of the lemma follows on substituting (3.13) into (3.12). □

Theorem 3.1 will follow from the following eigenvalue inequalities.

LEMMA 3.3. *Let* $\mathbf{D}$ *be an* $n \times n$ *diagonal matrix with positive entries, and let* $\mathbf{V}$ *be an* $n \times r$ *partial isometry, i.e.,* $\mathbf{V}^T\mathbf{V} = I_r$. *Then for all* $\alpha > 0$ *and* $1 \leqq i \leqq r$,

$$\lambda_i((\mathbf{V}^T\mathbf{D}^{-2}\mathbf{V})\mathbf{V}^T\mathbf{D}^{\alpha+2}\mathbf{V}) \geqq \lambda_i(\mathbf{V}^T\mathbf{D}^{\alpha}\mathbf{V}). \tag{3.14}$$

Proof. We use inequalities from the theory of operator-convex functions developed by Loewner. For $n \times n$ symmetric matrices $\mathbf{M}_1, \mathbf{M}_2$ write $\mathbf{M}_1 \geqq \mathbf{M}_2$ to mean that $\mathbf{M}_1 - \mathbf{M}_2$ is positive semidefinite. It is easy to see that $\mathbf{M}_1 \geqq \mathbf{M}_2$ implies that

$$\mathbf{W}^T\mathbf{M}_1\mathbf{W} \geqq \mathbf{W}^T\mathbf{M}_2\mathbf{W}, \tag{3.15}$$

for any rectangular matrix $\mathbf{W}$. The operator convexity inequalities we need are given in the following proposition.

PROPOSITION 3.1. *Let* $\mathbf{M}$ *be an* $n \times n$ *positive-definite symmetric matrix, and let* $\mathbf{V}$ *be an* $n \times r$ *partial isometry, i.e.,* $\mathbf{V}^T\mathbf{V} = \mathbf{I}_r$. *Then*

(1) *For* $1 \leqq \alpha < \infty$,

$$\mathbf{V}^T\mathbf{M}\mathbf{V} \geqq (\mathbf{V}^T\mathbf{M}^{-\alpha}\mathbf{V})^{-1/\alpha}. \tag{3.16}$$

(2) *For* $1 \leqq \alpha \leqq \infty$,

$$(\mathbf{V}^T\mathbf{M}^{\alpha}\mathbf{V})^{1/\alpha} \geqq \mathbf{V}^T\mathbf{M}\mathbf{V}. \tag{3.17}$$

(3) *For* $1 \leqq \alpha \leqq 2$,

$$(\mathbf{V}^T\mathbf{M}\mathbf{V})^{\alpha} \geqq \mathbf{V}^T\mathbf{M}^{\alpha}\mathbf{V}. \tag{3.18}$$

Proof. This is Corollary 4.2 in Ando [A], on taking $\Phi(\mathbf{M}) = \mathbf{V}^T\mathbf{M}\mathbf{V}$. □

To continue proving Lemma 3.3, let $\mathbf{A}_1 \approx \mathbf{A}_2$ mean that $\mathbf{A}_1$ is similar to $\mathbf{A}_2$. Then

$$(\mathbf{V}^T\mathbf{D}^{-2}\mathbf{V})\mathbf{V}^T\mathbf{D}^{\alpha+2}\mathbf{V} \approx (\mathbf{V}^T\mathbf{D}^{\alpha+2}\mathbf{V})^{1/2}(\mathbf{V}^T\mathbf{D}^{-2}\mathbf{V})(\mathbf{V}^T\mathbf{D}^{\alpha+2}\mathbf{V})^{1/2}. \tag{3.19}$$

Now Proposition 3.1 shows that, for $\alpha > 0$,

$$\begin{aligned}(\mathbf{V}^T\mathbf{D}^{\alpha+2}\mathbf{V})^{1/2}\mathbf{V}^T\mathbf{D}^{-2}\mathbf{V}(\mathbf{V}^T\mathbf{D}^{\alpha+2}\mathbf{V})^{1/2} &\geqq (\mathbf{V}^T\mathbf{D}^{\alpha+2}\mathbf{V})^{1/2}(\mathbf{V}^T\mathbf{D}^{\alpha+2}\mathbf{V})^{-2/(\alpha+2)} \\ &\quad \times (\mathbf{V}^T\mathbf{D}^{\alpha+2}\mathbf{V})^{1/2} \\ &= (\mathbf{V}^T\mathbf{D}^{\alpha+2}\mathbf{V})^{\alpha/(\alpha+2)} \geqq \mathbf{V}^T\mathbf{D}^{\alpha}\mathbf{V},\end{aligned} \tag{3.20}$$

where (1) was used in the first line, (2) in the second. Now recall (from (2.1)) that $\mathbf{M}_1 \geqq \mathbf{M}_2$ implies that

$$\lambda_i(\mathbf{M}_1) \geqq \lambda_i(\mathbf{M}_2), \qquad 1 \leqq i \leqq n.$$

Since (3.19) preserves eigenvalues and (3.20) decreases them, we have

$$\lambda_i((\mathbf{V}^T\mathbf{D}^{-2}\mathbf{V})\mathbf{V}^T\mathbf{D}^{\alpha+2}\mathbf{V}) \geqq \lambda_i(\mathbf{V}^T\mathbf{D}^{\alpha}\mathbf{V}), \qquad 1 \leqq i \leqq r. \qquad \square$$

Proof of Theorem 3.1. Since similarity preserves eigenvalues, Lemmas 3.2 and 3.3 give for $\alpha > 0$ that

$$\lambda_i(\mathbf{E}_r^T\mathbf{Y}(t+1)^{\alpha}\mathbf{E}_r) \geqq \lambda_i(\mathbf{E}_r^T\mathbf{Y}(t)^{\alpha}\mathbf{E}_r), \qquad 1 \leqq i \leqq r,$$

as required. □

4. Trace inequalities for generalized Toda flows. Our object is to prove Theorem 5, which asserts that for any nondecreasing real-valued function, the funtion

$$\mathrm{T}(t) = \mathrm{tr}\,(\mathbf{E}_r^T f(\mathbf{X}(t))\mathbf{E}_r)$$

is a nondecreasing function of t, when $\mathbf{X}(t)$ evolves according to the Toda flow.

We start with a general matrix inequality due to Mallows, which contains the crux of the proof.

LEMMA 4.1 (Mallows). *Let* $\mathbf{V}$ *be an* $n \times r$ *matrix which is a partial isometry, i.e.,* $\mathbf{V}^T\mathbf{V} = I_r$, *and suppose that* $\mathbf{A}$ *and* $\mathbf{B}$ *are real diagonal matrices such that*

$$\mathbf{A}_{ii} \geqq \mathbf{A}_{jj} \Leftrightarrow \mathbf{B}_{ii} \geqq \mathbf{B}_{jj}. \tag{4.1}$$

Then

$$\operatorname{tr}(\mathbf{V}^T\mathbf{A}\mathbf{B}\mathbf{V}) \geqq \operatorname{tr}((\mathbf{V}^T\mathbf{A}\mathbf{V})(\mathbf{V}^T\mathbf{B}\mathbf{V})). \tag{4.2}$$

Proof. Let $a_i = \mathbf{A}_{ii}$, $b_i = \mathbf{B}_{ii}$. Then

$$\sigma = \sum_{i=1}^{n}\sum_{j=1}^{n}(a_i - a_j)(b_i - b_j)\left(\sum_{p=1}^{r} v_{ip}v_{jp}\right)^2 \geqq 0$$

follows by hypothesis (4.1). Now

$$\sigma = \sum_{i,j=1}^{n}\sum_{p,q=1}^{r} v_{ip}v_{jp}v_{iq}v_{jq}(a_i - a_j)(b_i - b_j).$$

Since $\mathbf{V}$ is a partial isometry, then

$$\sum_{j=1}^{n} v_{jp}v_{jq} = \delta_{pq},$$

where $\delta_{pq} = 1$ if $p = q$ and 0 otherwise; whence

$$\begin{aligned}\sigma &= 2\sum_{i=1}^{n}\sum_{p,q=1}^{r} a_i b_i v_{ip}v_{iq}\delta_{pq} - 2\sum_{i,j=1}^{n}\sum_{p,q=1}^{r} a_i b_i v_{ip}v_{iq}v_{jp}v_{jq} \\ &= 2\operatorname{tr}(\mathbf{V}^T\mathbf{A}\mathbf{B}\mathbf{V}) - 2\operatorname{tr}((\mathbf{V}^T\mathbf{A}\mathbf{V})(\mathbf{V}^T\mathbf{B}\mathbf{V})).\end{aligned}$$

□

Proof of Theorem 5. The function $\mathbf{Z}(t) = f(\mathbf{X}(t))$ inherits the smoothness properties of $\mathbf{X}(t)$ independent of the nature of the function f, which may in fact be nondifferentiable or discontinuous. To see this, note that $\mathbf{X}(t) = \mathbf{O}(t)^T\mathbf{D}\mathbf{O}(t)$, where $\mathbf{O}(t)$ is orthogonal, $\mathbf{D}$ is a constant diagonal matrix since the Toda flow is isospectral, and $\mathbf{O}(t)$ is a C^∞-function of t (in fact, real-analytic). By definition, $f(\mathbf{X}(t)) = \mathbf{O}(t)^T f(\mathbf{D})\mathbf{O}(t)$, whence $\mathbf{Z}(t) = f(\mathbf{X}(t))$ is a smooth function of t. It evolves according to the differential equation

$$\dot{\mathbf{Z}}(t) = [\mathbf{Z}(t), \pi_s(\mathbf{X}(t))].$$

Consequently,

$$\mathrm{T}(t) = \operatorname{tr}(\dot{\mathbf{E}}_r f(\mathbf{X}(t))\mathbf{E}_r)$$

evolves according to the differential equation

$$\dot{\mathrm{T}}(t) = \operatorname{tr}(\mathbf{E}_r^T[f(\mathbf{X}(t)), \pi_s(\mathbf{X}(t))]\mathbf{E}_r). \tag{4.3}$$

The assertion of the theorem is that $\dot{\mathrm{T}}(t) \geqq 0$ for all t. This is a consequence of the following lemma.

LEMMA 4.2. *For any nondecreasing function* f, *any symmetric* $n \times n$ *matrix* $\mathbf{X}$, *and* $1 \leqq r \leqq n$,

$$\operatorname{Tr}(\mathbf{E}_r^T[f(\mathbf{X}), \pi_s(\mathbf{X})]\mathbf{E}_r) \geqq 0. \tag{4.4}$$

Proof. Represent $\mathbf{X}$ and $f(\mathbf{X})$ in block form:

$$\mathbf{X} = \begin{bmatrix}\mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C}\end{bmatrix}, \qquad f(\mathbf{X}) = \begin{bmatrix}\mathbf{D} & \mathbf{E} \\ \mathbf{E}^T & \mathbf{F}\end{bmatrix}$$

where $\mathbf{A} = \mathbf{E}_r^T \mathbf{X} \mathbf{E}_r$ and $\mathbf{D} = \mathbf{E}_r^T f(\mathbf{X}) \mathbf{E}_r$ are $r \times r$ submatrices. A calculation shows that

$$\operatorname{tr}(\mathbf{E}_r^T [f(\mathbf{X}), \pi_s(\mathbf{X})] \mathbf{E}_r) = 2 \operatorname{tr}(\mathbf{B}^T \mathbf{E}).$$

Now $\mathbf{X} = \mathbf{O}^T \mathbf{D} \mathbf{O}$ where $\mathbf{O}$ is orthogonal, $\mathbf{D}$ diagonal, and $f(\mathbf{X}) = \mathbf{O}^T f(\mathbf{D}) \mathbf{O}$. Write $\mathbf{O} = [\mathbf{V}, \mathbf{W}]$ in block form, where $\mathbf{V}$ is $n \times r$. Then $\mathbf{V}$ is a partial isometry $\mathbf{I}_r$ and

$$\mathbf{V}\mathbf{V}^T + \mathbf{W}\mathbf{W}^T = \mathbf{I}_n.$$

Also,

$$\mathbf{B} = \mathbf{V}^T \mathbf{D} \mathbf{W}, \qquad \mathbf{E} = \mathbf{V}^T f(\mathbf{D}) \mathbf{W}.$$

Now

$$\begin{aligned} \operatorname{tr}(\mathbf{B}^T \mathbf{E}) &= \operatorname{tr}((\mathbf{W}^T \mathbf{D} \mathbf{V})(\mathbf{V}^T f(\mathbf{D}) \mathbf{W})) \\ &= \operatorname{tr}(\mathbf{W}\mathbf{W}^T \mathbf{D} \mathbf{V} \mathbf{V}^T f(\mathbf{D})) \\ &= \operatorname{tr}((I_n - \mathbf{V}\mathbf{V}^T) \mathbf{D} \mathbf{V} \mathbf{V}^T f(\mathbf{D})) \\ &= \operatorname{tr}(\mathbf{V}^T \mathbf{D} f(\mathbf{D}) \mathbf{V}) - \operatorname{tr}((\mathbf{V}^T \mathbf{D} \mathbf{V})(\mathbf{V}^T f(\mathbf{D}) \mathbf{V})) \\ &\geqq 0, \end{aligned}$$

by Lemma 4.1, on taking $(\mathbf{A}, \mathbf{B}) = (\mathbf{D}, f(\mathbf{D}))$ and observing that (4.1) holds since f is nondecreasing. □

Lemma 4.2 shows that $\dot{\mathrm{T}}(t) \geqq 0$ in (4.3), which completes the proof of Theorem 5. □

Remark. The proof of Lemma 4.2 only required that $(\mathbf{D}, f(\mathbf{D}))$ satisfy condition (4.1), so that if extra restrictions are put on the eigenvalues $\mathbf{D}$, then the conditions on f may be relaxed. For example, if $\mathbf{D}$ is nonnegative, we need only require that $f(t)$ be nondecreasing on $[0, \infty)$. Thus we obtain, for example, that for $n \geqq 1$, we have

$$\operatorname{tr}(\mathbf{E}_r^T [\mathbf{X}^n, \pi_s(\mathbf{X})] \mathbf{E}_r) \geqq 0,$$

if n is odd and $\mathbf{X}$ is symmetric, and if n is even and $\mathbf{X}$ is a positive-definite symmetric matrix.

5. Subspace iteration and Ritz value inequalities. The *simple subspace iteration algorithm* [P, p. 290] takes as input a symmetric matrix $\mathbf{X}$ and an $n \times r$ partial isometry $\mathbf{V}_0$, which is an orthonormal basis of $\mathscr{S}_0$ and produces at step j an $n \times r$ partial isometry $\mathbf{V}_j$ which is a basis of $\mathscr{S}_j = \mathbf{X}^j \mathscr{S}_0$. The algorithm uses the fact that any $n \times r$ matrix $\mathbf{M}$ of full column rank r has a unique decomposition

$$\mathbf{M} = \hat{\mathbf{V}} \hat{\mathbf{R}} \tag{5.1}$$

where $\hat{\mathbf{V}}$ is an $n \times r$ partial isometry and $\hat{\mathbf{R}}$ is an upper triangular $r \times r$ matrix with positive-diagonal elements. This decomposition is obtained by Gram–Schmidt orthonormalization of the columns of $\mathbf{M}$. If the decomposition of $\mathbf{X}\mathbf{V}_j$ is $\mathbf{X}\mathbf{V}_j = \tilde{\mathbf{V}}\tilde{\mathbf{R}}$, then the algorithm sets $\mathbf{V}_{j+1} = \tilde{\mathbf{V}}$.

It is well known that simple subspace iteration is a projection of the QR-algorithm (see Watkins [W1, p. 434]). We include a proof for convenience.

THEOREM 5.1. *Let* $\mathbf{X}$ *be a positive-definite symmetric matrix, and* $\mathbf{V}_0$ *an* $n \times r$ *partial isometry. Choose any orthogonal matrix* $\mathbf{Q}_0$ *such that* $\mathbf{V}_0 = \mathbf{Q}_0^T \mathbf{E}_r$, *and let* $\mathbf{Y}(t)$ *denote the* QR-*flow with* $\mathbf{Y}(0) = \mathbf{Q}_0^T \mathbf{X} \mathbf{Q}_0$. *Then the simple subspace iteration iterates* $\mathbf{V}_j$ *satisfy*

$$\mathbf{V}_j = \mathbf{Q}_0 \bar{\mathbf{Q}}(j) \mathbf{E}_r, \qquad j = 0, 1, 2, \cdots \tag{5.2}$$

where $\mathbf{Y}(t) = \bar{\mathbf{Q}}(t)^T \mathbf{Y}(0)\bar{\mathbf{Q}}(t)$. *Consequently, for all* $j \geqq 0$,

$$\mathbf{V}_j^T \mathbf{X} \mathbf{V}_j = \mathbf{E}_r^T \mathbf{Y}(j)\mathbf{E}_r. \tag{5.3}$$

Proof. We prove (5.2) by induction on j. It is true for $j = 0$ by hypothesis. Now for all t,

$$\mathbf{Y}(t) = \mathbf{Q}(t)\mathbf{R}(t),$$

$$\mathbf{Y}(t+1) = \mathbf{R}(t)\mathbf{Q}(t) = \mathbf{Q}(t)^T \mathbf{Y}(t)\mathbf{Q}(t),$$

and since

$$\mathbf{Y}(t) = \bar{\mathbf{Q}}(t)^T \mathbf{Y}(0)\bar{\mathbf{Q}}(t)^T = \bar{\mathbf{Q}}(t)^T \mathbf{Q}_0^T \mathbf{X} \mathbf{Q}_0 \bar{\mathbf{Q}}(t),$$

this implies that

$$\bar{\mathbf{Q}}(j+1) = \bar{\mathbf{Q}}(j)\mathbf{Q}(j).$$

By the induction hypothesis

$$\begin{aligned}
\mathbf{X}\mathbf{V}_j &= \mathbf{S}\mathbf{Q}_0\bar{\mathbf{Q}}(j)\mathbf{E}_r \\
&= \mathbf{Q}_0\bar{\mathbf{Q}}(j)\mathbf{Y}(j)\mathbf{E}_r \\
&= \mathbf{Q}_0\bar{\mathbf{Q}}(j)\mathbf{Q}(j)\mathbf{R}(j)\mathbf{E}_r \\
&= \mathbf{Q}_0\bar{\mathbf{Q}}(j+1)\mathbf{R}(j)\mathbf{E}_r \\
&= (\mathbf{Q}_0\bar{\mathbf{Q}}(j+1)\mathbf{E}_r)(\mathbf{E}_r^T\mathbf{R}(j)\mathbf{E}_r) \quad \text{by (3.11)} \\
&= \hat{\mathbf{V}}\hat{\mathbf{R}}.
\end{aligned}$$

Since the decomposition $\hat{\mathbf{V}}\hat{\mathbf{R}}$ is unique, we have

$$\mathbf{V}_{j+1} = \hat{\mathbf{V}} = \mathbf{Q}_0\bar{\mathbf{Q}}(j+1)\mathbf{E}_r,$$

and the induction step is completed.

Now (5.3) follows directly on substituting (5.2) for $\mathbf{V}_j$. □

The monotonicity of Ritz values (1.1) follows directly from this result and Theorem 4.

COROLLARY 5.1a. *For any symmetric matrix* $\mathbf{X}$ *and subspace* $\mathcal{S}_0$, *if* $\mathcal{S}_1 = \mathbf{X}\mathcal{S}_0$, *then*

$$\lambda_i(\mathbf{X}, \mathcal{S}_1) \geqq \lambda_i(\mathbf{X}, \mathcal{S}_0), \qquad 1 \leqq i \leqq \dim(\mathcal{S}_0).$$

Hastie [H] suggested using simple subspace iteration as an algorithm to find a best rank r approximation $\hat{\mathbf{X}}$ to an $n \times n$ symmetric matrix $\mathbf{X}$, in the sense of minimizing the Frobenius norm $\|\mathbf{X} - \hat{\mathbf{X}}\|^2$ over all rank r matrices. (Here $\|\mathbf{X}\|^2 = \sum_{i,j} \mathbf{X}_{ij}^2$.) It is known [EY], [H] that a best rank r approximation is given by

$$\hat{\mathbf{X}} = \mathbf{X}\mathbf{V}\mathbf{V}^T, \tag{5.4}$$

where $\mathbf{V}$ is any $n \times r$ partial isometry whose rows consist of r left-eigenvectors of $\mathbf{X}$ corresponding to its r largest eigenvalues. $\hat{\mathbf{X}}$ need not be symmetric, and a best symmetric rank r approximation $\hat{\hat{\mathbf{X}}}$ to $\mathbf{X}$ is given by

$$\hat{\hat{\mathbf{X}}} = \mathbf{V}\mathbf{V}^T\mathbf{X}\mathbf{V}\mathbf{V}^T. \tag{5.5}$$

These best approximations are unique if $\lambda_r(\mathbf{X}) \neq \lambda_{r+1}(\mathbf{X})$. Hastie's algorithm uses simple subspace iteration starting with $\mathbf{V}_0$ and produces the rank r approximations $\hat{\mathbf{X}}_j$, $\hat{\hat{\mathbf{X}}}_j$ using (5.4) and (5.5), respectively, with $\mathbf{V} = \mathbf{V}_j$.

THEOREM 5.2. *The rank r approximations* $\hat{\mathbf{X}}_j$, $\hat{\hat{\mathbf{X}}}_j$ *to* $\mathbf{X}$, *produced using simple subspace iteration starting from any subspace* $\mathcal{S}_0$, *satisfy*

$$\|\hat{\mathbf{X}}_{j+1}-\mathbf{X}\|^2 \leqq \|\hat{\mathbf{X}}_j-\mathbf{X}\|^2, \tag{5.6}$$

$$\|\hat{\hat{\mathbf{X}}}_{j+1}-\mathbf{X}\|^2 \leqq \|\hat{\hat{\mathbf{X}}}_j-\mathbf{X}\|^2, \tag{5.7}$$

where $\|\cdot\|$ *is the Frobenius norm.*

Proof. Using Theorems 5.1 and 4 we have

$$\begin{aligned}\|\mathbf{X}-\hat{\mathbf{X}}_j\|^2 &= \operatorname{tr}((\mathbf{X}-\mathbf{X}\mathbf{V}_j\mathbf{V}_j^T)^T(\mathbf{X}-\mathbf{X}\mathbf{V}_j\mathbf{V}_j^T))\\ &= \operatorname{tr}(\mathbf{X}^T\mathbf{X})-\operatorname{tr}(\mathbf{V}_j^T\mathbf{X}^2\mathbf{V}_j)\\ &= \operatorname{tr}(\mathbf{X}^T\mathbf{X})-\operatorname{tr}(\mathbf{E}_r^T\mathbf{Y}(j)^2\mathbf{E}_r).\end{aligned}$$

By Theorem 4, $\operatorname{tr}(\mathbf{E}_r^T\mathbf{Y}(t)^2\mathbf{E}_r)$ is nondecreasing, hence (5.6) follows. Next

$$\begin{aligned}\|\mathbf{X}-\hat{\hat{\mathbf{X}}}\|^2 &= \operatorname{tr}((\mathbf{X}-\mathbf{V}_j\mathbf{V}_j^T\mathbf{S}\mathbf{V}_j\mathbf{V}_j^T)^T(\mathbf{X}-\mathbf{V}_j\mathbf{V}_j^T\mathbf{S}\mathbf{V}_j\mathbf{V}_j^T))\\ &= \operatorname{tr}(\mathbf{X}^T\mathbf{X})-\operatorname{tr}((\mathbf{V}_j^T\mathbf{X}\mathbf{V}_j)^2)\\ &= \operatorname{tr}(\mathbf{X}^T\mathbf{X})-\operatorname{tr}((\mathbf{E}_r^T\mathbf{Y}(j)\mathbf{E}_r)^2).\end{aligned}$$

By Theorem 4, $\operatorname{tr}((\mathbf{E}_r^T\mathbf{Y}(t)\mathbf{E}_r)^2)$ is nondecreasing, so that (5.7) follows. □

Finally, note that the relation of Hastie's algorithm to the QR-flow given in Theorem 5.1 shows that the iterates $\{\hat{\mathbf{X}}_j\}$ (respectively, $\{\hat{\hat{\mathbf{X}}}_j\}$) converge to a best rank r approximation $\hat{\mathbf{X}}$ (respectively, best symmetric rank r approximation $\hat{\hat{\mathbf{X}}}$) of $\mathbf{X}$ whenever $\mathbf{Y}(0)$ is a "generic" matrix, e.g., whenever all entries of $\mathbf{Y}(0)$ are nonzero.

Note added in proof. P. Deift, S. Rivera, C. Tomei, and D. Watkins have found an elegant proof of the Monotonicity Conjectures, which appears in this journal [DRTW].

Acknowledgment. I am indebted to Trevor Hastie for introducing me to the problems in this paper and to Colin Mallows for the essential idea used to prove Theorem 5. David Watkins pointed out the connection of my original results to subspace iteration, and Percy Deift supplied a simplification of the proof of Theorem 1.

REFERENCES

[A] T. ANDO, *Concavity of certain maps on positive definite matrices and applications to Hadamard products*, Linear Algebra Appl., 26 (1979), pp. 203–241.

[Be] R. BELLMAN, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.

[BS] J. BENDAT AND S. SHERMAN, *Monotone and convex operator functions*, Trans. Amer. Math. Soc., 79 (1955), pp. 58–71.

[By] R. BYERS, *A Hamiltonian* QR*-algorithm*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 212–229.

[C1] M. T. CHU, *The generalized Toda lattice, the* QR*-algorithm, and the centre manifold theory*, SIAM J. Algebraic Discrete Methods, 5 (1984), pp. 187–201.

[C2] ———, *On the global convergence of the Toda lattice for real normal matrices and its application to eigenvalue problems*, SIAM J. Math. Anal., 15 (1984), pp. 187–204.

[C3] ———, *Asymptotic analysis of the Toda lattice on diagonalizable matrices*, Nonlinear Anal., Theory, Methods, Appl., 9 (1985), pp. 193–201.

[D1] C. DAVIS, *A Schwarz inequality for convex operator functions*, Proc. Amer. Math. Soc., 8 (1957), pp. 42–44.

[D2] ———, *Notions generalizing convexity for functions defined on spaces of matrices*, in Convexity, V. Klee, ed., Proc. Symposium on Pure Mathematics No. 7, American Mathematical Society, Providence, RI, 1963, pp. 187–201.

[DLNT] P. DEIFT, L. C. LI, T. NANDA, AND C. TOMEI, *The Toda flow on a generic orbit is integrable*, Comm. Pure Appl. Math., 39 (1986), pp. 183–232.

[DLT] P. DEIFT, L. C. LI, AND C. TOMEI, *Matrix factorizations and integrable systems*, Comm. Pure Appl. Math., 42 (1989), pp. 443–521.

[DNT] P. DEIFT, T. NANDA, AND C. TOMEI, *Differential equations for the symmetric eigenvalue problem*, SIAM J. Numer. Anal., 20 (1983), pp. 1–22.

[DRTW] P. DEIFT, S. RIVERA, C. TOMEI, AND D. WATKINS, *A monotonicity property for Toda-type flows*, SIAM J. Matrix Anal. Appl., this issue, pp. 463–468.

[Do] W. DONOGHUE, *Monotone Matrix Functions and Analytic Continuation*, Springer-Verlag, Berlin, 1974.

[EY] C. ECKHART AND G. YOUNG, *The approximation of one matrix by another of lower rank*, Psychometrika, 1 (1936), pp. 211–218.

[F1] H. FLASCHKA, *On the Toda lattice* I, Phys. Rev. B, 9 (1974), pp. 1924–1925.

[F2] ———, *On the Toda lattice* II, Progr. Theoret. Phys., 51 (1974), pp. 703–716.

[Fr] J. FRANCIS, *The* QR*-transformation* I, II, Comput. J., 4 (1961), pp. 265–271; 5 (1962), pp. 332–345.

[G] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1977.

[H] T. HASTIE, *Pseudo-smoothers and Additive Model Approximations*, J. Roy. Stat. Soc., Ser. B, to appear.

[KM] M. KAC AND P. VON MOERBEKE, *On an explicitly solvable system of differential equations related to certain Toda lattices*, Adv. in Math., 16 (1975), pp. 160–169.

[Ko] B. KONSTANT, *The solution to a generalized Toda lattice and representation theory*, Adv. in Math., 34 (1979), pp. 195–338.

[Kr] F. KRAUS, *Über Konvexe Matrix funktionen*, Math. Z., 41 (1936), pp. 18–42.

[L] P. LAX, *Integrals of nonlinear equations of evolution and solitary waves*, Comm. Pure Appl. Math., 21 (1968), pp. 467–490.

[Lo] K. LOEWNER, *Über monotone Matrix funktionen*, Math. Z., 38 (1934), pp. 177–216.

[M] J. MOSER, *Finitely many mass points on a line under the influence of an exponential potential—an integrable system*, Lecture Notes in Physics 38, J. Moser, ed., Springer-Verlag, New York, 1975, pp. 467–497.

[MO] A. W. MARSHALL AND I. OLKIN, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.

[P] B. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[R] H. RUTISHAUSER, *Computational aspects of F. L. Bauer's simultaneous iteration method*, Numer. Math., 13 (1966), pp. 4–13.

[SV] M. SHUB AND T. VASQUEZ, *Some linearly induced Morse–Smale systems, the* QR*-algorithm and the Toda lattice*, in The Legacy of Sonya Kovalevskaya, L. Keen, ed., Contemporary Math. 64, American Mathematical Society, Providence, RI, 1987, pp. 181–194.

[Sy] W. W. SYMES, *The* QR*-algorithm and scattering for the finite nonperiodic Toda lattice*, Phys. D, 4 (1982), pp. 275–280.

[T] M. TODA, *Wave propagation in anharmonic lattices*, J. Phys. Soc. Japan, 23 (1967), pp. 501–506.

[W1] D. S. WATKINS, *Understanding the* QR*-algorithm*, SIAM Rev., 24 (1982), pp. 427–440.

[W2] ———, *Isospectral flows*, SIAM Rev., 26 (1984), pp. 379–391.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 463–468, July 1991

005

A MONOTONICITY PROPERTY FOR TODA-TYPE FLOWS*

P. A. DEIFT†, S. RIVERA†, C. TOMEI‡, AND D. S. WATKINS§

Abstract. If $(X(t))_r$ is the leading $r \times r$ submatrix of a matrix $X(t)$ undergoing a general Toda-type flow $\dot{X} = [X, B(f(X))]$, $B(f(X)) = (f(X))_- - (f(X))_-^T$, with f nondecreasing, two proofs that the eigenvalues of $(X(t))_r$ are nondecreasing in time are given. This property was conjectured by Lagarias in [L].

Key words. monotonicity, Toda flow, eigenvalues

AMS(MOS) subject classification. 15.25

1. Introduction.

The differential equations of Toda-type

$$\begin{aligned} \dot{X} &= [X, B(f(X))] \\ &= XB(f(X)) - B(f(X))X \end{aligned} \tag{1.1}$$

have been the object of substantial research in recent years ([T], [F], [M], [S1], [S2], [DNT], [DLNT], [W1]). Here $X = X(t)$ is a real, $n \times n$ symmetric matrix, f is a function, and $B(M) = M_- - M_-^T = -B(M)^T$, where M_- denotes the strictly lower part of M.

In this paper, we give two proofs of the following conjecture due to Lagarias ([L]).

THEOREM. *Let f be a nondecreasing, real valued function on $\mathbb{R}$, let X be a solution of* (1.1), *and let $(X)_r$ be the $r \times r$ submatrix $E_r^T X E_r$, where*

$$E_r = \begin{bmatrix} I_r \\ 0 \end{bmatrix}.$$

Then, the eigenvalues of $(X)_r$, arranged in nonincreasing order, are nondecreasing functions of $t \in \mathbb{R}$.

In [L], Lagarias proved some special cases of the above statement and applied his results to establish monotone convergence properties of an iterative method (simultaneous iteration—see [Wi, §§ 38, 39] and [H]), to obtain best possible rank r approximations for symmetric matrices.

2. Proofs.

A direct calculation shows that the solution to (1.1) is given by

$$X(t) = Q^T(t) X(0) Q(t), \tag{2.1}$$

where $Q(t)$ is the orthogonal transformation satisfying

$$\begin{aligned} \dot{Q}(t) &= Q(t) B(f(X(t))), \\ Q(0) &= I. \end{aligned} \tag{2.2a}$$

* Received by the editors February 27, 1989; accepted for publication (in revised form) February 27, 1990.

† Courant Institute of Mathematical Science, New York University, 251 Mercer St., New York, New York 10012. The first author's research was supported in part by National Science Foundation grant DMS-880230. The second author's research was supported in part by CONACYT-CIEA, Mexico.

‡ Pontificia Universidade Catolica, Rio de Janeiro, Brazil. This research was performed while the author was visiting the Department of Mathematics, Yale University, New Haven, Connecticut 06520. This author's research was supported in part by Conselho Nacional de Pesquisa, Brazil.

§ Department of Pure and Applied Mathematics, Washington State University, Pullman, Washington 99164.

Furthermore, $Q(t)$ is given explicitly via the factorization [S1]

$$e^{tf(X(0))} = Q(t)R(t), \tag{2.2b}$$

where $R(t)$ is upper triangular with positive diagonal.

We now give the first proof. Let V_r be the span of the first r standard vectors, $e_1, \cdots, e_r$. Then, by min-max, the ith eigenvalue of $(X(t))_r$ in nonincreasing order, $\lambda_1(t) \geqq \lambda_2(t) \geqq \cdots \geqq \lambda_r(t)$, is given by the formula

$$\lambda_i(t) = \max_{A_i \subset V_r} \min_{\substack{u \in A_i \\ u \neq 0}} \frac{(u, X(t)u)}{(u, u)}, \quad \text{where } \dim A_i = i,$$

$$= \max_{A_i \subset V_r} \min_{\substack{u \in A_i \\ u \neq 0}} \frac{(e^{tf(X(0))}(R(t))^{-1}u, X(0)e^{tf(X(0))}(R(t))^{-1}u)}{(e^{tf(X(0))}(R(t))^{-1}u, e^{tf(X(0))}(R(t))^{-1}u)}, \quad \text{by (2.1), (2.2b),}$$

$$= \max_{A_i \subset V_r} \min_{\substack{y \in A_i \\ y \neq 0}} \frac{(e^{tf(X(0))}y, X(0)e^{tf(X(0))}y)}{(e^{tf(X(0))}y, e^{tf(X(0))}y)}$$

as $(R(t))^{-1}$ is a bijection of the i-dimensional subspaces of V_r onto themselves. But

$$\frac{(e^{tf(X(0))}y, X(0)e^{tf(X(0))}y)}{(e^{tf(X(0))}y, e^{tf(X(0))}y)} = \frac{(y, e^{2tf(X(0))}X(0)y)}{(y, e^{2tf(X(0))}y)} \geqq \frac{(y, X(0)y)}{(y, y)},$$

as follows by (simultaneously) diagonalizing $X(0)$ and $f(X(0))$,

$$X(0) = U^T(0) \wedge U(0)$$

$$f(X(0)) = U^T(0)f(\Lambda)U(0),$$

and expanding the inequality

$$\sum_{j,k} ((g(\Lambda))_{jj} - (g(\Lambda))_{kk})((\Lambda)_{jj} - (\Lambda)_{kk})w_j^2 w_k^2 \geqq 0, \qquad U(0) \text{ orthogonal}, \tag{2.3}$$

where w_m is the mth coordinate of $w\,(=U(0)y)$, for the nondecreasing function $g = e^{2tf}$, $t > 0$. This inequality is a special case of a calculation of Mallows, presented in [L].

As

$$\lambda_i(0) = \max_{A_i \subset V_r} \min_{\substack{y \in A_i \\ y \neq 0}} \frac{(y, X(0)y)}{(y, y)},$$

and as the time $t = 0$ is not special, this proves the desired monotonicity.

For the second proof, note that

$$X(t) = U^T(t)\Lambda U(t) \tag{2.4}$$

where $U(t) = U(0)Q(t)$ satisfies the differential equation

$$\dot{U}(t) = U(t)B(f(X(t))). \tag{2.5}$$

From (2.1) and (2.2b), it follows that the solution $X(t)$ of (1.1) is real analytic in t, and in particular, $(X(t))_r$ is a real analytic, real, symmetric-matrix valued function of t. By the well-known result of Rellich (e.g., [K, § II.6]), its eigenvalues $\alpha_i = \alpha_i(t)$, and corresponding orthonormalized eigenvectors $v_i = v_i(t)$, $i = 1, \cdots, r$, can be taken to be real valued, real analytic functions of t on $\mathbb{R}$. In contrast to the $\lambda_i(t)$'s, the $\alpha_i(t)$'s may cross in time.

To prove the theorem it suffices to show that the functions $\alpha_i(t)$ are nondecreasing in t, as $\lambda_i(t)$, ordered as above, are given by

$$\lambda_i(t) = \min_{I_i = \{j_1, \cdots, j_{r-i+1}\} \subset \{1, \cdots, r\}} \max_{j \in I_i} \{\alpha_j(t)\}, \qquad 1 \leqq i \leqq r,$$

and hence will also be nondecreasing in t.

Let $\dot{h}$ denote (dh/dt) and let $(\cdot, \cdot)$ denote the real, Euclidean inner product. We will show that $\dot{\alpha}_i \geqq 0$. From $(X)_r v_i = \alpha_i v_i$, we obtain

$$\begin{aligned}
\dot{\alpha}_i &= (((X)_r)^{\cdot} v_i, v_i), \quad \text{as } (v_i, v_i) = 1, \\
&= ((E_r^T U^T \dot{\Lambda} U E_r)^{\cdot} v_i, v_i) \\
&= 2((E_r^T U^T \Lambda U B(f(X)) E_r) v_i, v_i), \quad \text{by (2.4)}, \\
&= 2(E_r^T U^T (\Lambda - \alpha_i) U B(f(X)) E_r v_i, v_i),
\end{aligned}$$

as $U^T U = I$ and $B(f(X))$ and hence $E_r^T B(f(X)) E_r$ is skew symmetric.

Now, from the definition of $B(M)$,

$$B(f(X)) = f(X) + R,$$

where R is upper triangular; also, by (2.4),

$$Uf(X) = f(\Lambda) U.$$

We then have

$$\frac{\dot{\alpha}_i}{2} = (E_r^T U^T (\Lambda - \alpha_i) f(\Lambda) U E_r v_i, v_i) + (E_r^T U^T (\Lambda - \alpha_i) U R E_r v_i, v_i). \tag{2.6}$$

The second term on the right-hand side is zero. Indeed, from the definition of α_i and v_i, $E_r^T U^T (\Lambda - \alpha_i) U E_r v_i = 0$. But, as R is upper triangular, $R E_r = E_r E_r^T R E_r$, and hence

$$E_r^T R^T U^T (\Lambda - \alpha_i) U E_r v_i = (E_r^T R^T E_r)(E_r^T U^T (\Lambda - \alpha_i) U E_r v_i) = 0.$$

For $w = U E_r v_i$, we have $\alpha_i = (w, \Lambda w)$ and (2.6) becomes

$$\frac{\dot{\alpha}_i}{2} = ((\Lambda - \alpha_i) f(\Lambda) w, w) = (w, \Lambda f(\Lambda) w) - (w, \Lambda w)(w, f(\Lambda) w),$$

which is nonnegative by Mallow's inequality (2.3) (note $(w, w) = 1$). Thus $\dot{\alpha}_i \geqq 0$, as desired.

3. Remarks and applications. (1) The theorem remains true for the eigenvalues of $(g(X))_r$, where $g(\cdot)$ is any nondecreasing function. The proofs are essentially unchanged.

(2) The theorem also shows that the eigenvalues of $F_r^T X(t) F_r$ are nonincreasing, where

$$F_r = \begin{bmatrix} 0 \\ I_r \end{bmatrix}.$$

Indeed, let $P = P^{-1} = P^T$ be the $n \times n$ matrix with ones on the antidiagonal, and zeros elsewhere. It is enough to show that the eigenvalues of $E_r^T P X(t) P E_r$ are nonincreasing. But a simple computation shows that $PB(M)P = -B(PMP)$. Hence $(PXP)^{\cdot} = [PXP, PB(f(X))P] = -[PXP, B(f(PXP))]$ and so $PX(t)P = X(-t, PX_0P)$, the solution of (1.1) with initial data PX_0P. The desired monotonicity now follows immediately.

(3) From the theorem, we know in particular that $\mu_1(t) \geqq \cdots \geqq \mu_{n-1}(t)$, the

eigenvalues of $E_{n-1}^T X(t) E_{n-1}$, are increasing in t. As we will show below, the μ_i's cannot cross the eigenvalues $\lambda_i(t) = \lambda_i$ of $X(t)$, $1 \leqq i \leqq n$. Also, in the case in which f is strictly increasing, a modification of the proof in [M], for example, shows that $X(t)$ converges to a diagonal matrix as $t \to \infty$. Thus, generically, a solution $X(t)$ gives rise to μ_i's that evolve as in Fig. 1.

Let $\sigma(M)$ denote the spectrum of M. Suppose $\sigma(X(0))$ is simple, and let $X(t)$ be the solution of (1.1) for an arbitrary f. To prove the no crossing property, it suffices to show that $\sigma(X(t)) \cap \sigma((X(t))_{n-1})$ is independent of t. To see this, we consider the evolution of $u(t) = U(t)e_n$, where $X(t) = U^T(t)\Lambda U(t)$, as in (2.4). Since $\dot{U} = UB(f(X))$, writing $B(f(X)) = -f(X) + L$, where L is lower triangular, and using $(e_n, Le_n) = (f(X))_{nn} = (f(\Lambda)u, u)$, we obtain

$$\dot{u} = -f(\Lambda)u + (f(\Lambda)u, u)u,$$

which can be solved explicitly ([M]) as

$$u(t) = e^{-tf(\Lambda)}u(0)/\| e^{-tf(\Lambda)}u(0)\|,$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^n$. Now, by Cramer's rule,

$$g(\lambda) \equiv \frac{\det((X(t)_{n-1} - \lambda))}{\det(X(t) - \lambda)} = [(X(t) - \lambda)^{-1}]_{nn}$$

$$= (u(t), (\Lambda - \lambda)^{-1}u(t)) = \sum_{i=1}^{n} \frac{u_i^2(t)}{\lambda_i - \lambda},$$

where $\Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$ and the u_i's are the coordinates of u. The residue $-u_i^2(t)$ at λ_i of the rational function $g(\lambda)$ is zero if and only if $\lambda_i \in \sigma((X(t))_{n-1})$, as $\sigma(X(t)) = \sigma(X(0))$ is simple. But, from the expression for $u(t)$ in terms of $u(0)$ one sees that $u_i(t) = 0$ if and only if $u_i(0) = 0$, and the no crossing property follows. Thus Fig. 1 is established.

We note in passing that the classical interlacing property of the sets $\{\lambda_i\}$ and $\{\mu_j\}$ is itself a consequence of the above observations. Also, knowledge of $\lim_{t\to\infty} X(t)$ can be used to obtain information about the μ_j's. For example, if $X(t) \to \operatorname{diag}(2, 4, 1, 3)$ as $t \to \infty$, then, arguing as above, we conclude that necessarily $\mu_1(t) \in (3, 4]$, $\mu_2(t) \equiv 2$, and $\mu_3(t) \equiv 1$.

(4) In [L], Lagarias considered simultaneous iteration ([Wi, §§ 38, 39], [H]) to obtain rank r approximations of symmetric matrices, as an adaptation of the QR algo-

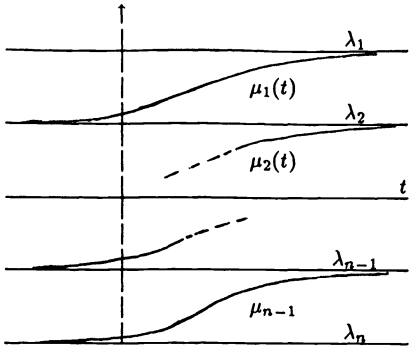


FIG. 1

rithm. By making use of the monotonicity results stated and proved for the QR algorithm (which corresponds to the evaluation at integral times of (1.1) for $f(x) = \ln x$ ([DNT], [S1], [S2])) in [L], he showed that the approximations are monotonically improving.

Lagarias's arguments naturally yield a family of approximating algorithms, one for each (increasing) function f. Start with an $n \times r$ partial isometry V_0 and a symmetric matrix S. Define V_{j+1} recursively to be the partial isometry in the formula $e^{f(S)}V_j = V_{j+1}R$, where R is an upper triangular $r \times r$ matrix with positive diagonal. Then

$$\hat{S}_j \equiv SV_jV_j^T \quad \text{and} \quad \hat{\hat{S}}_j \equiv V_jV_j^TSV_jV_j^T$$

give a sequence of approximations of S of rank at most r, the second being symmetric.

This algorithm has a continuous extension as follows. For an orthogonal matrix Q_0 such that $V_0 = Q_0E_r$, define $Y(0) = Q_0^TSQ_0$. Let $Y(t)$ solve (1.1) for a nondecreasing f, $Y(t) = Q(t)^TY(0)Q(t)$, $Q(0) = I$, as in (2.1), (2.2). Then, following [L], one can prove that $V_j = Q_0Q(j)E_r$, and the inequalities

$$\|S - \hat{S}_{j+1}\| \leqq \|S - \hat{S}_j\|, \qquad \|S - \hat{\hat{S}}_{j+1}\| \leqq \|S - \hat{\hat{S}}_j\|$$

follow by replacing Theorem 4 in [L] by the Theorem in § 1, above.

(5) In [L], Lagarias proved a weaker version of the general monotonicity theorem, viz., if $Y(t)$ is a solution of (1.1) for nondecreasing f, then trace $((Y(t))_r)$ is nondecreasing. The special case $f(x) = x$ of this weaker result was proved in [To] in the following (equivalent) form: trace $(DY(t))$ is nondecreasing for any diagonal matrix D with nonincreasing entries along the diagonal. Moreover, the map $Y \mapsto \text{trace}\,(DY)$, for $d_1 > d_2 > \cdots > d_n$, was shown to be a Morse function for the manifold of real, tridiagonal, symmetric $n \times n$ matrices with fixed simple spectrum. Later, Fried ([Fr]) showed that the map is actually a perfect Morse function, which, together with results in [To], allowed him to compute the cohomology ring of the manifold.

Another application of this monotonicity property gives an amusing proof of the classical Wielandt–Hoffman theorem: for real, $n \times n$ matrices A and B, there is an ordering of their eigenvalues $\{a_i\}$, $\{b_j\}$ such that

$$\sum_{i=1}^{n} (a_i - b_i)^2 \leqq \text{trace}\,((A - B)^2).$$

Indeed, by expanding both sides, it suffices to prove that trace $(AB) \leqq \sum a_ib_i$, and we can suppose that A is diagonal, $A = \text{diag}\,(a_1 \geqq a_2 \geqq \cdots \geqq a_n)$. Now, let $B(t)$ be the solution of (1.1) with, for instance, $f(x) = x$ and $B(0) = B$. Then, by the (weak) monotonicity result,

$$\text{trace}\,(AB(t)) \leqq \text{trace}\,(AB(\infty)) = \sum_{i=1}^{n} a_ib_i,$$

and this proves the result.

Acknowledgments. The authors would like to thank the referee for some helpful comments.

REFERENCES

[DNT] P. DEIFT, T. NANDA, AND C. TOMEI, *Differential equations for the symmetric eigenvalue problem*, SIAM J. Numer. Anal., 20 (1983), pp. 1–22.

[DLNT] P. DEIFT, L. C. LI, T. NANDA, AND C. TOMEI, *The Toda flow on a generic orbit is integrable*, Comm. Pure Appl. Math., 39 (1986), pp. 183–232.

[F] H. FLASCHKA, *The Toda lattice*, I, Phys. Rev. B, 9 (1974), pp. 1924–1925.
[Fr] D. FRIED, *The cohomology of an isospectral flow*, Proc. Amer. Math. Soc., 98 (1986), pp. 363–368.
[H] T. HASTIE, *Pseudo-smoothers and additive model approximations*, preprint.
[K] T. KATO, *Perturbation theory for linear operators*, Springer-Verlag, New York, 1966.
[L] J. LAGARIAS, *Monotonicity properties of the Toda flow, the* QR-*flow and subspace iteration*, SIAM J. Matrix Anal. Appl., this issue, pp. 449–462.
[M] J. MOSER, *Finitely many mass points on a line under the influence of an exponential potential—an integrable system*, Lecture Notes in Physics 38, J. Moser, ed., Springer-Verlag, New York, 1975, pp. 467–497.
[S1] W. W. SYMES, *Hamiltonian group actions and integrable systems*, Phys. D, 1 (1980), pp. 339–374.
[S2] ———, *The* QR-*algorithm and scattering for the finite nonperiodic Toda lattice*, Phys. D, 4 (1982), pp. 275–280.
[T] M. TODA, *Wave propagation in anharmonic lattices*, J. Phys. Soc. Japan, 23 (1967), pp. 501–506.
[To] C. TOMEI, *The topology of isospectral manifolds of tridiagonal matrices*, Duke Math. J., 51 (1984), pp. 981–996.
[W] D. S. WATKINS, *Isospectral flows*, SIAM Rev., 26 (1984), pp. 379–391.
[Wi] J. H. WILKINSON, *The algebraic eigenvalue problem*, Oxford University Press, New York, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 469–483, July 1991

006

STABLE SOLVERS AND BLOCK ELIMINATION FOR BORDERED SYSTEMS*

W. GOVAERTS†

Abstract. Linear systems with a fairly well conditioned matrix M of the form

$$\begin{matrix} \\ n \end{matrix}\begin{pmatrix} A & b \\ c & d \end{pmatrix}\begin{matrix} n \\ 1 \end{matrix},$$

for which a "black-box" solver for A is available, are considered. To solve systems with M, a mixed block elimination algorithm, called BEM, is proposed. It has the following advantages: (1) It is easier to understand and to program than the widely accepted deflated block elimination (DBE) proposed by Chan, yet allows the same broad class of solvers and has comparable accuracy. (2) It requires one less solve with A. (3) It allows a rigorous error analysis that shows why it may fail in exceptional cases (all other black-box methods known to us also fail in these cases).

BEM is also compared to iterative refinement of Crout block elimination (BEC) introduced by Pryce and Govaerts. BEC allows a more restricted class of solvers than BEM but is faster in cases where a solver is given not for A but for a matrix close to A, which is often the case in applications like numerical continuation theory.

Key words. bordered matrix, block elimination, black-box solver

AMS(MOS) subject classification. 65F30

1. Introduction and notation.

Let

$$M = \begin{matrix} \\ n \end{matrix}\begin{pmatrix} A & b \\ c & d \end{pmatrix}\begin{matrix} n \\ 1 \end{matrix}$$

be a bordered matrix. We want to solve

$$M\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}, \tag{1}$$

where x, f are n-vectors and y, g are scalars. In applications like numerical continuation theory, a solver for A is often available because A has special structure (banded, symmetric, sparse, or other). It is then advisable to use this solver to solve systems with M. Difficulties arise when A is nearly singular (in the continuation context this means that we are near a turning point; see Rheinboldt [13]).

Various authors (Keller [7], Moore [8]) solved bordered singular systems by altering A or the elimination strategy. Björck [2] suggests rescaling the last row of M in such a way that Gaussian elimination (further denoted by GE) with row interchanges on M does not pivot to the last row. The problem is then that of GE with a badly scaled matrix and Skeel [14] has shown that in most practical cases one iterative refinement leads to a stable algorithm. This is close (but not equivalent) to BEC + 1 (BEC is to be discussed further; see also Govaerts and Pryce [5]). We concentrate, however, on the case where a solver for A is given as a "black box," which in practice is often the case. The spirit is therefore that of Chan and Resasco [3], [4].

* Received by the editors February 6, 1989; accepted for publication (in revised form) March 1, 1990.

† Senior Research Associate of the Belgian National Fund of Scientific Research (N.F.W.O.); Department of Mathematics, University of Ghent, Krijgslaan 281, B-9000 Ghent, Belgium (govaerts@mathanal.rug.ac.be).

To be precise, we assume that a solver S for A is available, i.e., a map $S: R^n \to R^n$ such that $S(r)$ is an approximate solution to $As = r$.

S is called stable if, when it is applied in floating-point arithmetic of unit roundoff u, there exist a modest constant C_S, a matrix ΔA, and a vector Δr such that

$$(A+\Delta A)S(r) = (r+\Delta r),$$
$$\|\Delta A\| \leqq C_S u \|A\|, \qquad \|\Delta r\| \leqq C_S u \|r\|,$$

where $\|\ \|$ is the 2-norm and C_S will be called the stability constant of S.

Block elimination (BE) is a method to solve (1) by decomposing M blockwise. One way is to use the Crout factorization

$$\begin{pmatrix} A & b \\ c & d \end{pmatrix} = \begin{pmatrix} A & 0 \\ c & \delta \end{pmatrix} \begin{pmatrix} I & v \\ 0 & 1 \end{pmatrix} \tag{2}$$

followed by the solution of two block triangular systems.

This leads to the following algorithm.

ALGORITHM BEC.
1. Solve $Av = b$
2. Compute $\delta = d - cv$
3. Solve $Aw = f$
4. Compute $y = (g - cw)/\delta$
5. Compute $x = w - vy$

Another way is to use the Doolittle factorization

$$\begin{pmatrix} A & b \\ c & d \end{pmatrix} = \begin{pmatrix} I & 0 \\ \xi & 1 \end{pmatrix} \begin{pmatrix} A & b \\ 0 & \delta_1 \end{pmatrix}, \tag{3}$$

again followed by two solutions of block triangular systems.

This amounts to the following.

ALGORITHM BED.
1. Solve $\xi A = c$
2. Compute $\delta_1 = d - \xi b$
3. Compute $y = (g - \xi f)/\delta_1$
4. Solve $Ax = f - by$

Both algorithms provide perfectly satisfying answers if M, A are both well conditioned and the solver for A (and in BED, for A^T) is stable. If A is less well conditioned then it is generally a good idea to improve the obtained result by iterative refinement. If Alg is any algorithm that produces x_1, y_1 out of f, g, we define Alg + k($k = 0, 1, 2, \cdots$) as follows.

ALGORITHM Alg+k.
1. Compute x_1, y_1 out of f, g using Alg
2. For $i = 1, 2, \cdots, k$ do steps 3 to 5
3. Compute the residuals $f_1 = f - Ax_1 - by_1$ and $g_1 = g - cx_1 - dy_1$
4. Compute x_2, y_2 out of f_1, g_1 using Alg
5. Compute $x_1 = x_1 + x_2$, $y_1 = y_1 + y_2$

On first thinking, we might expect that:

(i) BEC and BED have roughly the same behaviour (in many treatments of Gaussian elimination, the difference between the Crout and Doolittle decompositions is hardly noticed).

(ii) If M is well conditioned and A tends to singularity, more and more iterations of BEC (respectively, BED) will be necessary to produce accurate values for x and y.

These assertions are both incorrect, and the behaviour of iterations of BEC and BED is far more complex. In [5] Govaerts and Pryce consider solvers based on an LU or QR decomposition. They show that BEC + 1 produces x and y accurately no matter how ill conditioned A is (except in rare cases of no practical interest). On the other hand, BED produces y accurately but requires several iterations to find x (if at all). As made clear in [5] the remarkable behaviour of BEC + 1 in this case depends on properties of matrix factorizations like LU and QR.

In § 2 we describe some experiments in the case of a solver based on the preconditioned conjugate gradient algorithm. They show that BEC + 1 no longer works in this case and also support the new algorithm BEM that we propose.

Section 3 gives an error analysis of BEM and shows that it usually produces x, y accurately if M is well conditioned and the solver is stable. It also highlights why exceptional cases may cause a failure. Propositions 3.1 and 3.3 further contain the basic ingredients to prove that in practically arising cases, BEM is stable.

Section 4 describes an "exceptional" situation. The aim is to compare the performance of BEM, BEC, a modified version of BEC, the deflated block elimination of Chan and Resasco [3], [4], and iterative refinements of these algorithms in a critical case.

Section 5 draws the final conclusions on the merits and disadvantages of the algorithms.

2. Tests of block elimination algorithms with a solver based on conjugate gradients. In the tests described in this section, A is an 80-by-80 symmetric nonnegative-definite matrix. It is constructed as

$$A = H_{1000}H_{999}\cdots H_2H_1 \operatorname{diag}(1.49, 1.48, \cdots, 0.71, 0)H_1H_2\cdots H_{999}H_{1000},$$

where each matrix H_i ($1 \leqq i \leqq 1000$) is a Householder elementary reflection matrix $H_i = I - 2h_ih_i^T$ and h_i is a normalized random vector. Except for rounding errors, A has singular values 1.49, 1.48, $\cdots$0.71, 0 and it is made nonsingular only by machine imprecision. Obviously, $\|A\| \simeq 1.49$.

Next, b, c, d, x, y are vectors and scalars with coefficients chosen uniformly random in [0, 1]. We then compute $f = Ax + by$ and $g = cx + dy$ and solve the resulting system of the form (1) by BEC, BED, and their iterations.

All computations are done in the PC-version of the Gauss programming language with no extra precision in the computation of residuals or updating the solutions. Here $u = 2^{-52} \approx 2.2\ 10^{-16}$. In all the examples M is well conditioned (2-norm condition number smaller than 200).

The solver for A is the preconditioned conjugate gradient algorithm in Axelsson and Barker [1, § 1.4] with the diagonal of A as a preconditioner. The stopping criterion is that the norm of the residual must be bounded by 10^{-14} times the norm of the computed solution. This ensures that the system with A is solved in a stable way (see [2]). It is to be remarked, however, that we had similar results with other stopping criteria, e.g., prescribing a fixed number of iterations.

Table 1 gives the logarithms of the relative errors of the computed x and y components by BEC + k and BED + k ($k = 0, 1, \cdots, 6$). For comparison, we also give the relative error in the solution by Gaussian elimination with row interchanges on the full matrix M.

The columns BEC-x, BEC-y, and BED-x apparently support the hypothesis that several iterations of BEC and BED are necessary to produce accurate values for x and y. Since A is very nearly singular it may even seem surprising that the algorithms converge

TABLE 1

Logarithms of relative errors in the computed x and y components by BEC, BED, *and their iterations using a preconditioned conjugate gradient solver.*

Number of iterations	BEC		BED	
	x	y	x	y
	−0.2348	−4.3704	−0.8711	−13.5273
1	−4.3612	−8.6889	−2.3433	$-\infty$
2	−9.0570	−13.4635	−6.7891	−15.2063
3	−13.5508	$-\infty$	−12.2967	−15.8083
4	−15.7204	−15.9106	−14.4728	−14.6622
5	−15.2564	−15.9106	−14.4763	−15.1094
6	−14.7037	−15.9106	−14.9900	−15.5073
Full GE	−14.7330		−14.7424	

at all; however, Jankowski and Wozniakowski [6] have shown that iterative refinement of almost any solution scheme to solve linear systems will ultimately converge to an accurate solution (within the bounds posed by the condition of the system and provided the solution scheme gives a solution with relative error smaller than one).

We can make two other observations:

(1) Without any iteration BED produces y accurately. This result is confirmed by many similar experiments and we shall prove it whenever A is solved in a stable way (§ 3).

(2) The relative error in the x-component of the solution by BEC + k + 1 is of the order of the relative error in the y-component of the solution by BEC + k (i.e., in the preceding iteration) for $k = 0, 1, \cdots$. Again, this is confirmed by many similar experiments and it will be proved in the important case where the y-component by BEC + k is accurate (§ 3).

To test this important case further we organize another experiment. The results are collected in Table 2. Here we perform BEC and two iterations starting with the accurate value for y and a zero vector for x. We also give the norm of the right-hand side vector in step 3 of BEC and the norm of the computed solution (the importance of these quantities will be clarified in § 3).

Again, two things are to be remarked:

(1) The first application of BEC already produces both x and y accurately. This is what we hoped for and it confirms the second observation concerning Table 1.

(2) In the first application of A (step 3 of BEC) the computed solution has the same size as the right-hand side (remember that $\|A\| \simeq 1.49$). This is surprising since A is nearly singular and for a random right-hand side vector the computed solution will typically have the size $u^{-1}\|A\|^{-1}$ times the size of the right-hand side. In § 3 we show that this observation is the key to the understanding of the algorithm.

The preceding experiments naturally lead us to first compute y by BED and to use this value, together with a zero vector as approximation to x, in one step of BEC. The resulting algorithm will be called BEM (block elimination mixed). It is given explicitly by the following algorithm.

TABLE 2

Logarithms of relative errors in the computed x and y components by BEC *and two iterations where a correct y and a zero vector for x are introduced (preconditioned conjugate gradient solver).*

	x	y	Norm of right-hand side in step 3	Norm of solution in step 3
Introduced	0	$-\infty$		
BEC	−14.0759	−15.4734	6.2112	5.2935
+1	−15.4510	−15.7744	1.9271E − 15	0.06403
+2	−15.4143	$-\infty$	8.3257E − 16	0.04805
Full GE	−15.1916			

ALGORITHM BEM

1. Solve $\xi A = c$
2. Compute $\delta_1 = d - \xi b$
3. Compute $y = (g - \xi f)/\delta_1$
4. Solve $Av = b$
5. Compute $\delta = d - cv$
6. Compute $f_1 = f - by$
7. Compute $g_1 = g - dy$
8. Solve $Aw = f_1$
9. Compute $y_1 = (g_1 - cw)/\delta$
10. Compute $x = w - vy_1$
11. Compute $y = y + y_1$

Remark that steps 1–3 of BEM are identical to steps 1–3 of BED. Steps 4–5 of BEM are identical to steps 1–2 of BEC. Steps 6–7 compute the residuals given y (from step 3) and a zero vector for x as first approximations to the solution. Steps 8–10 correspond to steps 3–5 of BEC applied to the new right-hand side components f_1, g_1. Finally, step 11 updates y.

Remark that steps 4–5 of BEM are interchangeable with steps 6–7 of BEM. Step 4 of BED is omitted to avoid one solve with A (if included, steps 6–7 have to be adapted and a step 12 is necessary to update x).

TABLE 3

Logarithms of relative errors in the computed x and y components by BEM *and two iterations with* BEC *(preconditioned conjugate gradient solver).*

	x	y	Norm of right-hand side in step 8(BEM), step 3(BEC)	Norm of solution in step 8(BEM), step 3(BEC)
Step 3	0	−15.8359		
Steps 10–11	−13.9947	−14.9328	6.4435	5.2883
+BEC	−15.1867	−15.8359	5.1164E − 14	0.5475
+BEC + 1	−15.5406	$-\infty$	2.3295E − 15	0.01477
Full GE	−15.3589			

Table 3 gives the result of a test with BEM (A, b, c, d, x, y, f, g, as before). Note that BEM produces x, y accurately, as we hoped, and that the right-hand side and computed solution in step 8 have the same order of magnitude (cf. the discussion of Fig. 2).

For completeness, Table 3 also shows the effect of two further iterations with BEC. The improvement so obtained is small and apparently not worth the effort.

3. Error analysis of BEM. Throughout this section we assume that M is well conditioned, i.e., $\kappa(M) = \|M\| \cdot \|M^{-1}\|$ is modest.

Proposition 3.1 and its Corollary 3.2 contain the analysis of steps 1–3 of BEM. The important result is that y, as computed in step 3 of BEM, is accurate even if A is very ill conditioned. This is consistent with the numerical evidence in Table 3 and also explains the observation (1) made in § 2 while discussing Table 1.

Proposition 3.3 is the backward error analysis of BEC. Its Corollary 3.4 draws the important conclusion: the accuracy of the solution obtained by BEC depends exclusively on the size of $\|\bar{w}\|$. This explains why we choose to represent this quantity in Tables 2 and 3.

Now the second part of BEM is precisely an application of BEC to a system transformed by Steps 1–3 and 6–7. Theorem 3.5 shows that in this transformed system, $\|\bar{w}\|$ is usually of order $\|A\|^{-1}\,\|M\|\,\|z\|$ (even for nearly singular A), and therefore BEM produces x, y accurately. This confirms the numerical results of Table 3. From the proof of Theorem 3.5 it is clear that the essential results (a modest $\|\bar{w}\|$ and accurate x, y) remain true if steps 1–3 of BEM are replaced by any method that produces y accurately. This explains the observations (1), (2) in the discussion of Table 2 in § 2.

All computations described in this paper are done in the same floating-point precision u. In general, $\bar{a}$ denotes the computed value of the quantity a (so $\bar{a}$ need not be close to a in any sense).

In the error analysis, we use the notation introduced by Pryce [12] for manipulating the relative error metric introduced by Olver [10] in the scalar case and generalized by Pryce [11] to the vector case. Throughout the analysis, $\theta_1, \theta_2, \cdots$ denote scalar or $n \times n$ matrix quantities close to the identity. In the scalar case the notation $\theta \in 1(\delta)$ where δ is a nonnegative constant, means $\theta = e^{\varepsilon}$ where $|\varepsilon| \leqq \delta$. In the matrix case, it means that θ is a product of a finite number of matrices $\exp(E_i)$ where $\sum_i \|E_i\| \leqq \delta$.

With this understanding we have

$$\mathrm{fl}(x \text{ op } y) = \theta(x \text{ op } y), \qquad \theta \in 1(u)$$

whenever x, y are scalars and "op" is one of the four basic operations. This remains true if x, y are vectors and "op" is a componentwise combination. It is also true when "op" denotes multiplication of a vector by a scalar.

Furthermore, there is a constant C_{IP} such that

$$\mathrm{fl}(x^T y) = x^T \theta y, \qquad \theta \in 1(C_{IP}u),$$

where θ is a diagonal matrix and $C_{IP} \leqq n$ (cf. [5]; in case of double precision accumulation we have $C_{IP} \simeq 1$).

The obvious bounds

$$\|e^{\theta}\| \leqq e^{\|\theta\|}, \qquad \|e^{\theta} - I\| \leqq \|\theta\|\, e^{\|\theta\|}$$

will often be used without notice.

PROPOSITION 3.1. *Let S be a stable solver for A^T with stability constant C_S. Let $\bar{y}$ be the result computed in step* 3 *of Algorithm* BEM. *Then $\bar{y}$ is the y-component of the*

exact solution of a system near $Mz = h$. *More precisely, there exist* ΔA, Δb, Δc, Δd, Δf, Δg, *and* x_∞ *such that*

$$\begin{pmatrix} A+\Delta A & b+\Delta b \\ c+\Delta c & d+\Delta d \end{pmatrix}\begin{pmatrix} x_\infty \\ \bar{y} \end{pmatrix}=\begin{pmatrix} f+\Delta f \\ g+\Delta g \end{pmatrix} \tag{4}$$

and

$$b+\Delta b=\theta_b b, \qquad \theta_b \in 1((1+C_{IP})u), \tag{4a}$$

$$d+\Delta d=\theta_d d, \qquad \theta_d \in 1(u), \tag{4b}$$

$$f+\Delta f=\theta_f f, \qquad \theta_f \in 1((2+C_{IP})u), \tag{4c}$$

$$g+\Delta g=\theta_g g, \qquad \theta_g \in 1(2u), \tag{4d}$$

$$\|\Delta A\| \leqq C_S u \|A\|, \tag{4e}$$

$$\|\Delta c\| \leqq C_S u \|c\|. \tag{4f}$$

Proof. We have

$$\bar{\xi}(A+\Delta A)=c+\Delta c, \qquad \|\Delta A\| \leqq C_S u\|A\|, \qquad \|\Delta c\| \leqq C_S u\|c\|, \tag{5}$$

$$\theta_1\bar{\delta}_1=d-\bar{\xi}\theta_2 b, \qquad \theta_1 \in 1(u), \qquad \theta_2 \in 1(C_{IP}u), \tag{6}$$

$$\theta_3\overline{(g-\xi f)}=g-\bar{\xi}\theta_4 f, \qquad \theta_3 \in 1(u), \qquad \theta_4 \in 1(C_{IP}u), \tag{7}$$

$$\theta_5\bar{y}=\overline{(g-\xi f)}/\bar{\delta}_1, \qquad \theta_5 \in 1(u). \tag{8}$$

Combining (6), (7), and (8), we obtain

$$\bar{y}=\frac{\theta_5^{-1}\theta_3^{-1}g-\bar{\xi}\theta_5^{-1}\theta_3^{-1}\theta_4 f}{\theta_1^{-1}d-\bar{\xi}\theta_1^{-1}\theta_2 b}. \tag{9}$$

So $\bar{y}$ is the exact y-component of the solution of

$$\begin{pmatrix} A+\Delta A & \theta_1^{-1}\theta_2 b \\ c+\Delta c & \theta_1^{-1}d \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}=\begin{pmatrix} \theta_5^{-1}\theta_3^{-1}\theta_4 f \\ \theta_5^{-1}\theta_3^{-1}g \end{pmatrix}, \tag{10}$$

from which the proposition follows. □

COROLLARY 3.2. *Let the assumptions of Proposition* 3.1 *be satisfied. Suppose, in addition, that* M *is nonsingular and*

$$uC_M\kappa(M)<1 \tag{11a}$$

where

$$C_M=(2+C_{IP}+2C_S)\exp((1+C_{IP})u). \tag{11b}$$

Then

$$|y-\bar{y}| \leqq C_y u\|z\| \tag{11c}$$

where

$$C_y=\frac{(C_h+C_M)\kappa(M)}{1-uC_M\kappa(M)} \tag{11d}$$

and

$$C_h=(4+C_{IP})\exp((2+C_{IP})u). \tag{11e}$$

Proof. By Proposition 3.1 we have

$$(M+\Delta M)(z+\Delta z)=h+\Delta h \tag{12}$$

where

$$z+\Delta z=\begin{pmatrix} x+\Delta x \\ y+\Delta y \end{pmatrix}=\begin{pmatrix} x_\infty \\ \bar{y} \end{pmatrix} \tag{13}$$

and

$$\Delta M=\begin{pmatrix} \Delta A & \Delta b \\ \Delta c & \Delta d \end{pmatrix}, \qquad \Delta h=\begin{pmatrix} \Delta f \\ \Delta g \end{pmatrix}.$$

Now standard perturbation arguments yield

$$\frac{\|\Delta z\|}{\|z\|}\leqq uC_y,$$

where we have used the bounds (4a)–(4f). From this, (11c) follows. □

PROPOSITION 3.3. *Let S be a stable solver for A with stability constant C_S and let $\bar{x}$, $\bar{y}$ be the components of* (1) *by* BEC. *Then $\bar{x}$, $\bar{y}$ exactly satisfy the matrix equality*

$$\begin{pmatrix} A+\Delta A & b+\Delta b \\ c+\Delta c & d+\Delta d \end{pmatrix}\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}=\begin{pmatrix} f+\Delta f \\ g \end{pmatrix}+\begin{pmatrix} T \\ U \end{pmatrix}\bar{w} \tag{14}$$

where

$$\|\Delta A\|\leqq(2+C_S)u\exp(2u)\|A\|, \tag{14a}$$

$$\|\Delta b\|\leqq C_S u\|b\|, \tag{14b}$$

$$\|\Delta c\|\leqq(5+C_{IP})u\exp((5+C_{IP})u)\|c\|, \tag{14c}$$

$$\|\Delta d\|\leqq 3u\exp(3u)\|d\|, \tag{14d}$$

$$\|\Delta f\|\leqq C_S u\|f\|, \tag{14e}$$

$$\|T\|\leqq(2C_S+(1+C_S u)\exp u)u\|A\|, \tag{14f}$$

$$\|U\|\leqq(4+2C_{IP})u\exp((6+C_{IP})u)\|c\|. \tag{14g}$$

Proof. The computed quantities $\bar{v}$, $\bar{w}$, $\bar{y}$, $\bar{x}$ satisfy

$$(A+\Delta_v A)\bar{v}=b+\Delta b, \qquad \|\Delta_v A\|\leqq C_S u\|A\|, \qquad \|\Delta b\|\leqq C_S u\|b\|, \tag{15}$$

$$(A+\Delta_w A)\bar{w}=f+\Delta f, \qquad \|\Delta_w A\|\leqq C_S u\|A\|, \qquad \|\Delta f\|\leqq C_S u\|f\|, \tag{16}$$

$$\theta_6\bar{y}=\frac{g-c\theta_7\bar{w}}{d-c\theta_8\bar{v}}, \qquad \theta_6\in 1(3u), \qquad \theta_7\in 1(C_{IP}u), \qquad \theta_8\in 1(C_{IP}u), \tag{17}$$

$$\theta_9\bar{x}=\bar{w}-\theta_{10}\bar{v}\bar{y}, \qquad \theta_9\in 1(u), \qquad \theta_{10}\in 1(u). \tag{18}$$

Eliminating $\bar{v}\bar{y}$ from (17) and (18) we get

$$\theta_6 d\bar{y}+\theta_6 c\theta_8\theta_{10}^{-1}\theta_9\bar{x}=g+c(\theta_6\theta_8\theta_{10}^{-1}-\theta_7)\bar{w}. \tag{19}$$

Combining (18), (15), and (16), we get

$$(A+\Delta_v A)\theta_{10}^{-1}\theta_g\bar{x}+(b+\Delta b)\bar{y}=f+\Delta f+[(\Delta_v A-\Delta_w A)+(A+\Delta_v A)(\theta_{10}^{-1}-I)]\bar{w}. \tag{20}$$

Now (19) and (20) may be rewritten as

$$\begin{pmatrix} A+\Delta A & b+\Delta b \\ c+\Delta c & d+\Delta d \end{pmatrix}\begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}=\begin{pmatrix} f+\Delta f \\ g \end{pmatrix}+\begin{pmatrix} T \\ U \end{pmatrix}\bar{w},$$

where bounds for $\|\Delta A\|$, $\|\Delta b\|$, $\|\Delta c\|$, $\|\Delta d\|$, $\|\Delta f\|$, $\|T\|$, and $\|U\|$ can be computed from the bounds in (15)–(18). □

COROLLARY 3.4. *Let the assumptions of Proposition* 3.3 *be satisfied and define*

$$C'_h=(5+2C_S+uC_S+2C_{IP})\exp((6+C_{IP})u), \tag{21a}$$

$$C'_M=(10+C_{IP}+2C_S)\exp((5+C_{IP})u). \tag{21b}$$

Assume that $uC'_M\kappa(M)<1$ *and define*

$$C_z=\frac{(C'_M+C_S)\kappa(M)}{1-uC'_M\kappa(M)}, \tag{21c}$$

$$C''_h=\frac{C'_h\kappa(M)}{1-uc'_M\kappa(M)}. \tag{21d}$$

Then

$$\|\bar{z}-z\|\leqq uC_z\|z\|+uC''_h\|\bar{w}\|. \tag{22}$$

So if M is well conditioned, then the accuracy of the computed solution $\bar{z}$ *is determined by the size of* $\bar{w}$.

Proof. Rewrite (14) as

$$(M+\Delta M)\bar{z}=\begin{pmatrix} f+\Delta f \\ g \end{pmatrix}+\begin{pmatrix} T \\ U \end{pmatrix}\bar{w},$$

where bounds for $\|\Delta M\|$, $\|\Delta f\|$, $\|T\|$, $\|U\|$ follow from (14a)–(14g).

The result now follows by standard perturbation arguments. □

THEOREM 3.5. *Let* $\bar{x}$ *be the x-component obtained by* BEM. *Let* $\bar{y}$ *be the y-component obtained by* BEM *with step* 11 *omitted, and* $\bar{y}_2$ *the y-component obtained by* BEM *with step* 11 *included. Assume that*

$$uC'_M\kappa(M)<1. \tag{23}$$

Then

$$\|\bar{x}-x\|\leqq uC''_z\|z\|+u^2CC''_hz\|\,\|(A+\Delta_wA)^{-1}\|\,\|M\|\,\|z\|, \tag{24a}$$

$$\|\bar{y}-y\|\leqq uC_y\|z\|, \tag{24b}$$

$$\|\bar{y}_2-y\|\leqq ue^u(C''_z+1)\|z\|+u^2CC''_he^u\|(A+\Delta_wA)^{-1}\|\,\|M\|\,\|z\|, \tag{24c}$$

where we have defined

$$C'''_h=2e^u+(1+uC_y)(4e^{2u}+C_S(1+e^u+2e^{2u})), \tag{25a}$$

$$C'_z=\frac{\kappa(M)(C'_M(1+uC_y)+C'''_h)}{1-uC'_M\kappa(M)}, \tag{25b}$$

$$C''_z=C'_z+C''_h, \tag{25c}$$

$$C=C_S+C_y+e^u+(1+uC_y)(2e^{2u}+C_S(1+e^u+2e^{2u})). \tag{25d}$$

Proof. Let $\bar{y}$ be the y-component computed in step 3 of BEM. Define

$$f_{1,0} = f - b\bar{y}, \tag{26a}$$

$$g_{1,0} = g - d\bar{y}, \tag{26b}$$

$$y_1 = y - \bar{y}. \tag{26c}$$

Then

$$M\begin{pmatrix} x \\ y_1 \end{pmatrix} = \begin{pmatrix} f_{1,0} \\ g_{1,0} \end{pmatrix}. \tag{27}$$

First note that

$$\overline{f_1} = \theta_{11}(f - \theta_{12} b\bar{y}), \qquad \theta_{11}, \theta_{12} \in 1(u), \tag{28a}$$

$$\overline{g_1} = \theta_{13}(g - \theta_{14} d\bar{y}), \qquad \theta_{13}, \theta_{14} \in 1(u). \tag{28b}$$

Applying Proposition 3.3 we get

$$(M + \Delta M)\begin{pmatrix} \bar{x} \\ \bar{y}_1 \end{pmatrix} = \begin{pmatrix} \overline{f_1} + \Delta\overline{f_1} \\ \overline{g_1} \end{pmatrix} + \begin{pmatrix} T \\ U \end{pmatrix}\bar{w} \tag{29}$$

with bounds for ΔM, $\Delta\bar{f}_1$, T, U as in (14a)–(14g). Put

$$\bar{f}_1 = f_{1,0} + \Delta f_1, \tag{30a}$$

$$\bar{g}_1 = g_{1,0} + \Delta g_1. \tag{30b}$$

Then by (26a) and (28a)

$$\|\Delta f_1\| \leqq ue^u \|f\| + 2ue^{2u}(1 + uC_y)\|M\|\|z\|, \tag{31a}$$

$$\|\Delta g_1\| \leqq ue^u \|g\| + 2ue^{2u}(1 + uC_y)\|M\|\|z\|. \tag{31b}$$

Now rewrite (34) as

$$(M + \Delta M)\begin{pmatrix} \bar{x} \\ \bar{y}_1 \end{pmatrix} = \begin{pmatrix} f_{1,0} \\ g_{1,0} \end{pmatrix} + \Delta h. \tag{32}$$

By straightforward computation we obtain from (11c), (14e), and (26a)–(31b)

$$\|\Delta h\| \leqq uC_h''' \|M\|\|z\| + uC_h' \|M\|\|\bar{w}\|. \tag{33}$$

Using (27), (32), (33), the assumption (23), and (11c) again, we obtain

$$\left\| \begin{pmatrix} \bar{x} \\ \bar{y}_1 \end{pmatrix} - \begin{pmatrix} x \\ y_1 \end{pmatrix} \right\| \leqq uC_z'' \|z\| + uC_h'' \|\bar{w}\|. \tag{34}$$

Clearly, the size of $\|\bar{w}\|$ is all-important. By the stability assumption we have

$$(A + \Delta_w A)\bar{w} = \overline{f_1} + \Delta\overline{f_1}, \tag{35}$$

$$\|\Delta_w A\| \leqq uC_S \|A\|, \tag{35a}$$

$$\|\Delta\overline{f_1}\| \leqq uC_S \|\overline{f_1}\|. \tag{35b}$$

By straightforward computations using (26a), (30a), (31a), and (28a) we find

$$\|\bar{w}\| \leqq \|x\| + Cu\|(A + A_w A)^{-1}\|\|M\|\|z\|. \tag{36}$$

Inserting (36) into (34) we get

$$\left\| \begin{pmatrix} \bar{x} \\ \bar{y}_1 \end{pmatrix} - \begin{pmatrix} x \\ y_1 \end{pmatrix} \right\| \leqq uC_z'' \|z\| + u^2 CC_h'' \|(A + \Delta_w A)^{-1}\| \|M\| \|z\|. \tag{37}$$

This implies (24a). Of course (24b) is just (11c).

To prove (24c) first remark that

$$\begin{aligned} \|\bar{y}_2 - y\| &= \|\overline{(\bar{y}_1 + \bar{y})} - y\| \\ &= \|\theta_{15}(\bar{y}_1 + \bar{y}) - y\|, \theta_{15} \in 1(u) \\ &= \|\theta_{15}(\bar{y}_1 - y_1) + \theta_{15}(y_1 + \bar{y}) - y\| \\ &\leqq \|\theta_{15}(\bar{y}_1 - y_1)\| + \|\theta_{15} y - y\| \\ &\leqq e^u \|\bar{y}_1 - y_1\| + ue^u \|y\|. \end{aligned}$$

Formula (24c) follows by inserting the bound in (37) for $|\bar{y}_1 - y_1|$ in this inequality.

DISCUSSION 3.6. (1) The error bounds in (24b) and (24c) suggest that step 11 of BEM can be omitted. This is indeed true for perfectly well conditioned M. Since in practice we deal with less extreme cases, we strongly recommend retaining step 11, whose computing cost is negligible anyway.

(2) The bound in (24a) shows that the accuracy of the x-component computed by BEM depends entirely on the size of $\|(A + \Delta_w A)^{-1}\|$. In particular, the x-component is accurate whenever $\|(A + \Delta_w A)^{-1}\| \lesssim u^{-1} \|M\|^{-1}$. This is the case that typically occurs in practice because roundoff errors in the computation of A and in the solution of systems with A tend to produce this bound.

(3) It is possible to construct highly artificial situations where BEM produces x accurately and $\|(A + \Delta_w A)^{-1}\|$ is arbitrarily large (provided there is no overflow or underflow in the computations). This may be achieved by choosing all components of A, b, c, d, x, y as appropriate integers in such a way that no roundoff error occurs. Typically, however, BEM will produce a completely nonaccurate x-component whenever $\|(A + \Delta_w A)^{-1}\| \gtrsim u^{-2} \|M\|^{-1}$. This is best seen from (35). Indeed, $\bar{f}_1 + \Delta\bar{f}_1$ will probably contain a vector of size at least $u\|M\|\|z\|$ in the singular direction of $(A + \Delta_w A)$. Therefore we expect

$$\|\bar{w}\| \gtrsim u^{-2} \|M\|^{-1} u \|M\| \|z\| \simeq u^{-1} \|z\|.$$

This means that $\bar{x}$ may have a relative error of order one.

(4) In the intermediate case

$$u^{-1} \|M\|^{-1} \leqq \|(A + \Delta_w A)^{-1}\| < u^{-2} \|M\|^{-1},$$

we infer from (24a) that x has a relative error of order less than one. In this case iterative refinement of BEM is in practice very satisfactory (cf. Jankowski and Wozniakowski [6]).

4. A series of experiments in an unusual situation. In this section we describe a series of experiments with four algorithms to solve bordered singular systems.

These methods are BEM, DBE, BEC + 1, and BEC2.

BEM (block elimination mixed) was introduced in § 2 and studied in § 3.

DBE (deflated block elimination) is the method introduced by Chan [3], [4]. We used the form proposed in [4].

BEC + 1 (block elimination (Crout)) is the BEC algorithm described in § 1 with one iterative refinement. It was studied by Govaerts and Pryce [5].

BEC2 is a modification of BEC + 1 in which step 5 of BEC is replaced by simply making all components of x zero. In the iteration, however, step 5 is retained.

As remarked before, in most practically occurring cases, the solver has norm bounded by $u^{-1}\|M\|^{-1}$. This is typically caused by roundoff error even if A is theoretically singular.

Tests with such solvers are described in [3] (DBE) and [5] (BEC + 1). Section 2 of this paper describes a test with BEM in the case of a conjugate gradient solver. These and similar experiments show that all four methods produce accurate results in the cases of practical interest (BEC + 1 and BEC2 only for solvers based on decompositions like LU or QR, not for solvers based on the conjugate gradient method).

Since our error analysis shows that in certain cases of little practical interest BEM may fail, it is of interest to know whether the other methods might do better. Since mildly pathological cases might also arise, we can further ask whether iterative refinement is useful in such cases.

To get insight into the critical cases we consider the ill-reputed matrix $A = W_n$.

$$W_n(i,j) = \begin{cases} 1 & \text{if } i = j, \\ -1 & \text{if } i > j, \\ 0 & \text{if } i < j. \end{cases}$$

If $2^n \gtrsim u^{-1}$, this triangular matrix has a unique small singular value of order 2^{-n}; the near null vector is $(2^{-n+1}, 2^{-n+2}, \cdots, 1)$. Moreover, small perturbations of the nonzero elements of W_n do not essentially change this behaviour and $\|(W_n + \Delta W_n)^{-1}\|$ is of order 2^n in this case (small random perturbations in all elements of W_n, however, tend to reduce $\|(W_n + \Delta W_n)^{-1}\|$ to order u^{-1}). The solver for W_n is forward elimination in all cases and $u \sim 10^{-16}$.

In all the experiments, b, c, d, x, y are chosen uniformly random in [0, 1] and f, g are computed in the same precision as $f = Ax + by$ and $g = cx + dy$. The resulting system is then solved by the four algorithms and for each of them two iterative improvements are performed as well. This is done for $n = 20, 40, 60, 80, 100, 120, 140$, and 160. Since the computed $\bar{y}$ is always accurate ($|\bar{y} - y|/\|z\|$ is of the order of u), only the logarithmic relative error $\log(\|\bar{x} - x\|/\|x\|)$ in the x-component is represented.

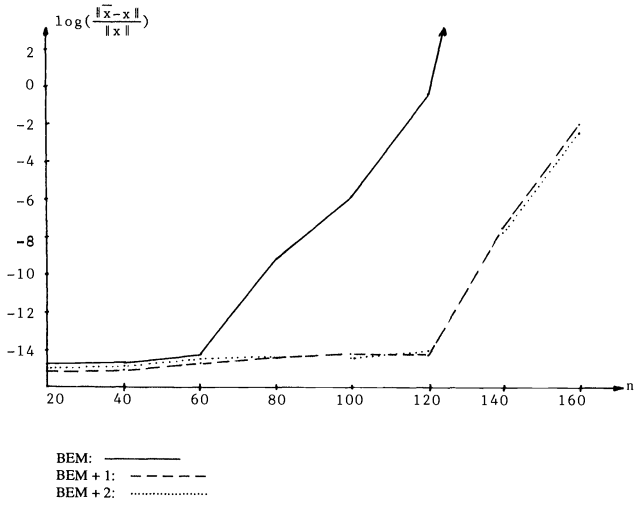


FIG. 1. BEM *with an ill-conditioned triangular A.*

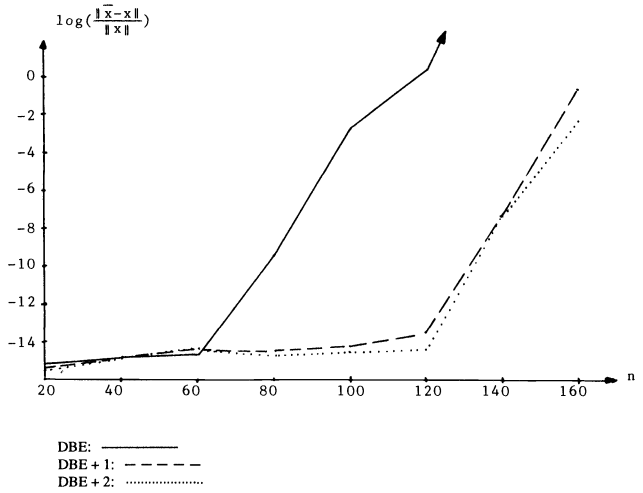


FIG. 2. DBE *with an ill-conditioned triangular* A.

The following should be noted.

(1) Figure 1 shows that BEM produces an accurate x-component for $n \lesssim 60$. Since $2^{60} \sim 10^{18}$, this confirms Discussion 3.6(2).

(2) Figure 1 also shows that BEM + 1 produces accurate results for $n \lesssim 120$ and that more iterations will not further improve the accuracy for higher n.

This is consistent with Discussion 3.6(2) and 3.6(3). Actually, the numerical results are even better than expected from theory. This might be due, however, to the special nature of A.

(3) Figure 2 shows that numerically, DBE behaves very much like BEM. In particular, it is not always a stable method. In the (admittedly rare) cases where it is not, it may be improved greatly by one iterative refinement.

(4) BEC + 1 as such is an inferior method in the case where $\|(A + \Delta_w A)^{-1}\| > u^{-1}\|M\|^{-1}$, since it does not improve by iterative refinement. The reason is obviously

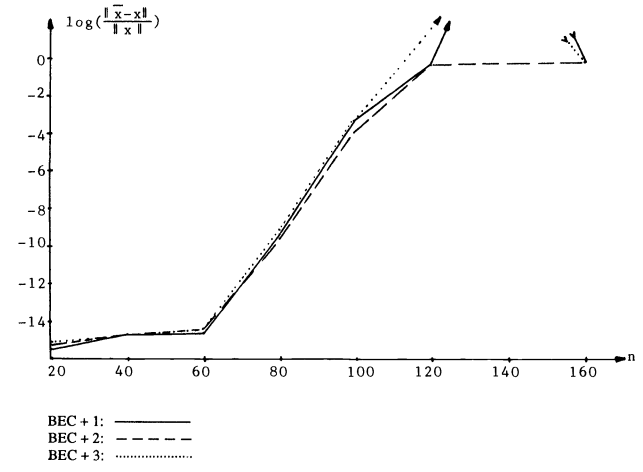


FIG. 3. BEC + k *with an ill-conditioned triangular* A.

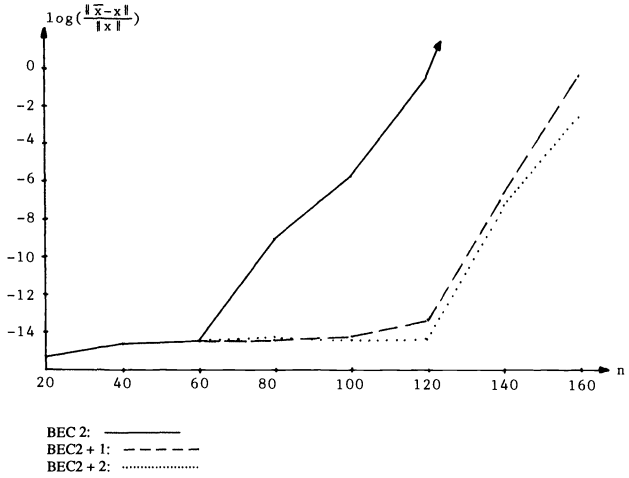


FIG. 4. BEC2 *with an ill-conditioned triangular* A.

that the large size of $\bar{x}$, computed in step 5 of BEC, causes catastrophic roundoff error in the computation of the residual (Fig. 3).

(5) BEC2 avoids this problem by simply omitting the computation of $\bar{x}$ in the first round of BEC. The results are then strikingly similar to those of BEM and DBE (Fig. 4).

5. Conclusions. BEC + 1 is implemented very easily. It has the further possible advantage that it only needs a solver for A, not for A^T. The computational cost is minimal: essentially three solves with A. Next, it fits well in applications like numerical continuation theory where a solver for a matrix close to A might be available. Therefore it is highly recommended in most practical cases.

BEC + 1 has the disadvantage that it requires a solver based on a decomposition like LU, QR, or a similar one. It fails, e.g., for a solver based on the conjugate gradient method for a symmetric positive-semidefinite matrix A.

BEC2 is an alternative to BEC + 1 and has the same properties. Its one advantage is that it can be improved by iterative refinement in some exceptional cases where BEC + 1 fails because $\|(A + \Delta_w A)^{-1}\|$ is excessive. Remark that BEC2 + 1 requires five solves with A.

Now let the solver be general, i.e., not necessarily based on an LU or QR decomposition. A solver for A^T is often also available in practice. Then BEM has the same cost as BEC + 1. A solver for a matrix close to A can be used only if BEM is iterated once. Remark that the cost of an iterative refinement is only one solve with A and that BEM + 1 is also more accurate in the cases with excessive $\|(A + \Delta_w A)^{-1}\|$.

DBE has roughly the same requirements as has BEM and similar performance as well. It uses, however, four solves with A and we think BEM also allows an easier implementation (see also Moore [9] for the error analysis of DBE).

Acknowledgments. The stimulating influence of J. D. Pryce (RMCS, England) led us to write this paper. We also thank L. Paquet (Mons, Belgium) for many critical remarks. The comments of two anonymous referees also led to a substantial improvement in the presentation, in particular, in that of § 2.

REFERENCES

[1] O. AXELSSON AND V. A. BARKER, *Finite Element Solution of Boundary Value Problems*, Academic Press, New York, 1984.

[2] A. BJÖRCK, *Iterative refinement and reliable computing*, in Advances in Reliable Numerical Computing, M. Cox and S. Hammarling, eds., Oxford University Press, Oxford, U.K., 1990.

[3] T. F. CHAN, *Deflation techniques and block-elimination algorithms for solving bordered singular systems*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 121–134.

[4] T. F. CHAN AND D. C. RESASCO, *Generalized deflated block-elimination*, SIAM J. Numer. Anal., 23 (1986), pp. 913–924.

[5] W. GOVAERTS AND J. D. PRYCE, *Block elimination with one iterative refinement solves bordered linear systems accurately*, BIT, 30 (1990), pp. 490–507.

[6] M. JANKOWSKI AND H. WOZNIAKOWSKI, *Iterative refinement implies numerical stability*, BIT, 17 (1977), pp. 303–311.

[7] H. B. KELLER, *The bordering algorithm and path following near singular points of higher nullity*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 573–582.

[8] G. MOORE, *The application of Newton's method to simple bifurcation and turning point problems*, Ph.D. thesis, University of Bath, Bath, U.K., 1979.

[9] ———, *Some remarks on the deflated block elimination method*, in Bifurcation: Analysis, Algorithms, Applications, T. Küpper, R. Seydel, and H. Troger, eds., Birkhauser-Verlag, Basel, Switzerland, 1987.

[10] F. W. J. OLVER, *A new approach to error arithmetic*, SIAM J. Numer. Anal., 15 (1978), pp. 368–393.

[11] J. D. PRYCE, *A new measure of relative error for vectors*, SIAM J. Numer. Anal., 21 (1984), pp. 202–215.

[12] ———, *Multiplicative error analysis of matrix transformation algorithms*, IMA J. Numer. Anal., 5 (1985), pp. 437–445.

[13] W. C. RHEINBOLDT, *Numerical Analysis of Parametrized Nonlinear Equations*, John Wiley, New York, 1986.

[14] R. D. SKEEL, *Iterative refinement implies numerical stability for Gaussian elimination*, Math. Comp., 35 (1980), pp. 817–832.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 484–496, July 1991

007

OBSERVABILITY OF LINEAR TIME-VARYING DESCRIPTOR SYSTEMS*

STEPHEN L. CAMPBELL† AND WILLIAM J. TERRELL†

Abstract. A characterization of observability for linear time-varying descriptor systems $E(t)x'(t) + F(t)x(t) = B(t)u(t)$, $y(t) = C(t)x(t)$, is given. E is not required to have constant rank. The characterization is designed to reduce symbolic computation and has potential advantages even when E is nonsingular. It is also shown that all observable analytic descriptor systems are smoothly observable even if they are not uniformly observable. Finally, the external behavior of time-varying descriptor systems is characterized.

Key words. descriptor, singular, observability, external behavior

AMS(MOS) subject classifications. 34A08, 93B07, 93B15, 93C15, 93C50

1. Introduction. In the last decade there has been increasing interest in the utilization of implicit differential equations

$$F(x', x, t) = 0, \tag{1}$$

known variously as descriptor, singular, or differential algebraic (DAE) systems [3]. There have been several motivations for this effort ranging from computational advantages to the fact that most physical systems are originally modeled in this form.

When confronted with an implicit system there are many questions that we can ask. In this paper we will be concerned with the observability of

$$E(t)x'(t) + F(t)x(t) = B(t)u(t), \tag{2a}$$

$$y(t) = C(t)x(t). \tag{2b}$$

Here, u is the control, y is the m-dimensional observation, and x is the n-dimensional state. For technical reasons we assume that E, F, B, C are infinitely differentiable although that will not usually be required in practice. We define *smooth*, then, to mean infinitely differentiable. The infinite differentiability is only used to make some of our conditions necessary as well as sufficient. Intervals are always assumed to be nontrivial.

Our goal is to develop characterizations and algorithms that can be reasonably rapid to apply. In one scenario for the use of our results, the researcher has formulated the equations (2) and wants to "quickly" know if the problem is observable. The coefficients E, F, B, C are assumed known functions. However, the researcher may want to try several different formulations of the underlying problem, perhaps by changing what the choices of control and observation are. Also, there may be design parameters in the problem description and observability will need to be tested for several values of these parameters. In this setting we want to reduce the amount of symbolic computation. Our goal is to develop procedures for which the only symbolic operation is simply differentiation. All other computation, including matrix multiplication, will then be done numerically.

* Received by the editors September 20, 1989; accepted for publication (in revised form) March 6, 1990. This research was partially supported by Air Force Office of Scientific Research grant AFOSR-87-0051.

† Department of Mathematics and Center for Research in Scientific Computation, North Carolina State University, Raleigh, North Carolina 27695-8205 (SLC@NCSUVM.BITNET and TERRELL@NCSUVM.BITNET).

With the obvious notation, (2) can be written as a single system

$$\tilde{E}x' + \tilde{F}x = \tilde{B}\begin{bmatrix} u \\ y \end{bmatrix}. \tag{3}$$

However, we shall see that there are advantages in keeping the pair of equations (2).

Our presentation is self-contained. The proofs make frequent use of results from [7].

1.1. Observability. The system (2) is *observable* on the interval $\mathcal{I}$ if knowledge of the output y and the control u on any subinterval $\tilde{\mathcal{I}}$ of $\mathcal{I}$ uniquely determines *smooth* solutions x of (2a) on $\tilde{\mathcal{I}}$. If C in (2b) is not full column rank on a dense set, then the additional information to determine x is gotten (at least theoretically) by differentiating (2b). Observability has been frequently discussed when $E(t) = I$ since the early work in [15], [19], [23], and [27] and in the descriptor case when E, F, B, C are constant matrices. However, ours is the first discussion of observability of time-varying descriptor systems. While our initial formulation is similar in spirit to that in [15], our assumptions, methods, and goals are different.

For linear time-invariant descriptor systems there is some variance in the definitions of observability depending on how the authors wish to deal with the potential impulsive behavior (for example, see [1], [12], [13], [14], [16], [20], [21], [26]). We shall assume that the controls u are sufficiently smooth and the initial conditions for the descriptor system consistent so that no impulsive behavior is present.

It was noted early in the observability literature that there are different forms of observability [23]. We have just defined *total observability*. In some problems a stronger type of observability is needed.

DEFINITION 1. The system (2) is *smoothly observable* (of order (k, l)) on the interval $\mathcal{I}$, if there exists smooth $K_i(t)$, $L_i(t)$ on $\mathcal{I}$ such that

$$x = \sum_{i=0}^{k} K_i(t) y^{(i)}(t) + \sum_{i=0}^{l} L_i(t)(Bu)^{(i)}(t). \tag{4}$$

DEFINITION 2. The system (2) is *uniformly observable* if it is smoothly observable of order $(n-1, n-1)$.

Uniform observability [23] is usually defined differently. We shall relate the two definitions later when we discuss the E nonsingular case.

Example 1. The system

$$x' = x + u, \tag{5}$$

$$y = \phi x, \tag{6}$$

where $\phi(t)$ is an infinitely differentiable function such that $\phi^{(i)}(0) = 0$ for $0 \leqq i < \infty$ and $\phi(t) \neq 0$ if $t \neq 0$, is observable but not smoothly observable on every interval $\mathcal{I}$ containing zero.

Example 2. The system

$$x' = x + u, \tag{7a}$$

$$y = t^2 x \tag{7b}$$

is smoothly observable of order (2, 1) since

$$x = (2t^2 + 2t + 2)^{-1}[y'' - y' + 2y - 4tu - t^2 u']. \tag{8}$$

However, (7) is not uniformly observable on any interval containing zero.

Using all of the information gotten in differentiating the output can be helpful, as indicated by the next example.

Example 3. Consider the system

$$x' = x + u, \tag{9}$$

$$y = tx. \tag{10}$$

Equation (10) implies that $x = t^{-1}y$, which is not smooth at $t = 0$. Differentiating (10) once and using (9) for x' gives

$$x = (1+t)^{-1}(-y' - tu), \tag{11}$$

which is not smooth at $t = -1$. However, if we use the differentiated equation *and* the original (10), and solve the overdetermined system that results, we get

$$x = y' - y + tu, \tag{12}$$

which is smooth on any interval.

There is a tradeoff here. By allowing extra differentiations of the inputs and outputs, we can obtain extra smoothness of the coefficients in the observation equation (4).

Our definition implies that any portions of the solution x which are completely determined by the control u are automatically observable.

Example 4. Let E be a constant matrix N which is nilpotent of index ν, and let $F = -I$. Then (2) is observable independent of B, C since

$$x = -\sum_{i=0}^{\nu-1} N^i (Bu)^{(i)}.$$

1.2. Terminology and background. The system of algebraic equations, $Ax = b$, written as

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{13}$$

is 1-*full* with respect to x_1 if (13) uniquely determines x_1 for any consistent b. From basic linear algebra we have Lemma 1.

LEMMA 1. *The following are equivalent for the system of algebraic equations* (13):

(1) *The system* (13) *is* 1-*full with respect to* x_1.

(2) *The submatrices*

$$A_1 = \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix}, \qquad A_2 = \begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix}$$

have disjoint ranges and A_1 *has full column rank.*

(3) *The row echelon form of* A *is*

$$\begin{bmatrix} I_{n\times n} & 0 \\ 0 & * \end{bmatrix} \tag{14}$$

where $*$ *is a possibly nonzero entry.*

(4) *The* x_1 *entry of any vector in the nullspace of* A *is zero.*

To obtain smooth observability we will need the next lemma.

LEMMA 2. *Suppose* A *in* (13) *is a smooth function of* t *defined on an interval* $\mathcal{I}$:

(1) *If* A *is* 1-*full with respect to* x_1 *for each* $t \in \mathcal{I}$ *and* A *has constant rank, then*

there is a smooth $\Theta(t)$ *such that* ΘA *has the form* (14) *and* A *is said to be smoothly* 1-*full* [7].

(2) *If* A *has constant rank on* $\mathscr{I}$ *and* A *is* 1-*full on a dense subset of* $\mathscr{I}$, *then* A *is* 1-*full for every* $t \in \mathscr{I}$.

Proof. The first statement is Lemma 3.1 of [7]. To prove the second statement, suppose that A has constant rank. Then the nullspace of A has a smooth basis

$$\left\{ \begin{bmatrix} x_{1i}(t) \\ x_{2i}(t) \end{bmatrix} \text{for } i = 0, \cdots, r \right\}.$$

By assumption, $x_{1i} = 0$ on a dense subset of $\mathscr{I}$. Hence $x_{1i} \equiv 0$ by continuity and A is 1-full for all $t \in \mathscr{I}$. □

We need to restrict the class of descriptor systems that we consider.

DEFINITION 3. The descriptor system $E(t)x' + F(t)x = f(t)$ is *solvable* on the interval $\mathscr{I}$ if

(1) For every sufficiently smooth f on $\mathscr{I}$, there is a solution to the descriptor system.

(2) Solutions are defined on all of $\mathscr{I}$.

(3) Solutions are uniquely determined by their values at any t_0 in $\mathscr{I}$.

This definition of solvability does not require E to have constant rank nor for it to be possible to carry out the usual inversion algorithms involving coordinate changes and differentiations [7], [24]. For a given f, the initial values at time t_0 form a proper submanifold if E is singular. More detailed exposition on the basic properties of DAEs may be found in [3], [6], [9], and [17].

2. Observability characterization. In this section we will develop our characterization of observability for (2). For simplicity, let $b(t) = B(t)u(t)$ and assume that the descriptor system (2a) is solvable. Differentiating the equation (2a) j times and the equation (2b) k times gives the system of equations

$$[\mathscr{F}_j \quad \mathscr{E}_j] \begin{bmatrix} x \\ \mathbf{x}_j \end{bmatrix} = \mathbf{b}_j, \tag{15a}$$

$$\mathscr{C}_k \begin{bmatrix} x \\ \mathbf{x}_{k-1} \end{bmatrix} = \mathbf{y}_k \tag{15b}$$

where

$$\mathscr{F}_j = \begin{bmatrix} F \\ F' \\ \vdots \\ F^{(j)} \end{bmatrix}, \quad \mathbf{y}_k = \begin{bmatrix} y \\ y' \\ \vdots \\ y^{(k)} \end{bmatrix}, \quad \mathbf{b}_j = \begin{bmatrix} b \\ b' \\ \vdots \\ b^{(j)} \end{bmatrix}, \quad \mathbf{x}_j = \begin{bmatrix} x' \\ \vdots \\ x^{(j+1)} \end{bmatrix},$$

$$\mathscr{E}_j = \begin{bmatrix} E & 0 & \cdot & \cdot & 0 \\ E' + F & E & 0 & \cdot & \cdot \\ E'' + 2F' & 2E' + F & E & \ddots & \cdot \\ * & * & * & \ddots & 0 \\ E^{(j)} + jF^{(j-1)} & * & * & * & E \end{bmatrix},$$

$$\mathscr{C}_k = \left[\begin{array}{c|ccc} C & 0 & \cdot & \cdot & 0 \\ C' & C & 0 & \cdot & \cdot \\ C'' & 2C' & C & \ddots & \cdot \\ * & * & * & \ddots & 0 \\ C^{(k)} & * & * & * & C \end{array}\right] = [\tilde{\mathscr{C}}_k | \hat{\mathscr{C}}_k].$$

These equations suggest the following result.

PROPOSITION 1. *The descriptor system* (2) *is observable on the interval* $\mathcal{I}$ *if and only if there are* j, k, *with* $k \leqq j+1$ *such that the matrix*

$$\mathcal{O}_{j,k} = \left[\begin{array}{c|c} \mathcal{F}_j & \mathcal{E}_j \\ \hline \hat{\mathcal{C}}_k & [\hat{\mathcal{C}}_k\ 0_{(k+1)m\times(j+1-k)n}] \end{array}\right] \tag{16}$$

is 1-*full with respect to* x *on a dense subset of* $\mathcal{I}$.

Proof. It is clear that the 1-fullness of $\mathcal{O}_{j,k}$ on a dense subset of $\mathcal{I}$ implies observability. The necessity of $\mathcal{O}_{j,k}$ being 1-full will follow from the proof of Proposition 3 where bounds on j, k are derived. □

PROPOSITION 2. *If* $\mathcal{O}_{j,k}$ *is* 1-*full on a dense subset of* $\mathcal{I}$ *and has constant rank, then* (2) *is smoothly observable of order* (k, j).

Proof. Under the given assumption, using Lemma 2, there is a smooth Θ such that

$$\Theta\mathcal{O}_{j,k} = \begin{bmatrix} I_{n\times n} & 0 \\ 0 & * \end{bmatrix}.$$

Then (4) holds where $[L_0, \cdots, L_j, K_0, \cdots, K_k]$ are the first n rows of Θ. □

Propositions 1 and 2 are the types of results we are seeking in that the only symbolic operations that need to be performed are the differentiation of the given coefficients. However, since E can have variable rank, calculations need to be carried out in a certain order to avoid incorrect rank determinations in verifying the observability condition. Also, we need more precise information on the needed values of j, k. Finally, ways to reduce the amount of computation have to be considered.

All of these concerns can be carried out simultaneously as we investigate the structure of $\mathcal{O}_{j,k}$. The key turns out to be the following fundamental result from [7].

THEOREM 1. *Suppose that* (2a) *is solvable on the interval* $\mathcal{I}$ *and that* E, F *are* $2n$-*times differentiable. Then*

$$\mathcal{E}_i \text{ has constant rank on } \mathcal{I} \text{ for } i = n+1, \tag{17}$$

$$\mathcal{E}_i \text{ is 1-full with respect to } x' \text{ for } i = n+1, \tag{18}$$

$$[\mathcal{F}_i\ \mathcal{E}_i] \text{ has full row rank for } 1 \leqq i \leqq n+1. \tag{19}$$

If the coefficients E, F are infinitely differentiable, then Theorem 1 provides sufficient as well as necessary conditions for solvability. If (19) holds, then the smallest value of i that satisfies the conditions (17), (18) of Theorem 1 is called the *index* ν of the descriptor system (2a). For time-invariant descriptor systems, the index is the same as the index of the pencil $\lambda E + F$. However, for time-varying solvable descriptor systems, the pencil $\lambda E + F$ need not be regular, and if the pencil is regular, its index need not be that of the descriptor system.

Theorem 1 is important since it assures us that if the descriptor system (2a) is solvable, then $\mathcal{E}_j$ will have constant rank even if E does not. Thus a computation concerning $\mathcal{E}_j$ can be well conditioned.

We now need the following technical lemma.

LEMMA 3. *Suppose that* (2a) *is a solvable index* ν *descriptor system. Then for any* $\ell \geqq 0$, *the row echelon form of* $[\mathcal{E}_{\nu+\ell} | \mathcal{F}_{\nu+\ell} | \mathbf{b}_{\nu+\ell}]$ *is*

$$\left[\begin{array}{c|c|c|c} I_{n(\ell+1)\times n(\ell+1)} & 0 & \tilde{Q}_1 & \tilde{b}_1 \\ 0 & R & \tilde{Q}_2 & \tilde{b}_2 \\ 0_{\rho\times n(\ell+1)} & 0 & \mathcal{M} & \tilde{b}_3 \end{array}\right], \tag{20}$$

where R and $\mathscr{M}$ have full row rank. Furthermore, the solutions of $\mathscr{M}x = \tilde{b}_3$ are independent of ℓ.

Proof. The independence of the solutions of $\mathscr{M}x = \tilde{b}_3$ follows from [7]. Index the $n \times n$ block entries of $\mathscr{E}_j$ by $0 \leqq r \leqq j$, $0 \leqq s \leqq j$. Thus, for example, $\mathscr{E}_{0,0} = E$. Let $r \geqq 1$, $s \geqq 1$. Then for any $j \geqq s$, the (r, s) $n \times n$ block entry of $\mathscr{E}_j$ is generated by the recursion

$$(\mathscr{E}_j)_{r,s} = (\mathscr{E}_j)_{r-1,s-1} + \frac{d}{dt}(\mathscr{E}_j)_{r-1,s}. \tag{21}$$

To see (21), note that if the $r-1$ block row of $\mathscr{E}_j$ is

$$+ \cdots + \mathscr{A}x^{(s-1)} + \mathscr{B}x^{(s)} + \cdots,$$

then, upon differentiation, we get that the coefficient of $x^{(s)}$ in the rth row of $\mathscr{E}_j$ is $\mathscr{A} + \mathscr{B}'$, which is (21). For $j \geqq 0$ partition $\mathscr{E}_{j+1}$ as

$$\mathscr{E}_{j+1} = \begin{bmatrix} E & 0_{n \times n(j+1)} \\ * & \hat{\mathscr{E}}_j \end{bmatrix}. \tag{22}$$

Suppose that ν is such that $\mathscr{E}_\nu$ is 1-full with respect to x' and of constant rank. We first show that $\mathscr{E}_{\nu+1}$ is 1-full with respect to x', x''. It will be of constant rank from [7]. The nullspace of $\mathscr{E}_{\nu+1}$ consists of the solutions of

$$\mathscr{E}_{\nu+1} \begin{bmatrix} z_0 \\ \vdots \\ z_{\nu+1} \end{bmatrix} = 0, \tag{23}$$

which implies that

$$\mathscr{E}_\nu \begin{bmatrix} z_0 \\ \vdots \\ z_\nu \end{bmatrix} = 0. \tag{24}$$

By the 1-fullness of $\mathscr{E}_\nu$, we have $z_0 = 0$. Hence (23) implies that

$$\hat{\mathscr{E}}_\nu \begin{bmatrix} z_1 \\ \vdots \\ z_{\nu+1} \end{bmatrix} = 0. \tag{25}$$

But by (21),

$$\hat{\mathscr{E}}_\nu = \mathscr{E}_\nu + \left[\frac{d}{dt}(\mathscr{E}_{\nu^*}) \mid 0_{(\nu+1)n \times n}\right] \tag{26}$$

where $\mathscr{E}_{\nu^*}$ denotes the last ν block columns of $\mathscr{E}_\nu$. Note that (24) implies that

$$\mathscr{E}_{\nu^*} \begin{bmatrix} z_1 \\ \vdots \\ z_\nu \end{bmatrix} = 0. \tag{27}$$

Equations (25) and (26) then imply that

$$\mathscr{E}_\nu \begin{bmatrix} z_1 \\ \vdots \\ z_{\nu+1} \end{bmatrix} + \left[\frac{d}{dt}(\mathscr{E}_{\nu^*}) \mid 0\right] \begin{bmatrix} z_1 \\ \vdots \\ z_{\nu+1} \end{bmatrix} = 0. \tag{28}$$

Since $\mathscr{E}_\nu$ is 1-full with constant rank, there exists smooth $[U_0, \cdots, U_\nu]$ with U_i, which are $n \times n$ such that

$$[U_0, \cdots, U_\nu]\mathscr{E}_\nu = [I\, 0 \cdots 0] \tag{29}$$

and hence

$$[U_0, \cdots, U_\nu]\mathscr{E}_{\nu^*} = [0 \cdots 0]. \tag{30}$$

Differentiating (30) yields

$$[U_0', \cdots, U_\nu']\mathscr{E}_{\nu^*} + [U_0, \cdots, U_\nu]\mathscr{E}_{\nu^*}' = 0. \tag{31}$$

Multiplying (28) by $[U_0, \cdots, U_\nu]$ and using (31), we have

$$z_1 - [U_0', \cdots, U_\nu'][\mathscr{E}_{\nu^*}|0]\begin{bmatrix} z_1 \\ \vdots \\ z_{\nu+1} \end{bmatrix} = 0$$

or $z_1 = 0$ by (27), and $\mathscr{E}_{\nu+1}$ is 1-full with respect to x' and x''. The $\mathscr{E}_{\nu+\ell}$ case now follows by a simple induction argument. □

Suppose then that (2a) is solvable and index ν. As noted in the proof of Lemma 3, $[\mathscr{E}_i\ \mathscr{F}_i\ 0|\mathbf{b}_i]$ is a leading submatrix of $[\mathscr{E}_j\ \mathscr{F}_j|\mathbf{b}_j]$ for every $i \leqq j$. In particular, $[\mathscr{E}_\nu\ \mathscr{F}_\nu]$ has constant rank and is 1-full with respect to x'. Performing a QR (or singular value decomposition) or using row operations on $[\mathscr{E}_\nu\ \mathscr{F}_\nu]$, as discussed in [4] and [5] gives

$$\left[\begin{array}{cc|c|c} I_{n\times n} & 0 & \bar{Q}_1 & \bar{b}_1 \\ 0 & H & \bar{Q}_2 & \bar{b}_2 \\ 0_{\rho\times n} & 0 & \mathscr{M} & \bar{b}_3 \end{array}\right]. \tag{32}$$

The equation

$$\mathscr{M}x = \bar{b}_3 \tag{33}$$

determines the solution manifold of (2a) at time t. Furthermore, (33) shows that there is a ρ-dimensional projection of x that is observable since it is given by b and its derivatives. Thus we have only to observe the solutions of $x' = -\bar{Q}_1 x + \bar{b}_1$ on an $(n-\rho)$-dimensional invariant submanifold. From the classical theory for observability, we then have that it suffices to take $k = n - \rho - 1$. We shall give a rigorous justification of this argument after summarizing it in the next proposition.

PROPOSITION 3. *Suppose that* (2a) *is solvable with index* ν. *Let* $\rho = n(\nu+1) - \operatorname{rank}(\mathscr{E}_\nu)$. *Note that* $0 \leqq \rho \leqq n$ *with* $\rho = 0$ *if and only if* E *is nonsingular. Then* (2) *is observable if and only if* $\mathscr{O}_{j,k}$ *is* 1-*full with respect to* x *on a dense subset of* $\mathscr{I}$ *where* (j, k) *are any pair of nonnegative integers satisfying* $k \geqq (n-\rho-1), j \geqq \nu + k - 1$. *In particular, since* $\nu \leqq n$, *we may take* $k = n-1, j = 2n-2$.

Proof. If we perform the time-varying coordinate changes, $x = S(t)\bar{x}$, and premultiplication by $T(t)$, the new derivative arrays are related to the old by

$$[\bar{\mathscr{F}}_j\ \bar{\mathscr{E}}_j] = \mathscr{T}_j[\mathscr{F}_j\ \mathscr{E}_j]\mathscr{S}_j, \tag{34}$$

$$[\bar{\mathscr{C}}_k\ 0] = [\mathscr{C}_k\mathscr{S}_k\ 0] = [\mathscr{C}_k\ 0]\mathscr{S}_j, \tag{35}$$

where

$$\mathscr{X}_i = \begin{bmatrix} X & 0 & \cdot & \cdot & 0 \\ X' & X & 0 & \cdot & \cdot \\ X'' & 2X' & X & \ddots & \cdot \\ * & * & * & \ddots & 0 \\ X^{(i)} & * & * & * & X \end{bmatrix} \quad \text{for } X = S, T.$$

Thus for a given k, j, the 1-fullness of $\mathcal{O}_{j,k}$ is unchanged by time-varying coordinate changes. Using the structure theorem for solvable linear DAEs developed in [7] we may assume that (2) has the form

$$x_1' + E_1(t)x_2' + G(t)x_1 = B_1(t)u, \tag{36a}$$

$$N(t)x_2' + x_2 = B_2(t)u, \tag{36b}$$

$$y = C_1x_1 + C_2x_2, \tag{36c}$$

where x_1 is n_1-dimensional and x_2 is n_2-dimensional. In general, N will have variable rank and nonsmooth nullspace and range [7]. However, the operator $N(d/dt) + I$ is an invertible operator of the space of infinitely differentiable functions onto itself. In particular, for each u there is only one solution of (36b). Let $\mathcal{M}_2$ be the $\mathcal{M}$ matrix for the derivative array $[\mathcal{E}_j\ \mathcal{F}_j]$ for (36b). But then rank $(\mathcal{M}_2) = n_2$ since $\mathcal{M}$ determines the solution manifold of a descriptor system and $\mathcal{M}_1 = 0$ for (36a). Thus $\rho = n_2$. Accordingly, we have that there exists smooth L_i such that

$$x_2 = \sum_{i=0}^{\rho-1} L_i(t)(B_2u)^{(i)} \tag{37}$$

and x_2 is already observable independent of C.

Thus observability of (36) reduces to considering (36a), (36c), which is a classical nonsingular observability problem in the form

$$x_1' = -Gx_1 + p_1, \tag{38a}$$

$$y = C_1x_1 + p_2. \tag{38b}$$

We know that the observability of (38) can be determined from the derivative array by no more than $n_1 - 1$ differentiations of (38b), which requires $n_1 - 2$ differentiations of G, p_1 in (38a) [23]. However, p_1 requires a differentiation of x_2 so that $n_1 - 1 = n - \rho - 1$ differentiations are needed. □

COROLLARY 1. *If the descriptor system satisfies the assumptions of Proposition* 3 *and* $\mathcal{O}_{j,k}$ *has constant rank, then* (2) *is smoothly observable.*

2.1. Nonsingular systems. Before continuing, we will briefly discuss what happens when E is nonsingular. In this case, we have that $\rho = 0$ and $\nu = 0$ so that

$$\mathcal{O}_{n-1,n-1} = \begin{bmatrix} [\mathcal{F}_{n-1} \quad \mathcal{E}_{n-1}] \\ [\mathcal{C}_{n-1}] \end{bmatrix}. \tag{39}$$

Since E is now assumed nonsingular, so is $\mathcal{E}_j$ for any j. Then $\mathcal{O}_{n-1,n-1}$ is 1-full with respect to x if and only if

$$\mathcal{W}_{n-1} = \tilde{C}_{n-1} - \hat{C}_{n-1}\mathcal{E}_{n-1}^{-1}\mathcal{F}_{n-1} \tag{40}$$

has full column rank. If we let $A = -E^{-1}F$ so that the differential equation is $x' = Ax + E^{-1}Bu$, then $\mathcal{W}_{n-1}$ is precisely the usual observability matrix [2]

$$\mathcal{W}_{n-1} = \begin{bmatrix} C \\ C' + CA \\ \vdots \\ C_{n-1} + C_{n-2}A \end{bmatrix} \tag{41}$$

where $C_i = C_{i-1}A + C_{i-1}'$ and $C_0 = C$. If $j > k - 1$, define

$$\mathcal{W}_k = \tilde{\mathcal{C}}_k - [\hat{\mathcal{C}}_k\ 0]\mathcal{E}_j^{-1}\mathcal{F}_j.$$

Uniform observability is defined in [23] to be that $\mathcal{W}_{n-1}$ has full rank for all $t \in \mathcal{I}$. Smooth observability will follow if $\mathcal{W}_j$ has full rank for all $t \in \mathcal{I}$ for some j that does not depend on t. Example 2 gives an example where $\mathcal{W}_j$ has full rank for all t but only for a $j > n$.

From a theoretical point of view, our approach is not saying anything new about the nonsingular case. However, there are other considerations that may make it preferable to utilize (15a), (15b) rather than (41). One such situation is when it is desirable to avoid the inversion of E. If E has some simple structure such as sparsity or bandedness, this structure can be lost in the inversion. Also, if E is time varying, then the inversion has to be done symbolically, and the resulting expressions differentiated repeatedly and multiplied symbolically. For even moderate-sized problems this can lead to a major expansion in the complexity of the expressions involved. In these types of problems it is much quicker to proceed numerically from the array $\mathcal{O}_{j,k}$, which has been computed with the minimal amount of expression expansion possible.

Many control problems have a structure that can be exploited in working with (15) [8].

Another advantage of working with $\mathcal{O}_{j,k}$ directly arises when we are dealing with problems where ranks, or perhaps even the index, change with parameter values. Many symbolic packages produce what are sometimes referred to as generic solutions. That is, given the equation $kx = 0$, the solution is given as $x = 0$ if k is a parameter. However, when dealing with descriptor systems, the case where $k = 0$ may be important. In the solution of the complicated nongeneric linear algebra problems that can occur in solving descriptor systems, this sort of behavior may be much more subtle.

2.2. Analytic systems. If the coefficients are real analytic, then we can make a stronger statement. The key is the following lemma.

LEMMA 4. *Suppose that $H(t)$ is an $m \times n$ real analytic matrix function defined on an open interval containing the closed bounded interval $\mathcal{I}$. Let*

$$H[j] = \begin{bmatrix} H \\ H' \\ \vdots \\ H^{(j)} \end{bmatrix}.$$

Suppose that $H[k]$ has full column rank at some $t_0 \in \mathcal{I}$. Then there is a j such that $H[j]$ has full column rank for all $t \in \mathcal{I}$.

Proof. Suppose that there is a k and a $t_0 \in \mathcal{I}$ such that $H[k]$ has full column rank at t_0. Then the real analyticity of H on an open set containing the closure of the interval $\mathcal{I}$ implies that $H[k]$ has full column rank at all but a finite number of points $t_1, \cdots, t_r$ in $\mathcal{I}$. Let $\mathcal{N}_{j,p}$ be the nullspace of $H[j]$ at time t_p. Clearly, $\mathcal{N}_{i,p} \subset \mathcal{N}_{j,p}$ if $i \geqq j$. Let $\mathcal{N}_p = \cap_{i \geq 0} \mathcal{N}_{i,p}$. If $\mathcal{N}_p \neq 0$, let v_p be a nonzero vector in $\mathcal{N}_p$. Then $\psi(t) = H(t)v_p$ is a real analytic function, all of whose derivatives vanish at t_p. Thus $\psi(t) \equiv 0$. But this implies that $H[j]v_p = 0$ for all t, j, which is a contradiction. Suppose then that $\mathcal{N}_p = 0$. But then $\mathcal{N}_{\mu_p,p} = 0$ for some μ_p, since $\mathcal{N}_p$ is the intersection of a nonincreasing chain of subspaces of a finite-dimensional vector space. Thus $H[\mu_p]$ will have full column rank. Let $\mu = \max\{k, \mu_1, \cdots, \mu_r\}$. Then $H[\mu]$ will have full column rank for all $t \in \mathcal{I}$. □

PROPOSITION 4. *The solvable system (2) with E, F, B, C real analytic is observable if and only if it is smoothly observable. Furthermore, it is smoothly observable if and only if $\mathcal{O}_{j,k}$ is 1-full with constant rank for some (j, k).*

Proof. Assume that (2) is solvable and that E, F, B, C are real analytic. It suffices to show that observable implies smoothly observable. The real analyticity implies [11]

that we may take $E_1 = 0$ in (36) and G, B_i, N, C_i are real analytic, as are the L_i in (37). Another analytic coordinate change gives $G = 0$ in (36a). Thus we can consider the nonsingular case. The proposition now follows by applying Lemma 4 to (41) with $A = 0$ and $C_i = C^{(i)}$. □

If we modify Example 1 by letting $y = t^s x$ where s is a nonnegative integer, and assume $0 \in \mathscr{I}$, we see that it is impossible to get a general upper bound for the amount of differentiation needed for smooth observability. However, having to perform a large number of extra differentiations in order to get smooth observability seems unlikely.

3. External behavior. In the systems theory literature the problem of representing the external behavior of a system, and of determining the external behavior given a representation, is frequently discussed ([22], [25], [28]). In this section we shall show how our characterizations of observability can be used to derive characterizations of the external behavior for (2). Our results do not follow from those of [22] since that paper makes a constant rank assumption at intermediate steps of the derivation and we allow these submatrices to have variable rank.

DEFINITION 4. The *external behavior* of (2) is the set $\Sigma_e = \{(y, u) \mid y, u$ are functions satisfying (2) for some state function $x(t)\}$.

An external description is sometimes called an input–output representation ([25], [28]).

DEFINITION 5. An *external description* of the system (2) is a set of equations

$$R(t, y, y', \cdots, y^{(\ell)}, u, u', \cdots, u^{(r)}) = 0 \tag{42}$$

with R continuous, such that the external behavior Σ_e of (2) is precisely the set of (y, u) satisfying (42).

Grimm [18] defines external behavior in terms of the Laplace transforms of y, u. Alternatively, using the notation of (3), the external behavior can be defined as the set of (u, y) such that $\begin{bmatrix} Bu \\ y \end{bmatrix}$ is in the range of $\tilde{E}(d/dt) + \tilde{F}$.

Suppose that $\mathcal{O}_{j,k}$ for (2) is 1-full with respect to the x variable and constant rank. Then

$$\begin{aligned} x &= \Psi(t, y, \cdots, y^{(k)}, u, \cdots, u^{(j)}) \\ &= \pi_1 \mathcal{O}_{j,k}^{\dagger} \begin{bmatrix} \mathbf{b}_j \\ \mathbf{y}_k \end{bmatrix}, \end{aligned} \tag{43}$$

where π_1 is the projection onto the first n coordinates. Define the functions R_1, R_2 by

$$R_1(t, y, \cdots, y^{(k)}, u, \cdots, u^{(j)}) = E(t)\frac{d}{dt}(\Psi) - F(t)\Psi - B(t)u, \tag{44}$$

$$R_2(t, y, \cdots, y^{(k)}, u, \cdots, u^{(j)}) = [I - \mathcal{O}_{j,k}\mathcal{O}_{j,k}^{\dagger}] \begin{bmatrix} \mathbf{b}_j \\ \mathbf{y}_k \end{bmatrix}. \tag{45}$$

PROPOSITION 5. *Suppose that the system* (2) *has* $\mathcal{O}_{j,k}$ 1*-full with respect to* x *and constant rank so that* (2) *is smoothly observable of order* (k, j). *Then the external behavior is characterized by*

$$R_1 = 0, \qquad R_2 = 0,$$

where R_1, R_2 *are given by* (44), (45).

Proof. That $(y, u) \in \Sigma_e$ implies (y, u) satisfies R_1 and R_2 is clear. Suppose then that (y, u) satisfies $R_1 = 0$, $R_2 = 0$. Since $R_2 = 0$, the equations (15) are algebraically consistent.

Since $\mathcal{O}_{j,k}$ is 1-full with respect to x and constant rank, there is a unique smooth $\bar{x}$ given by

$$\bar{x} = \pi_1 \mathcal{O}_{j,k}^{\dagger} \begin{bmatrix} \mathbf{b}_j \\ \mathbf{y}_k \end{bmatrix},$$

which satisfies $R_1 = 0$. There remains only to show that $y = C\bar{x}$. However, $R_2 = 0$ implies that (15a), (15b) are consistent and the first block equation in (15b) is $y = C\bar{x}$. □

3.1. Nonsingular case. To illustrate the previous result, consider (2) with E nonsingular and $A = -E^{-1}F$. From (39), (41) and (15a), (15b), we have

$$\mathbf{y}_k = \mathcal{W}_k x + \mathbf{b}_j. \tag{46}$$

By the observability assumption, $\mathcal{W}_k$ has full column rank for large enough k. The functions R_1, R_2 are defined by

(47a)

$$R_1(t, y, \cdots, y^{(k+1)}, u, \cdots, u^{(k)}) = \frac{d}{dt}[\mathcal{W}_k^{\dagger}(\mathbf{y}_k - \mathbf{b}_k)] - F(t)[\mathcal{W}_k^{\dagger}(\mathbf{y}_k - \mathbf{b}_k)] - B(t)u,$$

(47b)
$$R_2(t, y, \cdots, y^{(k+1)}, u, \cdots, u^{(k)}) = [I - \mathcal{W}_k \mathcal{W}_k^{\dagger}](\mathbf{y}_k - \mathbf{b}_k).$$

PROPOSITION 6. *If* (2) *with E nonsingular is smoothly observable on the interval* $\mathcal{I}$ *so that* $\mathcal{W}_k$ *has full column rank on* $\mathcal{I}$, *then the external behavior of* (2) *is characterized by the external description* $R_1 = 0$, $R_2 = 0$ *where the* R_i *are given by* (47).

Example 5. For the simple system

$$x' = bu, \qquad y = tx$$

we have

$$\mathcal{W}_k = \begin{bmatrix} t \\ 1 \end{bmatrix}, \qquad \mathcal{W}_k^{\dagger} = (1+t^2)^{-1}[t \quad 1].$$

Then

$$R_1 = \frac{d}{dt}[(1+t^2)^{-1}(ty + y' - tbu)] - bu$$

and

$$R_2 = \begin{bmatrix} 1-(1+t^2)^{-1}t^2 & -t(1+t^2)^{-1} \\ -t(1+t^2)^{-1} & 1-(1+t^2)^{-1} \end{bmatrix} \begin{bmatrix} y \\ y' - tbu \end{bmatrix}.$$

The equations R_1, R_2 simplify to

$$(t^2+1)y'' + (t^3 - t)y' + (1-t^2)y - [(t^4 + t^2 + 2)b + (t^3 + t)b']u - (t^3 + t)bu' = 0$$

and

$$y - ty' + t^2 bu = 0,$$

respectively.

Other 1-inverses [10] can be used besides $\mathcal{W}_k^{\dagger}$. For example, (8) used

$$\begin{bmatrix} t^2 \\ 2t + t^2 \\ 2 + 4t + t^2 \end{bmatrix}^{-} = (2t^2 + 2t + 2)^{-1}[2 \quad -1 \quad 1].$$

4. Conclusion. This paper has examined the observability of time-varying descriptor systems. Characterizations of different types of observability have been given in terms of rank conditions on arrays made up of derivatives only of the original coefficients. All algebra can be carried out numerically.

The concept of smooth observability, which is weaker than uniform observability, has been introduced. It is shown that every real-analytic system that is (totally) observable is smoothly observable, even if it is not uniformly observable. This is a new result even in the nonsingular case when E is invertible. These ideas have been used to develop a characterization of the external behavior of smoothly observable descriptor systems.

Several problems remain. One is a discussion of how to actually carry out these procedures in an efficient manner. In the characterization of the external behavior, an alternative characterization that does not require differentiating the computed Ψ would be desirable. Finally, it would be interesting to establish what the natural dual concept is to smooth observability.

REFERENCES

[1] V. A. ARMENTANO, *The pencil $(sE - A)$ and controllabillity-observability for generalized linear systems: A geometric approach*, SIAM J. Control Optim., 4 (1986), pp. 616–638.

[2] S. BARNETT, *An Introduction to Mathematical Control Theory*, Oxford University Press, Oxford, 1975.

[3] K. E. BRENAN, S. L. CAMPBELL, AND L. R. PETZOLD, *The Numerical Solution of Initial Value Problems in Ordinary Differential-Algebraic Equations*, Elsevier, Amsterdam, the Netherlands, 1989.

[4] S. L. CAMPBELL, *The numerical solution of higher index linear time varying singular systems of differential equations*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 334–348.

[5] ———, *Rank deficient least squares and the numerical solution of linear singular implicit systems of differential equations*, in Linear Algebra and Its Role in Systems Theory, Contemporary Mathematics, 47, American Mathematical Society, Providence, RI, 1985, pp. 51–63.

[6] ———, *Index two linear time varying singular systems of differential equations* II, Circuits Systems & Signal Process., 5 (1986), pp. 97–108.

[7] ———, *A general form for solvable linear time varying singular systems of differential equations*, SIAM J. Math. Anal., 18 (1987), pp. 1101–1115.

[8] ———, *Control problem structure and the numerical solution of linear singular systems*, Math. Control Signals Systems, 1 (1988), pp. 73–87.

[9] ———, *A computational method for general higher index singular systems of differential equations*, IMACS Trans. Sci. Comput., Numerical and Applied Mathematics, Vol 1.2, International Association for Mathematics and Computers in Simulation, C. Brezinski, ed., 1989, pp. 555–560.

[10] S. L. CAMPBELL AND C. D. MEYER, JR., *Generalized Inverses of Linear Transformations*, Pitman, Boston, 1979.

[11] S. L. CAMPBELL AND L. R. PETZOLD, *Canonical forms and solvable singular systems of differential-algebraic equations*, SIAM J. Algebraic Discrete Methods, 4 (1983), pp. 517–521.

[12] M. A. CHRISTODOULOU AND P. N. PARASKEVOPOULOS, *Solvability, controllability, and observability of singular systems*, J. Optim. Theory Appl., 45 (1985), pp. 53–72.

[13] D. COBB, *Controllability, observability, and duality in singular systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 1076–1082.

[14] L. DAI, *Singular Control Systems*, Lecture Notes in Control and Information Sciences, Vol. 118, Springer-Verlag, Berlin, New York, 1989.

[15] E. EMRE, *Observers for nonlinear time-varying systems*, in Proc. 21st Allerton Conference on Communication Control and Computing, University of Illinois, Urbana, IL, 1983, pp. 564–572.

[16] H. FRANKOWSKA, *On controllability, observability and optimality of linear descriptor systems*, preprint, 1989.

[17] E. GREIPENTROG AND R. MÄRZ, *Differential Algebraic Equations and Their Numerical Treatment*, Teubner-Texte zur Math., No. 88, Teubner, Leipzig, 1986.

[18] J. GRIMM, *Realization and canonicity for implicit systems*, SIAM J. Control. Optim., 26 (1988), pp. 1331–1347.

[19] E. KRIENDLER AND P. E. SARACHIK, *On the concepts of observability and controllability*, IEEE Trans. Automat. Control, 64 (1964), pp. 129–136.

[20] F. L. LEWIS, *Fundamental reachability, and observability matrices for discrete descriptor systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 502–505.

[21] B. G. MERTZIOS, M. A. CHRISTODOULOU, B. L. SYRMOS, AND F. L. LEWIS, *Direct controllability and observability time domain conditions for singular systems*, IEEE Trans. Automat. Control, 33 (1988), pp. 788–790.

[22] A. J. VAN DER SCHAFT, *Representing a nonlinear state space system as a set of higher-order differential equations in the inputs and outputs*, Systems Control Lett., 12 (1989), pp. 151–160.

[23] L. M. SILVERMAN AND H. E. MEADOWS, *Controllability and observability in time-variable linear systems*, SIAM J. Control, 5 (1967), pp. 64–73.

[24] L. M. SILVERMAN, *Inversion of multivariable systems*, IEEE Trans. Automat. Control, 14 (1969), pp. 270–276.

[25] E. D. SONTAG, *Bilinear realizability is equivalent to existence of a singular affine differential i/o equation*, Systems Control Lett., 11 (1988), pp. 181–187.

[26] E. L. YIP AND R. F. SINCOVEC, *Solvability, controllability, and observability of continuous descriptor systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 702–707.

[27] L. WEISS, *The concepts of differential controllability and differential observability*, J. Math. Anal. Appl., 10 (1965), pp. 442–449.

[28] J. C. WILLEMS, *Input-output and state-space representations of finite-dimensional linear time-invariant systems*, Linear Algebra Appl., 50 (1983), pp. 581–608.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 497–520, July 1991

008

IMMITANCE-TYPE THREE-TERM SCHUR AND LEVINSON RECURSIONS FOR QUASI-TOEPLITZ COMPLEX HERMITIAN MATRICES*

Y. BISTRITZ†‡, H. LEV-ARI†, AND T. KAILATH†

Abstract. A comprehensive analysis is made of Schur- and Levinson-type algorithms for Toeplitz and quasi-Toeplitz matrices that have half the number of multiplications and the same number of additions as the classical algorithms. Several results of this type have appeared in the literature under the label "split algorithms." In this approach the reduction in computation is obtained by a two-step procedure: (i) the first step is a variable (or "recursion-type") transformation from the classical (i.e., "scattering") variables to a new (so-called, "immitance") set of variables, which by itself reduces the number of multiplications at the cost of increasing the number of additions; (ii) the second step achieves control of the number of additions by converting the two-term recursions into the lesser known (for discrete orthogonal polynomials) three-term recursions. In the Toeplitz case the new variables turn out to be the odd and even parts of the classical variables, leading to the terminology of split algorithms, but this feature is lost in the quasi-Toeplitz case. Nevertheless, the network-theoretic interpretation of a transformation from scattering to immittance variables can still be maintained. Certain judicious choices of free parameters have to be made in each case in order to achieve the maximum computational reduction. It is shown how these results yield efficient procedures for determining the inertia of a quasi-Toeplitz matrix and the location of roots of its "predictor" polynomials from the immittance-type three-term recursions. In particular, connections with the Bistritz stability test, which was the motivation for our study of the Levinson and Schur algorithms in this paper, are noted.

Key words. Levinson algorithm, Schur algorithm, Toeplitz matrices, fast immittance-type recursions

AMS(MOS) subject classifications. primary 65F05, 65F30; secondary 15A06

1. Introduction.

Several recent papers have introduced computationally efficient (three-term) versions of the well-known Levinson algorithm and the somewhat less well known Schur algorithm. Bistritz has obtained several tests [1]–[4] for the root distribution of polynomials with respect to the unit circle that involve only the even (or odd) parts of the polynomials, and needed only half the number of multiplications (and the same number of additions) as the well-known Schur–Cohn test [5]. Since the Schur–Cohn test is essentially a reverse (degree-reducing) form of the Levinson algorithm, as well as a particular case of the Schur algorithm, it was reasonable to expect that similar reductions in computational complexity could also be obtained for both the Levinson and the Schur algorithms. Indeed, Delsarte and Genin derived one such computationally improved version for both of these algorithms: in [6] and [7] they presented the so-called "split Levinson" and "split Schur" algorithms for symmetric Toeplitz matrices with *real entries*. The adjective "split" arises from the ability to work with the odd and even (or symmetric and skew-symmetric) parts of the polynomials involved in the usual Levinson algorithm. Such improved algorithms were also obtained, in a slightly different context, by Bube and Burridge [23]. In our previous work [8], [9] we proved that: (i) the same approach applies not only to Toeplitz but also to certain *quasi-Toeplitz* (or Bezoutian) matrices, where the polynomials in the improved Levinson algorithm are not symmetric or skew-symmetric and *cannot* be viewed as an even/odd split of the polynomials in the usual

* Received by the editors September 4, 1987; accepted for publication (in revised form) March 6, 1990. This research was supported in part by the Air Force Office of Scientific Research, Air Force Systems Command contract AF-83-0228 and by U.S. Army Research Office contract DAALO3-86-K-0045.

† Information Systems Laboratory, Stanford University, Stanford, California 94305 (bistritz@genius.tau.ac.il, levari@northeastern.edu, and kailath@isl.stanford.edu).

‡ Present address, Department of Electrical Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel. This author's research was supported in part by a Chaim Weizmann postdoctoral fellowship award.

Levinson algorithm, and (ii) there are *three* computationally efficient and different-in-form versions of the Levinson algorithm. Finally, we note that the extension of the Bistritz stability test to the complex case (i.e., to polynomials with complex coefficients) was derived independently by Delsarte, Genin, and Kamp [10] and by Bistritz [2] and that Krishna and Morgera [11], [12] were perhaps the first to publish complex versions of the split Levinson algorithm.

This paper explores in detail the possibility of reducing the computational complexity of both the Levinson and the Schur algorithms by studying the effects of *variable* or *recursion-type transformations* and of introducing *three-term recursions*. We believe that the comparison of alternative computational procedures should not be carried out solely in terms of their computational requirements: other attributes, such as numerical robustness or suitability for parallel implementation, may be more significant in certain applications. Therefore, we consider in this paper *all* $O(N^2)$ *alternatives* to the conventional (scattering-type, two-term) recursions. Having established an explicit characterization of all efficient alternatives of the Schur and Levinson algorithms, we are in a position to *prove* that one so-called balanced immittance-type three-term version of the recursions (coinciding with the recursion in [11] and [12] in the Toeplitz case) has the lowest computational requirements. This advantage of the balanced version over all other alternatives has not been established in previous publications (i.e., in [6], [7], [11], [12], [20]), because no comparison to other alternatives was available. Moreover, we also prove that *all efficient three-term equivalents of the Schur/Levinson recursions are related to each other by scaling*.

We present our results in the somewhat generalized context of *quasi-Toeplitz matrices with complex entries*. We do so not only for the sake of extending results otherwise known for Toeplitz matrices, but mainly to establish the fact that the structural form of the recursions and the reduction in computational requirements depend not upon the special (persymmetry) property of Toeplitz matrices, but instead upon their so-called displacement structure. In contrast, the approaches used in previous publications rely heavily on the persymmetry property.

Before proceeding to a more specific outline of the background and the contributions of this paper we may suggest that the reader might also find it useful to scan the remarks in the concluding section of the paper.

1.1. The Levinson algorithm for quasi-Toeplitz matrices. The Levinson algorithm is a fast method that recursively solves, for $n = 1, \cdots, N$, the set of linear equations

$$[a_{n,n} \cdots a_{n,1} \quad 1]\mathbf{R}_{0:n} = R_n^e[0 \cdots 0 \quad 1] \tag{1a}$$

for the unknowns $\{a_{n,i}, R_n^e\}$, where $\mathbf{R}_{0:n}$ is the $(n+1)\times(n+1)$ leading submatrix of $\mathbf{R}_{0:N}$. The system matrix $\mathbf{R}_{0:N}$ is either a square Hermitian *Toeplitz* matrix, say

$$\mathbf{T}_{0:N} = \{c_{i-j}; 0 \leqq i, j \leqq N\}, \tag{1b}$$

or a square Hermitian *quasi-Toeplitz* matrix, i.e., one of the form

$$\mathbf{R}_{0:N} = \mathbf{H}\mathbf{T}_{0:N}\mathbf{H}^*, \tag{1c}$$

where $\mathbf{H}$ is a lower-triangular Toeplitz matrix of size $(N+1)\times(N+1)$, and the asterisk $(^*)$ denotes Hermitian transpose (complex conjugate for scalars). An alternative characterization of quasi-Toeplitz matrices is that they have *displacement inertia* (1, 1), as defined in [13], viz., $\mathbf{R}_{0:N}$ is such that the displacement matrix $\mathbf{R}_{0:N} - \mathbf{Z}\mathbf{R}_{0:N}\mathbf{Z}^*$ has one positive eigenvalue and one negative eigenvalue (and $N+1-2$ zero eigenvalues), where

$\mathbf{Z}$ is a matrix with unity elements on the first subdiagonal and zeros elsewhere. This means that

$$\mathbf{R}_{0:N} - \mathbf{Z}\mathbf{R}_{0:N}\mathbf{Z}^* = \mathbf{u}_0\mathbf{u}_0^* - \mathbf{v}_0\mathbf{v}_0^* \tag{1d}$$

for some column vectors $\mathbf{u}_0$, $\mathbf{v}_0$. For notational simplicity in further analysis we shall scale the matrix $\mathbf{R}_{0:N} = \{r_{i,j};\ 0 \leqq i, j \leqq N\}$ so that its top-left element is

$$r_{0,0} = 1. \tag{1e}$$

Therefore, in particular, Toeplitz matrices must satisfy $c_0 = 1$ and, consequently, the lower-triangular Toeplitz matrix $\mathbf{H}$ in (1c) must have unity diagonal elements. We may note that for certain choices of $\{\mathbf{u}_0, \mathbf{v}_0\}$ the matrix $\mathbf{R}_{0:N}$ becomes a so-called unit-circle Bezoutian, familiar from stability theory (see the discussion at the end of § 5).

Following [14], we say that $\mathbf{R}_{0:N}$ is *admissible* if there exists a scalar ρ such that

$$\mathbf{u}_0 - \rho\mathbf{v}_0 = [1 \quad 0 \cdots 0]^T. \tag{1f}$$

If $\mathbf{R}_{0:N}$ is admissible, then it is always possible to choose $\mathbf{v}_0(z)$ so that the corresponding ρ is real and nonnegative. We should also emphasize that by varying $\mathbf{H}$ in (1c), we obtain a *family* of quasi-Toeplitz matrices $\mathbf{R}_{0:N}$, all sharing the same reflection coefficients $\{k_n\}$. Some of these quasi-Toeplitz matrices are admissible and can be completely characterized by specifying the scalar coefficient $\rho \geqq 0$; others are nonadmissible and require a specification of $N + 1$ additional coefficients (see [14]).

Equation (1a) can be solved via the (generalized Levinson) recursions [14]

$$\begin{pmatrix} a_n(z) \\ b_n(z) \end{pmatrix} = L_n(z)\begin{pmatrix} a_{n-1}(z) \\ b_{n-1}(z) \end{pmatrix}, \qquad L_n(z) = \begin{pmatrix} z & -k_n \\ -k_n^* z & 1 \end{pmatrix}, \tag{2a}$$

where

$$a_n(z) := \sum_{i=0}^{n} a_{n,i} z^{n-i}, \tag{2b}$$

$b_n(z)$ is an auxiliary polynomial with coefficients $b_{n,i}$, viz.,

$$b_n(z) := \sum_{i=0}^{n} b_{n,i} z^i, \tag{2c}$$

and

$$a_0(z) = 1, \qquad b_0(z) = \rho. \tag{2d}$$

If $\mathbf{R}_{0:N}$ is not admissible, then its Levinson recursion is a further generalization of (2), which we shall not discuss in this paper, but which is indicated in [14]. For Toeplitz matrices, $\rho = 1$ and the recursions (2) become the well-known Levinson–Szegö recursions for the *orthogonal polynomials* $a_n(z)$ [15], with $b_n(z) \equiv a_n^{\#}(z) := z^n[a_n(z^{-*})]^*$, the *conjugate reverse polynomial of* $a_n(z)$. The *reflection coefficients* k_n are computed by certain inner-product formulas, which we discuss in further detail in § 4. We also recall here the readily verified fact that, by stacking the solutions of (1a) for $i = 0, 1, \cdots, N$, we can get the unique upper-diagonal-lower (UDL) triangular factorization of the inverse

of $\mathbf{R}_{0:N}$, $\mathbf{R}_{0:N}^{-1} = \mathbf{A}_{0:N}^{*}\mathbf{D}_{0:N}^{-1}\mathbf{A}_{0:N}$, where $\mathbf{D}_{0:N} = \text{diag}\{R_i^e;\ 0 \leqq i \leqq N\}$ and $\mathbf{A}_{0:N}$ is a lower-triangular matrix whose nth row contains the coefficients of $a_n(z)$, viz.,

$$\mathbf{A}_{0:N} = \begin{bmatrix} 1 & & & \\ a_{1,1} & 1 & & \\ \vdots & \vdots & \ddots & \\ a_{N,N} & a_{N,N-1} & \cdots & 1 \end{bmatrix}. \tag{3}$$

1.2. The Schur algorithm for quasi-Toeplitz matrices. The Schur algorithm is an alternative (and more direct) method for computing the reflection coefficients, which at the same time also determines the unique lower-diagonal-upper (LDU) triangular factorization of the matrix $\mathbf{R}_{0:N}$ itself, rather than its inverse [16]. It involves a recursion that we can rearrange (see Appendix B in [9]) in a form that is identical to the Levinson recursion (2a), viz.,

$$\begin{pmatrix} \tilde{u}_n(z) \\ \tilde{v}_n(z) \end{pmatrix} = L_n(z) \begin{pmatrix} \tilde{u}_{n-1}(z) \\ \tilde{v}_{n-1}(z) \end{pmatrix} \tag{4a}$$

where $u_n(z)$, $v_n(z)$ are power-series in z, viz.,

$$u_n(z) = \sum_{i=0}^{N} u_{n,i} z^i, \qquad v_n(z) = \sum_{i=0}^{N} v_{n,i} z^i, \tag{4b}$$

and $\tilde{u}_n(z)$ denotes conjugation of coefficients alone in the power series $u_n(z)$, i.e.,

$$\tilde{u}_n(z) := [u_n(z^*)]^*. \tag{4c}$$

Admissibility is not involved at all in the Schur recursion (4), which can be applied to every quasi-Toeplitz matrix.

The recursion starts with $u_0(z)$, $v_0(z)$. The coefficients $u_{0,i}$, $v_{0,i}$ of these polynomials are the elements of the column vectors $\mathbf{u}_0$, $\mathbf{v}_0$ in the displacement representation (1d) for $\mathbf{R}_{0:N}$. The representation (1d) of $\mathbf{R}_{0:N}$ is nonunique, as we can replace, for instance, the two-column matrix $[\mathbf{u}_0\ \mathbf{v}_0]$ by $[\mathbf{u}_0\ \mathbf{v}_0]\Theta(k)$, where

$$\Theta(k) := \frac{1}{\sqrt{1-|k|^2}} \begin{pmatrix} 1 & -k \\ -k^* & 1 \end{pmatrix}.$$

In particular, we can always select $\mathbf{u}_0$, $\mathbf{v}_0$ such that the first element of $\mathbf{v}_0$ vanishes, i.e.,

$$\mathbf{u}_0 := \begin{pmatrix} 1 \\ u_{0,1} \\ \vdots \\ u_{0,N} \end{pmatrix}, \qquad \mathbf{v}_0 := \begin{pmatrix} 0 \\ v_{0,1} \\ \vdots \\ v_{0,N} \end{pmatrix}, \tag{5}$$

where we use the convention (1e) that $r_{0,0} = 1$. In particular, for a Toeplitz matrix,

$$u_0(z) = \sum_{i=0}^{N} c_i z^i, \qquad v_0(z) = u_0(z) - 1, \tag{6}$$

which satisfies the constraint (5) with $u_{0i} = v_{0i}$ for $i > 0$. Moreover, the recursion (4a) imposes the same constraint upon all subsequent $v_n(z)$, i.e., $v_{n,n} = 0$ for all n. Note that, in addition, the first n coefficients of both $u_n(z)$ and $v_n(z)$ always equal zero.

The LDU factorization of $\mathbf{R}_{0:N}$ is obtained as follows: the nth diagonal element of the diagonal matrix $\mathbf{D}_{0:N}$ in $\mathbf{R}_{0:N} = \mathbf{L}_{0:N}\mathbf{D}_{0:N}\mathbf{L}_{0:N}^{*}$ is

$$d_n = u_{n,n} = R_n^e = \prod_{i=1}^{n} (1 - |k_i|^2), \qquad R_0^e = 1, \tag{7}$$

and $\mathbf{L}_{0:N}$ is a lower triangular matrix that has the coefficients of $u_n(z)/d_n$ as the elements in the nth column. The last equality in (7) is well known for Toeplitz matrices (see, e.g., [15]). In fact, it holds also for quasi-Toeplitz matrices because the $\{R_n^e\}$ corresponding to the quasi-Toeplitz matrix $\mathbf{R}_{0:N}$ of (1a), (1c) are independent of the matrix $\mathbf{H}$ and coincide with the $\{R_n^e\}$ that would appear in equations (1c) with the Toeplitz matrix $\mathbf{T}_{0:N}$. This is so because the lower-triangular matrix $\mathbf{H}$ must have unity diagonal elements in order to conform with the scaling convention (1e).

The computational costs of the Schur and Levinson algorithms are similar (see Table 2 for a summary of operation counts). The Schur algorithm is, however, more advantageous for *parallel* computation because it does not involve inner products (see, e.g., [16]). We shall now see how further computational reductions can be achieved for both algorithms.

1.3. Variable transformations and three-term recursions. Our approach to the problem of reducing computational requirements is different from that of Delsarte and Genin [6], [7] and that of Krishna and Morgera [11], [12], and it follows the method used in [8] and [9][1]: first we make a suitable variable transformation and then we convert the resulting two-term recursion into a three-term form. Thus consider first a linear transformation of the recursions, viz.,

$$\begin{pmatrix} f_n(z) \\ g_n(z) \end{pmatrix} := T_n \begin{pmatrix} a_n(z) \\ b_n(z) \end{pmatrix}, \qquad \begin{pmatrix} x_n(z) \\ y_n(z) \end{pmatrix} := T_n \begin{pmatrix} \tilde{u}_n(z) \\ \tilde{v}_n(z) \end{pmatrix}, \tag{8a}$$

which results in a modified set of two-term recursions. Namely,

$$\begin{pmatrix} f_n(z) \\ g_n(z) \end{pmatrix} = T_n L_n(z) T_{n-1}^{-1} \begin{pmatrix} f_{n-1}(z) \\ g_{n-1}(z) \end{pmatrix} \tag{8b}$$

and similarly for the Schur recursion. Note that the effect of the (nonsingular) matrices T_n is to transform the degree one polynomial matrices $L_n(z)$ into another set of matrices of the same nature. Thus, the modified recursions (8b) require $O(N^2)$ operations *for every choice of the transformation matrices* $\{T_n; 0 \leqq n \leqq N\}$.

An alternative form of the recursions is obtained by eliminating $g_n(z)$ altogether from (8b). This results in a three-term recursion, i.e., $f_n(z)$ is determined from $f_{n-1}(z)$ and $f_{n-2}(z)$, rather than from $f_{n-1}(z)$ and $g_{n-1}(z)$. The three-term version of the recursion may, in general, involve *polynomial division*, which significantly raises the computational requirements. We show in § 2 that the only way to avoid this additional computation is to choose

$$T_n = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \psi_n & 0 \\ 0 & \nu_n \end{pmatrix}, \tag{9a}$$

---

[1] It might be noted that the work in [6], [7], [11], and [12] deals only with the Toeplitz case, for which the reduction is obtained by working with the symmetric (or skew-symmetric) parts of the polynomial $a_n(z)$, leading to the name "split Levinson," introduced in [6]. However, in the quasi-Toeplitz case this symmetry is not available, though equivalent computational reductions can still be obtained.

where $\{\psi_n, \nu_n\}$ are complex scalars, and the ratio

$$\eta_n := \nu_n/\psi_n \tag{9b}$$

is constrained by the recursion

$$\eta_n = \frac{\eta_{n-1} + k_n}{1 + k_n^* \eta_{n-1}}, \qquad \eta_0 = 1. \tag{9c}$$

Since this recursion completely determines η_n in terms of the reflection coefficients $\{k_i;\ 1 \leqq i \leqq n\}$ and since $f_n(z) = \psi_n[a_n(z) + \eta_n b_n(z)]$, it follows that all efficient three-term equivalents of the Schur/Levinson recursions are related to each other by scaling (note that the η_i all have unit modulus). A suitable choice of the scaling coefficients $\{\psi_n\}$ may reduce the number of nontrivial coefficients in the recursion. As will be seen in § 2, the most efficient version involves a single complex multiplication[2] per recursion step, per coefficient. This unique, computationally efficient, version has the form

$$f_{n+1}^B(z) = (\delta_n z + \delta_n^*) f_n^B(z) - z f_{n-1}^B(z). \tag{10}$$

For real covariances, $\delta_n = \delta_n^*$, and this recursion reduces to the so called *balanced* recursion of [8] and [9]. Previous publications pointed out that scaling the $f_n(z)$ polynomials produces $O(N^2)$ equivalents of the balanced recursion, but did not show that every three-term immittance-type equivalent of the balanced recursion is produced in this manner.

It turns out that, in addition to the balanced recursion, there are only four distinct versions of the recursion with *two nontrivial coefficients*. They consist of two pairs: the monic/comonic pair and the dual-codual pair. The *monic/comonic recursions* have the form

$$f_{n+1}^M(z) = \left(z + \frac{\eta_n}{\eta_{n-1}}\right) f_n^M(z) - \lambda_n z f_{n-1}^M(z), \tag{11a}$$

$$f_{n+1}^{CM}(z) = \left(\frac{\eta_{n-1}}{\eta_n} z + 1\right) f_n^{CM}(z) - \lambda_n^* z f_{n-1}^{CM}(z), \tag{11b}$$

and reduce, for real covariances, to the monic recursion of [8] and [9]. The *dual/codual recursions* have the form

$$\lambda_{n+1} f_{n+1}^D(z) = \left(z + \frac{\eta_n}{\eta_{n-1}}\right) f_n^D(z) - z f_{n-1}^D(z), \tag{12a}$$

$$\lambda_{n+1}^* f_{n+1}^{CD}(z) = \left(\frac{\eta_{n-1}}{\eta_n} z + 1\right) f_n^{CD}(z) - z f_{n-1}^{CD}(z) \tag{12b}$$

and reduce, for real covariances, to the dual recursion of [8] and [9]. Though more computationally expensive than (10), we introduce (11)–(12) for completeness and because they may have other applications (e.g., (10)–(12) may have different degrees of numerical robustness).

---

[2] More precisely, the equivalent of a single complex multiplication, i.e., a total of four real multiplications. We remark also that in the real case we have three distinct versions of the recursion with a single nontrivial coefficient.

Of course, the recursions (10)–(12) also hold for $x_n(z)$ in the Schur algorithm. For instance, the balanced recursion for $x_n(z)$ is

$$x^B_{n+1}(z) = (\delta_n z + \delta_n^*) x^B_n(z) - z x^B_{n-1}(z),$$

and similarly for the other four versions.

The analysis in this paper extends the results of [8] and [9], including the useful *transmission line* interpretation. The ratio $v_n(z)/u_n(z)$ in the Schur algorithm is bounded by unity (for $|z| < 1$) and can be interpreted as the scattering function of a transmission line consisting of a cascade of (uniform) sections with different characteristic impedances. On the other hand, the ratio $x_n(z)/y_n(z)$ is positive-real (for $|z| < 1$) and can be interpreted as the impedance (or admittance) function of a related transmission line. For this reason we shall say that the recursions (2), (4) are of the *scattering type*, whereas the transformed recursions (i.e., those for (f_n, g_n) or for (x_n, y_n)) are of the *immittance type*.[3] Indeed, if we denote $s_n(z) := \tilde{v}_n(z)/\tilde{u}_n(z)$, then $c_n(z) := y_n(z)/x_n(z)$ is given by

$$c_n(z) = [1 - \eta_n s_n(z)]/[1 + \eta_n s_n(z)],$$

which we recognize as the well-known Cayley transform, mapping bounded functions into positive-real functions and vice versa.

The derivation of the three-term immittance-type recursions (10)–(12) is carried out in § 2. In order to propagate these recursions, beginning with the given covariance $\mathbf{R}_{0:N}$, we must also have formulas for computing the coefficients $\{\lambda_i, \delta_i\}$, similar to those used in the scattering-type formulation of the Levinson and the Schur algorithms to compute the reflection coefficients $\{k_i\}$; these calculations, which require the same number of multiplications as in the scattering-type recursions, are derived in §§ 3 and 4. We also present, in § 4, the relations required to reconstruct $\{g_n(z)\}$ from the three-term recursion for $\{f_n(z)\}$; this makes it possible to reconstruct the predictor polynomials $\{a_n(z)\}$ when necessary. Finally, § 5 briefly considers the relation between the immittance-type parameters $\{\delta_n\}$ and the *inertia* (i.e., the number of positive, null, and negative eigenvalues) of quasi-Toeplitz matrices. In particular, we show that any quasi-Toeplitz matrix $\mathbf{R}_{0:N}$ is congruent to a *tridiagonal* (Jacobi) matrix ∇_N whose nontrivial elements are the parameters $\{\delta_n\}$ (see (49)). Consequently, both matrices have the same inertia. This congruence relationship also appears in recent work of Delsarte and Genin (see, e.g., [20]). We present, in § 5, an efficient computational procedure for determining the inertia of ∇_N. We also show how to apply this procedure to locate the roots of the polynomial $a_N(z)$ of (2a) with respect to the unit circle.

2. Transformed recursions and three-term forms. We introduced in [8] and [9] the general linear transformation (8a), viz.,

$$\begin{pmatrix} f_n(z) \\ g_n(z) \end{pmatrix} := T_n \begin{pmatrix} a_n(z) \\ b_n(z) \end{pmatrix}, \tag{13}$$

where T_n is any constant nonsingular 2×2 matrix. This results in a transformed two-term recursion for $f_n(z)$, $g_n(z)$, namely,

$$\begin{pmatrix} f_n(z) \\ g_n(z) \end{pmatrix} = \begin{pmatrix} \alpha_n(z) & \beta_n(z) \\ \gamma_n(z) & \delta_n(z) \end{pmatrix} \begin{pmatrix} f_{n-1}(z) \\ g_{n-1}(z) \end{pmatrix}, \tag{14a}$$

[3] Bode coined the term *immittance* to denote both *impedance* and *admittance* [17].

where

$$\begin{pmatrix} \alpha_n(z) & \beta_n(z) \\ \gamma_n(z) & \delta_n(z) \end{pmatrix} := T_n \begin{pmatrix} z & -k_n \\ -k_n^* z & 1 \end{pmatrix} T_{n-1}^{-1}, \tag{14b}$$

from which we can obtain a three-term recursion for $f_n(z)$. The same transformation can be applied to the Schur recursions (4). The corresponding *transformed Schur recursions are obtained by replacing* here (and in the remainder of § 2) $a_n(z)$, $b_n(z)$ by $\tilde{u}_n(z)$, $\tilde{v}_n(z)$ and, similarly, $f_n(z)$, $g_n(z)$ by $x_n(z)$, $y_n(z)$.

Since the three-term recursion for $f_n(z)$ does not involve $g_n(z)$, it should not depend upon the elements in the second row of the transformation matrix T_n. We may, therefore, assume any particular form for the second row of T_n, for instance,

$$T_n := \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \psi_n & 0 \\ 0 & \nu_n \end{pmatrix}, \qquad \psi_n \neq 0 \neq \nu_n \tag{15a}$$

without affecting at all the three-term recursion for $f_n(z)$. Thus,

$$f_n(z) = \psi_n [a_n(z) + \eta_n b_n(z)], \tag{15b}$$

where

$$\eta_n := \nu_n / \psi_n. \tag{15c}$$

When the underlying covariances are real-valued, the choice $\eta_n = 1$ leads to the simplest recursions, which we have already analyzed in [9].

Following the general technique for converting two-term recursions into three-term recursions (see [9]) we obtain

$$f_{n+1}(z) = \left\{ \alpha_{n+1}(z) + \frac{\beta_{n+1}(z)\delta_n(z)}{\beta_n(z)} \right\} f_n(z) - \frac{\beta_{n+1}(z)\Delta_n(z)}{\beta_n(z)} f_{n-1}(z), \tag{16a}$$

where

$$\Delta_n(z) := \alpha_n(z)\delta_n(z) - \gamma_n(z)\beta_n(z) = \frac{\psi_n \nu_n}{\psi_{n-1}\nu_{n-1}}(1 - |k_n|^2)z. \tag{16b}$$

Also,

$$g_n(z) = \frac{\delta_n(z) f_n(z) - \Delta_n(z) f_{n-1}(z)}{\beta_n(z)}. \tag{16c}$$

Since $\alpha_n(z)$, $\beta_n(z)$, $\gamma_n(z)$, $\delta_n(z)$ are all polynomials of degree one (see explicit expressions in Appendix A), it follows that the three-term recursion involves *rational coefficients*, which significantly complicates the computation. Thus a simplified three-term recursion for $f_n(z)$ is possible if and only if these coefficients become polynomials in z of degree one, or less. We show in Appendix A that the rational coefficient $\beta_{n+1}(z)/\beta_n(z)$ becomes a constant if and only if

$$\frac{\eta_{n+1} - k_{n+1}}{1 - \eta_{n+1} k_{n+1}^*} = \eta_n, \tag{17a}$$

which can also be written as an ascending recursion in η_n, viz.,

$$\eta_{n+1} = \frac{\eta_n + k_{n+1}}{1 + \eta_n k_{n+1}^*}, \qquad \eta_0 := 1. \tag{17b}$$

Our choice of the initial condition $\eta_0 = 1$ is motivated by the observation that for real reflection coefficients $\{k_n\}$, this choice leads to $\eta_n = 1$ for all n, while for complex reflection coefficients it still yields

$$|\eta_n| = 1. \tag{17c}$$

With the ratio $\eta_n = \nu_n/\psi_n$ being constrained as in (17), the three-term recursion simplifies to (see Appendix A for derivation)

$$f_{n+1}(z) = \frac{\psi_{n+1}}{\psi_n}\frac{\eta_{n+1}-k_{n+1}}{\eta_n - k_n}\left\{\left(\frac{\eta_{n-1}}{\eta_n}z+1\right)f_n(z) - \frac{\psi_n}{\psi_{n-1}}(1-|k_n|^2)zf_{n-1}(z)\right\} \tag{18a}$$

and the auxiliary expression for $g_n(z)$ becomes

$$(z-1)g_n(z) = (\zeta_n z + \zeta_n^*)f_n(z) - \frac{\psi_n}{\psi_{n-1}}(1-\eta_n k_n^*)(\zeta_n+\zeta_n^*)zf_{n-1}(z), \tag{18b}$$

where

$$\zeta_n := \frac{1+\eta_n k_n^*}{1-\eta_n k_n^*}. \tag{18c}$$

These expressions reduce, for $\eta_n = 1$ (for real k_n), to (16a), (16b) in [8] and [9].

To initialize the three-term recursion for $\{f_n(z)\}$, one needs to know both $f_0(z)$ *and* $f_1(z)$. The definition (8), combined with the two-term recursion (2), implies that

$$f_0(z) = \psi_0\{a_0(z)+b_0(z)\}, \qquad f_1(z) = \psi_1(\eta_1 - k_1)\{za_0(z)+b_0(z)\}, \tag{19}$$

but it will be more convenient to have the initial conditions on $f_{-1}(z)$ and $f_0(z)$. Both (18a) and (18b) imply that

$$zf_{-1}(z) = \frac{\psi_{-1}}{1-|k_0|^2}\{(1+\eta_{-1}k_0^*z)a_0(z) + \eta_{-1}(\eta_{-1}z+k_0)b_0(z)\}.$$

These initial conditions involve the undefined quantities ψ_0, ψ_{-1}, η_{-1}, and k_0. We show in Appendix B that a consistent choice for these quantities is

$$\psi_0 = (1-k_0)^{-1}, \qquad \psi_{-1} = 1+k_0, \qquad \eta_{-1} = 1, \tag{20}$$

where k_0 is not subject to any constraints. In particular, if we choose $k_0 = -1$, then

$$\psi_0 = \tfrac{1}{2}, \qquad \psi_{-1} = 0, \qquad \eta_{-1} = 1, \tag{21a}$$

which results in

$$2zf_{-1}(z) = (1-z)\{a_0(z) - b_0(z)\}, \qquad 2f_0(z) = a_0(z) + b_0(z). \tag{21b}$$

This choice is motivated by the observation that the Levinson recursions for Toeplitz matrices are initialized with $a_0(z) = 1 = b_0(z)$, which reduces (21b) to $z\,f_{-1}(z) = 0$ and $f_0(z) = 1$.

A further reduction in complexity can be achieved by appropriately choosing the scaling factors ψ_n. There is a single choice that leads to recursions with *one* nontrivial coefficient, and four choices that lead to recursions with *two* nontrivial coefficients:

(1) *Balanced recursion*, obtained by choosing ψ_n to satisfy the constraints

$$\frac{\psi_{n+1}^B}{\psi_{n-1}^B}\frac{\eta_{n+1}-k_{n+1}}{\eta_n - k_n}(1-|k_n|^2) = 1, \qquad n \geqq 0 \tag{22a}$$

and

$$(\psi_n^B)^* = \eta_n \psi_n^B, \qquad n \geqq 0, \tag{22b}$$

resulting in

$$f_{n+1}^B(z) = (\delta_n z + \delta_n^*) f_n^B(z) - z f_{n-1}^B(z), \qquad n \geqq 0. \tag{23}$$

The special form of the multiplier of $f_n^B(z)$ with

$$\delta_n = \frac{\psi_{n+1}^B(\eta_{n+1} - k_{n+1})\eta_{n-1}}{\psi_n^B(\eta_n - k_n)\eta_n}$$

is established in Appendix B. As a consequence of the initialization (20), $\delta_0 = (1 - |k_0|^2)^{-1}(\psi_{-1}/\psi_0)^* = 1$. The remaining $\{\delta_n\}$ are related to each other by a recursion derived from the constraint on ψ_n^B, viz.,

$$\delta_n \delta_{n-1} \lambda_n = 1, \qquad n \geqq 1, \qquad \delta_0 = 1, \tag{24a}$$

where (note that $\eta_n^* = \eta_n^{-1}$ by (17c))

$$\lambda_n := (\eta_{n-1} + k_n)(\eta_{n-1} - k_{n-1})^*, \qquad n \geqq 1. \tag{24b}$$

The reason for the name "balanced" for (23) is that the recursions for ascending and descending indices are, essentially, identical.

Note that the balanced recursions involve only four real multiplications (i.e., the equivalent of one complex multiplication) per recursion step, per coefficient. In fact, the balanced recursion can be carried out as two interlacing three-term recursions that involve only real arithmetic [2]. Decomposing $f_n(z)$ into two real polynomials, viz.,

$$f_n(z) = S_n(z) + jA_n(z) \tag{25}$$

and separating the real and imaginary parts of (23), we obtain

$$S_{k+1}(z) = \delta_k^I(z+1)S_k(z) + \delta_k^R(z-1)A_k(z) - zS_{k-1}(z), \tag{26a}$$

$$A_{k+1}(z) = \delta_i^I(z+1)A_k(z) - \delta_k^R(z-1)S_k(z) - zA_{k-1}(z), \tag{26b}$$

where δ_k^R and δ_k^I denote the real and imaginary parts of δ_k, respectively. The recursions (26) involve four real multiplications and eight real additions per recursion step, per polynomial coefficient. In the Toeplitz case, $S_n(z)$ and $A_n(z)$ are, respectively, symmetric and skew-symmetric, and only half of their coefficients need to be computed.

(2) *Monic recursion*, obtained by choosing ψ_n to satisfy the constraint

$$\psi_n^M(1 - \eta_n k_n^*) = 1, \qquad n \geqq 0 \tag{27}$$

resulting in

$$f_{n+1}^M(z) = \left(z + \frac{\eta_n}{\eta_{n-1}}\right) f_n^M(z) - \lambda_n z f_{n-1}^M(z), \qquad n \geqq 0. \tag{28}$$

Note that as a consequence of the initialization (20), $\lambda_0 = \psi_0(1 - |k_0|^2)/(\eta_{-1}\psi_{-1}) = 1$ also. This recursion involves *two complex multiplications* per recursion step, per coefficient, and is therefore, in general, inferior to the balanced recursion. The reason for the name "monic" for (28) is that, with the appropriate initialization, $\{f_n^M(z)\}$ in the Levinson recursion are monic polynomials.

(3) *Comonic recursion*, obtained by choosing ψ_n to satisfy the constraint

$$\psi_n^{CM}(\eta_n - k_n) = 1, \qquad n \geqq 0 \tag{29}$$

resulting in the comonic recursion (11b), which has the same computational complexity as the monic recursion (28). The reason for the name "comonic" is that the $f_n^{CM}(z)$ in the Levinson recursion for *Toeplitz matrices* are comonic polynomials. However, the same property does not hold for other (quasi-Toeplitz) matrices.

(4) *Dual recursion*, obtained by choosing ψ_n to satisfy the constraint

$$\eta_n \psi_n^D (1 - |k_n|^2) = \eta_{n-1} \psi_{n-1}^D, \qquad n \geqq 0 \tag{30}$$

resulting in

$$\lambda_{n+1} f_{n+1}^D(z) = \left(z + \frac{\eta_n}{\eta_{n-1}} \right) f_n^D(z) - z f_{n-1}^D(z), \qquad n \geqq 0. \tag{31}$$

This recursion also involves *two complex multiplications* per recursion step, per coefficient.

(5) *Codual recursion*, obtained by choosing ψ_n to satisfy the constraint

$$\psi_n^{CD} (1 - |k_n|^2) = \psi_{n-1}^{CD}, \qquad n \geqq 0 \tag{32}$$

resulting in the codual recursion (12b), which has the same computational complexity as the dual recursion (31).

Remark. Note that, in view of (15b), $f_n^B(z)$, $f_n^M(z)$, $f_n^{CM}(z)$, $f_n^D(z)$, $f_n^{CD}(z)$ are all proportional to $a_n(z) + \eta_n b_n(z)$ and, therefore, to each other. The coefficients of proportionality can be determined by comparing the leading coefficients in these polynomials. Since $f_n^M(z)$ is monic, it follows, for instance, that $f_n^B(z)/f_n^M(z) = \prod_{i=0}^{n-1} \delta_i$. We show in Appendix B that

$$f_n^M(z) = \xi_{n-1} f_n^B(z), \qquad f_n^{CM}(z) = \xi_{n-1}^* f_n^B(z), \tag{33a}$$

$$f_n^D(z) = \xi_n^{-1} f_n^B(z), \qquad f_n^{CD}(z) = \xi_n^{-*} f_n^B(z), \tag{33b}$$

where

$$\xi_n := \prod_{i=0}^{n} \delta_i^{-1}. \tag{33c}$$

The same proportionality coefficients also relate the various versions of $x_n(z)$ in the immittance-type Schur algorithm.

The recursions just described are incomplete because we have not given methods for computing the coefficients λ_n, δ_n in them from the given matrix $\mathbf{R}_{0:N}$. There are two generic methods of doing this—what we call the Schur-type, where these coefficients are computed as certain ratios, and the Levinson-type, where their computation involves certain inner products. Besides the fact that Schur-type algorithms are better adapted to parallel computation, we also note that the functions propagated in the Schur-type recursions yield the Cholesky factors of $\mathbf{R}_{0:N}$, while those in the Levinson-type recursion yield the factors of $\mathbf{R}_{0:N}^{-1}$. We could also use the Schur recursions to compute the coefficients and then, under the assumption of admissibility, use the coefficients to compute the polynomials in the Levinson recursions.

3. Immittance-type Schur algorithms. We first review the Schur method for computing the scattering-type reflection coefficients $\{k_n\}$. From (4) we note that since $[z^{-n} v_n(z)]_{z=0} = 0$, it follows that k_n is the ratio of two known coefficients, viz.,

$$k_n = \left. \frac{v_{n-1}(z)}{z u_{n-1}(z)} \right|_{z=0} = \frac{v_{n-1,n}}{u_{n-1,n-1}}. \tag{34}$$

TABLE 1
Immittance-type three-term Schur recursions for real covariances.

Balanced	Monic	Dual
$zx_{-1}(z) = \frac{1}{2}(1-z)[u_0(z) - v_0(z)]$		
$x_0(z) = \frac{1}{2}[u_0(z) + v_0(z)]$		
$\delta_0 = 1 = \lambda_0$		
for $n = 0, 1, 2, \cdots, N-1$ **do**		
$\delta_n = [zx^B_{n-1}(z)/x^B_n(z)]\|_{z=0}$ (*)	$\lambda_n = [x^M_n(z)/zx^M_{n-1}(z)]\|_{z=0}$ (*)	$\tilde{x}^D_{n+1}(z) := (z+1)x^D_n(z) - zx^D_{n-1}(z)$
$x^B_{n+1}(z) = \delta_n(z+1)x^B_n(z) - zx^B_{n-1}(z)$	$x^M_{n+1}(z) = (z+1)x^M_n(z) - \lambda_n zx^M_{n-1}(z)$	$\lambda_{n+1} = [z^{-(n+1)}\tilde{x}^D_{n+1}(z)]\|_{z=0}$ $x^D_{n+1}(z) = \lambda^{-1}_{n+1}\tilde{x}^D_{n+1}(z)$

(*) Skip this step for $n = 0$.

A similar approach can be used for the three-term immittance-type Schur recursion and it yields, for instance, the expression $\delta_n^* = [zx^B_{n-1}(z)/x^B_n(z)]_{z=0}$. Combining such expressions with the recursions and initial conditions (21)–(32), we obtain a family of complete Schur algorithms, which we summarize below.

3.1. Real covariances. The analysis of § 2 yields *three* computationally efficient sets of recursions, which are summarized in Table 1. Note that all three versions begin with the same initial conditions $x_{-1}(z)$, $x_0(z)$. Also, all three versions require a *single real multiplication and two real additions* per recursion per each coefficient of $x_{n+1}(z)$, as compared to *two real multiplications and two real additions* for the scattering-type Schur algorithm (4) for real quasi-Toeplitz covariances.

3.2. Complex covariances. The analysis of § 2 yields a *single* computationally efficient recursion (the balanced version), viz.,

$$x^B_{n+1}(z) = (\delta_n z + \delta_n^*)x^B_n(z) - zx^B_{n-1}(z), \qquad n \geqq 0, \tag{35a}$$

where

$$x^B_0(z) := \tfrac{1}{2}\{\tilde{u}_0(z) + \tilde{v}_0(z)\}, \qquad zx^B_{-1}(z) := \tfrac{1}{2}(1-z)\{\tilde{u}_0(z) - \tilde{v}_0(z)\} \tag{35b}$$

and, with the notation $x^B_n(z) = \sum_{i=0}^N x^B_{n,i}z^i$,

$$\delta_n^* := \left.\frac{zx^B_{n-1}(z)}{x^B_n(z)}\right|_{z=0} = \frac{x^B_{n-1,n-1}}{x^B_{n,n}}. \tag{35c}$$

We emphasize again that even though (35a) seems to involve two nontrivial coefficients, namely, δ_n and δ_n^*, it can be carried out by two real three-term recursions, similar to (26), and requires, in fact, only *four real multiplications and eight real additions* per recursion step per each coefficient of $x^B_{n+1}(z)$. This is half the number of multiplications and the same number of additions as compared to the scattering-type Schur algorithm (4) for complex covariances. The relative efficiency of the immittance-type algorithm over the scattering-type one is, therefore, the same for both real and complex covariances and for all quasi-Toeplitz matrices (see Table 2).

TABLE 2
Computation counts.

	Scattering		Immittance	
	Mult.	Add.	Mult.	Add.
Schur				
Real	$O(N^2)$	$O(N^2)$	$O(0.5N^2)$	$O(N^2)$
Complex	$O(4N^2)$	$O(4N^2)$	$O(2N^2)$	$O(4N^2)$
Levinson: quasi-Toeplitz				
Real	$O(1.5N^2)$	$O(1.5N^2)$	$O(N^2)$	$O(1.5N^2)$
Complex	$O(6N^2)$	$O(6N^2)$	$O(4N^2)$	$O(6N^2)$
Levinson: Toeplitz				
Real	$O(N^2)$	$O(N^2)$	$O(0.5N^2)$	$O(N^2)$
Complex	$O(4N^2)$	$O(4N^2)$	$O(2N^2)$	$O(4N^2)$

We use the notation $m = O(aN^p)$ to mean that m is a polynomial of degree p in N, viz., $m = aN^p + bN^{p-1} + \cdots$.

4. Immittance-type Levinson algorithms. The scattering-type Levinson recursions (2) involve the reflection coefficients $\{k_n\}$, which are usually computed via an inner-product formula[4]

$$k_n = \frac{1}{R^e_{n-1}} \mathbf{a}_{n-1}[c_1 \cdots c_n]^T, \tag{36}$$

where $\mathbf{a}_{n-1} := [a_{n-1,n-1} \cdots a_{n-1,1} \quad 1]$ is a row vector consisting of the coefficients of the polynomial $a_{n-1}(z)$, and R^e_n is updated by

$$R^e_n = (1 - |k_n|^2) R^e_{n-1}, \qquad R^e_0 = c_0 = 1. \tag{37}$$

Similarly, the immittance-type Levinson recursions (10)–(12) involve the recursion coefficients $\{\delta_n\}$, $\{\lambda_n\}$, which can also be computed via suitable inner-product formulas, as we presently show.

Let $\mathbf{R}_{0:N}$ be the quasi-Toeplitz covariance associated with the pair $\mathbf{u}_0$, $\mathbf{v}_0$ via (5). If this covariance is *admissible*, i.e., if $u_0(z) = 1 + \rho v_0(z)$ for some scalar ρ, then, as we have shown in Appendix B of [9] (see also [22]),

$$\mathbf{a}_n[1 \quad u_{0,1} \cdots u_{0,n}]^T = 0, \tag{38a}$$

$$\mathbf{b}_n[1 \quad u_{0,1} \cdots u_{0,n}]^T = \rho R^e_n \tag{38b}$$

for all $n \geqq 1$, where $\mathbf{a}_n$, $\mathbf{b}_n$ are row vectors consisting of the coefficients of $a_n(z)$, $b_n(z)$, respectively. Consequently,

$$\tau_n = \psi_n \eta_n \rho R^e_n, \qquad n \geqq 1, \tag{39a}$$

where

$$\tau_n := \mathbf{f}_n[1 \quad u_{0,1} \cdots u_{0,n}]^T \tag{39b}$$

[4] This expression for k_n has to be slightly modified for quasi-Toeplitz matrices (see [9]).

and $\mathbf{f}_n$ is the row vector consisting of the coefficients of the polynomial $f_n(z)$. We may extend (39a) to $n = 0$ and *define* τ_0 as

$$\tau_0 := \rho\psi_0\eta_0 R_0^e = \frac{\rho}{2}. \tag{39c}$$

It turns out that the recursion coefficients $\{\delta_n\}$, $\{\lambda_n\}$ can always be computed as ratios of subsequent τ_n. For instance, the coefficients of the balanced recursion (23) are given by (see Appendix B)

$$\delta_n := \frac{\psi_{n+1}}{\psi_n}\frac{\eta_{n+1}-k_{n+1}}{\eta_n-k_n}\frac{\eta_{n-1}}{\eta_n} = \frac{\psi_{n-1}\eta_{n-1}}{\psi_n\eta_n(1-|k_n|^2)},$$

which, by comparison with (39a), implies that

$$\delta_n = \frac{\tau_{n-1}^B}{\tau_n^B}, \qquad n \geqq 1. \tag{40}$$

Combining such expressions with the recursions and initial conditions (21)–(32) we obtain a family of complete Levinson algorithms, which we summarize below.

4.1. Real covariances. The analysis of § 2 yields *three* computationally efficient sets of recursions, which are summarized in Table 3. Note that all three versions begin with the same initial conditions $zf_{-1}(z) = \frac{1}{2}(1 - z)(1 - \rho)$, $f_0(z) = \frac{1}{2}(1 + \rho)$, which are the initial conditions presented in [9] (but differ from those in [8] in also allowing $\rho = 0$). All three recursions require a *single real multiplication and two real additions* per recursion per each coefficient of $f_{n+1}(z)$, as compared to *two real multiplications and two real additions* for the scattering-type Levinson algorithm for real quasi-Toeplitz covariances. The computation of the recursion coefficients via (39b) requires one inner-product and one division per recursion, which is the same as in the scattering-type Levinson algorithm. In the Toeplitz case the (conjugate) symmetry of the polynomial $f_n(z)$ results in a further reduction of the computational requirements (see Table 2).

TABLE 3
Immittance-type three-term Levinson recursions for real covariances.

Balanced	Monic	Dual
$zf_{-1}(z) = \frac{1}{2}(1-z)(1-\rho)$		
$f_0(z) = \frac{1}{2}(1+\rho)$		
$\tau_0 = \frac{1}{2}\rho, \quad \delta_0 = 1 = \lambda_0$		
for $n = 0, 1, 2, \cdots, N-1$ **do**		
$f_{n+1}^B(z)$ $= \delta_n(z+1)f_n^B(z) - zf_{n-1}^B(z)$	$f_{n+1}^M(z)$ $= (z+1)f_n^M(z) - \lambda_n zf_{n-1}^M(z)$	$\tilde{f}_{n+1}^D(z)$ $:= (z+1)f_n^D(z) - zf_{n-1}^D(z)$
$\sum_i f_{n,i}^B z^i := f_n^B(z)$	$\sum_i f_{n,i}^M z^i := f_n^M(z)$	$\sum_i \tilde{f}_{n,i}^D z^i := \tilde{f}_n^D(z)$
$\tau_n^B = \sum_i f_{n,i}^B u_{0,i}$ (*)	$\tau_n^M = \sum_i f_{n,i}^M u_{0,i}$ (*)	$\lambda_{n+1} = 2\rho^{-1}\sum_i \tilde{f}_{n+1,i}^D u_{0,i}$
$\delta_n = \tau_{n-1}^B/\tau_n^B$ (*)	$\lambda_n = \tau_n^M/\tau_{n-1}^M$ (*)	$f_{n+1}^D(z) = \lambda_{n+1}^{-1}\tilde{f}_{n+1}^D(z)$

(*) Skip this step for $n = 0$.

4.2. Complex covariances. The analysis of § 2 yields a *single* computationally efficient recursion (the balanced version), viz.,

$$f_{n+1}^{B}(z)=(\delta_n z+\delta_n^{*})f_n^{B}(z)-zf_{n-1}^{B}(z), \qquad n\geqq 0, \tag{41a}$$

where

$$f_0^{B}(z):=\tfrac{1}{2}(1+\rho), \qquad zf_{-1}^{B}(z):=\tfrac{1}{2}(1-z)(1-\rho), \tag{41b}$$

and

$$\delta_n:=\frac{\tau_{n-1}^{B}}{\tau_n^{B}}, \qquad n\geqq 1, \tag{41c}$$

where τ_n is computed via the inner-product (39b), viz.,

$$\tau_n:=\mathbf{f}_n[1 \quad u_{0,1}\cdots u_{0,n}]^{T}, \qquad n\geqq 1, \qquad \tau_0=\frac{\rho}{2}. \tag{41d}$$

We emphasize that (41a) can be carried out by (26) and requires, like the corresponding Schur algorithm, only *four real multiplications and eight real additions* per recursion step per each coefficient of $f_{n+1}^{B}(z)$. This is half the number of multiplications and the same number of additions as compared to the scattering-type Levinson recursion for quasi-Toeplitz complex covariances. If the covariance matrix is not Toeplitz, the inner product formula has the same efficiency as in the scattering-type formulation for both the complex and real cases. Consequently, the relative efficiency of the immittance-type Levinson algorithms is the same for both real and complex quasi-Toeplitz covariances (a factor of 1.5, see Table 2) and is less than the (factor of two) relative efficiency of the corresponding Schur algorithms.

In the Toeplitz case, however, the symmetry in $f_n(z)$ can be exploited to simplify the computation of τ_n, viz.,

$$\begin{aligned}\tau_n=&[u_{0,0}+u_{0,n} \quad u_{0,1}+u_{0,n-1}\cdots u_{0,[n/2]}+u_{0,[(n+1)/2]}][S_{n,0} \quad S_{n,1}\cdots S_{n,[n/2]}]^{T}\\ &+j[u_{0,0}-u_{0,n} \quad u_{0,1}-u_{0,n-1}\cdots u_{0,[n/2]}-u_{0,[(n+1)/2]}][A_{n,0} \quad A_{n,1}\cdots A_{n,[n/2]}]^{T},\end{aligned} \tag{42}$$

where $S_n(z)$ and $A_n(z)$ are the real polynomials obtained from the real and imaginary parts of the coefficients of $f_n(z)$ (see (26)), and $[x]$ denotes the integer part of a real number x. If the Toeplitz covariance matrix is real, all $A_n(z)$ vanish, and the simplified formula (42) can be used for the three efficient versions of the Levinson algorithm for real Toeplitz matrices, as already mentioned in [8] and [9]. In summary, since both the inner-product formula (42) and the recursions (41a) have half the complexity of the corresponding scattering-type equivalents, the relative efficiency of the immittance-type algorithm is the same (i.e., a factor of 2) for both real and complex Toeplitz covariances, and is comparable to the corresponding efficiencies of the Schur algorithms (see Table 2).

4.3. Recovery of the orthogonal polynomials. The orthogonal polynomial $a_n(z)$ can always be recovered from $f_n(z)$ and $g_n(z)$ by inverting the recursion-type transformation (8), viz.,

$$a_n(z)=(2\psi_n)^{-1}\{f_n(z)+g_n(z)\},$$

which suggests the more convenient expression

$$a_n(z)=\frac{f_n(z)+g_n(z)}{f_n(\infty)+g_n(\infty)}, \tag{43}$$

where $f_n(\infty)$ indicates the leading coefficient of the polynomial $f_n(z)$.

In order to recover $g_n(z)$ from $\{f_i(z);\ 0 \leqq i \leqq n\}$, we observe that the balanced version of (18b) is

$$(z-1)g_n^B(z)=(\zeta_n z+\zeta_n^*)f_n^B(z)-2\mu_n z f_{n-1}^B(z), \tag{44a}$$

where

$$\mu_n := \frac{1+\zeta_n}{2\delta_n}. \tag{44b}$$

This is so because

$$\delta_n^{-1}=\frac{\eta_n\psi_n^B}{\eta_{n-1}\psi_{n-1}^B}(1-|k_n|^2)=\frac{\psi_n^B}{\psi_{n-1}^B}\frac{1-\eta_n k_n^*}{1-\eta_n^* k_n}(1-|k_n|^2),$$

and combining this with the identity (A.5b) simplifies the coefficient of $zf_{n-1}^B(z)$ in (18b) to

$$\frac{\psi_n^B}{\psi_{n-1}^B}(1-\eta_n k_n^*)(\zeta_n+\zeta_n^*)=\frac{2}{\delta_n(1-\eta_n k_n^*)}=\frac{1+\zeta_n}{\delta_n}.$$

An alternative expression for $g_n(z)$ can be obtained by using the balanced recursion (41a) to eliminate $zf_{n-1}^B(z)$ from (44a). This results in

$$(z-1)g_n^B(z)=2\mu_n f_{n+1}^B(z)-(z+1)f_n^B(z), \tag{45a}$$

which also implies that the coefficient μ_n can be computed directly via the expression

$$\mu_n=\frac{f_n^B(1)}{f_{n+1}^B(1)}. \tag{45b}$$

Note that the coefficient μ_n is *real*, even though both ζ_n and δ_n are, in general, complex. This follows from comparing (A.5a) with (B.3). As a consequence, the evaluation of $(z-1)g_n^B(z)$ via (45) involves a single multiplication of a complex-valued vector by a real scalar. In comparison, using (44) for the same purpose involves an additional multiplication of a complex-valued vector by a complex scalar. Furthermore, using (44) requires us to also compute ζ_n itself, in addition to μ_n. It follows from (45) and the balanced recursion (23) that

$$\frac{2\delta_n}{1+\zeta_n}=\mu_n^{-1}=2\delta_n^R-\mu_{n-1}, \tag{46}$$

which provides the real multiplier in the right-hand side of (44) and also, since δ_n is known, gives ζ_n itself for (44).

5. Inertia and stability. As is well known, a Hermitian Toeplitz matrix is positive definite if and only if the magnitude of its reflection coefficients is strictly less than one. More generally, the inertia of a Hermitian Toeplitz matrix coincides with the inertia of the diagonal matrix $\mathbf{D}_{0:N} = \operatorname{diag}\{R_n^e;\ 0 \leqq n \leqq N\}$ and can therefore be conveniently determined from the reflection coefficients via the relation (7). The same holds for quasi-Toeplitz matrices because all matrices congruent to a common Toeplitz matrix (as in (10)) share the same inertia [14].

The *real* coefficients $\{\mu_n\}$ that were introduced in (44)–(46) contain the same information as the $\{R_n^e\}$, because (see Appendix B for proof)

$$\frac{\mu_0}{\mu_n}=2\prod_{i=1}^{n}|\delta_i|^2\cdot(1-|k_i|^2). \tag{47}$$

In other words, the matrices $\mathbf{R}_{0:N}$, diag $\{R_n^e;\ 0 \leqq n \leqq N\}$, and μ_0^{-1} diag $\{\mu_n;\ 0 \leqq n \leqq N\}$ are all congruent to each other and, consequently, have the same inertia. Therefore, in particular, a quasi-Toeplitz matrix is positive definite if and only if the ratios $\{\mu_n/\mu_0\}$ are all strictly positive. Notice that this result is independent of the choice of initial conditions.

The most convenient way to compute the $\{\mu_n\}$ coefficients is via the recursion (46), viz.,

$$\mu_n^{-1} = 2\delta_n^R - \mu_{n-1}, \tag{48}$$

which is the immittance-type "analogue" of the recursive relation $R_n^e = R_{n-1}^e(1 - |k_n|^2)$. This relation also implies that there is yet another matrix with the same inertia as $\mathbf{R}_{0:N}$. Indeed, (48) implies that

$$\begin{bmatrix} 1 & & & & \\ \mu_0 & 1 & & & \\ & \mu_1 & 1 & & \\ & & \ddots & \ddots & \\ & & & \mu_{N-1} & 1 \end{bmatrix} \begin{bmatrix} \mu_0^{-1} & & & & \\ & \mu_1^{-1} & & & \\ & & \ddots & & \\ & & & & \mu_N^{-1} \end{bmatrix} \begin{bmatrix} 1 & \mu_0 & & & \\ & 1 & \mu_1 & & \\ & & 1 & \ddots & \\ & & & \ddots & \mu_{N-1} \\ & & & & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \mu_0^{-1} & 1 & & & \\ 1 & 2\,\mathrm{Re}\,\delta_1 & 1 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & 2\,\mathrm{Re}\,\delta_N \end{bmatrix} := \nabla_N, \tag{49}$$

which proves that (with $\mu_0 > 0$) the tridiagonal matrix ∇_N is congruent to diag $\{\mu_n^{-1};\ 0 \leqq n \leqq N\}$ and hence to $\mathbf{R}_{0:N}$.

We note that the principal minors of ∇_N are given by $\{\sigma_n;\ 0 \leqq n \leqq N\}$ where

$$\sigma_{-1} := 1, \qquad \sigma_0 = 2, \qquad \sigma_n = 2\delta_n^R \sigma_{n-1} - \sigma_{n-2}, \qquad n = 1, \cdots, N, \tag{50}$$

and by comparison with (48),

$$\mu_n = \sigma_{n-1}/\sigma_n. \tag{51}$$

Therefore, the inertia of $\mathbf{R}_{0:N}$ can be determined from the signs of the μ_n computed by (48) or from the sign *changes* in the σ_n sequence computed by (50). In conclusion, the matrix $\mathbf{R}_{0:N}$ is strongly regular if and only if all $\sigma_n \neq 0$ (equivalent to all $|k_n| \neq 1$); in such a case the number of its negative and positive eigenvalues is ν_N and $N - \nu_N$ where

$$\nu_N = n_-\{\mu_N, \cdots, \mu_1\} = \mathrm{Var}\,\{\sigma_N, \cdots, \sigma_0\}, \tag{52}$$

where n_- and Var stand, respectively, for the number of negative terms and the number of sign variations in the indicated sequences.

The magnitudes of the reflection coefficients $\{k_n\}$ of a Toeplitz matrix $\mathbf{T}_{0:N}$ also provide information about the location of the roots of the corresponding orthogonal polynomials $\{a_n(z)\}$ with respect to the unit circle. As is well known (see, e.g., [19]), $a_N(z)$ has all its roots strictly inside the unit circle if and only if $\mathbf{T}_{0:N}$ is *positive definite*, which by the foregoing discussion is equivalent to $\nabla_N > 0$. When $\mathbf{T}_{0:N}$ is indefinite, the magnitudes of $\{k_n\}$ determine the number of roots of $a_N(z)$ *inside* and *outside* the unit circle.

To be more specific, assume that we wish to locate the roots of $p(z)$, a given polynomial of degree N, with respect to the unit circle, and that $p(z)$ and $p^{\#}(z)$ do not have a common divisor, where $p^{\#}(z) := z^n[p(z^{-*})]^*$ denotes the conjugate reverse polynomial of $p(z)$. Then the Schur–Cohn test amounts to carrying out Levinson's recursion for Toeplitz matrices *in reverse order*, viz.,

$$za_{n-1}(z) = \frac{a_n(z) + k_n a_n^{\#}(z)}{1 - |k_n|^2}, \qquad k_n = -\frac{a_n(0)}{a_n^{\#}(0)} \tag{53a}$$

with the initialization $a_N(z) = p(z)$. This determines the reflection coefficients k_N, $k_{N-1}, \cdots, k_1$; the classical result of Cohn is that the number of roots of $p(z)$ inside (respectively, outside) the unit circle equals the number of positive (respectively, negative) P_n, $1 \leqq n \leqq N$, where

$$P_n := \prod_{i=n}^{N} (1 - |k_i|^2). \tag{53b}$$

If we use the balanced immittance-type recursions (10), then we shall have the coefficients $\{\delta_n\}$ or, equivalently, $\{\mu_n\}$ instead of the reflection coefficients $\{k_n\}$. The identity (47) implies that

$$\operatorname{sgn} P_n = \operatorname{sgn} \frac{\mu_{n-1}}{\mu_N} \quad \text{for } n = N, N-1, \cdots, 1 \tag{54a}$$

and consequently, the number of roots of $p(z)$ inside (respectively, outside) the unit circle equals the number of positive (respectively, negative) elements in the sequence

$$\left\{ \frac{\mu_{N-1}}{\mu_N}, \frac{\mu_{N-2}}{\mu_N}, \cdots, \frac{\mu_0}{\mu_N} \right\}. \tag{54b}$$

In particular, $p(z)$ is a stable polynomial (i.e., it has all its roots within the unit circle) if and only if all μ_n are positive (which will happen if and only if ∇_N is positive definite).

The balanced polynomials $f_n^B(z)$ are related to the orthogonal polynomials $a_n(z)$ of the Schur–Cohn test via the Toeplitz version of (15b), viz.,

$$f_n^B(z) = \psi_n^B[a_n(z) + \eta_n a_n^{\#}(z)],$$

where η_n and ψ_n^B have to satisfy the constraints (17) and (22), respectively. These constraints leave the parameters η_N, ψ_N^B, and ψ_{N-1}^B partially undetermined. Nevertheless, observe that $\{\mu_n\}$ are determined via (45b), viz.,

$$\mu_n = \frac{f_n^B(1)}{f_{n+1}^B(1)}, \qquad n = N, N-1, \cdots, 0 \tag{54c}$$

and that (54a)–(54c) hold *regardless* of the freedom in selecting the initialization.

The symmetric polynomials $f_n^B(z)$ are determined by propagating the balanced recursion (10) in reversed order, viz.,

$$zf_{n-1}^B(z) = (\delta_n z + \delta_n^*) f_n^B(z) - f_{n+1}^B(z), \qquad 1 \leqq n \leqq N-1, \tag{55a}$$

where

$$\delta_n = \left(\frac{f_{n+1}^B(0)}{f_n^B(0)} \right)^*. \tag{55b}$$

This recursion is initialized by $f_N^B(z)$ and $f_{N-1}^B(z)$ which, in turn, are determined by the parameters η_N, ψ_N^B, and ψ_{N-1}^B. Note that these three parameters determine all η_n for

$n < N$ (via (17a)), as well as all ψ_n^B for $n < N-1$ (via (22a)). The only constraint imposed on our initialization is (22b), i.e., $(\psi_N^B)^* = \eta_n \psi_N^B$, and $(\psi_{N-1}^B)^* = \eta_{N-1}\psi_{N-1}^B$. A particular choice that is consistent with this constraint is $\eta_N = 1$, $\psi_N^B = 1$, which results in

$$f_N^B(z) = p(z) + p^{\#}(z), \tag{56a}$$

where we used the fact that for Toeplitz matrices, $b_n(z) = a_n^{\#}(z)$, and where we set $a_N(z) = p(z)$, as in the Schur–Cohn procedure. Furthermore, we still maintain the property $|\eta_n| = 1$; in particular, $\eta_n = 1$ for matrices with real-valued elements.

Further simplification of the initial conditions for (55) may be achieved by a judicious choice of ψ_{N-1}^B, leading to a simplified expression for $f_{N-1}^B(z)$. An even simpler approach is to initialize (55) with $n = N$ rather than with $n = N-1$. This requires us to introduce the polynomial $f_{N+1}^B(z)$, which depends, via (55), on $f_N^B(z)$, $f_{N-1}^B(z)$, and the completely unconstrained parameter δ_N. The flexibility in selecting $f_{N-1}^B(z)$ and δ_N makes it possible to obtain a relatively simple expression for $f_{N+1}^B(z)$. Indeed, letting

$$\psi_{N-1}^B = 2\nu \frac{1-|k_N|^2}{1-k_N}, \qquad \delta_N = \nu \frac{1+k_N^*}{1-k_N^*} + \lambda,$$

where λ, ν are arbitrary positive constants, we find that

$$f_{N+1}^B(z) = q(z) + q^{\#}(z), \qquad q(z) := [\lambda(z+1) + \nu(z-1)]p(z). \tag{56b}$$

Moreover, these choices result in $\mu_N = (2\lambda)^{-1} > 0$, so that (54b) can be replaced by the sequence

$$\{\mu_{N-1}, \mu_{N-2}, \cdots, \mu_0\}. \tag{57a}$$

The number of negative elements in this sequence (i.e., the number of roots outside the unit circle) is also given by the number of sign changes in the sequence

$$\{f_N^B(1), f_{N-1}^B(1), \cdots, f_0^B(1)\}, \tag{57b}$$

which is always real-valued because: (i) $f_N^B(1) = p(1) + p^{\#}(1) = 2 \operatorname{Re} p(1)$ is real, (ii) the remaining $f_n^B(1)$ are obtained via $f_{n-1}^B(1) = \mu_{n-1} f_n^B(1)$, and (iii) μ_n are real.

Since μ_N (respectively, $f_{N+1}^B(1)$) does not appear in (57a) (respectively, (57b)), we can allow the limiting case $\lambda \to 0$ (with $\nu = 1$). This results in $q(z) = (z-1)p(z)$ so that

$$f_N^B(z) = p(z) + p^{\#}(z) = \frac{q(z) - q^{\#}(z)}{z-1}.$$

This is precisely the initialization that arises when the root-location procedure of Bistritz [1], [2] is applied to the *augmented polynomial* $q(z)$, which has the same root-distribution as $p(z)$ and an additional zero at $z = 1$. According to [1] and [2], the number of roots of $q(z)$ outside the unit circle equals the number of sign changes in the sequence

$$\{f_{N+1}^B(1), f_N^B(1), \cdots, f_0^B(1)\}.$$

Since the initial $f_{N+1}^B(1) = 0$ accounts for the zero at $z = 1$, we conclude that the remaining elements of this sequence determine the root-distribution of the polynomial $p(z)$. This coincides with our criterion (57b).

6. Concluding remarks. The Levinson and Schur algorithms showed how the Toeplitz (and quasi-Toeplitz) structure of linear equations could be used to provide an order of magnitude reduction in the amount of computation, from $O(N^3)$ to $O(N^2)$.

Normally we would not be too concerned with further reductions that do not affect the order of magnitude; however, the work of Bistritz [1]–[4] showed an alternative structure that achieves a reduction of exactly one-half in the number of multiplications. Such an improvement cannot be accidental, and that has been the motivation for the studies reported in this paper and our earlier paper [8], [9]. The first results of Bistritz (on stability tests) and of Delsarte and Genin (on the split Levinson algorithm) obtained this reduction in the amount of computation by carefully exploiting the persymmetry property of Toeplitz matrices. We were not completely satisfied with this approach because our earlier work on the Levinson algorithm showed that the algorithm could be generalized to close-to-Toeplitz matrices, and that this generalization was very simple for the class of (admissible) quasi-Toeplitz matrices, amounting essentially to a change in the initial conditions (see (2d)). Even though such non-Toeplitz matrices were not persymmetric and do not yield immittance symmetric polynomials, we were able to obtain a corresponding reduction in the number of multiplications, which seems to indicate that the persymmetry property does not fully explain the improved efficiency (this notwithstanding the fact that, at least in retrospect, the Levinson recursions for admissible quasi-Toeplitz matrices can be obtained from the usual Levinson algorithm by using the congruence (1c) and some algebraic manipulation).

The key to reduction in computation is really the proper use of two additional degrees of freedom that were always known but never fully exploited. These are:

(i) The possibility of linear transformations of the variables propagated in the Levinson and Schur algorithms, and especially the transformations (well known in circuit theory) between wave variables and immittance (voltage, current) variables;

(ii) The use of the three-term recursions (already noted in the classical work of Geronimus [18], [19]).

This is the approach developed in the present paper and in [8] and [9]. We may note that besides enabling a simple extension for Toeplitz to quasi-Toeplitz systems, our approach has also served to delimit the whole set of efficient Levinson and Schur algorithms.

Returning to the complexity reduction, we may remark that for us the main interest is not so much the reduction itself, which need not be significant in actual applications (e.g., studies of robustness and stability still need to be made), but more the reasons for the exact factor of two of reduction and the scope for its extension beyond the Toeplitz case. The simplicity of the two-step approach used in this paper showed that the same reduction in complexity could also be achieved for Hermitian quasi-Toeplitz matrices. How much further can they go? This is hard to say. However, our method of proof has recently enabled us to show that the reduction does *not* extend to *non-Hermitian* Toeplitz and quasi-Toeplitz matrices [21].

Appendix A. Derivation of general three-term recursions. It follows from (14b) that

$$\alpha_n(z)=\frac{\psi_n}{2\psi_{n-1}}\left\{(1-\eta_n k_n^*)z+\frac{\eta_n-k_n}{\eta_{n-1}}\right\}, \tag{A.1a}$$

$$\beta_n(z)=\frac{\psi_n}{2\psi_{n-1}}\left\{(1-\eta_n k_n^*)z-\frac{\eta_n-k_n}{\eta_{n-1}}\right\}, \tag{A.1b}$$

$$\gamma_n(z)=\frac{\psi_n}{2\psi_{n-1}}\left\{(1+\eta_n k_n^*)z-\frac{\eta_n+k_n}{\eta_{n-1}}\right\}, \tag{A.1c}$$

$$\delta_n(z)=\frac{\psi_n}{2\psi_{n-1}}\left\{(1+\eta_n k_n^*)z+\frac{\eta_n+k_n}{\eta_{n-1}}\right\}. \tag{A.1d}$$

Consequently,

$$\frac{\beta_{n+1}(z)}{\beta_n(z)} = \frac{\psi_{n+1}\psi_{n-1}\eta_{n-1}}{\psi_n^2\eta_n} \cdot \frac{\eta_n(1-\eta_{n+1}k_{n+1}^*)z-(\eta_{n+1}-k_{n+1})}{\eta_{n-1}(1-\eta_n k_n^*)z-(\eta_n-k_n)}.$$

This expression is independent of z if and only if

$$\frac{\eta_{n-1}(1-\eta_n k_n^*)}{\eta_n - k_n} = \mu$$

where μ is a constant independent of n. Therefore, $\{\eta_n\}$ are recursively determined via the recursion

$$\eta_n = \frac{\mu\eta_{n-1}+k_n}{1+k_n^*\mu\eta_{n-1}}, \tag{A.2}$$

which involves only two undetermined constants (μ and η_0). To be consistent with the real case, where $\eta_n = 1$ (see, e.g., [8], [9]), we choose $\eta_0 = 1$, $\mu = 1$, which results in the recursive relation (17). Note that we always get $|\eta_n| = 1$ with this choice of η_0, μ.

Incorporating the constraint (17) into the expressions (A.1) simplifies them to

$$\alpha_n(z) = \frac{\psi_n}{2\psi_{n-1}}(1-\eta_n k_n^*)(z+1), \tag{A.3a}$$

$$\beta_n(z) = \frac{\psi_n}{2\psi_{n-1}}(1-\eta_n k_n^*)(z-1), \tag{A.3b}$$

$$\gamma_n(z) = \frac{\psi_n}{2\psi_{n-1}}(1-\eta_n k_n^*)(\zeta_n z-\zeta_n^*), \tag{A.3c}$$

$$\delta_n(z) = \frac{\psi_n}{2\psi_{n-1}}(1-\eta_n k_n^*)(\zeta_n z+\zeta_n^*), \tag{A.3d}$$

where

$$\zeta_n := \frac{1+\eta_n k_n^*}{1-\eta_n k_n^*}. \tag{A.4}$$

The expressions (18a), (18b) for $f_{n+1}(z)$ and $g_n(z)$ are obtained by substituting (A.3) into (16a), (16c) and using the following easily established identities:

$$\frac{1+\zeta_n}{1+\zeta_n^*} = \frac{\eta_{n-1}}{\eta_n}, \tag{A.5a}$$

$$|\eta_n - k_n|^2(\zeta_n+\zeta_n^*) = 2(1-|k_n|^2). \tag{A.5b}$$

Appendix B. Properties of recursion coefficients. The constraint (22a), which characterizes the balanced recursion, implies that

$$\eta_{n+1}\frac{\psi_{n+1}^B}{(\psi_{n+1}^B)^*} = \eta_{n-1}\frac{\psi_{n-1}^B}{(\psi_{n-1}^B)^*}$$

where we have used (17) and the fact that $|\eta_n| = 1$. Therefore, (22b) follows for all n if we assume that it holds for $n = -1, 0$. Thus we must have

$$(\psi_{-1}^B)^* = \eta_{-1}\psi_{-1}^B, \qquad (\psi_0^B)^* = \psi_0.$$

It will be convenient, though by no means necessary, to have *the same initial conditions*

for all versions of the recursions. Taking this approach we conclude, via (27) and (29), that

$$\psi_0 = (1-k_0)^{-1} = (1-k_0^*)^{-1},$$

so that k_0 must be real and, consequently,

$$\eta_{-1} = \frac{1-k_0}{1-k_0^*} = 1.$$

This also proves that ψ_{-1} is real. In fact, (30) and (32) imply that

$$\psi_{-1} = \psi_0(1-|k_0|^2) = 1+k_0.$$

In summary, we can initialize all versions of the recursions with the same set of initial constants, viz.,

(B.1a) $$\psi_0 = (1-k_0)^{-1}, \qquad \psi_{-1} = 1+k_0$$

and

(B.1b) $$\eta_{-1} = 1 = \eta_0, \qquad \delta_0 = 1 = \lambda_0.$$

The only undetermined parameter is k_0, which can take any real value *except unity*. In particular, the initial conditions (21) are obtained by choosing $k_0 = -1$. Other simple choices of k_0, such as $k_0 = 0$, are also feasible.

Returning to establish the form (23) of the balanced recursions we denote the coefficient of $zf_n(z)$ in (18a) by δ_n and substitute into the latter (22a), (22b), viz.,

(B.2) $$\delta_n = \frac{\psi_{n+1}^B(\eta_{n+1}-k_{n+1})\eta_{n-1}}{\psi_n^B(\eta_n - k_n)\eta_n} = \frac{\psi_{n-1}^B\eta_{n-1}}{\psi_n^B(1-|k_n|^2)\eta_n} = \frac{(\psi_{n-1}^B)^*}{(\psi_n^B)^*(1-|k_n|^2)},$$

so that the coefficient of $f_n(z)$ in (23) is, indeed, $\delta_n z + \delta_n^*$. This also implies that

(B.3) $$\delta_n^* = \frac{\eta_n}{\eta_{n-1}}\delta_n$$

and, consequently, that

(B.4) $$\xi_n^* = \eta_n^{-1}\xi_n$$

where $\xi_n := \prod_{i=1}^{n}\delta_i^{-1}$, as in (33c). Similarly,

$$\lambda_n^* = (\delta_n^{-1}\delta_{n-1}^{-1})^* = \frac{\eta_{n-1}}{\eta_n}\frac{\eta_{n-2}}{\eta_{n-1}}\delta_n^{-1}\delta_{n-1}^{-1},$$

namely,

(B.5) $$\lambda_n^* = \frac{\eta_{n-2}}{\eta_n}\lambda_n.$$

The relations (33a), (33b) are established by a comparison of the leading coefficients in the polynomials $f_n^B(z)$, $f_n^M(z)$, etc. Since $f_0(z)$ is the same for all versions of the recursion and since $f_n^M(z)$ is monic, it follows, for instance, that

$$\frac{f_n^B(z)}{f_n^M(z)} = \prod_{i=0}^{n-1}\delta_i = \xi_{n-1}^{-1}$$

as well as

$$\frac{f_n^D(z)}{f_n^M(z)} = \prod_{i=1}^{n} \lambda_i^{-1} = \prod_{i=1}^{n} \delta_i \delta_{i-1} = \xi_n^{-1} \xi_{n-1}^{-1},$$

which implies that

$$\frac{f_n^D(z)}{f_n^B(z)} = \xi_n^{-1}.$$

The rest of the relations in (33) can be obtained in a similar manner.

In order to establish (47), we observe from (18b) that $\mu_{n-1} = f_{n-1}^B(1)/f_n^B(1)$ is given by

$$\mu_{n-1} = \frac{\psi_{n-1}^B}{\psi_n^B(1-\eta_n k_n^*)} = \psi_{n-1}^B \frac{\eta_{n-1}}{\psi_n^B(\eta_n - k_n)}$$

where the second equality invokes (17a). Consequently, we obtain, using (B.2),

$$\frac{\mu_{n-1}}{\mu_n} = \frac{\psi_{n-1}^B}{\psi_n^B} \cdot \frac{\psi_{n+1}^B \eta_{n-1}(\eta_{n+1} - k_{n+1})}{\psi_n^B \eta_n (\eta_n - k_n)} = \delta_n^*(1 - |k_n|^2) \cdot \delta_n$$

and, therefore,

$$\frac{\mu_m}{\mu_n} = \prod_{i=m+1}^{n} |\delta_i|^2 (1 - |k_i|^2). \tag{B.6}$$

This result does not depend on the choice of initialization.

REFERENCES

[1] Y. BISTRITZ, *Zero location with respect to the unit circle of discrete-time linear system polynomials*, Proc. IEEE, 72 (1984), pp. 1131–1142.

[2] ———, *A circular stability test for general polynomials*, Systems Control Lett., 7 (1986), pp. 89–97.

[3] ———, *Z-domain continued-fraction expansions for stable discrete systems polynomials*, IEEE Trans. Circuits and Systems, 32 (1985), pp. 1162–1166.

[4] ———, *A new unit circle stability criterion*, in Proc. 1983 Internat. Symposium on the Mathematical Theory of Networks and Systems, Beer-Sheva, Israel, 1983, pp. 69–87; Also in Lecture Notes in Control and Information Science, Vol. 58, Springer-Verlag, Berlin, New York, 1984, pp. 69–87.

[5] A. COHN, *Über die Anzahl der Wurzeln einer algebraischen Gleichung in einem Kreise*, Math. Z., 14 (1922), pp. 110–148.

[6] P. DELSARTE AND Y. GENIN, *The split Levinson algorithm*, IEEE Trans. Acoust. Speech Signal Process., 34 (1986), pp. 470–478.

[7] ———, *On the splitting of classical algorithms in linear prediction theory*, IEEE Trans. Acoust. Speech Signal Process., 35 (1987), pp. 645–653.

[8] Y. BISTRITZ, H. LEV-ARI, AND T. KAILATH, *Immittance-domain Levinson algorithms*, in Proc. 1986 IEEE Internat. Conference on Acoustics, Speech and Signal Processing, Tokyo, Japan, April 1986, pp. 253–256.

[9] ———, *Immittance-domain Levinson algorithms*, IEEE Trans. Inform. Theory, 35 (1989), pp. 675–682.

[10] P. DELSARTE, Y. GENIN, AND Y. KAMP, *Application of the index theory of pseudo-lossless functions to the Bistritz stability test*, Philips J. Res., 39 (1984), pp. 226–241.

[11] S. D. MORGERA AND H. KRISHNA, *The Szegö recurrence and the role of Hermitian and skew-Hermitian polynomial spaces*, manuscript.

[12] H. KRISHNA AND S. D. MORGERA, *The Levinson recurrence and fast algorithms for solving Toeplitz systems of linear equations*, IEEE Trans. Acoust. Speech Signal Process., 35 (1987), pp. 839–848.

[13] T. KAILATH, S.-Y. KUNG, AND M. MORF, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407; Also in Bull. Amer. Math. Soc., 1 (1979), pp. 769–773.

[14] H. LEV-ARI AND T. KAILATH, *Lattice-filter parameterization and modeling of nonstationary processes*, IEEE Trans. Inform. Theory, 30 (1984), pp. 2–16.
[15] J. D. MARKEL AND A. H. GRAY, JR., *Linear Prediction of Speech*, Springer-Verlag, New York, 1978.
[16] T. KAILATH, *A theorem of I. Schur and its impact on modern signal processing*, in I. Schur Methods in Operator Theory and Signal Processing, Operator Theory: Advances and Applications, Vol. 18, I. Gohberg, ed., Birkhäuser-Verlag, Basel, Switzerland, 1986, pp. 9–30.
[17] H. W. BODE, *Network Analysis and Feedback Amplifier Design*, Van Nostrand, New York, 1945.
[18] YA. L. GERONIMUS, *Polynomials orthogonal on a circle and their applications*, Amer. Math. Soc. Trans., No. 104, 1954 (Russian Publication, 1948).
[19] ———, *Orthogonal Polynomials*, Consultant's Bureau, New York, 1961.
[20] P. DELSARTE AND Y. GENIN, *Multichannel singular predictor polynomials*, IEEE Trans. Circuits and Systems, 35 (1988), pp. 190–200.
[21] Y. BISTRITZ, H. LEV-ARI, AND T. KAILATH, *Immittance versus scattering domains fast algorithms for non-Hermitian Toeplitz and quasi-Toeplitz matrices*, Linear Algebra Appl., 124 (1989), pp. 847–888.
[22] Y. BISTRITZ AND T. KAILATH, *Inversion and factorization of non-Hermitian quasi-Toeplitz matrices*, Linear Algebra Appl., 98 (1988), pp. 77–121.
[23] K. P. BUBE AND R. BURRIDGE, *The one-dimensional inverse problem of reflection seismology*, SIAM Rev., 25 (1983), pp. 497–559.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 521–533, July 1991

009

NESTED EPSILON DECOMPOSITIONS OF LINEAR SYSTEMS: WEAKLY COUPLED AND OVERLAPPING BLOCKS*

M. E. SEZER† AND D. D. ŠILJAK‡

Abstract. A graph-theoretic algorithm is proposed for decomposition of a linear system of equations into subsystems having a prescribed size of mutual interactions. The algorithm generates a whole range of nested decompositions with an apparent trade-off between levels of coupling and sizes of the subsystems. Both disjoint and overlapping subsystems are considered. Having a linear time complexity for a selected strength of coupling, the algorithm is suitable for conditioning large systems and achieving fast convergence rates in block-iterative computations via parallel multiprocessor schemes.

Key words. linear systems, block partitions, weak coupling, overlapping blocks, block-iterative solutions, bigraphs, VLSI circuits

AMS(MOS) subject classification. 65

1. Introduction. In solving large systems of linear equations, parallelism can be introduced by a suitable preprocessing of the equations. Powerful graph-theoretic algorithms are available for restructuring the equations into desirable forms for parallel computations. Efficiency of these algorithms depends heavily on the degree of *sparsity* of the coefficient matrix (e.g., Duff (1981a)). The objective of this paper is to present a graph-theoretic method for decomposing *dense* matrices into weakly-coupled blocks, which can then be assigned to individual processors. The proposed method allows for choice of the threshold ε of the interconnection strengths between the blocks, which determines the amount and frequency of communication between the processors.

The proposed algorithm for epsilon decompositions is remarkably simple. The idea initiated in (Sezer and Šiljak (1986)) is to associate a graph with the given matrix, disconnect the edges of the graph which correspond to elements of the matrix with absolute values less than a prescribed threshold ε, and identify the disconnected subgraphs of the resulting graph. The subgraphs represent the blocks of the matrix which have mutual couplings smaller than ε. If digraphs are used, then symmetric permutations are obtained to sort out the rows and columns and arrive at a partitioned matrix. By applying bigraphs we determine more general nonsymmetric permutations to get epsilon decompositions of a given matrix. In either case, the essential part of the algorithm is enumeration of disconnected components of a graph that has linear time complexity.

An important feature of epsilon decompositions is their *nestedness*. When we obtain a decomposition of a given matrix for one value of epsilon, then for a larger value of epsilon a decomposition is performed on the diagonal blocks only. This results in great computational savings. A full range of epsilon decompositions can be obtained with different epsilon values within each block providing for considerable flexibility in setting up multiprocessor schemes for parallel computations in blockwise solutions of linear problems.

An additional freedom is provided by including *overlapping decompositions*. This type of decomposition has been introduced in the context of dynamic systems (Šiljak (1979), Ikeda and Šiljak (1980), Šiljak (1991)) and has been used for graph-theoretic partitions of matrices into weakly coupled blocks (Arabacioglu, Sezer, and Oral (1986)).

* Received by the editors October 11, 1988; accepted for publication (in revised form) March 22, 1990. This research was supported by National Science Foundation grant ECS-8813273.

† Bilkent University, 06572 Maltepe, Ankara, Turkey.

‡ School of Engineering, Santa Clara University, Santa Clara, California 95053 (dsiljak@scu.bitnet).

The idea is to decompose a given matrix into overlapping diagonal blocks in such a way that when the matrix is expanded into a larger space the diagonal blocks appear as disjoint. When such a decomposition is performed using the epsilon technique, the expanded matrix has diagonal blocks with mutual couplings smaller than a chosen threshold. In this way, we can take advantage of weak coupling in the expanded space which otherwise would not be available in the original space. For example, we can establish stability by diagonal dominance of overlapping blocks where disjoint diagonal blocks fail to be dominant (Ohta and Šiljak (1985)).

The organization of the paper is as follows. In the next section we define nested epsilon decompositions of linear equations and describe a procedure for generating all such decompositions using bigraphs (nonsymmetric permutations) and digraphs (symmetrical permutations). Utility of epsilon decompositions in block-iterative Jacobi schemes is indicated. In § 3, we introduce expansions of linear systems using binary transformation matrices and use bigraphs to interpret the expansion process. Using a simple example we show how a matrix can be expanded into a larger matrix having noninteracting blocks. In § 4, we outline a general procedure for overlapping epsilon decompositions that is based upon bigraphs. Finally, in § 5, a VLSI circuit is used to illustrate how the expansion-decomposition algorithm can be applied to partition a matrix into blocks having mutual coupling smaller than a prescribed threshold.

2. Nested epsilon decompositions. Let us consider a system of linear equations

$$\mathbf{S}: Ax = b, \tag{2.1}$$

where $A = (a_{ij})$ is a given nonsingular $n \times n$ matrix, b is a given $n \times 1$ vector, and x is an $n \times 1$ vector of unknowns. Our interest is to determine permutation matrices P and Q, which produce an equivalent system

$$\bar{\mathbf{S}}: \bar{A}\bar{x} = \bar{b}, \tag{2.2}$$

where

$$\bar{A} = PAQ, \qquad \bar{b} = Pb, \tag{2.3}$$

and the new matrix $\bar{A}$ has a block partition

$$\bar{A} = \begin{bmatrix} \bar{A}_{11} & \varepsilon\bar{A}_{12} \cdots \varepsilon\bar{A}_{1N} \\ \varepsilon\bar{A}_{21} & \bar{A}_{22} \cdots \varepsilon\bar{A}_{2N} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \varepsilon\bar{A}_{N1} & \varepsilon\bar{A}_{N2} \cdots \bar{A}_{NN} \end{bmatrix}. \tag{2.4}$$

In (2.4), ε is a prescribed (fixed) number and elements of each submatrix $\bar{A}_{ij}$, $i \neq j$, are all less than one in absolute value. In other words, a choice of the threshold ε induces a partition (2.4) having N diagonal blocks, and off-diagonal blocks have elements smaller than ε. The number N is not fixed and depends on the choice of the threshold ε. This choice is guided by our desire to have weakly coupled and, at the same time, nonsingular diagonal blocks of $\bar{A}$.

For convenience, we rewrite $\bar{A}$ of (2.4) in a compact form

$$\bar{A} = \bar{A}_D + \varepsilon\bar{A}_C, \tag{2.5}$$

with

$$\bar{A}_D = \text{diag}\,\{\bar{A}_{11}, \bar{A}_{22}, \cdots, \bar{A}_{NN}\}, \qquad \bar{A}_C = (\bar{A}_{ij}). \tag{2.6}$$

To come up with a procedure which produces (2.4), we associate a bigraph $\mathbf{B} = (\mathcal{X}, \mathcal{Y}; \mathcal{E})$ with the matrix A of (2.1) such that $|\mathcal{X}| = |\mathcal{Y}| = n$, and $(x_j, y_i) \in \mathcal{E}$ if and only if $a_{ij} \neq 0$, $i, j = 1, 2, \cdots, n$. Consistent with our assumption of nonsingularity of A, the bigraph $\mathbf{B}$ contains a perfect matching. For a selected value of ε we form a subgraph $\mathbf{B}^\varepsilon = (\mathcal{X}, \mathcal{Y}; \mathcal{E}^\varepsilon)$ by removing the edges of $\mathbf{B}$ that correspond to those elements a_{ij} of A such that $|a_{ij}| < \varepsilon$. Let us assume, for a moment, that $\mathbf{B}^\varepsilon$ has a perfect matching. Obviously this implies that each component of $\mathbf{B}^\varepsilon$ is a bigraph, which itself contains a perfect matching. In this case, each component $\mathbf{B}_i^\varepsilon$ of $\mathbf{B}^\varepsilon$ identifies a block $\bar{A}_{ii}$ of the matrix $\bar{A}_D$, which is generically nonsingular. The permutation matrices P and Q, which relate the decomposed system $\bar{\mathbf{S}}$ to the original $\mathbf{S}$, are obtained automatically. By regrouping the terms of $\bar{A}$, which were thrown away in forming $\mathbf{B}^\varepsilon$ from $\mathbf{B}$ according to P and Q, we recover the second term $\varepsilon\bar{A}_C$ in the epsilon decomposition (2.5).

If for a fixed epsilon, $\mathbf{B}^\varepsilon$ does not contain a perfect matching (either some of the components are not bigraphs, or they do not contain a perfect matching), then an obvious remedy is to reduce epsilon and try another decomposition. This reduction process adds edges at each step and may result in a satisfactory decomposition before the process reaches a connected bigraph $\mathbf{B}$.

Example 2.7. Let us use a simple example to illustrate the above arguments. Consider a matrix

$$A = \begin{bmatrix} 0 & 1 & 0.3 & 0 \\ 1 & 0 & 0 & 0.2 \\ 0.2 & 0 & 0.1 & 0.1 \\ 0.1 & 0.4 & 1 & 0 \end{bmatrix}. \tag{2.8}$$

Obviously, for $\varepsilon > 0.2$, the bigraph $\mathbf{B}^\varepsilon$ has x_4 as an isolated node and, therefore, lacks a perfect matching. On the other hand, for $\varepsilon_1 = 0.2$, the bigraph $\mathbf{B}^{0.2}$ shown in Fig. 1(a) has a perfect matching indicated by heavy lines. The two components of $\mathbf{B}^{0.2}$ contain the nodes $\{x_1, x_4; y_3, y_2\}$ and $\{x_2, x_3; y_1, y_4\}$ resulting in a matrix

$$\bar{A} = \left[\begin{array}{cc:cc} 0.2 & 0.1 & 0 & 0.1 \\ 1 & 0.2 & 0 & 0 \\ \hdashline 0 & 0 & 1 & 0.3 \\ 0.1 & 0 & 0.4 & 1 \end{array}\right]. \tag{2.9}$$

We can now increase epsilon in the lower diagonal block to $\varepsilon_2 = 0.5$ to get a finer decomposition into three components, which have two values of epsilon, namely, $\varepsilon_1 = 0.2$ and $\varepsilon_2 = 0.5$.

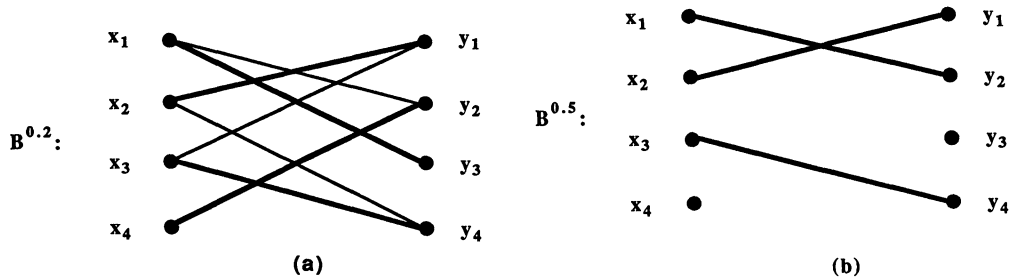


FIG. 1. *Bigraphs for Example* 2.7.

Let us now consider a situation where we choose $\varepsilon_1 = 0.5$ to start with, and obtain the bigraph $\mathbf{B}^{0.5}$ shown in Fig. 1(b). Although $\mathbf{B}^{0.5}$ has no complete matching, we can identify a subgraph of $\mathbf{B}^{0.5}$, which contains a perfect matching, and remove it from the original bigraph $\mathbf{B}$. Then, we check the remainder of $\mathbf{B}$ for a perfect matching which, if present, would correspond obviously to an epsilon smaller than 0.5. Referring to Fig. 1(b), the subgraph we remove from $\mathbf{B}$ is determined by the vertices $\{x_1, x_2, x_3; y_2, y_1, y_4\}$. In the remaining subgraph we can get a perfect matching by using $\varepsilon_2 = 0.1$. The corresponding partitioned matrix $\bar{A}$ is

$$\bar{A} = \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0.2 \\ 0 & 1 & 0.3 & 0 \\ 0.1 & 0.4 & 1 & 0 \\ \hline 0.2 & 0 & 0.1 & 0.1 \end{array}\right]. \tag{2.10}$$

The upper diagonal block is obtained from the removed subgraph and is itself divided into three components identified in Fig. 1(b).

We note an important property of epsilon decompositions, which is that vertices that are connected in $\mathbf{B}^{\varepsilon_1}$ are also connected in $\mathbf{B}^{\varepsilon_2}$ whenever $\varepsilon_2 < \varepsilon_1$. This means that the epsilon decompositions are inherently *nested*. This nestedness property is observed not only for the entire graph, but also for each component of the graph independently, which results in considerable computational savings; for ε_1 we decompose only the blocks corresponding to ε_2, not the overall system. Furthermore, nestedness provides us with the flexibility of choosing two or more different values of epsilon in a single decomposition. The general structure of $\bar{A}$ for K values of epsilon, $\varepsilon_1 > \varepsilon_2 > \cdots > \varepsilon_K$, is

$$\bar{A} = \bar{A}_0 + \varepsilon_1\bar{A}_1 + \varepsilon_2\bar{A}_2 + \cdots + \varepsilon_K\bar{A}_K, \tag{2.11}$$

where $\bar{A}_0, \bar{A}_1, \bar{A}_2, \cdots, \bar{A}_K$ are all partitioned matrices with compatible blocks, $\bar{A}_0$ is block diagonal, and any nonzero block of $\bar{A}$ appears in one and only one matrix $\bar{A}_k$, $k = 0, 1, 2, \cdots, K$, and none of the elements of any $\bar{A}_k$ is larger than one in absolute value. A typical situation is illustrated in Fig. 2 for two values of ε.

We note that in solving linear equations (2.1) by a block-iterative method on parallel processors, the nested epsilon decomposition (2.11) induces a hierarchy of processors. For speeding up the solution process, the grouping of the processors according to the amount of intercommunication should be compatible with the decomposition in (2.11). In this way, the effect of the term $\varepsilon_k\bar{A}_k$ is computed more frequently than the effect of the term $\varepsilon_{k+1}\bar{A}_{k+1}$.

To illustrate this idea, let us consider the case of Fig. 2, which is a decomposition:

$$\bar{A} = \bar{A}_0 + \varepsilon_1\bar{A}_1 + \varepsilon_2\bar{A}_2, \qquad \varepsilon_1 > \varepsilon_2. \tag{2.12}$$

A two-level block-iterative procedure for solution of (2.2) proceeds as follows:

$$\text{Fast Iterations:} \quad \bar{A}_0\bar{x}^{(k_1+1,k_2)} = \bar{b}^{(k_2)} - \varepsilon_1\bar{A}_1\bar{x}^{(k_1,k_2)}, \tag{2.13}$$

$$\text{Slow Iterations:} \quad \bar{A}_0\bar{x}^{(0,k_2+1)} = \bar{b} - (\varepsilon_1\bar{A}_1 + \varepsilon_2\bar{A}_2)\bar{x}^{(\infty,k_2)}, \tag{2.14}$$

where the superscripts k_1 and k_2 refer to the fast and slow iteration steps, respectively, $\bar{x}^{(\infty,k_2)}$ is the limit of the fast iterations at the k_2th slow iteration step, and

$$\bar{b}^{(k_2)} = \begin{cases} \bar{b}, & k_2 = 0, \\ \bar{b} - \varepsilon_2\bar{A}_2x^{(\infty,k_2-1)}, & k_2 \geqq 0. \end{cases} \tag{2.15}$$

Since $(\bar{A}_0 + \varepsilon\bar{A}_1)$ is block diagonal, the fast iterations are decoupled. This means that only the processors that are assigned to the individual blocks of $(\bar{A}_0 + \varepsilon_1\bar{A}_1)$ should

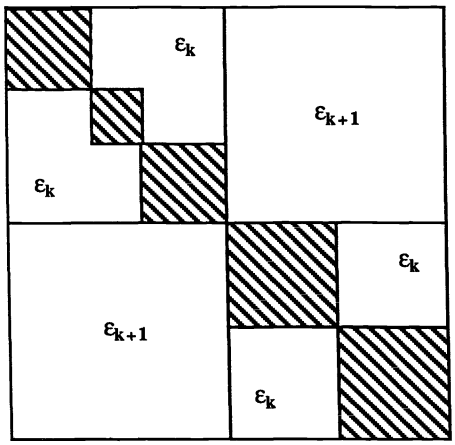


FIG. 2. *Nested epsilon decompositions.*

exchange the corresponding parts of $\bar{x}^{(k_1,k_2)}$ during fast iterations. Communication of all processors is required only once at every slow iteration step. For a sufficiently small ε_2, slow iterations converge in few steps resulting in communication overhead, which is mainly due to the local exchanges; it should be far less than in the single-level block-iterative scheme.

We also note that a multilevel scheme involving a hierarchy of processors can take advantage of sparsity of A. If A is sparse, then so are $\bar{A}_0$, $\bar{A}_1$, and $\bar{A}_2$, resulting in a reduction of both the communication overhead and the number of operations involved in the backward-forward substitutions at every iteration step.

An important special case of epsilon decompositions takes place when we use *symmetric permutations*, that is, $Q = P^T$ in (2.3). The ordering of the columns is the same as the ordering of the rows of $\bar{A}$, and we can associate a digraph $\mathbf{D} = (\mathscr{X}; \mathscr{E})$ with the matrix A of $\mathbf{S}$ in the standard way (Harary (1969)) to achieve a decomposition. An important advantage in this case is that we do not need to be concerned with obtaining a perfect matching in the components of $\mathbf{B}^\varepsilon$, nor with the components of $\mathbf{B}^\varepsilon$ being bigraphs themselves. However, the diagonal blocks $\bar{A}_{ii}$ of $\bar{A}$ should be checked for generic nonsingularity using known algorithms (Duff (1981b)), because all one gets from $\mathbf{D}^\varepsilon$ is connectivity of its components.

An essential feature of epsilon decompositions is that they are conducive to *convergence* of iterative procedures for solving linear equations. For example, if we use the block Jacobi method (e.g., Hageman and Young (1981)), the iteration matrix $J = (J_{ij})$ corresponding to $\bar{A}$ of (2.4) has the blocks

$$J_{ij} = \begin{cases} 0, & i = j, \\ -\varepsilon A_{ii}^{-1} A_{ij}, & i \neq j. \end{cases} \tag{2.16}$$

The iterative process is convergent if the matrix $W = (w_{ij})$, with

$$w_{ij} = \begin{cases} 1, & i = j, \\ -\varepsilon \| A_{ii}^{-1} A_{ij} \|, & i \neq j, \end{cases} \tag{2.17}$$

is an M-matrix (e.g., Robert (1969)). The smaller the value of ε, the better is the chance for the matrix W to be an M-matrix, and the faster is the convergence of the iterative process. These facts follow readily from a stability analysis of the iterative process via

vector Lyapunov functions (Sezer and Šiljak (1988), Kaszkurewicz, Bhaya, and Šiljak (1989), Šiljak (1991)).

3. Expansions of linear equations. A considerable increase in flexibility of the epsilon decomposition can be achieved by allowing diagonal blocks of the coefficient matrix to overlap. By expanding the matrix, the overlapping blocks become disjoint with coupling smaller than the threshold value of epsilon. To illustrate this idea, let us consider a simple example.

Example 3.1. Let A be given as

$$A = \begin{bmatrix} * & * & \odot \\ \odot & * & \odot \\ \odot & * & * \end{bmatrix}, \tag{3.2}$$

where $\odot$ denotes an element with magnitude smaller than a given ε and * an element larger than ε. The corresponding bigraph $\mathbf{B}^\varepsilon$ is connected and A has no disjoint ε decomposition. It has, however, an overlapping ε decomposition as indicated in (3.2). By repeating the second row and splitting the second column, we obtain an expanded matrix

$$\tilde{A} = \begin{bmatrix} * & * & 0 & \odot \\ \odot & * & 0 & \odot \\ \odot & 0 & * & \odot \\ \odot & 0 & * & * \end{bmatrix}, \tag{3.3}$$

which has an ε decomposition where the overlapping blocks of A appear as disjoint. The expansion process and its justification are explained next.

With the system $\mathbf{S}$ of (2.1), we associate another system of equations

$$\tilde{\mathbf{S}}: \tilde{A}\tilde{x} = \tilde{b}, \tag{3.4}$$

where $\tilde{A} = (\tilde{a}_{ij})$ is an $\tilde{n} \times \tilde{n}$ matrix, and $\tilde{b}$ and $\tilde{x}$ are $\tilde{n} \times 1$ vectors. Our crucial assumption is that the order of $\tilde{\mathbf{S}}$ is larger than the order of $\mathbf{S}$, that is, $\tilde{n} > n$.

We denote the solution sets of $\mathbf{S}$ and $\tilde{\mathbf{S}}$ as Φ and $\tilde{\Phi}$, and state the following definition.

DEFINITION 3.5. A system $\tilde{S}$ is said to be a *left expansion* (*right expansion*) of the system $\mathbf{S}$ if there exists an $\tilde{n} \times n$ matrix V ($n \times \tilde{n}$ matrix U) with full rank such that $V\Phi \subseteq \tilde{\Phi}$ ($U\tilde{\Phi} \subseteq \Phi$).

A sufficient condition for $\tilde{\mathbf{S}}$ to be an expansion of $\mathbf{S}$ is given by the following theorem.

THEOREM 3.6. *A system* $\tilde{\mathbf{S}}$ *is a left expansion of* $\mathbf{S}$ *if there exist* $\tilde{n} \times n$ *matrices* V *and* $\tilde{V}$ *with full column rank such that*

$$\tilde{A}\tilde{V} = VA, \qquad \tilde{b} = Vb \tag{3.7}$$

and $\tilde{\mathbf{S}}$ *is a right expansion of* $\mathbf{S}$ *if there exist* $n \times \tilde{n}$ *matrices* U *and* $\tilde{U}$ *such that*

$$\tilde{U}\tilde{A} = AU, \qquad \tilde{U}\tilde{b} = b. \tag{3.8}$$

Proof. Suppose (3.7) holds and let φ be a solution of $\mathbf{S}$. Then, $A\varphi = b \Rightarrow VA\varphi = Vb \Rightarrow \tilde{A}\tilde{V}\varphi = \tilde{b}$, so that $\tilde{V}\varphi$ is a solution of $\tilde{\mathbf{S}}$ and, therefore, $\tilde{V}\Phi \subseteq \tilde{\Phi}$. The second part of the theorem is proved likewise. □

At this point we should mention that expansions of linear equations have been considered by Calvet and Titli (1989) in the context of linear quadratic control. Their expansion procedure is based upon the expansion scheme of Ikeda and Šiljak (1980), which is devised for linear dynamic systems. This fact severely restricts the procedure to

symmetric partitioning of the coefficient matrix; if the kth equation is repeated, then the kth unknown has to be split into two parts as well. We also note that the condition of Proposition 2.1 in the paper by Calvet and Titli (1989) is not necessary and sufficient, but only sufficient.

Of special interest in this paper are left expansions which are obtained using binary matrices V and $\tilde{V}$, because such expansions have useful graph-theoretic interpretations. We do not consider right expansions of the same type since, by duality, their properties can be derived from those of left expansions.

In generating appropriate left expansions $\tilde{\mathbf{S}}$ of $\mathbf{S}$, we make use of the bigraph $\mathbf{B} = (\mathscr{X}, \mathscr{Y}; \mathscr{E})$ associated with the matrix A. For this purpose, we need the definition of a left expansion $\tilde{\mathbf{B}}$ of $\mathbf{B}$, which we formulate using the following.

PROCEDURE 3.9. Let $\mathbf{M}$ be a one-to-one correspondence between the nodes $\mathscr{X}$ and $\mathscr{Y}$ of $\mathbf{B}$ with $\mathbf{M}(x_i) = y_{q_i}$; $i, q_i = 1, 2, \cdots, n$. Let each x_i be associated with an integer $k_i \geqq 1$ such that $k_i > 1$ for at least one i. Let $\tilde{\mathbf{B}} = (\tilde{\mathscr{X}}, \tilde{\mathscr{Y}}; \tilde{\mathscr{E}})$, where $\tilde{\mathscr{X}}$, $\tilde{\mathscr{Y}}$, and $\tilde{\mathscr{E}}$ are defined as follows:

$$\tilde{\mathscr{X}} = \bigcup_{i=1}^{n} \{\tilde{x}_i^{(1)}, \tilde{x}_i^{(2)}, \cdots, \tilde{x}_i^{(k_i)}\},$$
$$\tilde{\mathscr{Y}} = \bigcup_{i=1}^{n} \{\tilde{y}_{q_i}^{(1)}, \tilde{y}_{q_i}^{(2)}, \cdots, \tilde{y}_{q_i}^{(k_i)}\}. \tag{3.10}$$

For each $(x_j, y_{q_i}) \in \mathscr{E}$, $\tilde{\mathscr{E}}$ contains exactly k_i edges $(\tilde{x}_j^{(p_l)}, \tilde{y}_{q_i}^{(l)})$, $l = 1, 2, \cdots, k_i$, $p_l \in \{1, 2, \cdots, k_j\}$.

Using the above procedure, we obtain $\tilde{\mathbf{B}}$ from $\mathbf{B}$ by simply repeating nodes and edges of $\mathbf{B}$. Which nodes and edges are repeated is uniquely determined by our choice of $\mathbf{M}$, k_i, and p_l. Once we obtain a $\tilde{\mathbf{B}}$ from $\mathbf{B}$ using Procedure 3.9, we can produce the corresponding left expansion $\tilde{\mathbf{S}}$. For this we need Theorem 3.11.

THEOREM 3.11. *Let* $\tilde{\mathbf{B}}$ *be obtained from* $\mathbf{B}$ *using Procedure* 3.9. *Then there exists a matrix* $\tilde{A}$ *corresponding to* $\tilde{\mathbf{B}}$, *which is the matrix of a left expansion* $\tilde{\mathbf{S}}$ *of* $\mathbf{S}$.

Proof. The proof is by construction. Let $\tilde{A} = (\tilde{A}_{ij})$ which is an $n \times n$ block matrix. The block $\tilde{A}_{q_i j} = (\tilde{a}_{rs}^{q_i j})$ is a $k_i \times k_j$ matrix with elements defined as

$$\tilde{a}_{rs}^{q_i j} = \begin{cases} a_{q_i j}, & s = p_r, \\ 0 \quad \text{otherwise.} \end{cases} \tag{3.12}$$

In other words, rows and columns of $\tilde{A}_{q_i j}$ are associated with the vertices $\{\tilde{y}_{q_i}^{(1)}, \cdots, \tilde{y}_{q_i}^{(k_i)}\}$ and $\{\tilde{x}_j^{(1)}, \cdots, \tilde{x}_j^{(k_j)}\}$, respectively, and the element in the position (r, s) is nonzero and is $a_{q_i j}$ if and only if $(\tilde{x}_j^{(s)}, \tilde{y}_{q_i}^{(r)}) \in \tilde{\mathscr{E}}$. Now, let $\tilde{V}$ be the matrix obtained from the $n \times n$ identity matrix I_n by repeating the ith row k_i times, and let V be obtained from I_n by repeating the q_ith row k_i times. Obviously, $\tilde{A}\tilde{V} = VA$, and the proof follows upon setting $\tilde{b} = Vb$. □

We note that Definition 3.5 does not presume existence or uniqueness of solutions of $\mathbf{S}$ and $\tilde{\mathbf{S}}$. We describe here a class of expansions that preserve generic nonsingularity of A.

THEOREM 3.13. *Let* $\mathbf{M}$ *in Procedure* 3.9 *be a perfect matching in* $\mathbf{B}$. *Then, for any choice of the integers* $k_1, k_2, \cdots, k_n$, *there exists a corresponding left expansion* $\tilde{\mathbf{B}}$ *of* $\mathbf{B}$ *which also contains a perfect matching.*

Proof. By assumption $(x_i, y_{q_i}) \in \mathscr{E}$, $i = 1, 2, \cdots, n$. In defining edges $(\tilde{x}_i^{(p_l)}, \tilde{y}_{q_i}^{(l)})$ of $\tilde{\mathbf{B}}$, choose $p = l$, $l = 1, 2, \cdots, k_i$. Then, obviously, $\tilde{\mathbf{M}}$ defined as $\tilde{\mathbf{M}}(\tilde{x}_i^{(l)}) = \tilde{y}_{q_i}^{(l)}$ is a perfect matching in $\tilde{\mathbf{B}}$. □

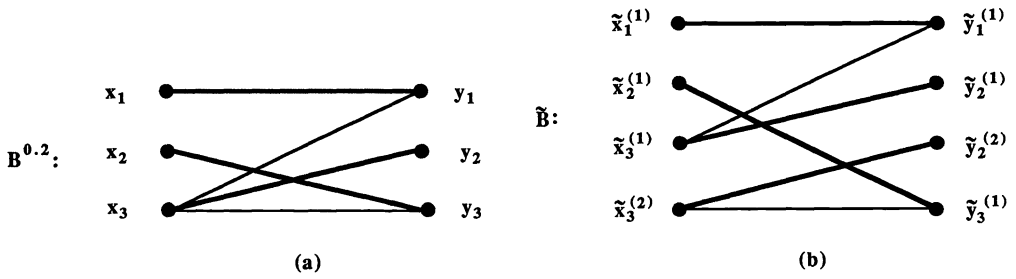


FIG. 3. *Bigraphs for Example* 3.14.

Example 3.14. To illustrate the concept of expansions and their use in decompositions of linear equations, let us consider a matrix

$$A = \begin{bmatrix} a_{11} & 0 & a_{13} \\ 0 & 0 & a_{23} \\ 0 & a_{32} & a_{33} \end{bmatrix}. \tag{3.15}$$

The corresponding bigraph **B** is shown in Fig. 3(a), where a (unique) perfect matching is indicated by heavy lines. Since B is connected, A cannot be permuted into a block-diagonal matrix. An expansion $\tilde{\mathbf{B}}$ of **B** which has two disjoint components can easily be constructed. For this, we follow Procedure 3.9, where we choose **M** to be the perfect matching indicated in Fig. 3(a), and $k_1 = k_2 = 1$, $k_3 = 2$. The resulting bigraph $\tilde{\mathbf{B}}$ in Fig. 3(b) has two components each containing a perfect matching, as indicated. The expanded matrix $\tilde{A}$ and the expansion matrices $\tilde{V}$ and V are obtained from $\tilde{\mathbf{B}}$ as

$$\tilde{A} = \left[\begin{array}{cc:cc} a_{11} & a_{13} & \multicolumn{2}{c}{0} \\ 0 & a_{23} & \multicolumn{2}{c}{} \\ \hdashline \multicolumn{2}{c:}{} & a_{32} & a_{33} \\ \multicolumn{2}{c:}{0} & 0 & a_{23} \end{array}\right],$$

$$\tilde{V} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad V = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}. \tag{3.16}$$

It is interesting to note that if we replace the zeros in A of (3.15) by small numbers of order ε, so that A becomes a *full matrix*, the expanded matrix $\tilde{A}$ would retain the epsilon decomposition along the same lines as in (3.16). We next consider how to perform an expanded epsilon decomposition on a given matrix.

4. Overlapping epsilon decomposition. In obtaining overlapping decompositions of a matrix A, we require that all diagonal blocks $\tilde{A}_{ii}$ of $\tilde{A}$ be smaller than A, and that no $\tilde{A}_{ii}$ is further expandable into two or more diagonal blocks. The reason for the first requirement is obvious. The second requirement guarantees that the decomposition is maximal.

To simplify the decomposition procedure, we first provide the following result.

THEOREM 4.1. *Let* **M** *in Procedure* 3.9 *be a perfect matching and consider a left expansion* $\tilde{\mathbf{B}}$ *as in Theorem* 3.13. *Let* **B** *contain a simple alternating cycle. If a component of* $\tilde{\mathbf{B}}$ *contains a replica of any one of the vertices appearing in the cycle, then it contains at least one replica of each of the other vertices in the cycle.*

Proof. Without loss of generality, assume that $\mathbf{M}(x_i) = y_i$ and that $\mathbf{B}$ has a cycle $\{(x_1, y_1), (x_2, y_1), (x_2, y_2), \cdots, (x_p, y_p), (x_1, y_p)\}$. This can always be achieved by a bipartite labeling of the vertices of $\mathbf{B}$. Let $\mathbf{B}_s$ be the subgraph of $\mathbf{B}$ which consists only of the vertices and the edges that appear in the cycle, and consider an expansion $\tilde{\mathbf{B}}_s$ of $\mathbf{B}_s$. Any component of $\tilde{\mathbf{B}}_s$ which includes a replica of x_1, should also include a replica of y_1 because $(x_1, y_1) \in \mathscr{E}_s$. On the other hand, since each replica of y_1 is connected to a replica of x_2 (by construction of $\tilde{\mathbf{B}}_s$), the same component includes a replica of x_2 and, therefore, a replica of y_2. Proceeding in this way, we observe that any component of $\tilde{\mathbf{B}}_s$, which contains a replica of x_1, also contains at least one replica of each x_i and y_i, $i = 1, 2, \cdots, p$. The proof then follows from the fact that addition of vertices and edges to complete $\mathbf{B}_s$ to $\mathbf{B}$ only results in larger components in $\tilde{\mathbf{B}}$. □

Theorem 4.1 simply states that cycles of $\mathbf{B}$ cannot be split in the process of expansion; they should be treated as a whole. This suggests that the cycles can be condensed to form an acyclic bigraph $\mathbf{B}^*$, which can be used as a basis for expansion. The condensation process consists of collapsing the x and y nodes of each cycle into a single pair of supernodes. The lines from the x nodes of one cycle to y nodes of another are similarly collapsed into a single superline between the corresponding supernodes.

PROCEDURE 4.2.

(1) Identify and condense successively the cycles of $\mathbf{B}$ to form the condensation $\mathbf{B}^* = (\mathscr{X}^*, \mathscr{Y}^*; \mathscr{E}^*)$ of $\mathbf{B}$.

(2) Identify the unique perfect matching $\mathbf{M}^*$ of $\mathbf{B}^*$, and let $\mathbf{M}^*(x_i^*) = y_{q_i}^*$, $i = 1, 2, \cdots, n^*$, $n^* = |\mathscr{X}^*|$.

(3) For some unflagged x_i^* (first time at this step, all x_i^* are unflagged), construct the sets $\mathscr{X}_i^*$ and $\mathscr{Y}_i^*$ as follows:
   (a) $x_i^* \in \mathscr{X}_i^*$, $y_{q_i}^* \in \mathscr{Y}_i^*$.
   (b) For all $x_j^* \in \mathscr{X}_i^*$ add $y_{q_j}^*$ to $\mathscr{Y}_i^*$.
   (c) For all $y_j^* \in \mathscr{Y}_i^*$ add all x_k^* such that $(x_k^*, y_j^*) \in \mathscr{E}^*$ to $\mathscr{X}_i^*$.

(4) If $\mathscr{X}_i^* = \mathscr{X}^*$, then stop. No suitable left expansion of $\mathbf{B}^*$ exists. Otherwise, flag all $x_j^* \in \mathscr{X}_i^*$. If x_i^* are not all flagged then go to step 3.

(5) Eliminate all $\mathscr{X}_i^*$ and their corresponding $\mathscr{Y}_i^*$, such that $\mathscr{X}_i^* \subset \mathscr{X}_j^*$ for some $j = 1, 2, \cdots, n^*$. The remaining sets have the property that

$$\cup \mathscr{X}_i^* = \mathscr{X}^*, \qquad \cup \mathscr{Y}_i^* = \mathscr{Y}^*.$$

(6) For each x_j^*, let k_j be the number of remaining sets $\mathscr{X}_i^*$ which include x_j^*, $j = 1, 2, \cdots, n^*$. Expand $\mathbf{B}^*$ into $\tilde{\mathbf{B}}^*$ using Procedure 3.9 under the conditions of Theorem 3.11.

(7) Decondense $\tilde{\mathbf{B}}^*$, and form $\tilde{\mathbf{B}}$.

It is easy to see how Procedure 4.2 can be used to obtain an overlapping epsilon decomposition of a given matrix A. All one has to do is select ε, form $\mathbf{B}^\varepsilon$, and expand $\mathbf{B}^\varepsilon$ to get $\tilde{\mathbf{B}}^\varepsilon$. In generating $\tilde{\mathbf{B}}^\varepsilon$, Procedure 4.2 provides the expansion matrices $\tilde{V}$ and V, which are then used to get the expanded matrix $\tilde{A}$ having the epsilon decomposition:

$$\tilde{A} = \tilde{A}_D + \varepsilon \tilde{A}_C. \tag{4.3}$$

Here, $\tilde{A}_D = \text{diag}\,\{\tilde{A}_{11}, \tilde{A}_{22}, \cdots, \tilde{A}_{NN}\}$ where the blocks $\tilde{A}_{ii}$ correspond to the components of $\tilde{\mathbf{B}}^\varepsilon$, and the elements of $\tilde{A}_C$ have magnitudes smaller than one. The disjoint decomposition (4.3) of $\tilde{A}$ is what is meant by an overlapping epsilon decomposition of the original matrix A.

It is understood that Procedure 4.2 requires $\mathbf{B}^\varepsilon$ to have a perfect matching, that is, the matrix A should be generically nonsingular after the removal of epsilon elements. Then, generic nonsingularity of $\tilde{A}_D$ and, therefore, of $\tilde{A}$, is automatic.

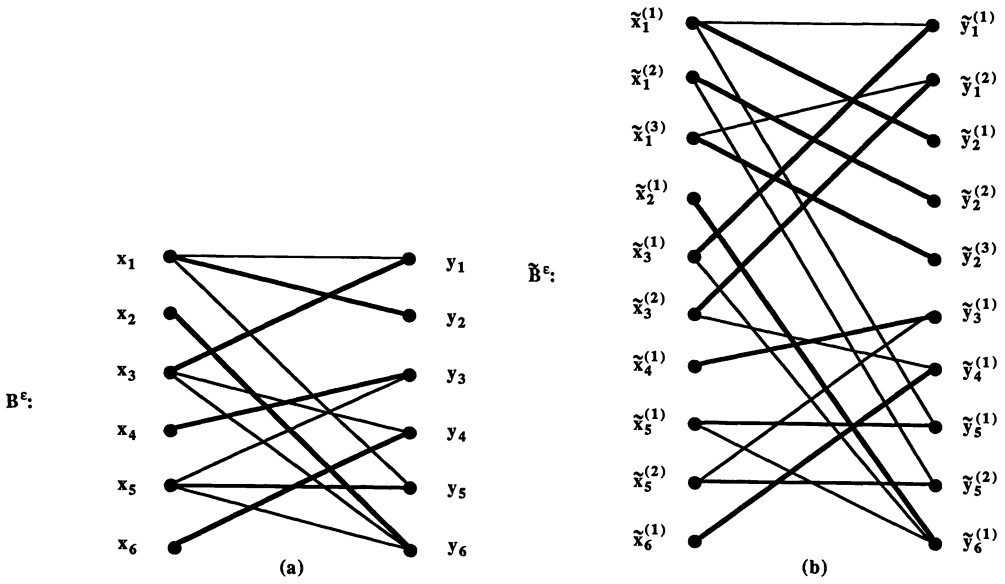


FIG. 4. *Bigraphs for Example* 4.4.

Example 4.4. Let us consider a matrix

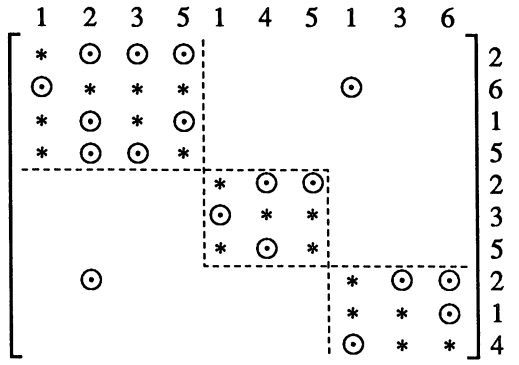
$$A = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{bmatrix} * & \odot & * & \odot & \odot & \odot \\ * & \odot & \odot & \odot & \odot & \odot \\ \odot & \odot & \odot & * & * & \odot \\ \odot & \odot & * & \odot & \odot & * \\ * & \odot & \odot & \odot & * & \odot \\ \odot & * & * & \odot & * & \odot \end{bmatrix} \end{array} \begin{matrix} 1 \\ 2 \\ 3 \\ 4, \\ 5 \\ 6 \end{matrix} \tag{4.5}$$

where $\odot$ denotes an element with magnitude smaller than a given ε, and $*$ an element larger than ε. The bigraph $\mathbf{B}^\varepsilon$ is given in Fig. 4(a). We apply Procedure 4.2 to $\mathbf{B}^\varepsilon$ as follows:

(1) Since $\mathbf{B}^\varepsilon$ is acyclic, we set $\mathbf{B}^* = \mathbf{B}^\varepsilon$ and drop superscript $*$.
(2) The unique perfect matching $\mathbf{M}$ is indicated by heavy lines in Fig. 4(a).
(3) $\mathscr{X}_1 = \{x_1\}$, $\mathscr{Y}_1 = \{y_2\}$.
(4) Flag x_1. Not all x_i are flagged.
(3) $\mathscr{X}_2 = \{x_2, x_3, x_5, x_1\}$, $\mathscr{Y}_2 = \{y_6, y_1, y_5, y_2\}$.
(4) Flag x_2, x_3, and x_5. Not all x_i are flagged.
(3) $\mathscr{X}_4 = \{x_4, x_5, x_1\}$, $\mathscr{Y}_4 = \{y_3, y_5, y_2\}$.
(4) Flag x_4. Not all x_i are flagged.
(3) $\mathscr{X}_6 = \{x_6, x_3, x_1\}$, $\mathscr{Y}_6 = \{y_4, y_1, y_2\}$.
(4) Flag x_6. All x_i are flagged.
(5) Discard $\mathscr{X}_1$.
(6) $k_1 = 3$, $k_2 = k_4 = k_6 = 1$, $k_3 = k_5 = 2$. Expanded bigraph $\tilde{\mathbf{B}}$ is shown in Fig. 4(b).
(7) Since no condensations took place, $\tilde{\mathbf{B}}$ is the final product. It has three components, each containing a perfect matching.

Using the expanded bigraph $\tilde{\mathbf{B}}$, we form the expanded matrix

(4.6)



Finally, we note that no diagonal block of $\tilde{A}$ is further expandable.

5. VLSI circuit simulation. Efficient computer simulation of VLSI circuits containing a large number of dynamic and nonlinear components is a challenging problem with both conceptual and numerical difficulties. Due to a very large number of state variables (typically, over 10,000), standard circuit simulators like SPICE (Nagel (1975)) may require excessive computer time. Recently, a new simulation method based on model reduction using moment matching has been proposed by Pillage (1989). The method requires a solution of a resistive d.c. network for every piecewise linear region of the nonlinear components, and for as many times as there are moments to be matched. Since a typical VLSI circuit consists of several subnetworks, which are interconnected by components having very small or very large values, they are ideal candidates for application of (overlapping) epsilon decompositions.

To illustrate the application of epsilon decomposition to network equations, let us consider the simple network of Fig. 5, where N_1 and N_2 are two subnetworks driven by a π-circuit. Suppose that the network elements are normalized so that most of them have values close to unity except R_1, R_2, and G, which are of the order ε. The network equation can be written (e.g., Chua, Desoer, and Kuh (1987)) for N_1 and N_2 as follows:

(5.1)

$$\begin{aligned} &\text{Terminal equations:} && M_l V_l + N_l I_l = 0, \\ &\text{Circuit equations:} && F_l V_l + f_l v_l = g_l v_l^*, \\ &\text{Cut-set equations:} && H_l I_l + h_l i_l^* = 0; \end{aligned}$$

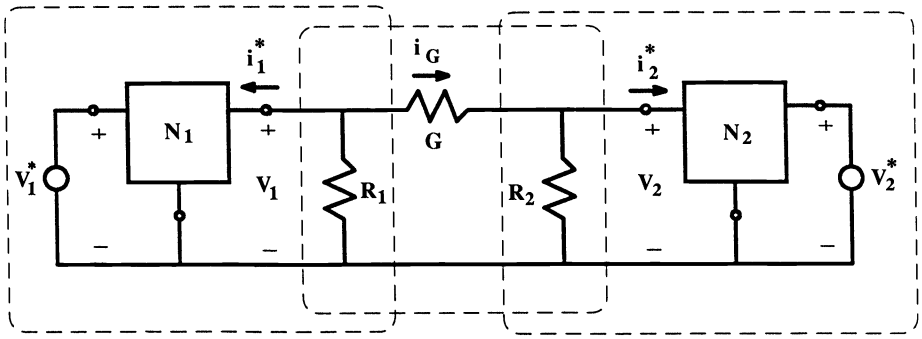


FIG. 5. *Overlapping decomposition.*

$l = 1, 2$, and for the interconnecting π-circuit as

$$\text{Terminal equations:} \quad v_1 - R_1 i_1 = 0, \tag{5.2}$$
$$v_2 - R_2 i_2 = 0,$$
$$i_G - G v_G = 0;$$

$$\text{Circuit equation:} \quad v_G - v_1 + v_2 = 0; \tag{5.3}$$

$$\text{Cut-set equations:} \quad i_1 + i_1^* + i_G = 0, \tag{5.4}$$
$$i_2 + i_2^* - i_G = 0.$$

If the networks N_1 and N_2 are irreducible and contain n_1 and n_2 components, respectively, then (5.1) and (5.2) form a set of $2(n_1 + n_2) + 8$ equations, where the coefficient matrix is irreducible. However, keeping in mind that R_1, R_2, and G are of the order of ε, the graph-theoretic procedure of § 5 gives an overlapping epsilon decomposition indicated in Fig. 5, which results in an expanded matrix:

$$\tilde{A} = \begin{array}{c|cccc|cccc|cccc|cc}
 & V_1 & I_1 & i_1^* & v_1 & i_1 & i_G & v_1 & v_G & v_2 & i_G & i_2 & v_2 & i_2^* & I_2 & V_2 \\
\hline
 & M_1 & N_1 & & & & & & & & & & & & & \\
 & F_1 & & & f_1 & & & & & & & & & & & \\
 & & H_1 & h_1 & & & & & & & & & & & & \\
 & & & & 1 & -R_1 & & & & & & & & & & \\
 & & & 1 & & 1 & 1 & & & & & & & & & \\
 & & & & & & 1 & & G & & & & & & & \\
\hline
 & & & & & -R_1 & & 1 & & & & & & & & \\
 & & & & & & & 1 & 1 & -1 & & & & & & \\
 & & & & & & & & & 1 & & -R_2 & & & & \\
\hline
 & & & & & & & & G & & 1 & & & & & \\
 & & & & & & & & & & -1 & 1 & & 1 & & \\
 & & & & & & & & & & & -R_2 & 1 & & & \\
 & & & & & & & & & & & & & h_2 & H_2 & \\
 & & & & & & & & & & & & f_2 & & & F_2 \\
 & & & & & & & & & & & & & & N_2 & M_2
\end{array} \tag{5.5}$$

The matrix $\tilde{A}$ consists of three square blocks of sizes $2n_1 + 4$, 3, and $2n_2 + 4$.

To compare the efficiency of a block-iterative method based on an overlapping decomposition such as (5.5) and that of a direct method, let us assume that for a network containing n components, LU-decomposition has a time bound $k_d n^d$, while backward-forward substitutions take additional $k_s n^s$ time. If the network consists of K weakly coupled subnetworks, each containing n components, then a direct method requires $k_d(Kn)^d + k_s(Kn)^s$ time. When a block-iterative method with parallel processor is used, which converges in N steps, it requires $k_d n^d + N k_s n^s$ time for completion. It has been reported (Atalar (1989)) that for a typical VLSI circuit, $d \approx 1.5$ and $s \approx 1.2$ (due to sparsity of the network equations). Thus, for $K = 100$, up to $N = 100^{1.2} \approx 250$ iterations are due to substitutions alone. Because of sparsity of the off-diagonal blocks of $\tilde{A}$ in (5.5), the communication overhead is negligible. If ε is sufficiently small, few (typically 10) iterations are enough for convergence. As a result, even if a dominating factor is the operation count involved in substitutions, a block iterative solution based on an overlapping decomposition proves to be considerably more efficient than a direct solution procedure.

6. Conclusions. We have presented a decomposition algorithm that can be used to generate a whole range of nested partitions of a given matrix. The most attractive property of the algorithm is its simplicity in offering a trade-off between size of diagonal blocks and level of their mutual coupling. The decomposition scheme is suitable for partitioning of linear equations for block-iterative computations on parallel-pipeline multiprocessor architectures. In this application, assignments of subsystems having appropriate size and mutual coupling to individual processors is essential for a fast convergence of the solution process. Other applications of the proposed decomposition scheme are considered in partitioning and clustering of models in fields as diverse as economics (Simon and Ando (1961)) and computer systems (Courtois (1977)), aggregation of Markov chains (Kemeny and Snell (1960)), and decentralized control and estimation (Sezer and Šiljak (1986), Šiljak (1991)).

REFERENCES

I. M. ARABACIOGLU, M. E. SEZER, AND O. H. ORAL (1986), *Overlapping decomposition of large scale systems into weakly coupled subsystems*, Computational and Combinatorial Methods in System Theory, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, pp. 135–147.

A. ATALAR (1989), private communication with M. E. Sezer.

J. L. CALVET AND A. TITLI (1989), *Overlapping vs partitioning in block-iteration methods: Application in large-scale system theory*, Automatica, 25, pp. 137–145.

L. O. CHUA, C. A. DESOER, AND E. K. KUH (1987), *Linear and Nonlinear Circuits*, McGraw-Hill, New York.

P. J. COURTOIS (1977), *Decomposability*, Academic Press, New York.

I. S. DUFF (1981a), *Sparse Matrices and Their Uses*, Academic Press, New York.

——— (1981b), *On algorithms for obtaining a maximum transversal*, ACM Trans. Math. Software, 7, pp. 315–330.

L. A. HAGEMAN AND D. M. YOUNG (1981), *Applied Iterative Methods*, Academic Press, New York.

F. HARARY (1969), *Graph Theory*, Addison-Wesley, Reading, MA.

M. IKEDA AND D. D. ŠILJAK (1980), *Overlapping decompositions, expansions, and contractions of dynamic systems*, Large Scale Systems, 1, pp. 29–38.

E. KASZKUREWICZ, A. BHAYA, AND D. D. ŠILJAK (1990), *On the convergence of parallel asynchronous block-iterative computations*, Linear Algebra Appl., 131, pp. 139–160.

J. G. KEMENY AND J. L. SNELL (1960), *Finite Markov Chains*, Van Nostrand-Reinhold, Princeton, NJ.

L. W. NAGEL (1975), SPICE2: *A computer program to simulate semiconductor circuits*, Tech. Report ERL-M520, University of California, Berkeley, CA.

Y. OHTA AND D. D. ŠILJAK (1985), *Overlapping block diagonal dominance and existence of Liapunov functions*, J. Math. Anal. Appl., 112, pp. 396–410.

L. PILLAGE (1989), *Asymptotic waveform evaluation for timing analysis*, Res. Report CMUCAD-89-34, Carnegie-Mellon University, Pittsburgh, PA.

F. ROBERT (1969), *Blocs-H-matrices et convergence des methodes iteratives classiques par blocs*, Linear Algebra Appl., 2, pp. 223–265.

M. E. SEZER AND D. D. ŠILJAK (1986), *Nested ε decompositions and clustering of complex systems*, Automatica, 22, pp. 321–331.

——— (1988), *Robust stability of discrete systems*, Internat. J. Control, 48, pp. 2055–2063.

D. D. ŠILJAK (1979), *Overlapping decentralized control*, in Handbook of Large Scale Systems Engineering Applications, M. G. Singh and A. Titli, eds., North-Holland, New York, pp. 145–166.

——— (1991), *Decentralized Control of Complex Systems*, Academic Press, Cambridge, MA.

H. A. SIMON AND A. ANDO (1961), *Aggregation of variables in dynamic systems*, Econometrica, 29, pp. 111–138.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 534–540, July 1991

010

A VARIANT OF THE GOHBERG–SEMENCUL FORMULA INVOLVING CIRCULANT MATRICES*

GREGORY AMMAR† AND PAUL GADER‡

Abstract. The Gohberg–Semencul formula expresses the inverse of a Toeplitz matrix as the difference of products of lower triangular and upper triangular Toeplitz matrices. In this paper the idea of cyclic displacement structure is used to show that the upper triangular matrices in this formula can be replaced by circulant matrices. The resulting computational savings afforded by this modified formula is discussed.

Key words. Toeplitz matrix, circulant matrix, Gohberg–Semencul formula, displacement, cyclic displacement, fast Fourier transform

AMS(MOS) subject classifications. 65F05, 15A09, 15A23

1. Introduction. Let $M = [\mu_{j-k}]_{j,k=0}^{n-1}$ be a real symmetric positive-definite Toeplitz matrix of order n. There are several well-known $O(n^2)$ algorithms for solving the linear system of equations $Mx = b$, and more recently, several $O(n \log^2 n)$ algorithms have been developed. See, for example, [16], [12], [11], [9], [1], and [2] and the references contained therein. Algorithms from both of these classes often rely, either implicitly or explicitly, on the Gohberg–Semencul formula [8], which provides a decomposition of M^{-1} into the sum of products of lower triangular and upper triangular Toeplitz matrices. Although we will consider M to be a real positive-definite Toeplitz matrix, formulas presented by Gohberg and Semencul apply to the inverse of any invertible Toeplitz matrix.

Given $x \in \mathbb{R}^n$, let $L(x)$ denote the lower triangular Toeplitz matrix whose first column is x. Let $e_0, e_1, \cdots, e_{n-1}$ be the columns of the identity matrix I of order n, and let $Z_n = L(e_1)$ be the *downshift matrix* of order n. Since the Toeplitz matrix M is positive definite, the last column of M^{-1} can be written as $M^{-1}e_{n-1} = r/\delta_{n-1}$, where $\delta_{n-1} > 0$ and $r = [\rho_j]_{j=0}^{n-1}$ with $\rho_{n-1} = 1$. The *Gohberg–Semencul formula* is then given by

$$\delta_{n-1}M^{-1} = L(r_1)L(r_1)^T - L(r_0)L(r_0)^T, \tag{1}$$

where $r_0 = Z_n r$ and $r_1 = [\rho_{n-j-1}]_0^{n-1}$. See, for example, [7] and [10].

Many algorithms for the solution of a Toeplitz system can be considered as a two-phase procedure:

Phase 1: The computation of r and δ_{n-1}.

Phase 2: The computation of $M^{-1}b$ using the Gohberg–Semencul formula (1).

Algorithms for Phase 1 include the Levinson–Durbin algorithm (see, e.g., [9]), the split Levinson algorithm [3], and the generalized Schur algorithm [1], [2]. If n is sufficiently large, Phase 2 can be efficiently implemented in $O(n \log n)$ operations using fast Fourier transform techniques [11].

We note in passing that the components of r are the coefficients of the monic *Szegö polynomial* $\chi_{n-1}(\lambda) = \sum_{j=0}^{n-1} \rho_j \lambda^j$ of degree $n-1$ determined by M, and δ_{n-1} is the

* Received by the editors October 31, 1988; accepted for publication (in revised form) April 15, 1990.

† Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115 (ammar@math.niu.edu). The research of this author was supported in part by National Science Foundation grant DMS-8704196.

‡ Algorithms and Robot-Vision Department, Environmental Research Institute of Michigan, P.O. Box 8618, Ann Arbor, Michigan 48107 (gader@ssdd3260d.erim.org). The research of this author was supported in part by the Institute for Mathematics and Its Applications with funds provided by the National Science Foundation.

norm of χ_{n-1} in the inner product determined by M. Furthermore, the Gohberg–Semencul formula is a manifestation of the Christoffel–Darboux formula for Szegö polynomials. See, for example, [13] and [1] for more details.

The notion of displacement structure underlies many techniques for solving Toeplitz systems of equations, and the Gohberg–Semencul formula fits naturally into this framework. Moreover, displacement structure can be used to extend algorithms for Toeplitz matrices to other classes of matrices [12], [5]. In fact, every square matrix can be written as the sum of products of upper triangular and lower triangular Toeplitz matrices. Furthermore, the number of terms in this sum is small if the matrix is "close" to a Toeplitz matrix in the sense that its *displacement rank* is small. Matrices of small displacement rank can then be treated using extended versions of algorithms for Toeplitz matrices.

The notion of displacement structure has been generalized to include circulant matrices and other group matrices in [6]. In § 2 we show how the circulant displacement representation of the inverse of a Toeplitz matrix can be used to derive the following factorization of the inverse of the positive definite Toeplitz matrix M.

PROPOSITION 1.

$$\delta_{n-1} M^{-1} = L(r_1)C(r_1)^T - L(r_0)C(r_1), \tag{2}$$

where $C(r)$ denotes the circulant matrix whose first column is r.

In § 3 we discuss the computational savings resulting from the use of this formula. We obtain a computational savings of more than 35 percent over the Gohberg–Semencul formula in the second phase of a Toeplitz solver when n is a power of two.

Toeplitz inversion formulas involving circulant matrices have also been presented by Lerer and Tismenetsky [14].

2. A circulant Gohberg–Semencul formula. In the following, all matrices are assumed to be real and $n \times n$. If the *displacement* of a square matrix A is given by the sum of α outer products,

$$A - ZAZ^T = \sum_{m=1}^{\alpha} x_m y_m^T$$

where $x_m, y_m \in \mathbb{R}^n$, then

$$A = \sum_{m=1}^{\alpha} L(x_m)L(y_m)^T.$$

This is the *displacement representation* of A developed in [5] and [12]. The usefulness of this representation for Toeplitz matrices stems from the fact that a Toeplitz matrix and its inverse have displacement rank $\alpha \leqq 2$. In particular, the Gohberg–Semencul formula follows from the fact that

$$M^{-1} - ZM^{-1}Z^T = \frac{1}{\delta_{n-1}}(r_1 r_1^T - r_0 r_0^T).$$

The circulant analogue of displacement structure is based on replacing the downshift matrix Z above with the *cyclic downshift matrix* $E = C(e_1)$. In particular, we will need the following result from [6].

PROPOSITION 2. *If the cyclic displacement of A is given as the sum*

$$A - EAE^T = \sum_{m=1}^{\alpha} x_m y_m^T, \tag{3}$$

then

$$A = C_l + \sum_{m=1}^{\alpha} L(x_m)C(y_m)^T, \tag{4}$$

where C_l is the circulant matrix with the same last row as that of A.

Given (3) and (4), the derivation of the corresponding analogue of the Gohberg–Semencul formula is straightforward. Let M be a real positive-definite Toeplitz matrix and $A = \delta_{n-1}M^{-1}$. Then the cyclic displacement of A is given by

$$\begin{aligned} A - EAE^T &= A - ZAZ^T - e_0 r_0^T - r_0 e_0^T - e_0 e_0^T \\ &= r_1 r_1^T - r_0 r_0^T - e_0(e_0 + r_0)^T - r_0 e_0^T \end{aligned}$$

since $E = Z + e_0 e_{n-1}^T$ and $ZAe_{n-1} = r_0$. Proposition 2 then yields

$$A = L(r_1)C(r_1)^T - L(r_0)C(r_0)^T - C(r_1) - L(r_0) + C(r_1)$$

because $C(e_0 + r_0)^T = C(r_1)$ and $L(e_0) = C(e_0) = I$. Hence,

$$\begin{aligned} A &= L(r_1)C(r_1)^T - L(r_0)(C(r_0) + I)^T \\ &= L(r_1)C(r_1)^T - L(r_0)C(r_1), \end{aligned} \tag{5}$$

which is the desired formula.

3. Computational implications. We now show how the analogue of the Gohberg–Semencul formula derived in § 2 leads to a more efficient way to calculate $M^{-1}b$. The increased computational efficiency is due to the fact that multiplication by a circulant matrix is roughly twice as fast as multiplication by a triangular Toeplitz matrix of the same size. In fact, the efficient multiplication of a Toeplitz matrix and a vector is achieved by embedding the Toeplitz matrix in a circulant matrix of twice the size.

We first recall some fundamental facts regarding fast Fourier transforms and efficient circulant-vector multiplication. Let $\omega_n = e^{2\pi i/n}$ denote the principal nth root of unity. The discrete Fourier transform (DFT) of the n-vector x is defined by $y = F_n x$, where $nF_n = [\omega_n^{-jk}]_{j,k=0}^{n-1}$, and the inverse discrete Fourier transform (IDFT) of y is $x = W_n y$, where $W_n \equiv F_n^{-1} = [\omega_n^{jk}]_0^{n-1} = n\bar{F}_n$. The computation of $F_n x$ and $W_n x$ can be performed in $O(n \log n)$ arithmetic operations using any one of many well-known techniques, collectively called fast Fourier transforms (FFTs). Let $\tau(n)$ denote the amount of computation required to perform one real FFT of order n.

Recall that the circulant-vector product $z = C(x)y$ is equal to the cyclic convolution of the vectors x and y, which we denote by $x * y$. Moreover, $z = x * y$ if and only if $F_n z = (F_n x) \cdot (F_n y)$, where $x \cdot y$ denotes the componentwise product of x and y. Consequently, $z = W_n((F_n x) \cdot (F_n y))$, so z can be computed in $3\tau(n) + O(n)$ arithmetic operations.

Let us now write the Gohberg–Semencul formula (1) as

$$\delta_{n-1}M^{-1} = A = T_1^T T_1 - T_0 T_0^T,$$

where $T_1^T = L(r_1)$ and $T_0 = L(r_0)$. Let $u = T_0^T b$, $v = T_1 b$, $r = T_0 u$, and $s = T_1^T v$. Then $Ab = s - r$. Note that

$$\begin{bmatrix} T_0 & T_1 \\ T_1 & T_0 \end{bmatrix} = C\left(\begin{bmatrix} r_0 \\ e_0 \end{bmatrix}\right), \qquad \begin{bmatrix} T_0^T & T_1^T \\ T_1^T & T_0^T \end{bmatrix} = C\left(\begin{bmatrix} 0 \\ r_1 \end{bmatrix}\right)$$

are circulant matrices of order $2n$. The following convolution formulas can therefore be used to calculate r and s. The symbols $\times$ denote n-vectors that are irrelevant in the computation:

$$\begin{bmatrix} u \\ \times \end{bmatrix} := \begin{bmatrix} 0 \\ r_1 \end{bmatrix} * \begin{bmatrix} b \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} \times \\ v \end{bmatrix} := \begin{bmatrix} r_0 \\ e_0 \end{bmatrix} * \begin{bmatrix} b \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} r \\ \times \end{bmatrix} := \begin{bmatrix} r_0 \\ e_0 \end{bmatrix} * \begin{bmatrix} u \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} s \\ \times \end{bmatrix} := \begin{bmatrix} 0 \\ r_1 \end{bmatrix} * \begin{bmatrix} 0 \\ v \end{bmatrix}.$$

In terms of FFTs, the computations can be performed as follows:

$$t := F_{2n}\begin{bmatrix} b \\ 0 \end{bmatrix}, \qquad p := F_{2n}\begin{bmatrix} r_0 \\ e_0 \end{bmatrix}, \qquad q := F_{2n}\begin{bmatrix} 0 \\ r_1 \end{bmatrix},$$

$$\begin{bmatrix} u \\ \times \end{bmatrix} := W_{2n}(q \cdot t), \qquad \begin{bmatrix} \times \\ v \end{bmatrix} := W_{2n}(p \cdot t),$$

$$z := F_{2n}\begin{bmatrix} u \\ 0 \end{bmatrix}, \qquad w := F_{2n}\begin{bmatrix} 0 \\ v \end{bmatrix},$$

$$\begin{bmatrix} s - r \\ \times \end{bmatrix} := W_{2n}(q \cdot w - p \cdot z).$$

Thus, $x = M^{-1}b$ can be computed in $8\tau(2n) + O(n)$ computations. This is the implementation described in [11]. However, one of these FFTs can be eliminated using the observation that $p = \bar{q}$. This follows from the fact that

$$\begin{bmatrix} r_0 \\ e_0 \end{bmatrix} = K_{2n}\begin{bmatrix} 0 \\ r_1 \end{bmatrix}$$

and $F_n K_n = \bar{F}_n$, where $K_n = [e_0, e_{n-1}, \cdots, e_1]$ denotes the *reflection matrix* of order n. Thus, the implementation of the Gohberg–Semencul formula requires at most $7\tau(2n) + O(n) = 14\tau(n) + O(n)$ arithmetic operations.

Let us now write (2) as

$$\delta_{n-1} M^{-1} = A = T_1^T C_1 - T_0 C_0,$$

where T_1 and T_0 are as above, $C_0 = C(r_1)$ and $C_1 = C_0^T = C(K_n r_1)$. Define $u = C_0 b$, $v = C_1 b$, $r = T_0 u$, and $s = T_1^T v$. Then the following convolution formulas can be used to calculate r and s:

$$u := r_1 * b, \qquad v := (K_n r_1) * b,$$

$$\begin{bmatrix} r \\ \times \end{bmatrix} := \begin{bmatrix} r_0 \\ e_0 \end{bmatrix} * \begin{bmatrix} u \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} s \\ \times \end{bmatrix} := \begin{bmatrix} 0 \\ r_1 \end{bmatrix} * \begin{bmatrix} v \\ 0 \end{bmatrix}.$$

Note that $F_n(K_n r_1) = \overline{F_n r_1}$, so in terms of FFTs, we have

$$t := F_n b, \qquad p := F_n r_1, \qquad u := W_n(p \cdot t), \qquad v := W_n(\bar{p} \cdot t),$$

$$q := F_{2n}\begin{bmatrix} 0 \\ r_1 \end{bmatrix}, \qquad z := F_{2n}\begin{bmatrix} u \\ 0 \end{bmatrix}, \qquad w := F_{2n}\begin{bmatrix} 0 \\ v \end{bmatrix}, \tag{6}$$

$$\begin{bmatrix} s - r \\ \times \end{bmatrix} := W_{2n}(q \cdot w - \bar{q} \cdot z).$$

These computations require $4\tau(n) + 4\tau(2n) + O(n) = 12\tau(n) + O(n)$. However, we can reduce this operation count further as follows.

Define the permutation matrix P_{2n} by $P_{2n}^T y = \begin{bmatrix} y_0' \\ y_1' \end{bmatrix}$, where $y_0' = [\eta_{2j}]_0^{n-1}$ and $y_1' = [\eta_{2j+1}]_0^{n-1}$ are the *even-* and *odd-indexed parts* of $y = [\eta_j]_0^{2n-1}$, respectively. Let $y = F_{2n}\begin{bmatrix} x_0 \\ x_1 \end{bmatrix}$, where x_0 and x_1 are n-vectors. Then it is easy to see that

$$\begin{bmatrix} y_0' \\ y_1' \end{bmatrix} = P_{2n}^T y = \begin{bmatrix} F_n & F_n \\ F_n D_n' & -F_n D_n' \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} F_n(x_0 + x_1) \\ F_n D_n'(x_0 - x_1) \end{bmatrix},$$

where $D_n' = \operatorname{diag}[\omega_{2n}^{-k}]_{k=0}^{n-1}$. Thus,

$$P_{2n}^T F_{2n} \begin{bmatrix} x \\ 0 \end{bmatrix} = \begin{bmatrix} F_n x \\ F_n D_n' x \end{bmatrix},$$

$$P_{2n}^T F_{2n} \begin{bmatrix} 0 \\ x \end{bmatrix} = \begin{bmatrix} F_n x \\ -F_n D_n' x \end{bmatrix}.$$

It is shown in [2], and easily verified, that $F_n D_n' x$, where $x \in \mathbb{R}^n$, can be computed using one complex FFT of order $n/2$, which requires $\tau(n) + O(n)$ operations.

These formulas allow us to reduce the computation of the FFTs in display (6) as follows:

1. The jth component of p is the $2j$th component of q, so only q needs to be calculated, saving $\tau(n)$.

2. z can be obtained from $p \cdot t$ and u using $\tau(n) + O(n)$ operations, saving roughly $\tau(n)$ operations. The same observation holds for the computation of w from $\bar{p} \cdot t$ and v.

Consequently, the computation of $x = M^{-1}b$ using our modification of the Gohberg–Semencul formula requires at most $5\tau(n) + 2\tau(2n) + O(n)$ or $9\tau(n) + O(n)$ computations. This represents a computational savings of $\frac{5}{14}$ over the implementation of the Gohberg–Semencul formula as described above, neglecting the $O(n)$ terms.

Numerical experiments show that a savings of $\frac{5}{14}$ (about 36 percent) in CPU time is indeed achieved. The results are summarized in Table 1, which shows average CPU times for the implementations of the Gohberg–Semencul formula (GS) and our circulant variant of the Gohberg–Semencul formula (CGS), as described above. Also displayed are the ratios of the average time used by CGS to those of GS. The experiments were performed on the VAX 11/750 at Northern Illinois University.

Our modified formula (2) can be used to achieve over 35 percent computational savings in the second phase of any two-phase Toeplitz solver. Of course, since efficient

TABLE 1
Timing comparison (CPU *seconds*).
Gohberg–Semencul (GS) *versus circulant variant* (CGS).

n	CGS	GS	CGS/GS
64	0.130	0.202	0.646
128	0.265	0.414	0.639
256	0.558	0.877	0.636
512	1.184	1.855	0.638
1024	2.499	3.908	0.639
2048	5.284	8.256	0.640
4096	11.167	17.394	0.642
8192	23.695	36.825	0.643
16384	50.858	79.182	0.642

TABLE 2
Operation counts.

Algorithm	Number of real arithmetic operations
1a. The Levinson–Durbin algorithm	$2n^2$
1b. The split Levinson algorithm	$\frac{3}{2}n^2$
1c. The generalized Schur algorithm	$8n \log_2^2 n - 2n \log_2 n$
2a. The Gohberg–Semencul formula (1)	$28n \log_2 n$
2b. The circulant Gohberg–Semencul formula (2)	$18n \log_2 n$

algorithms for Phase 1 require a higher-order amount of computation than $O(n \log n)$, the amount of computation for Phase 2 relative to Phase 1 will decrease as n increases. Closer inspection shows that this ratio is not insignificant for moderately sized n. Moreover, the computational work of Phase 2 relative to that of Phase 1 decreases slowly as n increases.

In Table 2 we list the number of real arithmetic operations (multiplications and additions) required by the algorithms for Phase 1 mentioned in the Introduction, as well as operation counts for the implementations of formulas (1) and (2) described above. We have neglected $O(n)$ terms in these operation counts. For Algorithms 1a, 2a, and 2b, we have assumed that $n = 2^\nu$, and that the Fourier transforms are performed using split-radix FFT algorithms [4], [15], which require $\tau(n) = 2n \log_2 n + O(n)$ real arithmetic operations.

From Table 2 it is easy to see that the amount of computation for Phase 2 compared with Phase 1 is significant for moderately sized n. For example, if $n = 128$ the Gohberg–Semencul formula requires 86 percent of the work required by the Levinson–Durbin algorithm, and 102 percent of that of the split Levinson algorithm. If $n = 2{,}048$ the Gohberg–Semencul formula requires over 32 percent of the computation of the generalized Schur algorithm, and the formula (2) will result in about a 10 percent overall savings in the solution of $Mx = b$. Even for $n = 2^{24}$, the 35 percent savings in Phase 2 results in a 5 percent overall savings.

The value of our more efficient formula increases dramatically in situations in which $M^{-1}b_j$ is to be obtained for several different vectors b_j. One instance of this situation is in the iterative improvement of solutions. Another instance in which several systems of equations with the same Toeplitz coefficient matrix arises is in the calculation of multistep predictors in time series analysis. (The Yule–Walker equations are used to calculate single-step predictors; multistep predictors are calculated using the same matrix and different right-hand sides.) In these cases the improved efficiency in the computations in Phase 2 will be of great benefit.

REFERENCES

[1] G. S. AMMAR AND W. B. GRAGG, *The generalized Schur algorithm for the superfast solution of Toeplitz systems*, in Rational Approximation and Its Applications in Mathematics and Physics, J. Gilewicz, M. Pindor, and W. Siemaszko, eds., Lecture Notes in Mathematics 1237, Springer-Verlag, Berlin, 1987, pp. 315–330.

[2] ———, *Superfast solution of real positive definite Toeplitz systems*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 61–76.

[3] P. DELSARTE AND Y. V. GENIN, *The split Levinson algorithm*, IEEE Trans. Acoust. Speech Signal Process., 34 (1986), pp. 470–478.

[4] P. DUHAMEL, *Implementation of "split-radix" FFT algorithms for complex, real, and real-symmetric data*, IEEE Trans. Acoust. Speech Signal Process., 34 (1986), pp. 285–295.

[5] B. FRIEDLANDER, M. MORF, T. KAILATH, AND L. LJUNG, *New inversion formulas for matrices classified in terms of their distance from Toeplitz matrices*, Linear Algebra Appl., 27 (1979), pp. 31–60.

[6] P. GADER, *Displacement operator based decompositions of matrices using circulants or other group matrices*, Linear Algebra Appl., 139 (1990), pp. 111–131.

[7] I. C. GOHBERG AND I. A. FEL'DMAN, *Convolution Equations and Projection Methods for Their Solution*, American Mathematical Society, Providence, RI, 1974.

[8] I. C. GOHBERG AND A. A. SEMENCUL, *On the inversion of finite Toeplitz matrices and their continuous analogs*, Mat. Issled, 7 (1972), pp. 201–223. (In Russian.)

[9] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1984.

[10] I. S. IOHVIDOV, *Hankel and Toeplitz Matrices and Forms: Algebraic Theory*, Birkhäuser, Boston, MA, 1982.

[11] J. R. JAIN, *An efficient algorithm for a large Toeplitz set of linear equations*, IEEE Trans. Acoust. Speech Signal Process., 27 (1979), pp. 612–615.

[12] T. KAILATH, S. Y. KUNG, AND M. MORF, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.

[13] T. KAILATH, A. VIEIRA, AND M. MORF, *Inverses of Toeplitz operators, innovations, and orthogonal polynomials*, SIAM Rev., 20 (1978), pp. 106–119.

[14] L. LERER AND M. TISMENETSKY, *Generalized Bezoutian and the inversion problem for block matrices*, I. *General scheme*, Integral Equations Operator Theory, 9 (1986), pp. 790–819.

[15] H. V. SORENSEN, D. L. JONES, M. T. HEIDEMAN, AND C. S. BURRUS, *Real-valued fast Fourier transform algorithms*, IEEE Trans. Acoust. Speech Signal Process., 35 (1987), pp. 849–863.

[16] W. TRENCH, *An algorithm for the inversion of finite Toeplitz matrices*, SIAM J. Appl. Math., 12 (1964), pp. 515–522.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 541–548, July 1991

011

PROPERTIES OF THE INVERSE OF THE GAUSSIAN MATRIX*

M. J. C. GOVER†

Abstract. The Gaussian matrix is a symmetric Toeplitz matrix and in addition the elements in its first row form a pattern. This enables a specific formula to be obtained, in the nonsingular case, for the elements in the first row of the inverse. Recurrence formulae are then obtained which enable this inverse to be obtained in $\frac{1}{2}n^2$ flops, as against $2n^2$ flops, for a general symmetric Toeplitz matrix using the Trench algorithm.

Key words. Gaussian matrix, Toeplitz matrix, inverse

AMS(MOS) subject classifications. 15A09, 65F09, 62

1. Introduction. It is well known that in applied problems involving matrices, the elements of these matrices often form patterns and many of these special matrices are named after mathematicians of the past such as Toeplitz, Hilbert, Markov, Gauss, Hankel, and Vandermonde. All the above types of matrices occur in statistical applications and are, in fact, related as follows:

(i) If M and G are Markovian and Gaussian matrices, respectively, then they are special cases of an nth order symmetric Toeplitz matrix T, where

$$[T]_{i+1,j+1} = [T]_{ij}, \qquad i,j = 1,2,\cdots,n-1.$$

This matrix T is defined by its first row $[t_{11},\cdots,t_{1n}]$ and M and G have first rows

$$[1,a,\cdots,a^{n-1}] \quad \text{and} \quad [1,a^{1^2},\cdots,a^{(n-1)^2}], \tag{1.1}$$

respectively.

The interest in the Markovian and Gaussian matrices in statistics is because M is derived from the covariance function $e^{-|x|}$ and G is derived from the Gaussian covariance function $e^{-x^2/2}$.

(ii) A Hankel matrix L with its rows reversed becomes a Toeplitz matrix. Specifically, if the reverse unit matrix is $J = [\delta_{i,n-j+1}]$, where δ_{ik} is the Kronecker delta, then $LJ = T$.

The Hilbert matrix $H = [h_{ij}]$ is a special case of a Hankel matrix and is defined by

$$h_{ij} = \frac{1}{i+j-1}, \qquad i,j = 1,2,\cdots,n. \tag{1.2}$$

(iii) If the jth column of the Vandermonde matrix V is $[1, a_j, a_j^2, \cdots, a_j^{n-1}]$, then VV^T is a Hankel matrix.

Properties of most of the above-named matrices are well known and can be found in various papers and books such as [1], [2], [6], [7], and [10]. For statistical applications, [5], [8], [9], and [11] will be found useful.

Because of the special pattern of the generating elements of the Hilbert, Markovian, and Gaussian matrices, it might be expected that special efficient numerical techniques could be developed for them and that the elements of their inverses in the nonsingular case could be found explicitly. For the Hilbert and Markovian matrices some results are well known.

* Received by the editors May 22, 1989; accepted for publication (in revised form) May 17, 1990.

† Department of Mathematics, University of Bradford, Bradford, West Yorkshire, BD7 1DP, United Kingdom (mts617@uk.ac.bradford.cyber2).

The Markovian matrix M defined in (1.1) has a tridiagonal inverse, $M^{-1} = [\mu_{ij}]$, given by

$$\mu_{ij} = \begin{cases} \dfrac{1}{1-a^2}, & i=j=1,n, \\ \dfrac{1+a^2}{1-a^2}, & i=j=2,3,\cdots,n-1, \\ -\dfrac{a}{1-a^2}, & |i-j|=1, \\ 0 & \text{otherwise.} \end{cases} \tag{1.3}$$

Thus its inverse is found in $O(1)$ flops and the solution of $M\mathbf{x} = \mathbf{b}$ can be found in $O(n)$ flops.

The elements of the inverse of the Hilbert matrix H, defined by (1.2), can also be given explicitly as

$$[H^{-1}]_{ij} = \frac{(-1)^{i+j}(n+i-1)!(n+j-1)!}{[(i-1)!(j-1)!]^2(n-i)!(n-j)!(i+j-1)}, \qquad i,j=1,2,\cdots,n. \tag{1.4}$$

It is well known [12] that H is very ill conditioned so that solving $H\mathbf{x} = \mathbf{b}$ accurately is difficult.

The above results can all be found in [4] and [9], where it is stated that there do not seem to be explicit formulae for the elements of the inverse of a Gaussian matrix.

Formulae equivalent to the results of Theorems 2.1 and 2.4 can be found in [3] for the case of a positive-definite Gaussian matrix, although this matrix is not referred to by name. The derivations in this paper, however, are different and use only very elementary ideas.

The purpose of this paper is to obtain, in an elementary way, some results on the less well known Gaussian matrix. We begin by obtaining its determinant, from which we can obtain criteria for nonsingularity. We then give an explicit formula for the elements in the first row of the inverse of a nonsingular Gaussian matrix. However, as far as efficient calculation is concerned, this form is not used. Some recurrence formulae are obtained, both for this first row and for the interior elements that enable the inverse to be found in approximately $\frac{1}{2}n^2$ flops, as opposed to the usual $2n^2$ flops required for the inverse of a general symmetric Toeplitz matrix of order n using the Trench method [13].

2. Gaussian matrices. We consider the Gaussian matrix G_n based on the Gaussian covariance function $e^{-x^2/2}$. Thus it is a symmetric Toeplitz matrix with the first row

$$[1, a, a^4, a^9, \cdots, a^{(n-1)^2}]. \tag{2.1}$$

We can easily see that $a = 0$ gives $G_n = I$ and $a = \pm 1$ gives two real symmetric Toeplitz matrices of rank one. However, the first theorem in this section will show that there are other values of a for which G_n is singular.

After finding a recurrence relationship between the minors of G_n and G_{n-1}, we obtain an explicit form for these minors and hence for the elements in the first row of G_n^{-1}.

This explicit form is not particularly suitable for numerical calculation and hence two recurrence formulae are obtained that give an efficient numerical method for finding the first row of G_n^{-1} in $O(n)$ flops, as opposed to $O(n^2)$ flops for a general Toeplitz matrix inverse.

Finally, a recurrence formula is derived so that the internal elements of G_n^{-1} can be found in approximately $\frac{1}{2}n^2$ flops. The complete inverse is thus also found in approximately $\frac{1}{2}n^2$ flops compared with $2n^2$ flops for a general symmetric Toeplitz matrix.

We begin by finding an explicit formula for the determinant of G_n.

THEOREM 2.1. *The determinant of G_n of order n is*

$$|G_n| = \prod_{k=1}^{n-1} (1-a^{2k})^{n-k}. \tag{2.2}$$

Proof. If

$$G_n = \begin{bmatrix} 1 & a & a^4 & \cdots & a^{(n-1)^2} \\ a & 1 & a & \cdots & a^{(n-2)^2} \\ a^4 & a & 1 & \cdots & a^{(n-3)^2} \\ \vdots & \vdots & \vdots & & \vdots \\ a^{(n-1)^2} & a^{(n-2)^2} & a^{(n-3)^2} & \cdots & 1 \end{bmatrix},$$

then carrying out the column operations

$$\text{column } i - a^{2i-3} \text{ column } (i-1), \qquad i = n, n-1, \cdots, 2$$

on $|G_n|$ gives

$$|G_n| = \begin{vmatrix} 1 & 0 & 0 & \cdots & 0 \\ a & 1-a^2 & a(1-a^2) & \cdots & a^{(n-2)^2}(1-a^2) \\ a^4 & a(1-a^4) & (1-a^4) & \cdots & a^{(n-3)^2}(1-a^4) \\ \vdots & \vdots & \vdots & & \vdots \\ a^{(n-1)^2} & a^{(n-2)^2}(1-a^{2n-2}) & a^{(n-3)^2}(1-a^{2n-2}) & \cdots & 1-a^{2n-2} \end{vmatrix}$$

$$= \prod_{k=1}^{n-1} (1-a^{2k}) |G_{n-1}|.$$

Continuing this reduction, ending with $|G_1| = 1$, we obtain (2.2). □

We can now easily obtain criteria for the singularity of G_n.

COROLLARY. *G_n is singular if a is an lth root of unity for some l, $l = 2, 4, \cdots, 2n-2$.*

Proof. If $|G_n| = 0$, then, from (2.2), $a^{2k} = 1$ for some k, $k = 1, 2, \cdots, n-1$. □

Remark 2.1. Theorem 2.1 is proved in different ways in [3] and [9].

Remark 2.2. Since the leading principal minor of order k is $|G_k|$, it is easily seen that if $|G_n| \neq 0$, then so is $|G_k|$, $k = 1, 2, \cdots, n-1$, so that if G_n is nonsingular, then it is strongly nonsingular. Also, in the statistical case, if $|a| < 1$ and a is real, then G_n is positive definite.

We next determine a recurrence formula for the minors of the first row of G_n, before finding an explicit form of these minors.

THEOREM 2.2. *If $|G_n^{(1,j)}|$ is the minor formed by omitting the first row and jth column from $|G_n|$, then*

$$|G_n^{(1,j)}| = a \prod_{\substack{k=1 \\ k \neq j-1}}^{n-1} (1-a^{2k}) |G_{n-1}^{(1,j-1)}|, \qquad j = 2, 3, \cdots, n \tag{2.3}$$

with $|G_n^{(1,1)}| = |G_{n-1}|$.

Proof. First, it is obvious that $G_n^{(1,1)} = G_{n-1}$. Now to find $G_n^{(1,j)}$, $j = 2, 3, \cdots, n$, consider the matrix $G_{n-1,n}$ formed from G_n by omitting its first row. Thus

$$G_{n-1,n} = \begin{bmatrix} a & 1 & a & \cdots & a^{(n-2)^2} \\ a^4 & a & 1 & \cdots & a^{(n-3)^2} \\ a^9 & a^4 & a & \cdots & a^{(n-4)^2} \\ \vdots & \vdots & \vdots & & \vdots \\ a^{(n-1)^2} & a^{(n-2)^2} & a^{(n-3)^2} & \cdots & 1 \end{bmatrix}. \tag{2.4}$$

If we now consider a formal determinant $|G_{n-1,n}|$ and carry out the row operations

$$\text{row } i - a^{2i-1} \text{ row } (i-1), \qquad i = n-1, n-2, \cdots, 2,$$

we obtain

$$\begin{aligned} |G_{n-1,n}| &= \begin{vmatrix} a & 1 & a & \cdots & a^{(n-1)^2} \\ 0 & a(1-a^2) & 1-a^4 & \cdots & a^{(n-3)^2}(1-a^{2n-2}) \\ 0 & a^4(1-a^2) & a(1-a^4) & \cdots & a^{(n-4)^2}(1-a^{2n-2}) \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & a^{(n-2)^2}(1-a^2) & a^{(n-3)^2}(1-a^4) \cdots & & 1-a^{2n-2} \end{vmatrix} \\ &= a \prod_{k=1}^{n-1} (1-a^{2k}) |G_{n-2,n-1}|. \end{aligned} \tag{2.5}$$

In order to find $|G_n^{(1,j)}|$ we omit the jth column from $|G_{n,n-1}|$ in (2.5). This is equivalent to removing the $(j-1)$th column from $|G_{n-2,n-1}|$ and omitting the factor $(1 - a^{2j-2})$ from (2.5), which proves the theorem. □

From the previous theorem we can obtain an explicit form of $|G_n^{(1,j)}|$ and hence an explicit form of the elements in the first row of G_n^{-1}.

THEOREM 2.3. *If* $|G_n^{(1,j)}|$ *is defined as in Theorem* 2.2 *then, with* $\prod_{k=\alpha}^{\alpha-1} = 1$ *for* $\alpha = 1$ *or* n,

$$|G_n^{(1,j)}| = \frac{a^{j-1} \prod_{k=1}^{n-j} (1-a^{2k})^{n-k-1} \prod_{k=n-j+1}^{n-1} (1-a^{2k})^{n-k}}{\prod_{k=1}^{j-1} (1-a^{2k})}, \tag{2.6}$$

$$j = 1, 2, \cdots, n.$$

Proof. Setting $j = 1$ in (2.6) gives

$$|G_n^{(1,1)}| = \prod_{i=1}^{n-1} (1-a^{2i})^{n-i-1} = \prod_{i=1}^{n-2} (1-a^{2i})^{n-i-1}.$$

Since we know that $|G_n^{(1,1)}| = |G_{n-1}|$, this agrees with (2.2) when n is replaced by $n-1$.

For $j = 2, 3, \cdots, n$, repeated use of (2.3) gives

$$|G_n^{(1,j)}| = a \prod_{\substack{k=1 \\ k \neq j-1}}^{n-1} (1-a^{2k}) a \prod_{\substack{k=1 \\ k \neq j-2}}^{n-2} (1-a^{2k}) \cdots a \prod_{\substack{k=1 \\ k \neq 1}}^{n-j+1} (1-a^{2k}) |G_{n-j+1}^{(1,1)}|$$

and with $|G_{n-j+1}^{(1,1)}| = |G_{n-j}|$, found from (2.2) with n replaced by $n-j$, we obtain (2.6) as required. □

The elements in the first (defining) row of G_n^{-1} are $(-1)^{j-1} |G_n^{(1,j)}| / |G_n|$ and so we have the following result immediately from the Theorems 2.1 and 2.3.

THEOREM 2.4. *If* $[G_n^{-1}]_{ij} = \beta_{ij}$ *and* $\prod_{k=1}^{0} = 1$, *then*

$$\beta_{1j} = \frac{(-1)^{j-1} a^{j-1}}{\prod_{k=1}^{j-1}(1-a^{2k}) \prod_{k=1}^{n-j}(1-a^{2k})}, \qquad j = 1, 2, \cdots, n. \tag{2.7}$$

Remark 2.3. Since we have already stated that a is nonzero, then all β_{1j}, and in particular β_{11}, are nonzero.

Remark 2.4. In [9] it was stated that it did not appear possible to find an explicit form of β_{1j}. Equation (2.7) is also given implicitly in [3].

Although (2.7) gives an explicit way of calculating β_{1j}, it is certainly not the most efficient since a simple recurrence formula can be found as given below, the proof of which follows immediately from Theorem 2.4.

THEOREM 2.5. *If* $[G_n^{-1}]_{1j} = \beta_{1j}$, *then*

$$\beta_{1,j+1} = \frac{-a(1-a^{2n-2j})}{(1-a^{2j})} \beta_{1j}, \qquad j = 2, 3, \cdots, n-1 \tag{2.8}$$

with

$$\beta_{11} = 1 \Big/ \prod_{k=1}^{n-1} (1-a^{2k}). \tag{2.9}$$

In fact, even this result can be improved for about half of the coefficients with use of the next result.

THEOREM 2.6. *Using the notation of the previous theorem, we have*

$$\beta_{1j} = (-1)^{n-1} a^{2j-n-1} \beta_{1,n-j+1}, \qquad j = \left[\frac{n+3}{2}\right], \cdots, n. \tag{2.10}$$

Remark 2.5. The use of Theorems 2.5 and 2.6 enables the first row of G_n^{-1} to be found in $O(n)$ flops as opposed to the usual $O(n^2)$ flops for general Toeplitz matrices. This is because it is calculated directly without having to carry out the iterations in the Levinson algorithm [13], which determine not only the first row of G_n^{-1} but also the first rows of G_k^{-1}, $k = 1, 2, \cdots, n-1$.

Example 2.1. Let $n = 7$ and $a = 1/\sqrt{2}$. Then the first row of G_n is

$$\left[1, \frac{1}{\sqrt{2}}, \frac{1}{4}, \frac{1}{2^4\sqrt{2}}, \frac{1}{2^8}, \frac{1}{2^{12}\sqrt{2}}, \frac{1}{2^{18}}\right]. \tag{2.11}$$

To find the first row of G_n^{-1} we first use Theorem 2.5 for $j = 1, 2, 3, 4$. From (2.9),

$$\beta_{11} = 1 \Big/ \frac{1}{2}\frac{3}{4}\frac{7}{8}\frac{15}{16}\frac{31}{32}\frac{63}{64}.$$

From (2.8),

$$\beta_{12} = -\frac{63}{64\sqrt{2}} \Big/ \frac{1}{2}\beta_{11}, \qquad \beta_{13} = -\frac{31}{32\sqrt{2}} \Big/ \frac{3}{4}\beta_{12},$$

$$\beta_{14} = -\frac{15}{16\sqrt{2}} \Big/ \frac{7}{8}\beta_{13}.$$

For the other first row elements we can use (2.10) to obtain

$$\beta_{15} = \frac{1}{2}\beta_{13}, \qquad \beta_{16} = \frac{1}{4}\beta_{12}, \qquad \beta_{17} = \frac{1}{8}\beta_{11},$$

and the first row of G_n^{-1} is given by

$$(2.12)\qquad \frac{1}{3^4.5.7^2.31}[2^{21}, -2^{15}63\sqrt{2}, 2^{12}.651, -2^{10}.1395\sqrt{2}, 2^{11}651, -2^{13}63\sqrt{2}, 2^{18}].$$

Remark 2.6. It is worth noting that, although there is a wide range in the values of the elements of the first row of G_n in (2.11), all the elements in (2.12) are of a similar order.

Since G is both symmetric and persymmetric, then so is G_n^{-1} and so only just over a quarter of its elements need to be found. Explicitly, we require β_{ij} for $i = 1, 2, \cdots, [(n+1)/2], j = i, i+1, \cdots, n-i+1$, which is $[(n+1)/2][(n+2)/2]$ elements compared with the total of n^2.

From Remark 2.3, $\beta_{11} \neq 0$, and so the following well-known equation [13] can be used to calculate β_{ij} for the second and subsequent rows:

$$(2.13)\qquad \beta_{i+1,j+1} = \beta_{ij} + \frac{1}{\beta_{11}}(\beta_{1,i+1}\beta_{1,j+1} - \beta_{1,n-i+1}\beta_{1,n-j+1}).$$

Thus each element requires three flops, giving a total of $\frac{3}{4}n^2 + o(n^2)$ to find G_n^{-1}. This can, however, be improved upon by using the following result.

THEOREM 2.7. *If* $G_n^{-1} = [\beta_{ij}]$, $i, j = 1, 2, \cdots, n$, *then*

$$(2.14)\qquad \beta_{i,j+i-1} = \sum_{k=0}^{i-2}(1 - a^{2n-2j-4k})\frac{\beta_{1,k+1}}{\beta_{11}}\beta_{1,j+k} + \frac{\beta_{1i}}{\beta_{11}}\beta_{1,j+i-1},$$
$$i = 2, 3, \cdots, \left[\frac{n+1}{2}\right], \qquad j = 1, 2, \cdots, n-2i+2.$$

Proof. If $i = 2$, then from (2.13) we have

$$(2.15)\qquad \beta_{2,j+1} = \beta_{1j} + \frac{1}{\beta_{11}}(\beta_{12}\beta_{1,j+1} - \beta_{1n}\beta_{1,n-j+1}).$$

Since from (2.10),

$$\beta_{1,n} = (-1)^{n-1}a^{n-1}\beta_{11} \quad \text{and} \quad \beta_{1,n-j+1} = (-1)^{n-1}a^{n-2j+1}\beta_{1j},$$

then (2.15) becomes

$$\beta_{2,j+1} = \beta_{1j} + \frac{1}{\beta_{11}}(\beta_{12}\beta_{1,j+1} - a^{2n-2j}\beta_{11}\beta_{1j}),$$

which agrees with (2.14) when $i = 2$.

Now suppose that (2.14) is true for $i = 1, 2, \cdots, l-1$. Again from (2.13)

$$(2.16)\qquad \beta_{l,j+l-1} = \beta_{l-1,j+l-2} + \frac{1}{\beta_{11}}(\beta_{1l}\beta_{1,j+l-1} - \beta_{1,n-l+2}\beta_{1,n-j-l+3}).$$

If we write $\beta_{l-1,j+l-2}$ as the right-hand side of (2.14) with i replaced by $l-1$ and use (2.10) to give different forms of $\beta_{1,n-l+2}$ and $\beta_{1,n-j-l+3}$, then (2.16) becomes

$$\beta_{l,j+l-1} = \sum_{k=0}^{l-3}(1 - a^{2n-2j-4k})\frac{\beta_{1,k+1}}{\beta_{11}}\beta_{1,j+k} + \frac{\beta_{1,l-1}}{\beta_{11}}\beta_{1,j+l-2}$$
$$+ \frac{1}{\beta_{11}}(\beta_{1l}\beta_{1,j+l-1} - a^{2n-2j+8-4l}\beta_{1,l-1}\beta_{1,j+l-2}).$$

Collecting together the terms in $\beta_{1,l-1}\beta_{1,j+l-2}$ and combining into the summation gives (2.14). Thus the theorem is proved by induction. □

Remark 2.7. It should be noted that (2.14) holds for all $i, j = 1, 2, 3, \cdots, n$, although for calculation purposes only the range of i and j given in (2.14) is required.

We now show that use of (2.14) enables each required internal element of G_n^{-1} to be calculated from the previous row, using two rather than three flops, using (2.13).

We begin by supposing that $1 - a^{2k}$, $k = 1, 2, \cdots, n-1$, and β_{1i}/β_{11}, $i = 2, 3, \cdots, [(n+1)/2]$ are known.

If we set

$$\beta_{i-1,j+i-2} = \sum_{k=0}^{i-3} (1 - a^{2n-2j-4k}) \frac{\beta_{1,k+1}}{\beta_{11}} \beta_{1,j+k} + \frac{\beta_{1,i-1}}{\beta_{11}} \beta_{1,j+i-2}$$

$$= X_{i-1,j} + Y_{i-1,j},$$

where the sum on the right-hand side is denoted by $X_{i-1,j}$ and the second term by $Y_{i-1,j}$, then it is easy to see that with $X = 0$ when $i = 2$,

$$\beta_{i,j+i-1} = X_{i-1,j} + (1 - a^{2n-2j-4i+8}) Y_{i-1,j} + \frac{\beta_{1i}}{\beta_{11}} \beta_{1,j+i-1}, \tag{2.17}$$

which involves two sums and products as required.

We have thus found G_n^{-1} in $\frac{1}{2}n^2 + o(n^2)$ flops using (2.9), (2.10), and (2.17).

Example 2.1 (continued). To find the other elements of G_n^{-1}, where G_n is defined by (2.11), we use (2.17) to give

$$[\beta'_{22}, \beta'_{23}, \beta'_{24}, \beta'_{25}, \beta'_{26}] = [2^{10}5985, -2^6.72261\sqrt{2}, 2^5.166005, -2^5\sqrt{2}.80073, 2^8.7875],$$

$$[\beta'_{33}, \beta'_{34}, \beta'_{35}] = [2^3.1158129, -2.3056445\sqrt{2}, 2^2.1048761],$$

$$[\beta'_{44}] = [10363455],$$

$$\beta_{ij} = \frac{1}{3^4.5.7^2.31} \beta'_{ij}.$$

3. Conclusion. In § 2 a special matrix occurring in statistics, called Gaussian, has been considered. It is a Toeplitz matrix and, in addition, has a pattern in its generating row. This extra condition enables us to find an explicit formula for the elements of the first row of the inverse and to improve on the efficient Trench method for determining this inverse by a factor of four.

Acknowledgment. The author is grateful to one of the referees for bringing [3] to his attention.

REFERENCES

[1] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Second Edition, The Johns Hopkins University Press, Baltimore, MD, 1989.

[2] M. J. C. GOVER, *Analysis and application of Toeplitz-like matrices*, Department of Mathematics, Ph.D. thesis, University of Bradford, Bradford, U.K., 1984.

[3] W. B. GRAGG, *Positive definite Toeplitz matrices, the Arnoldi process for isometric operators and Gaussian quadrature on the unit circle*, in Numerical Methods in Linear Algebra, E. S. Nikolaev, ed., Moscow University Press, Moscow, U.S.S.R., 1982, pp. 16–32. (In Russian.)

[4] R. T. GREGORY AND D. L. KARNEY, *A Collection of Matrices for Testing Computational Algorithms*, John Wiley, New York, 1969.

[5] F. A. GRAYBILL, *Matrices with Applications in Statistics*, Wadsworth, Belmont, CA, 1983.
[6] G. HEINIG AND K. ROST, *Algebraic Methods for Toeplitz-Like Matrices and Operators*, Akademie-Verlag, Berlin, GDR, 1984.
[7] A. S. HOUSEHOLDER, *The Numerical Treatment of a Single Nonlinear Equation*, McGraw-Hill, New York, 1971.
[8] J. R. MAGNUS AND H. NEUDECKER, *Matrix Differential Calculus*, John Wiley, New York, 1988.
[9] K. S. MILLER, *Some Eclectic Matrix Theory*, Krieger, Malabar, FL, 1987.
[10] P. A. ROEBUCK AND S. BARNETT, *A survey of Toeplitz and related matrices*, Internat. J. Systems Sci., 9 (1978), pp. 921–934.
[11] S. R. SEARLE, *Matrix Algebra Useful for Statistics*, John Wiley, New York, 1982.
[12] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, U.K., 1965.
[13] S. ZOHAR, *Toeplitz matrix inversion: The algorithm of W. F. Trench*, J. Assoc. Comput. Mach., 16 (1969), pp. 592–601.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 549–551, July 1991

012

NEW STOPPING CRITERIA FOR SOME ITERATIVE METHODS FOR A CLASS OF UNSYMMETRIC LINEAR SYSTEMS*

D. J. EVANS† AND C. LI†

Abstract. When the iterative procedure $x_{k+1} = Gx_k + g$ for the linear system $Ax = b$ is considered, one of the important items for the method is the stopping criterion. Usually one kind of norm is used as a measure, and if the norm of the pseudoresidual vector $\delta_k = Gx_k + g - x_k$ is small, then the iterative procedure is terminated. However, this does not guarantee that the norm of the error vector $\varepsilon_k = x_k - x^*$ is small. In this short note it is shown that if there exists a nonsingular matrix Z such that ZGZ^{-1} is skew-symmetric, then $\|\varepsilon_k\|_Z \leqq \|\delta_k\|_Z$ where $\|y\|_Z = \|Zy\|_2$. The relative error bound is also given.

Key words. stopping criterion, pseudoresidual vector, error vector, relative error vector, symmetrizable and skew-symmetrizable

AMS(MOS) subject classifications. 65F10, 6JN20

1. Introduction.

Suppose we have a linear system,

$$Ax = b \quad \text{with } A = M - N, \tag{1.1}$$

where A and M are nonsingular. The iterative procedure

$$x_{k+1} = Gx_k + g, \tag{1.2}$$

is convergent if and only if $S(G)$, the spectral radius of G, is less than unity where x_0 is the initial guess, and

$$G = M^{-1}N, \qquad g = M^{-1}b. \tag{1.3}$$

If we let the *error vector* ε_k and *pseudoresidual vector* δ_k be defined, respectively, by

$$\varepsilon_k = x_k - x^*, \qquad \delta_k = Gx_k + g - x_k, \tag{1.4}$$

then it can be shown that

$$\varepsilon_k = (G - I)^{-1}\delta_k. \tag{1.5}$$

Here x^* is the exact solution of (1.1).

For the iterative procedure (1.2), the iterations are to be terminated whenever some measure of the error vector ε_k becomes sufficiently small, i.e., whenever

$$\|\varepsilon_k\|_\beta \leqq \varepsilon, \tag{1.6}$$

where $\|\cdot\|_\beta$ denotes some vector norm and ε is the desired accuracy. Since x^* is not known in advance, $\|\varepsilon_k\|_\beta$ cannot be computed directly, so an alternative choice must be made. Usually we use the pseudoresidual δ_k to approximate ε_k. But sometimes, even though $\|\delta_k\|_\beta$ is small, $\|\varepsilon_k\|_\beta$ may be very large.

However, if the iterative procedure (1.2) is symmetrizable, i.e., there exists a nonsingular matrix Z such that $Z(I - G)Z^{-1}$ is symmetric positive definite (SPD), then we

* Received by the editors February 9, 1988; accepted for publication (in revised form) June 8, 1990.

† Parallel Algorithms Research Centre, Department of Computer Studies, Loughborough University of Technology, Loughborough, Leicestershire, United Kingdom. The work of the second author was supported by the Chinese Academy of Science, the People's Republic of China, and by an Overseas Research Student Scholarship of the United Kingdom.

have (cf. Hageman and Young [1981])

$$\|\varepsilon_k\|_Z \leqq \frac{1}{(1-S(G))}\|\delta_k\|_Z, \tag{1.7}$$

where $\|y\|_Z = \|Zy\|_2$. It is obvious that if

$$\frac{1}{(1-S(G))}\|\delta_k\|_Z \leqq \varepsilon, \tag{1.8}$$

then $\|\varepsilon_k\|_Z \leqq \varepsilon$. Hence, for the symmetrizable case, inequality (1.8) as a stopping criterion is quite safe according to (1.7), which, in fact, is used consistently in the literature (Hageman and Young [1981]).

To our knowledge (also see Elman [1982]) for the nonsymmetrizable procedure (1.2) there is no simple relationship between ε_k and δ_k given in the literature. It is normal to use

$$\|\delta_k\|_\beta \leqq \varepsilon, \tag{1.9}$$

instead of (1.6), as a stopping criterion. However, when it is satisfied, we still cannot guarantee that (1.6) is satisfied.

In this short note, first we want to show that if the iterative procedure (1.2) is *skew-symmetrizable*, i.e., there exists a nonsingular matrix Z such that ZGZ^{-1} is a skew-symmetric matrix, then $\|\varepsilon_k\|_Z \leqq \varepsilon$ on the condition that $\|\delta_k\|_Z \leqq \varepsilon$. Hence, for such a case, if the pseudoresidual vector is small, then we can guarantee that the error vector is small, based on the concerned norm. Second, the relative error is also considered. Finally, we give some remarks concerning the application of the results.

2. Main results. Let G be the iterative matrix of the procedure (1.2).

THEOREM 1. *If the iterative procedure* (1.2) *is derived from the system* (1.1), Z *is the skew-symmetrization matrix, and if* ε_k *and* δ_k *are defined by* (1.4), *then*

$$\|\varepsilon_k\|_Z \leqq \|\delta_k\|_Z, \tag{2.1}$$

where $\|y\|_Z = \|Zy\|_2$.

Proof. By (1.5) we have

$$Z\varepsilon_k = Z(G-I)^{-1}Z^{-1}Z\delta_k = (G'-I)^{-1}Z\delta_k,$$

where $G' = ZGZ^{-1}$ is skew-symmetric. Thus we have

$$\|\varepsilon_k\|_Z \leqq \|(I-G')^{-1}\|_2\|\delta_k\|_Z \leqq \|\delta_k\|_Z,$$

since the minimum eigenvalue of the symmetric matrix $(I-G')^T(I-G') = I-G'^2$ is not less than unity, thus concluding the proof.

Now we consider the relative error. Note that since x^* satisfies $(I-G)x^* = g$, then

$$\|g\|_Z \leqq \|I-G'\|_2\|x^*\|_Z.$$

It follows from $\|I-G'\|_2 = [1+S^2(G')]^{1/2} = [1+S^2(G)]^{1/2}$ that

$$\|x^*\|_Z \geqq \|g\|_Z/[1+S^2(G)]^{1/2}. \tag{2.2}$$

Combining (2.2) and (2.1) gives Theorem 2.

THEOREM 2. *Under the conditions of Theorem* 1, *we have*

$$\frac{\|x_k-x^*\|_Z}{\|x^*\|_Z} \leqq [1+S^2(G)]^{1/2}\frac{\|\delta_k\|_Z}{\|g\|_Z}. \tag{2.3}$$

The left-hand side of the inequality above is usually called the relative error.

Note that $[1 + S^2(G)]^{1/2}$ plays the role of the condition number. The pseudoresidual vector may be magnified by as much as a factor of $[1 + S^2(G)]^{1/2}$.

It follows from (2.3) that if

$$[1+S^2(G)]^{1/2}\|\delta_k\|_Z/\|g\|_Z \leqq \varepsilon, \tag{2.4}$$

then $\|\varepsilon_k\|_Z/\|x^*\|_Z \leqq \varepsilon$. Thus (2.4) can be used as a relative error stopping criterion.

However, for the use of (2.4) it is necessary to know some information about the spectral radius $S(G)$ of G as in the symmetrizable case. If we have to use the relative error and $S(G)$ (or even an approximation) is not available, the following,

$$\|\delta_k\|_Z/\|x_k\|_Z \leqq \varepsilon, \tag{2.5}$$

can be an alternative choice since x_k may be near to x^* and $\|x_k\|_Z$ may be approximately $\|x^*\|_Z$.

3. Concluding remarks. If M and N, defined in (1.1), satisfy

$$M=\tfrac{1}{2}(A+A^T), \qquad N=\tfrac{1}{2}(A-A^T), \tag{3.1}$$

then it can be shown that procedure (1.2) is skew-symmetrizable with the skew-symmetrization matrix $Z = M^{1/2}$, the *square root* of M, if M is SPD. In certain cases, a number of problems of the form $Mx = y$ can be solved with less computational effort than the original system $Ax = b$, such as the large sparse system arising from elliptic problems (cf. Widlund [1978]) and the systems arising from the least squares problems (cf. Li [1989]). In such a case, the generalized conjugate gradient method of Concus and Golub [1976] and Widlund [1978], and the Chebyshev acceleration in the nonsymmetrizable case (cf. Chapter 12 of Hageman and Young [1981]) are quite applicable. It is believed that the new stopping criteria can make these methods more attractive.

Acknowledgments. The authors would like to thank the referees for their valuable comments.

REFERENCES

P. CONCLUS AND G. H. GOLUB [1976], *A generalized conjugate gradient method for nonsymmetric systems of linear equations*, Lecture Notes in Economics and Mathematical Systems, Vol. 134, R. Glowinski and J. L. Lions, eds., Springer-Verlag, New York, pp. 56–65.

H. C. ELMAN [1982], *Iterative methods for large sparse, nonsymmetric systems of linear equations*, Ph.D. thesis, Yale University, New Haven, CT.

L. A. HAGEMAN AND D. M. YOUNG [1981], *Applied Iterative Methods*, Academic Press, New York.

C. LI [1989], *Iterative methods for a class of large, sparse and nonsymmetric linear systems*, Ph.D. thesis, Loughborough University of Technology, Loughborough, Leicestershire, U.K.

O. WIDLUND [1978], *A Lanczos method for a class of nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 15, pp. 801–812.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 552–564, July 1991

013

FAST QR DECOMPOSITION OF VANDERMONDE-LIKE MATRICES AND POLYNOMIAL LEAST SQUARES APPROXIMATION*

LOTHAR REICHEL†

Dedicated to Richard S. Varga on the occasion of his 60th birthday.

Abstract. Let f and g be functions defined at the real and distinct nodes x_k, and consider the inner product $(f, g) := \sum_{k=1}^{m} f(x_k)g(x_k)w_k^2$ with positive weights w_k^2. The present paper discusses the computation of orthonormal polynomials $\pi_0, \pi_1, \cdots, \pi_{n-1}$, $n \leq m$, with respect to this inner product, and the use of these polynomials in a fast scheme for computing a QR decomposition of the transpose of Vandermonde-like matrices. Two methods are compared for computing the recurrence coefficients for the polynomials π_j and their values at the nodes x_k: the Stieltjes procedure and a method in which an inverse eigenvalue problem for a tridiagonal symmetric matrix is solved by an algorithm proposed by Rutishauser, Gragg, and Harrod. The latter method is found to generally yield higher accuracy than the Stieltjes procedure if n is close to m, and roughly the same accuracy otherwise. This method for solving an inverse eigenvalue problem is applied in an algorithm for computing a QR decomposition of the transpose of $n \times m$ Vandermonde-like matrices. The algorithm so obtained requires only $O(mn)$ arithmetic operations. This operation count compares favorably with the $O(mn^2)$ arithmetic operations necessary for the QR decomposition if the structure of Vandermonde-like matrices is ignored.

Key words. orthogonal polynomial, polynomial least squares approximation, Vandermonde-like matrix, QR decomposition

AMS(MOS) subject classifications. 65F25, 65D05

1. Introduction.

Let $\{x_k\}_{k=1}^m$ be a set of distinct nodes on the real axis, and let $\{w_k^2\}_{k=1}^m$ be a set of positive weights. Introduce the inner product

$$(f,g) := \sum_{k=1}^{m} f(x_k)g(x_k)w_k^2 \tag{1.1}$$

for functions f and g defined at the nodes. Let $\{\pi_j\}_{j=0}^{m-1}$ be a family of *orthonormal* polynomials with respect to this inner product, where we assume that π_j is of degree j and has a positive leading coefficient. The π_j satisfy a three-term recurrence relation

$$\begin{aligned} &\beta_0\pi_0(x) = 1, \qquad \beta_1\pi_1(x) = (x-\alpha_1)\pi_0(x), \\ &\beta_j\pi_j(x) = (x-\alpha_j)\pi_{j-1}(x) - \beta_{j-1}\pi_{j-2}(x), \qquad j = 2, 3, \cdots, m-1, \end{aligned} \tag{1.2}$$

where the coefficients α_j and $\beta_j > 0$ satisfy

$$\begin{aligned} &\beta_0 = (1,1)^{1/2}, \qquad \beta_1 = ((x\pi_0, x\pi_0) - \alpha_1^2)^{1/2}, \\ &\alpha_j = (x\pi_{j-1}, \pi_{j-1}), \qquad j = 1, 2, \cdots, m, \\ &\beta_j = ((x\pi_{j-1}, x\pi_{j-1}) - \alpha_j^2 - \beta_{j-1}^2)^{1/2}, \qquad j = 2, 3, \cdots, m-1. \end{aligned} \tag{1.3}$$

In the *Stieltjes procedure* we combine formulas (1.2) and (1.3) in order to compute coefficients α_j and β_j, and the values of the polynomials π_j at the nodes x_k for increasing values of j as follows. First β_0 and π_0 are computed, and then we can determine α_1 and β_1 from (1.3), and the value of π_1 at the nodes from (1.2). Now we can compute α_2 and

* Received by the editors May 1, 1989; accepted for publication (in revised form) May 23, 1990. This research was supported in part by IBM Bergen Scientific Centre, Air Force Office of Scientific Research grant AFOSR-87-0102, and National Science Foundation grant DMS-8704196.

† Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506.

β_2 from (1.3), and π_2 at the nodes from (1.2), and so on. The Stieltjes procedure has recently been discussed by Gautschi [10], [12]. It has also been advocated by Forsythe [9].

However, the Stieltjes procedure can be sensitive to roundoff errors (see below) and we therefore propose a different method for computing the π_j. This method is suggested by a matrix interpretation of the Stieltjes procedure. Such an interpretation shows that the Stieltjes procedure for a discrete inner product (1.1) is equivalent to the *Lanczos procedure* for determining a symmetric tridiagonal matrix which contains the recursion coefficients and which is orthogonally similar to the diagonal matrix $\Lambda = \operatorname{diag}[x_1, x_2, \cdots, x_m]$ (see de Boor and Golub [5], Gragg and Harrod [15], or the recent survey paper by Boley and Golub [4]). In fact, the Lanczos procedure solves the following inverse eigenvalue problem: determine a symmetric tridiagonal matrix $T = [t_{jk}]_{j,k=1}^{m}$ of order m, and an orthogonal matrix $Q = [q_{jk}]_{j,k=1}^{m}$ of order m, such that

$$\begin{aligned} TQ^T &= Q^T\Lambda, \\ \Lambda &= \operatorname{diag}[x_1, x_2, \cdots, x_m], \\ Q\mathbf{e}_1 &= \frac{[w_1, w_2, \cdots, w_m]^T}{(\sum_{j=1}^{m} w_j^2)^{1/2}}. \end{aligned} \tag{1.4}$$

Here and below, $\mathbf{e}_j = [0, \cdots, 0, 1, 0, \cdots, 0]^T$ denotes the jth axis vector of appropriate dimension. The matrices T and Q are uniquely determined by (1.4), and, moreover,

$$\begin{aligned} t_{jj} &= \alpha_j, \qquad 1 \leqq j \leqq m, \\ t_{j+1,j} &= t_{j,j+1} = \beta_j, \qquad 1 \leqq j < m, \\ q_{kj} &= \pi_{j-1}(x_k) w_k, \qquad 1 \leqq j, k \leqq m \end{aligned} \tag{1.5}$$

(see [5], [15], or [4] for details). Hence, given a set of nodes $\{x_k\}_{k=1}^{m}$ and a set of positive weights $\{w_k^2\}_{k=1}^{m}$, we seek to compute the unique symmetric tridiagonal matrix T with positive subdiagonal elements and the unique orthogonal matrix Q, which satisfy (1.4). Then the recurrence coefficients and the values of the orthogonal polynomials at the nodes are determined by (1.5).

It is known that the Lanczos procedure for solving the inverse eigenvalue problem (1.4) can be sensitive to roundoff errors, and that computations in finite precision arithmetic can yield matrices Q that are far from orthogonal (see [5], [15]). Equivalently, the polynomials computed by the Stieltjes procedure in finite precision arithmetic can be far from orthonormal with respect to the inner product (1.1). This is illustrated in § 2. We therefore propose to determine the polynomials π_j by solving the inverse eigenvalue problem (1.4) by an algorithm described by Rutishauser [20] and more recently by Gragg and Harrod [15]. This algorithm proceeds by applying a carefully chosen sequence of Givens rotations to the diagonal matrix Λ in order to transform it into tridiagonal form. We therefore refer to this algorithm as the Givens rotation (GR) algorithm. Like the Stieltjes procedure, it determines the recursion coefficients in (1.2) and (1.3) in $O(m^2)$ arithmetic operations. Details and computed examples illustrating the accuracy of this scheme are presented in § 2.

Let $h(x)$ be a continuous function which is explicitly known at the nodes x_k, $1 \leqq k \leqq m$. Consider the approximation of h by polynomials and measure the approximation error by the seminorm $\|h\|_2 := (h, h)^{1/2}$. It is well known that the polynomial

$$r_{n-1}(x) := \sum_{j=0}^{n-1} (h, \pi_j)\pi_j(x), \qquad n \leqq m, \tag{1.6}$$

is the best polynomial approximant to h of degree less than n with respect to the seminorm $\| \ \|_2$. Introduce the vectors and diagonal matrix

$$\begin{aligned} \mathbf{c} &:= [(h,\pi_0),(h,\pi_1),\cdots,(h,\pi_{n-1})]^T, \\ \mathbf{h} &:= [h(x_1),h(x_2),\cdots,h(x_m)]^T, \\ D &:= \operatorname{diag}[w_1,w_2,\cdots,w_m], \end{aligned} \tag{1.7}$$

where $w_j := (w_j^2)^{1/2}$. Then the vector $\mathbf{c}$ of coefficients of $r_{n-1}(x)$ can be computed by the GR algorithm as

$$\mathbf{c} = Q^T D\mathbf{h} \tag{1.8}$$

in $O(mn)$ arithmetic operations. This operation count uses the fact that only the n first columns of Q have to be determined (see Algorithm 2.1 of § 2). We remark, in passing, that recently Elhay et al. [8] presented several algorithms for modifying the polynomials π_j when a weight w_k^2 is modified in the inner product (1.1), for instance, when a node x_k is added or deleted.

In applications in which the polynomial r_{n-1} has to be differentiated or integrated, it can be convenient to express r_{n-1} in a polynomial basis different from $\{\pi_j\}_{j=0}^{n-1}$. This gives rise to the solution of overdetermined linear systems of equations with matrices V^T, where V is a Vandermonde-like matrix defined as follows. Following Higham [16], [17], we say that a matrix $V = [v_{jk}]_{1\leq j\leq n, 1\leq k\leq m}$ is *Vandermonde-like*, if $v_{jk} = p_{j-1}(x_k)$, where p_{j-1} is a polynomial of precisely degree $j-1$. In the present paper, we are primarily interested in the case where $p_{j-1}(x) = x^{j-1}$ for all j (V then is a "classical" Vandermonde matrix) and in the case where the p_{j-1} satisfy a three-term recurrence relation

$$\begin{aligned} &b_0 p_0(x) = 1, \qquad b_1 p_1(x) = (x-a_1)p_0(x), \\ &b_j p_j(x) = (x-a_j)p_{j-1}(x) - b_{j-1}p_{j-2}(x), \qquad j = 2,3,\cdots,m-1, \end{aligned} \tag{1.9}$$

with given recurrence coefficients a_j and $b_j > 0$. Vandermonde-like matrices with elements $v_{jk} = p_{j-1}(x_k)$, where the $p_{j-1}(x)$ satisfy a three-term recurrence relation, were first studied by Gautschi [11], who has shown that these matrices often have a smaller condition number than classical Vandermonde matrices (see [11], [13]).

Example 1.1. The choice $b_0 := b_1 := \sqrt{2}$, $b_j := 1$ for $j \geqq 2$, and $a_j = 0$ for $j \geqq 0$, yields $p_0(x) = 2^{-1/2}T_0(x/2)$ and $p_j(x) = T_j(x/2)$ for $j \geqq 1$, where $T_j(x) = \cos(j \arccos x)$ is the usual Chebyshev polynomial.

Example 1.2. The choice $b_0 = 2$, $b_j = 2j/(4j^2-1)^{1/2}$ for $j \geqq 1$, and $a_j = 0$ for $j \geqq 0$, yields the orthonormal Legendre polynomials for the interval $[-2, 2]$, i.e., the p_j defined by (1.9) satisfy

$$\int_{-2}^{2} p_j(x)p_k(x)\,dx = \begin{cases} 1, & j = k, \\ 0, & j \neq k. \end{cases}$$

Consider the overdetermined *dual Vandermonde-like system* of equations

$$DV^T\mathbf{c}' = D\mathbf{h}, \tag{1.10}$$

where V is an $n \times m$ Vandermonde-like matrix whose elements $v_{jk} = p_{j-1}(x_k)$ are defined by polynomials p_{j-1} that satisfy a three-term recurrence relation (1.9), and where the diagonal weighting matrix D is given by (1.7). Introduce the right triangular $n \times n$ matrix $R = [r_{jk}]_{j,k=1}^{n}$, defined by

$$p_{k-1}(x) = \sum_{j=1}^{k} r_{jk}\pi_{j-1}(x), \qquad 1 \leqq k \leqq n, \tag{1.11}$$

i.e., R expresses the given polynomial basis $\{p_j\}_{j=0}^{n-1}$ in terms of the computed orthonormal basis $\{\pi_j\}_{j=0}^{n-1}$. From $b_j > 0$ and $\beta_j > 0$ for all j, it follows that $r_{jj} > 0$ for all j. By (1.5), (1.7), and (1.11) we obtain

$$DV^T = QR, \tag{1.12}$$

i.e., (1.12) is the unique QR decomposition of DV^T such that $r_{jj} > 0$ for all j. The least squares solution of (1.10) can now be written, using (1.12),

$$\mathbf{c}' = (VD^2V^T)^{-1}VD\mathbf{h} = R^{-1}Q^TD\mathbf{h} = R^{-1}\mathbf{c}, \tag{1.13}$$

where $\mathbf{c}$ is defined by (1.8). The recurrence relations (1.2) and (1.9) make it possible to determine the elements of R, as well as of R^{-1}, in $O(n^2)$ arithmetic operation. We therefore can compute the solution $\mathbf{c}'$ of (1.10) in $O(mn)$ arithmetic operations by first determining the vector $\mathbf{c} := Q^TD\mathbf{h}$ by the GR algorithm, and then computing $\mathbf{c}' = R^{-1}\mathbf{c}$. This solution method requires only $O(m)$ storage locations (see § 3 for details and computed examples). This operation and storage count is also valid if Q were to be determined by the Stieltjes procedure. However, numerical experiments show that the sensitivity to roundoff errors of the Stieltjes procedure make it unsuitable for the present application unless $m \gg n$ (see § 3 for an illustrative example). Also note that in order for this operation and storage count to hold, it suffices that the polynomials p_{j-1} defining V satisfy a recurrence relation in which the number of terms is bounded independently of j.

For comparison, we note that $O(mn^2)$ arithmetic operations and $O(mn)$ storage locations are required to determine a QR decomposition of a general $m \times n$ matrix (see, e.g., [14, Chap. 5]). We remark that the data used by our fast algorithm is different from the data for schemes for the QR decomposition of a general matrix; the latter require the matrix elements, while our scheme requires the nodes x_k, weights w_k^2, and recursion coefficients (1.9). This affects the sensitivity to perturbations.

The factorization (1.12) can also be used for the solution of *primal Vandermonde-like systems*

$$V\mathbf{c}'' = \mathbf{d}, \tag{1.14}$$

with an $m \times m$ Vandermonde-like matrix V, and with $\mathbf{c}''$, $\mathbf{d} \in R^m$. We can choose $w_j^2 = 1$ for all j. This yields $D = I$ in (1.12), and we can compute the solution of (1.14) as $\mathbf{c}'' = QR^{-T}\mathbf{d}$ in $O(m^2)$ arithmetic operations if the polynomials p_{j-1} defining V satisfy a recurrence relation in which the number of terms is bounded independently of j.

Vandermonde-like linear systems of equations (1.10) and (1.14) arise in polynomial interpolation and polynomial least squares approximation problems, as well as in numerical quadrature. They may also be a part of more complicated approximation schemes (see [1] for an example). If $m = n$ and the polynomials defining V satisfy a recurrence relation with a bounded number of terms, then there are also other methods available that require only $O(m^2)$ arithmetic operations for the solution of (1.10) and (1.14), such as the Björck–Pereyra algorithms (see [3], [14, Chap. 4.6]), and modifications thereof [16], [17], [19]. These algorithms factor a square Vandermonde-like matrix into triangular matrices. The case $m \geqq n$ is also discussed by Demeure [6], who uses the Stieltjes procedure to determine an orthogonal matrix Q, and computes an upper triangular matrix R^{-1} such that $V^TR^{-1} = Q$.

We finally remark that algorithms analogous to those described in the present paper can also be derived if all the nodes x_k lie on the unit circle in the complex plane. The GR algorithm is then replaced by an algorithm described in [2] for the calculation of an inverse eigenvalue problem for unitary upper Hessenberg matrices (see [18]).

2. Least squares approximation. Let $h(x)$ be a function defined at the nodes x_k. Given a vector $\mathbf{h} = [h(x_1), h(x_2), \cdots, h(x_m)]^T$, we seek to compute the vector $\mathbf{c} = [c_1, c_2, \cdots, c_n]^T$ defined by (1.7). This section describes how $\mathbf{c}$ can be determined using the GR algorithm. We first outline the GR algorithm. Computed examples comparing our approach with the Stieltjes procedure conclude this section.

The GR algorithm is defined recursively. Introduce the symmetric tridiagonal Jacobi matrix

$$T_l = \begin{bmatrix} \alpha_{1l} & \beta_{1l} & & & 0 \\ \beta_{1l} & \alpha_{2l} & \beta_{2l} & & \\ & \beta_{2l} & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_{l-1,l} \\ 0 & & & \beta_{l-1,l} & \alpha_{ll} \end{bmatrix} \in R^{l\times l},$$

containing the recurrence coefficients for polynomials orthogonal with respect to the inner product

$$(f,g)_l := \sum_{k=1}^{l} f(x_k)g(x_k)w_k^2.$$

Then T_l has spectral resolution

$$T_l = Q_l^T \Lambda_l Q_l,$$

where $Q_l \in R^{l\times l}$, $Q_l^T Q_l = I$, $\Lambda_l = \operatorname{diag}[x_1, x_2, \cdots, x_l]$, and

$$Q_l \mathbf{e}_1 = \frac{[w_1, w_2, \cdots, w_l]^T}{\beta_{0l}}, \qquad \beta_{0l} = \left(\sum_{k=1}^{l} w_k^2 \right)^{1/2}.$$

Now consider the following matrix that has T_l as a submatrix:

$$T'_{l+2} = \left[\begin{array}{c|ccc|c} \alpha_0 & \beta_{0l} & 0 & 0 & w_{l+1} \\ \hline \beta_{0l} & & & & 0 \\ 0 & & T_l & & \\ 0 & & & & 0 \\ \hline w_{l+1} & 0 & & 0 & x_{l+1} \end{array}\right] \in R^{(l+2)\times(l+2)},$$

where $\alpha_0 \in R$ is arbitrary. The GR algorithm proceeds by carrying out an orthogonal similarity transformation of T'_{l+2} in order to obtain a symmetric tridiagonal matrix, which we denote by T''_{l+2}. The similarity transformation is done by applying a sequence of Givens rotations

$$G_{jk} := I + (\mathbf{e}_j \mathbf{e}_j^T + \mathbf{e}_k \mathbf{e}_k^T)(\gamma - 1) + (\mathbf{e}_j \mathbf{e}_k^T - \mathbf{e}_k \mathbf{e}_j^T)\sigma \in R^{(l+2)\times(l\times 2)}, \qquad j<k,$$

where $\mathbf{e}_j$, $\mathbf{e}_k \in R^{l+2}$, $\gamma^2 + \sigma^2 = 1$, $-1 \leqq \gamma \leqq 1$, and $\sigma \geqq 0$. The Givens rotations are chosen to "chase" the element w_{l+1}, in position $(l+2, 1)$ of T'_{l+2} along the last row to position $(l+2, l+1)$. By symmetry, the element w_{l+1} in position $(1, l+2)$ of T'_{l+2} is chased along the last column to position $(l+1, l+2)$. Moreover, the rotations are selected so that the element α_0 of T'_{l+2} is neither used nor changed. The desired "chasing" is obtained by applying l Givens rotations for rotation in the planes $(2, l+2)$, $(3, l+$

2), $\cdots$, $(l+1, l+2)$. See [15] for a detailed description. We obtain the symmetric tridiagonal matrix

$$T''_{l+2} = \left[\begin{array}{c|cc} \alpha_0 & \beta_{0,l+1} & 0 \\ \hline \beta_{0l+1} & & \\ 0 & T_{l+1} & \\ 0 & & \end{array}\right] \in R^{(l+2)\times(l+2)},$$

whose trailing principal $(l+1) \times (l+1)$ submatrix T_{l+1} contains the recurrence coefficients for polynomials orthonormal with respect to the inner product

$$(f,g)_{l+1} := \sum_{k=1}^{l+1} f(x_k)g(x_k)w_k^2.$$

The spectral resolution of T_{l+1} is

$$T_{l+1} = Q_{l+1}^T \Lambda_{l+1} Q_{l+1},$$

where $Q_{l+1}^T Q_{l+1} = I$, $\Lambda_{l+1} = \operatorname{diag}[x_1, x_2, \cdots, x_{l+1}]$, and

$$Q_{l+1}\mathbf{e}_1 = \frac{[w_1, w_2, \cdots, w_{l+1}]^T}{\beta_{0,l+1}}, \qquad \beta_{0,l+1} = \left(\sum_{k=1}^{l+1} w_k^2\right)^{1/2}.$$

Thus we can compute $\{T_{l+1}, Q_{l+1}, \beta_{0,l+1}\}$ from $\{T_l, Q_l, \beta_{0,l}, x_{l+1}, w_{l+1}^2\}$ in $O(l)$ arithmetic operations. This operation count assumes that Q_l and Q_{l+1} are stored in factored form as a product of Givens rotations.

Starting with

$$T'_2 = T''_2 = \begin{bmatrix} \alpha_0 & w_1 \\ w_1 & x_1 \end{bmatrix}, \qquad T_1 = [x_1],$$

we compute T'_{l+2}, T''_{l+2}, T_{l+1}, and Q_{l+1} for $l = 1, 2, \cdots, m-1$, until we obtain $T := T_m$ and $Q := Q_m$. It follows that we can determine T and Q from $\{x_k\}_{k=1}^m$ and $\{w_k^2\}_{k=1}^m$ in $O(m^2)$ arithmetic operations. This operation count assumes that we determine the whole matrix $Q \in R^{m\times m}$. However, if the degree $n-1$ of the desired polynomial r_{n-1} is less than $m-1$, then only the recurrence coefficients $\{\alpha_j\}_{j=1}^{n-1}$ and $\{\beta_j\}_{j=1}^{n-1}$, and Fourier coefficients $c_j = (h, \pi_{j-1})$ for $1 \leqq j \leqq n$ have to be computed. This can be done by using the first n columns of Q only, and then requires just $O(mn)$ arithmetic operations, as is demonstrated by the following algorithm.

ALGORITHM 2.1. Polynomial least squares approximation.
Number of nodes: $m \geqq 2$. Highest degree of approximating polynomial: $n - 1 < m$.
Input: m, n, $\{x_k\}_{k=1}^m$, $\{w_k^2\}_{k=1}^m$, $\{h(x_k)\}_{k=1}^m$.
Output: $\{\alpha_j\}_{j=1}^{n-1}$, $\{\beta_j\}_{j=0}^{n-1}$, $c_j := (h, \pi_{j-1})$, $1 \leqq j \leqq n$.

```
* τ is used as a temporary variable   *
* l = 0, initialize T''_2   *
  α_1 := x_1; β_0 := w_1; c_1 := w_1 h(x_1);
* l = 1, add pair {x_2, w_2^2}, compute T''_3   *
  τ := β_0; β_0 := (β_0^2 + w_2^2)^{1/2}; γ := τ/β_0; σ := −w_2/β_0;
  τ := γσ(α_1 − x_2); β_1 := |τ|;
  c_2 := sign(τ)(σc_1 + γw_2 h(x_2)); c_1 := γc_1 − σw_2 h(x_2);
  α_2 := σ^2 α_1 + γ^2 x_2; α_1 := γ^2 α_1 + σ^2 x_2;
* l > 1, add pair {x_{l+1}, w_{l+1}^2}, compute T''_{l+2}   *
  for l = 2, 3, ···, m − 1 do
```

$\tau := \beta_0$; $\beta_0 := (\beta_0^2 + w_{l+1}^2)^{1/2}$; $\gamma := \tau/\beta_0$; $\sigma := -w_{l+1}/\beta_0$;
$c_{l+1} := \sigma c_1 + \gamma w_{l+1} h(x_{l+1})$; $c_1 := \gamma c_1 - \sigma w_{l+1} h(x_{l+1})$;
* the variables v_1, v_2, and v_3 contain nonzero elements of the last row *
* of $G^T T''_{l+2} G$, where G is a product of (at least one) Givens rotations *
$v_1 := \gamma\sigma(\alpha_1 - x_{l+1})$; $v_2 := \sigma\beta_1$; $v_3 := \gamma^2 x_{l+1} + \sigma^2\alpha_1$;
$\alpha_1 := \gamma^2\alpha_1 + \sigma^2 x_{l+1}$; $\beta_1 := \gamma\beta_1$;
for $k = 1, 2, \cdots, \min\{l-2, n-1\}$ **do**
  $\tau := \beta_k$; $\beta_k := (\beta_k^2 + v_1^2)^{1/2}$; $\gamma := \tau/\beta_k$; $\sigma := -v_1/\beta_k$;
  $\tau := \sigma c_{k+1} + \gamma c_{l+1}$; $c_{k+1} := \gamma c_{k+1} - \sigma c_{l+1}$; $c_{l+1} := \tau$;
  $\tau := \alpha_{k+1}$; $\alpha_{k+1} := \gamma^2\tau + \sigma^2 v_3 - 2\gamma\sigma v_2$;
  $v_1 := (\gamma - \sigma)(\gamma + \sigma)v_2 + \gamma\sigma(\tau - v_3)$; $v_3 := \gamma^2 v_3 + 2\gamma\sigma v_2 + \sigma^2\tau$;
  $v_2 := \sigma\beta_{k+1}$; $\beta_{k+1} := \gamma\beta_{k+1}$;
if $l - 1 < n$ **then**
  $\tau := \beta_{l-1}$; $\beta_{l-1} := (\beta_{l-1}^2 + v_1^2)^{1/2}$; $\gamma := \tau/\beta_{l-1}$; $\sigma := -v_1/\beta_{l-1}$;
  $\tau := (\gamma - \sigma)(\gamma + \sigma)v_2 + \gamma\sigma(\alpha_l - v_3)$; $\beta_l := |\tau|$;
  $\tau := \operatorname{sign}(\tau)(\sigma c_l + \gamma c_{l+1})$; $c_l := \gamma c_l - \sigma c_{l+1}$; $c_{l+1} := \tau$;
  $\alpha_{l+1} := \sigma^2\alpha_l + 2\gamma\sigma v_2 + \gamma^2 v_3$; $\alpha_l := \gamma^2\alpha_l - 2\gamma\sigma v_2 + \sigma^2 v_3$; □

The operation count for the algorithm as presently coded in FORTRAN is $20mn + O(m)$ multiplications or divisions and $10mn + O(m)$ additions or subtractions. Furthermore, $mn + O(m)$ square root computations are required. It might be possible to reduce this operation count slightly by a more careful implementation. Having determined the recurrence coefficients α_j and β_j, and the Fourier coefficients c_j, by Algorithm 2.1, we can evaluate the polynomial $r_{n-1}(x) = \sum_{j=1}^{n} c_j \pi_{j-1}(x)$ for arbitrary $x \in R$ by Clenshaw's recursion formula (see, e.g., Smith [21] or Higham [17]).

We turn to some computed examples that compare Algorithm 2.1 with the Stieltjes procedure. All computations of the present paper have been carried out on an IBM 3090 VF computer in single or double precision arithmetic, i.e., with 6 or 15 significant digits, respectively.

Example 2.1. Let $\{x_k^{(m)}\}_{k=1}^{m}$, $m = 2, 3, 4, \cdots$ be sets of equidistant nodes in $[-2, 2]$ defined by

$$x_k^{(m)} := 2 - 4\frac{k-1}{m-1}, \qquad 1 \leqq k \leqq m. \tag{2.1}$$

Let all the weights w_k^2 be unity, and define the function $h(x) := \exp(x)$. We let c_j denote the approximation of the Fourier coefficient (h, π_{j-1}) obtained by Algorithm 2.1 in single precision arithmetic, and we let d_j denote the corresponding approximation obtained in double precision arithmetic by the subroutine DQRDC of LINPACK [7]. This subroutine computes a QR decomposition of V^T without using the structure of the matrix. Introduce the vectors $\mathbf{c} := [c_1, c_2, \cdots, c_m]^T$ and $\mathbf{d} := [d_1, d_2, \cdots, d_m]^T$. The norm

$$\|\mathbf{c} - \mathbf{d}\|_\infty := \max_{1 \leqq j \leqq m} |c_j - d_j| \tag{2.2}$$

yields an estimate of the largest error in the computed coefficients c_j.

We also compute Fourier coefficients by the Stieltjes procedure. The Stieltjes procedure has been implemented for the computation of monic orthogonal polynomials as described in [9], [5], and [15, p. 323]. The monic orthogonal polynomials are then normalized to yield orthonormal polynomials. These computations are carried out in single precision arithmetic, and due to roundoff errors, we only obtain approxi-

mations of orthonormal polynomials, denoted by $\tilde{\pi}_j$. We evaluate the inner products $s_j := (h, \tilde{\pi}_{j-1})$, $1 \leqq j \leqq m$, in single precision arithmetic, and define the vector $\mathbf{s} := [s_1, s_2, \cdots, s_m]^T$.

Figure 2.1 shows $m \to \log_{10} \|\mathbf{c} - \mathbf{d}\|_\infty$ (continuous curve) and $m \to \log_{10} \|\mathbf{s} - \mathbf{d}\|_\infty$ (dashed curve) for $2 \leqq m \leqq 100$. The figure shows that if all m Fourier coefficients are required, then Algorithm 2.1 yields higher accuracy than the Stieltjes procedure. The poor accuracy obtained with the Stieltjes procedure depends on the fact that the computed polynomials $\tilde{\pi}_j$ are far from orthonormal for j close to m. Computation of fewer than m coefficients is discussed in Example 2.4 below.

Example 2.2. This example differs from Example 2.1 only in the selection of nodes $x_k^{(m)}$. We now choose

$$x_k^{(m)} := -2 + 4\left(\frac{k-1}{m-1}\right)^2, \qquad 1 \leqq k \leqq m. \tag{2.3}$$

There are more nodes close to $x = -2$ than to $x = 2$. For instance, we have $x_k^{(m)} < x_{k+1}^{(m)}$ for all k and $x_{\lfloor m/2 \rfloor}^{(m)} < -1$. Figure 2.2 shows the error in the computed Fourier coefficients. The error is measured in the same way as in Example 2.1, and shows that clustering of nodes at $x = -2$ makes the Stieltjes procedure perform worse than in Example 2.1. The error obtained with Algorithm 2.1 is roughly the same as in the previous example.

Example 2.3. This example differs from Examples 2.1–2.2 only in the selection of nodes. We now let the nodes $x_k^{(m)}$ be zeros of Chebyshev polynomials for $[-2, 2]$, i.e.,

$$x_k^{(m)} := 2\cos\left(\pi\frac{2k-1}{2m}\right), \qquad 1 \leqq k \leqq m. \tag{2.4}$$

For these nodes the errors in the coefficient vector $\mathbf{c}$ obtained by Algorithm 2.1 and the coefficient vector $\mathbf{s}$ computed by the Stieltjes procedure are roughly the same.

A comparison of Figs. 2.1–2.3 suggests that the error in the Fourier coefficient vector $\mathbf{c}$ obtained by Algorithm 2.1 is fairly independent of the distribution of the nodes x_k, but the error in the coefficient vector $\mathbf{s}$ obtained by the Stieltjes procedure is not. Moreover, the error in the vector $\mathbf{c}$ is never much larger than the error in the vector $\mathbf{s}$, but it can be much smaller. Similar behavior has been observed in numerous other numerical experiments with other distributions of nodes x_k and other functions h. In particular, if the analytic function h in Examples 2.1–2.3 is replaced by the very nonsmooth function

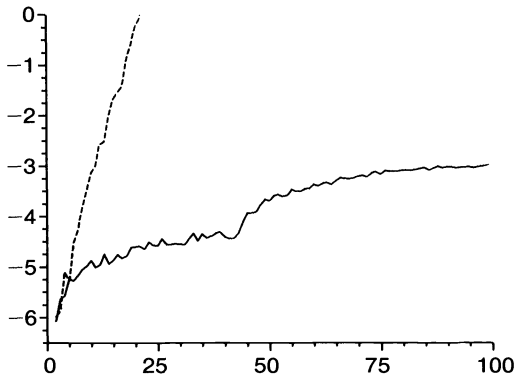


FIG. 2.1. $\log_{10} \|\mathbf{c} - \mathbf{d}\|_\infty$ (*continuous curve*), $\log_{10} \|\mathbf{s} - \mathbf{d}\|_\infty$ (*dashed curve*), $h(x) = \exp(x)$, *equidistant nodes* (2.1).

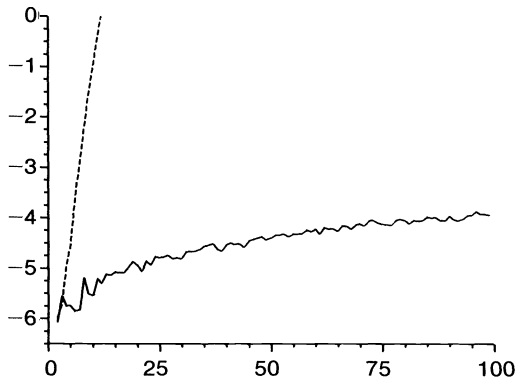


FIG. 2.2. $\log_{10} \|\mathbf{c} - \mathbf{d}\|_\infty$ (*continuous curve*), $\log_{10} \|\mathbf{s} - \mathbf{d}\|_\infty$ (*dashed curve*), $h(x) = \exp(x)$, *nodes* (2.3).

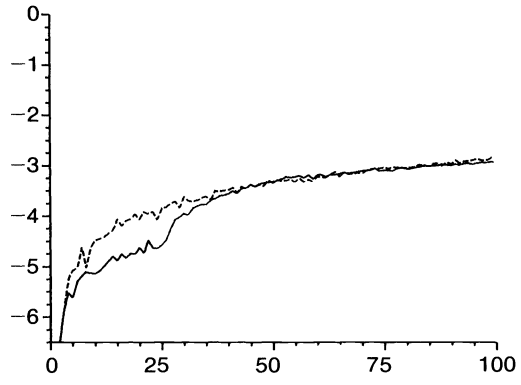


FIG. 2.3. $\log_{10} \|\mathbf{c} - \mathbf{d}\|_\infty$ (*continuous curve*), $\log_{10} \|\mathbf{s} - \mathbf{d}\|_\infty$ (*dashed curve*), $h(x) = \exp(x)$, *nodes* (2.4).

$h(x_k) := (-1)^{k+1}$ for $1 \leqq k \leqq m$, then the graphs look almost the same as in Figs. 2.1–2.3. We now show an example when fewer than m Fourier coefficients are computed.

Example 2.4. In the computed examples above, the differences $|s_j - d_j|$ are largest for indices j close to m. This suggests that it may be possible to compute the first n Fourier coefficients of a function $h(x)$ accurately by the Stieltjes procedure, provided that the degree $n-1$ of the polynomial r_{n-1} is sufficiently much smaller than the number of nodes m. We illustrate this by modifying Example 2.2, which is the one of the above examples in which the Stieltjes procedure performed most poorly. Let the function $h(x)$, the nodes $x_k^{(m)}$, the weights w_k^2, and the vectors $\mathbf{c}, \mathbf{s}, \mathbf{d} \in R^m$ be the same as in Example 2.2, but now measure the length of vectors $\mathbf{u} \in R^m$ by the seminorm

$$\|\mathbf{u}\|_{\infty,m/2} := \max_{1 \leqq j \leqq m/2} |u_j|.$$

Figure 2.4 shows that $\|\mathbf{s} - \mathbf{d}\|_{\infty,m/2}$ is somewhat smaller than $\|\mathbf{c} - \mathbf{d}\|_{\infty,m/2}$ for $m \leqq 46$. Hence, in this example the Stieltjes procedure determines the first n Fourier coefficients accurately, if n is sufficiently much smaller than m. The accuracy of the first n Fourier coefficients depends on the distribution of the m nodes. Numerical experiments suggest that the Stieltjes procedure yields high accuracy if among the m nodes there are n nodes that are distributed roughly like the zeros of an nth degree Chebyshev polynomial for the interval between the smallest and largest nodes.

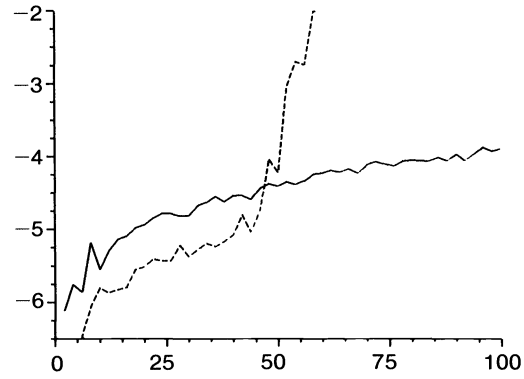


FIG. 2.4. $\log_{10} \|\mathbf{c} - \mathbf{d}\|_{\infty,m/2}$ (*continuous curve*), $\log_{10} \|\mathbf{s} - \mathbf{d}\|_{\infty,m/2}$ (*dashed curve*), $h(x) = \exp(x)$, *nodes* (2.3), $m = 2, 4, 6, \cdots, 100$.

Turning to Algorithm 2.1, we find that the differences $|c_j - d_j|$, $1 \leqq j \leqq m$, are in the Examples 2.1–2.3 largest for small values of j. In particular, for the nodes, weights, and function h of the present example, we have that $\|\mathbf{c} - \mathbf{d}\|_\infty = \|\mathbf{c} - \mathbf{d}\|_{\infty,m/2}$.

In the examples shown, as well as in many other computed examples, Algorithm 2.1 performs sometimes much better, but never much worse than the Stieltjes procedure. We therefore find it to be a viable method for computing Fourier coefficients when the inner product is discrete. The next section applies Algorithm 2.1 to the QR decomposition of Vandermonde-like matrices.

3. QR decomposition of Vandermonde-like matrices. We consider the remaining details of an algorithm for computing a QR decomposition of the $m \times n$ matrix DV^T and discuss applications to the solution of linear systems of equations. In Algorithm 2.1 the right-hand side vector $\mathbf{h} = [h(x_1), h(x_2), \cdots, h(x_m)]^T$ is multiplied by a sequence of Givens rotations, which form Q^T. These rotations are defined by the σs and γs in the algorithm. If $n < m$, then only the Givens rotations forming the first n columns of Q are computed. Thus, we obtain Q by storing the Givens rotations used by Algorithm 2.1. An economical storage scheme is described by Stewart [22].

We turn to the computation of the right triangular matrix $R = [r_{jk}]_{j,k=1}^n$ in (1.12). Assume that V has elements $v_{jk} = p_{j-1}(x_k)$, where the p_{j-1} are defined by (1.9). Having determined the sets of coefficients $\{\alpha_j\}_{j=1}^{n-1}$ and $\{\beta_j\}_{j=0}^{n-1}$ by Algorithm 2.1, we can compute the elements of R in $O(n^2)$ arithmetic operations, as described by Algorithm 3.1 below. This algorithm is derived by combining (1.11) with the recurrence formulas for the π_j and p_j.

ALGORITHM 3.1. Computation of right triangular matrix $R = [r_{jk}]_{j,k=1}^n$ from recurrence coefficients for the polynomials p_j and π_j.
Input: n, $\{\alpha_j\}_{j=1}^{n-1}$, $\{\beta_j\}_{j=0}^{n-1}$, $\{a_j\}_{j=1}^{n-1}$ $\{b_j\}_{j=0}^{n-1}$.
Output: elements r_{jk}, $1 \leqq j \leqq k \leqq n$, of matrix R.

$r_{11} := \beta_0 / b_0$;
$r_{22} := r_{11}\beta_1 / b_1$; $r_{12} := r_{11}(\alpha_1 - a_1)/b_1$;
for $j := 2, 3, \cdots, n-1$ **do**
  $r_{j+1,j+1} := r_{jj}\beta_j / b_j$;
  $r_{j,j+1} := (r_{j-1,j}\beta_{j-1} + r_{jj}(\alpha_j - a_j))/b_j$;
  for $k := 2, 3, \cdots, j-1$ **do**
    $r_{k,j+1} := (r_{k-1,j}\beta_{k-1} + r_{kj}(\alpha_k - a_j) + r_{k+1,j}\beta_k - r_{k,j-1}b_{j-1})/b_j$;
  $r_{1,j+1} := (r_{1,j}(\alpha_1 - a_j) + r_{2,j}\beta_1 - r_{1,j-1}b_{j-1})/b_j$;

□

The algorithm requires $\frac{5}{2}n^2 + O(n)$ multiplications or divisions and $2n^2 + O(n)$ additions or subtractions.

If the a_l and α_l, as well as the b_l and β_l, are interchanged in Algorithm 3.1, then the algorithm computes the elements of R^{-1}. This follows from the fact that interchanging the coefficients corresponds to interchanging p_{j-1} and π_{j-1} in (1.11). Hence, we can compute the elements of R^{-1} in $O(n^2)$ arithmetic operations. When computing R^{-1} in the manner described, the elements of R^{-1} are determined columnwise. This makes it possible to compute $\mathbf{c}' := R^{-1}\mathbf{c}$, for any $\mathbf{c} \in R^n$, using only $O(n)$ storage locations, i.e., without storing R^{-1}. In view of the fact that $\mathbf{c} := Q^T D\mathbf{h}$ can be computed using $O(m)$ storage locations by Algorithm 2.1, we find that (1.10) can be solved using only $O(m)$ storage locations.

If V is a classical Vandermonde matrix, then the polynomials p_{k-1} in (1.11) are replaced by monomials. It is straightforward to modify Algorithm 3.1 accordingly. In order to solve primal Vandermonde-like systems (1.14), we have to store the Givens rotations defining Q. This requires $O(m^2)$ storage locations. The following example illustrates the application of our scheme to the solution of dual Vandermonde-like linear systems of equations (1.10).

Example 3.1. The main advantage of our solver for Vandermonde-like linear systems of equations, when compared with other fast solvers presented in [3], [16], [17], and [19], is that our scheme is applicable to the least squares solution of overdetermined Vandermonde-like systems. It is of interest to compare the accuracy of our solver with the accuracy of other fast solution methods when V is a square matrix. This is the purpose of the present example. Let $m \geqq 2$ and let the set of equidistant nodes $\{x_k^{(m)}\}_{k=1}^m$ be defined by (2.1). Assume that all the weights are unity, i.e., $D = I$ in (1.10). Let the square Vandermonde-like matrix $V = [v_{jk}]_{j,k=1}^m$ be defined by $v_{jk} = p_{j-1}(x_k^{(m)})$, where p_{j-1} are the scaled Chebyshev polynomials of Example 1.1. We consider the solution of

$$V^T\mathbf{c}' = \mathbf{h}, \qquad \mathbf{h} := [\exp(x_1^{(m)}), \exp(x_2^{(m)}), \cdots, \exp(x_m^{(m)})]^T. \tag{3.1}$$

Figure 3.1 compares the accuracy obtained with our scheme with the accuracy achieved by a fast solution scheme for Vandermonde-like linear systems of equations with square matrices recently proposed by Higham [16], [17]. The latter method factors V^T into triangular matrices. Both Higham's scheme and ours require $O(m^2)$ arithmetic operations, but the constant multiplying m^2 is smaller for Higham's method.

Let $\hat{\mathbf{c}}$ denote the computed solution of (3.1) obtained by the QR decomposition method of the present paper using single precision arithmetic, and let $\hat{\mathbf{d}}$ denote the solution

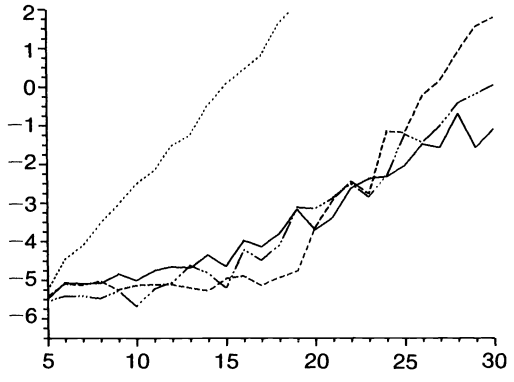


FIG. 3.1. *Errors in computed solutions of* (3.1).

of (3.1) computed in double precision arithmetic by the LINPACK [7] subroutines DQRDC and DQRSL. These subroutines solve an overdetermined linear system of equations by QR decomposition of the matrix, without using the structure of the matrix. Introduce the Euclidean norm $\|\mathbf{u}\|_2 := (\sum_{j=1}^{m} u_j^2)^{1/2}$ for $\mathbf{u} = [u_1, u_2, \cdots, u_m]^T \in R^m$. Then $\|\hat{\mathbf{c}} - \hat{\mathbf{d}}\|_2$ yields an estimate of the error in $\hat{\mathbf{c}}$. Figure 3.1 shows the growth of $\log_{10} \|\hat{\mathbf{c}} - \hat{\mathbf{d}}\|_2$ with m (continuous curve).

We now replace Algorithm 2.1 by the Stieltjes procedure in the computation of the QR decomposition of V^T and solve (3.1). Let $\hat{\mathbf{s}}$ denote the computed solution obtained in this manner using single precision arithmetic. Figure 3.1 shows the growth of $\log_{10} \|\hat{\mathbf{s}} - \hat{\mathbf{d}}\|_2$ with m (dotted curve). Clearly, QR decomposition based on Algorithm 2.1 yields higher accuracy than if the Stieltjes procedure is used.

The remaining error curves in Fig. 3.1 are for a scheme proposed by Higham [16], [17]. Let $\hat{\mathbf{a}}$ denote the computed solution of (3.1) obtained by Algorithm 1 in [16] using single precision arithmetic. The dashed curve of Fig. 3.1 shows $m \to \log_{10} \|\hat{\mathbf{a}} - \hat{\mathbf{d}}\|_2$. Let $S_m := \{2 - 4(k-1)/(m-1)\}_{k=1}^{m}$. Higham [17] proposed that the nodes be ordered to satisfy

$$
\begin{aligned}
&|x_1^{(m)}| = \max_{x \in S_m} |x|, \\
&\prod_{j=1}^{k-1} |x_k^{(m)} - x_j^{(m)}| = \max_{x \in S_m} \prod_{j=1}^{k-1} |x - x_j^{(m)}|, \qquad 2 \leqq k \leqq m,
\end{aligned} \tag{3.2}
$$

because this ordering corresponds to partial pivoting during the factorization of V^T into triangular matrices. Assume for the moment that the nodes $x_k^{(m)}$ satisfy (3.2) with $x_1^{(m)} := 2$, and let $\hat{\mathbf{b}}$ denote the computed solution of (3.1) obtained by Algorithm 1 in [16] using single precision arithmetic. The dash-triple-dotted curve of Fig. 3.1 shows $m \to \log_{10} \|\hat{\mathbf{b}} - \hat{\mathbf{d}}\|_2$.

Figure 3.1 and other computed examples suggest that when solving systems of equations with $m \times m$ Vandermonde-like matrices by the QR decomposition scheme of the present paper, the accuracy achieved is often roughly the same as the accuracy obtained by methods based on computing a triangular factorization of V^T.

4. Conclusion. Polynomials orthonormal with respect to a discrete inner product of the form (1.1) can for many problems be computed accurately by using an algorithm proposed by Rutishauser, Gragg, and Harrod for the solution of an inverse eigenvalue problem for a symmetric tridiagonal matrix. This algorithm can also be used to rapidly compute a QR decomposition of the transpose of Vandermonde-like matrices. This decomposition gives rise to new fast solution methods for Vandermonde-like overdetermined systems of equations.

Acknowledgments. This paper was completed while visiting the Department of Mathematics, Massachusetts Institute of Technology. I would like to thank Nick Trefethen for making this visit possible and enjoyable. Also, I would like to thank Bill Gragg and Nick Trefethen for discussions. Comments by a referee improved the presentation.

REFERENCES

[1] M. ALMACANY, C. B. DUNHAM, AND J. WILLIAMS, *Discrete Chebyshev approximation by interpolating rationals*, IMA J. Numer. Anal., 4 (1984), pp. 467–477.

[2] G. S. AMMAR, W. B. GRAGG, AND L. REICHEL, *Constructing a unitary Hessenberg matrix from spectral*

data, in Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, G. H. Golub and P. Van Dooren, eds., Springer-Verlag, New York, 1990, pp. 385–386.

[3] Å. BJÖRCK AND V. PEREYRA, *Solution of Vandermonde systems of equations*, Math. Comp., 24 (1970), pp. 893–903.

[4] D. BOLEY AND G. H. GOLUB, *A survey of matrix inverse eigenvalue problems*, Inverse Problems, 3 (1987), pp. 595–622.

[5] C. DE BOOR AND G. H. GOLUB, *The numerically stable reconstruction of a Jacobi matrix from spectral data*, Linear Algebra Appl., 21 (1978), pp. 245–260.

[6] C. J. DEMEURE, *Fast* QR *factorization of Vandermonde matrices*, Linear Algebra Appl., 122–124 (1989), pp. 165–194.

[7] J. J. DONGARRA, J. R. BUNCH, C. B. MOLER, AND G. W. STEWART, LINPACK *Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.

[8] S. ELHAY, G. H. GOLUB, AND J. KAUTSKY, *Updating and downdating of orthogonal polynomials with data fitting applications*, Report NA-89-04, Computer Science Department, Stanford University, Stanford, CA, 1989.

[9] G. E. FORSYTHE, *Generation and use of orthogonal polynomials for data-fitting with a digital computer*, J. Soc. Indust. Appl. Math., 5 (1957), pp. 74–88.

[10] W. GAUTSCHI, *On generating orthogonal polynomials*, SIAM J. Sci. Statist. Comput., 3 (1982), pp. 289–317.

[11] ———, *The condition of Vandermonde-like matrices involving orthogonal polynomials*, Linear Algebra Appl., 52/53 (1983), pp. 293–300.

[12] ———, *Orthogonal polynomials—constructive theory and applications*, J. Comput. Appl. Math., 12/13 (1985), pp. 61–76.

[13] W. GAUTSCHI AND G. INGLESE, *Lower bounds for the condition number of Vandermonde matrices*, Numer. Math., 52 (1988), pp. 241–250.

[14] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[15] W. B. GRAGG AND W. J. HARROD, *The numerically stable reconstruction of Jacobi matrices from spectral data*, Numer. Math., 44 (1984), pp. 317–335.

[16] N. J. HIGHAM, *Fast solution of Vandermonde-like systems involving orthogonal polynomials*, IMA J. Numer. Anal., 8 (1988), pp. 473–486.

[17] ———, *Stability analysis of algorithms for solving confluent Vandermonde-like systems*, SIAM J. Matrix Anal. Appl., 11 (1989), pp. 23–41.

[18] L. REICHEL, G. S. AMMAR, AND W. B. GRAGG, *Discrete least squares approximation by trigonometric polynomials*, Math. Comp., to appear.

[19] L. REICHEL AND G. OPFER, *Chebyshev-Vandermonde systems*, Math. Comp., to appear.

[20] H. RUTISHAUSER, *On Jacobi rotation patterns*, in Proc. Symposia in Applied Mathematics, Vol. 15, Experimental Arithmetic, High Speed Computing and Mathematics, American Mathematical Society, Providence, RI, 1963, pp. 219–239.

[21] F. J. SMITH, *An algorithm for summing orthogonal polynomial series and their derivatives with application to curve-fitting and interpolation*, Math. Comp., 19 (1976), pp. 33–36.

[22] G. W. STEWART, *The economical storage of plane rotations*, Numer. Math., 25 (1976), pp. 137–138.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 565–576, July 1991

014

FACTORIZATION PROBLEMS FOR NONMONIC MATRIX POLYNOMIALS*

MAITE GASSÓ† AND VICENTE HERNÁNDEZ‡

Abstract. In this paper, it is proven that the problem of the nonlinear factorization of a nonmonic matrix polynomial, $L(\lambda) = L_1(\lambda)L_2(\lambda)\cdots L_k(\lambda)$, where $L_2(\lambda), \cdots, L_k(\lambda)$ are regular, is related to the strict equivalence between two appropriate pencils. These pencils are then obtained from the comonic companion matrices of matrix polynomials $M(\lambda), M_1(\lambda), \cdots, M_k(\lambda)$ associated with $L(\lambda), L_1(\lambda), \cdots, L_k(\lambda)$.

Key words. matrix polynomials, factorization problems, companion pencils, strict equivalence

AMS(MOS) subject classifications. 15A23, 15A22

1. Introduction.

Consider the matrix polynomial

$$L(\lambda) = \sum_{j=0}^{n} A_j \lambda^j, \tag{1.1}$$

where $A_j \in C^{m\times m}$, $j = 0, 1, \cdots, n$, and the factorization problem given by

$$L(\lambda) = L_1(\lambda)L_2(\lambda)\cdots L_k(\lambda), \tag{1.2}$$

$$L_i(\lambda) = \sum_{j=0}^{n_i} A_{ij}\lambda^j, \qquad i = 1, \cdots, k,$$

$$\sum_{i=1}^{k} n_i = n. \tag{1.3}$$

This problem has been studied in [2]–[7] in the case where $L(\lambda)$, $L_i(\lambda)$, $i = 1, 2, \cdots, k$, are monic matrix polynomials. Two different approaches appear in these papers. One is based on the spectral properties of $L(\lambda)$, $L_i(\lambda)$, $i = 1, 2, \cdots, k$, [2], [3], [5], [7], and the other on a similarity condition between appropriate matrices obtained from the companion matrices of $L(\lambda)$, $L_i(\lambda)$, $i = 1, \cdots, k$ [4], [6]. For example, in Theorem 3.2 of [2, p. 85], it has been proved that if a matrix polynomial $L(\lambda)$ admits a factorization

$$L(\lambda) = L_2(\lambda)L_1(\lambda),$$

where $L_1(\lambda)$ and $L_2(\lambda)$ are monic, then the standard triples of $L(\lambda)$ can be obtained in terms of standard triples of $L_1(\lambda)$ and $L_2(\lambda)$. This is an example of the first approach and the above result gives only a sufficient condition. A necessary and sufficient condition has since been obtained by Hernández and Incertis [4] using the second approach.

In the nonmonic case, it has been proved [1] that the factorization problem defined by

$$L(\lambda) = (A_n\lambda - T_1)(I\lambda - T_2)\cdots(I\lambda - T_n) \tag{1.4}$$

* Received by the editors November 23, 1988; accepted for publication (in revised form) May 30, 1990.
† Departamento de Matemática Aplicada, Universidad Politécnica de Valencia, P.O. Box 22012, 46071 Valencia, Spain.
‡ Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, P.O. Box 22012, 46071 Valencia, Spain (dcnt@fi.upv.es).

is related to the strict equivalence between the pencils $C_L(\lambda)$, $J(\lambda)$, where $C_L(\lambda)$ is the companion pencil of $L(\lambda)$:

$$C_L(\lambda) = \operatorname{diag}(I, \cdots, I, A_n)\lambda - C_L,$$

with

$$C_L = \begin{bmatrix} 0 & I & 0 & \cdots & 0 \\ 0 & 0 & I & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & I \\ -A_0 & -A_1 & -A_2 & \cdots & -A_{n-1} \end{bmatrix}; \tag{1.5}$$

$J(\lambda)$ is the upper bidiagonal pencil defined by

$$J(\lambda) = \operatorname{diag}(I, \cdots, I, A_n)\lambda - \begin{bmatrix} T_n & I & & & \\ & T_{n-1} & I & & \\ & & \cdot & \cdot & \\ & & & T_2 & I \\ & & & & T_1 \end{bmatrix};$$

and $\operatorname{diag}(I, \cdots, I, A_n)$ is a block-diagonal matrix formed by matrices $I, \cdots, I, A_n$.

In this paper we extend this result to the more general factorization problem given by (1.1)–(1.3) with the regularity condition

$$\det L_i(\lambda) \not\equiv 0, \qquad i = 2, \cdots, k. \tag{1.6}$$

Note that we do not assume anything about the regularity of $L(\lambda)$. Particular cases of (1.6) are

$$L_i(\lambda) = \sum_{j=0}^{n_i - 1} A_{ij}\lambda^j + I\lambda^{n_i}, \qquad i = 2, \cdots, k \tag{1.7}$$

and

$$L_i(\lambda) = I + \sum_{j=1}^{n_i} A_{ij}\lambda^j, \qquad i = 2, \cdots, k. \tag{1.8}$$

Associated with factorization (1.1)–(1.3), (1.7), we consider two pencils: the companion pencil of $L(\lambda)$, $C_L(\lambda)$; and the upper-bidiagonal pencil defined by the companion matrices of $L_1(\lambda), \cdots, L_k(\lambda)$:

$$E(\lambda) = \operatorname{diag}(I, \cdots, I, A_n)\lambda - \begin{bmatrix} C_{L_k} & J_k & & & \\ & C_{L_{k-1}} & J_{k-1} & & \\ & & \cdot & \cdot & \\ & & & C_{L_2} & J_2 \\ & & & & C_{L_1} \end{bmatrix},$$

where J_i, $i = 2, \cdots, k$, are block-matrices with the identity in the southwest corner and zeros elsewhere.

Related to factorization (1.1)–(1.3), (1.8), we consider the comonic companion pencil of $L(\lambda)$

$$R_L(\lambda) = \operatorname{diag}(I, \cdots, I, A_0)\lambda - R_L, \tag{1.9}$$

$$R_L = \begin{bmatrix} 0 & I & 0 & \cdot & 0 \\ 0 & 0 & I & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & I \\ -A_n & -A_{n-1} & -A_{n-2} & \cdot & -A_1 \end{bmatrix},$$

and the upper bidiagonal pencil defined by the comonic companion matrices of $L_1(\lambda), \cdots, L_k(\lambda)$:

$$F(\lambda) = \operatorname{diag}(I, \cdots, I, A_0)\lambda - \begin{bmatrix} R_{L_k} & J_k & & & \\ & R_{L_{k-1}} & J_{k-1} & & \\ & & \cdot & \cdot & \\ & & & R_{L_2} & J_2 \\ & & & & R_{L_1} \end{bmatrix}. \tag{1.10}$$

In this paper we prove that the strict equivalence between $C_L(\lambda)$ and $E(\lambda)$ is a necessary and sufficient condition for the existence of factorization (1.1)–(1.3), (1.7). From this result we obtain that $L(\lambda)$ can be factorized into the form (1.1)–(1.3), (1.8) if and only if $R_L(\lambda)$ and $F(\lambda)$ are strictly equivalent. As a consequence, we prove that factorization (1.1)–(1.3), (1.6) is related to the strict equivalence between two pencils of the type (1.9), (1.10). These pencils are defined by matrix polynomials $M(\lambda), M_1(\lambda), \cdots, M_k(\lambda)$ obtained from $L(\lambda), L_1(\lambda), \cdots, L_k(\lambda)$.

2. Necessary condition for nonlinear factorization. First, we prove the following necessary condition for the existence of factorization (1.1)–(1.3), (1.7), with $k = 2$. A unit lower block-triangular matrix is a lower block-triangular matrix where all the diagonal blocks are equal to the identity matrix.

THEOREM 1. *Let* $L(\lambda) = \sum_{j=0}^{n} A_j\lambda^j$ *be a nonmonic matrix polynomial (not necessarily regular) with* $A_i \in C^{m\times m}$, $i = 0, 1, \cdots, n$. *If* $L(\lambda)$ *can be factorized into the form*

$$L(\lambda) = L_p(\lambda)L_q(\lambda), \tag{2.1}$$

$$L_p(\lambda) = \sum_{j=0}^{p} B_j\lambda^j,$$

$$L_q(\lambda) = \sum_{j=0}^{q-1} C_j\lambda^j + I\lambda^q, \tag{2.2}$$

then there exist unit lower block-triangular matrices $P, Q \in C^{mn\times mn}$ *such that*

$$PC_L(\lambda)Q = E(\lambda)$$

with

$$C_L(\lambda) = \operatorname{diag}(I, \cdots, I, A_n)\lambda - C_L,$$

$$E(\lambda) = \operatorname{diag}(I, \cdots, I, A_n)\lambda - \begin{bmatrix} C_{L_q} & J_q \\ 0 & C_{L_p} \end{bmatrix},$$

where C_L, C_{L_p}, C_{L_q} *are the companion matrices of* $L(\lambda)$, $L_p(\lambda)$, $L_q(\lambda)$, *and*

$$J_q = \begin{bmatrix} 0 & 0 \cdots 0 \\ \cdot & \cdots \\ \cdot & \cdots \\ 0 & 0 \cdots 0 \\ I & 0 \cdots 0 \end{bmatrix} \in C^{mq \times mp}. \tag{2.3}$$

First, we prove the following proposition, which will be needed in the proof of Theorem 1.

PROPOSITION 1. *If* $L(\lambda) = \sum_{j=0}^{n} A_j \lambda^j$ *can be factorized into the form* (2.1), (2.2), *then*

$$\sum_{j=0}^{n} A_j X E^j = (I - A_n)[\underbrace{0, \cdots, 0}_{mq} \mid \underbrace{B_0, B_1, \cdots, B_{p-1}}_{mp}], \tag{2.4}$$

where

$$X = [X_1, 0] = [\underbrace{I, 0, \cdots, 0}_{mq} \mid \underbrace{0, \cdots, 0}_{mp}] \tag{2.5}$$

and

$$E = \begin{bmatrix} C_{L_q} & J_q \\ 0 & C_{L_p} \end{bmatrix} \begin{matrix} mq \\ mp \end{matrix}. \tag{2.6}$$

$\qquad\qquad mq \quad mp$

Proof. From (2.6) we obtain that

$$E^j = \begin{bmatrix} C_{L_q}^j & \sum_{h=0}^{j-1} C_{L_q}^{j-1-h} J_q C_{L_p}^h \\ 0 & C_{L_p}^j \end{bmatrix}, \qquad j = 1, 2, \cdots, n,$$

and from (2.5),

$$\sum_{j=0}^{n} A_j X E^j = \left[\sum_{j=0}^{n} A_j X_1 C_{L_q}^j, \sum_{j=1}^{n} A_j X_1 \left(\sum_{h=0}^{j-1} C_{L_q}^{j-1-h} J_q C_{L_p}^h \right) \right]. \tag{2.7}$$

We consider the standard triple of $L_q(\lambda)$ given by (X_1, C_{L_q}, Y_1):

$$Y_1 = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ I \end{bmatrix} mq.$$

By Proposition 5.4.1 of [3],

$$\sum_{j=0}^{n} A_j X_1 C_{L_q}^j = 0. \tag{2.8}$$

In order to obtain (2.4), in view of (2.7), (2.8), we only need to prove that

$$\sum_{j=1}^{n} A_j X_1 \left(\sum_{h=0}^{j-1} C_{L_q}^{j-1-h} J_q C_{L_p}^{h} \right) = \sum_{h=1}^{n} \sum_{j=h}^{n} A_j X_1 C_{L_q}^{j-h} J_q C_{L_p}^{h-1} \tag{2.9}$$

$$= (I - A_n)[B_0, B_1, \cdots, B_{p-1}].$$

We use an argument similar to the proof of the above-mentioned proposition. If $L(\lambda)$ can be factorized into the form (2.1), (2.2), then

$$L(\lambda) L_q(\lambda)^{-1} = L_p(\lambda).$$

From (2.2) and the resolvent form of $L_q(\lambda)$ given by (X_1, C_{L_q}, Y_1), we obtain

$$\left(\sum_{j=0}^{n} A_j \lambda^j \right) X_1 (I\lambda - C_{L_q})^{-1} Y_1 = \sum_{j=0}^{p} B_j \lambda^j.$$

For $|\lambda|$ large enough,

$$\left(\sum_{j=0}^{n} A_j \lambda^j \right) X_1 \left(\sum_{j=0}^{\infty} C_{L_q}^{j} \lambda^{-j-1} \right) Y_1 = \sum_{j=0}^{p} B_j \lambda^j.$$

Equalizing the coefficients of the positive powers of λ, we get

$$\begin{aligned}
B_0 &= A_1 X_1 Y_1 + A_2 X_1 C_{L_q} Y_1 + \cdots + A_n X_1 C_{L_q}^{n-1} Y_1, \\
B_1 &= A_2 X_1 Y_1 + A_3 X_1 C_{L_q} Y_1 + \cdots + A_n X_1 C_{L_q}^{n-2} Y_1, \\
&\vdots \\
B_{p-1} &= A_p X_1 Y_1 + A_{p+1} X_1 C_{L_q} Y_1 + \cdots + A_n X_1 C_{L_q}^{n-p} Y_1, \\
B_p &= A_{p+1} X_1 Y_1 + A_{p+2} X_1 C_{L_q} Y_1 + \cdots + A_n X_1 C_{L_q}^{n-p-1} Y_1, \\
0 &= A_{p+2} X_1 Y_1 + A_{p+3} X_1 C_{L_q} Y_1 + \cdots + A_n X_1 C_{L_q}^{n-p-2} Y_1, \\
&\vdots \\
0 &= A_n X_1 Y_1.
\end{aligned} \tag{2.10}$$

The matrix J_q, given by (2.3), can be factorized into the form

$$J_q = Y_1 X_2, \qquad X_2 = [I, 0, \cdots, 0].$$
$$mp$$

Then, from (2.10), we deduce that

$$\sum_{j=1}^{n} A_j X_1 C_{L_q}^{j-1} J_q = \left(\sum_{j=1}^{n} A_j X_1 C_{L_q}^{j-1} Y_1 \right) X_2 = B_0 [I, 0, \cdots, 0] = [B_0, 0, \cdots, 0],$$

$$\sum_{j=2}^{n} A_j X_1 C_{L_q}^{j-2} J_q C_{L_p} = \left(\sum_{j=2}^{n} A_j X_1 C_{L_q}^{j-2} Y_1 \right) X_2 C_{L_p} = B_1 [0, I, 0, \cdots, 0] = [0, B_1, 0, \cdots, 0],$$

$$\vdots$$

$$\sum_{j=p}^{n} A_j X_1 C_{L_q}^{j-p} J_q C_{L_p}^{p-1} = \left(\sum_{j=p}^{n} A_j X_1 C_{L_q}^{j-p} Y_1 \right) X_2 C_{L_p}^{p-1}$$

$$= B_{p-1}[0, \cdots, 0, I] = [0, \cdots, 0, B_{p-1}],$$

$$\sum_{j=p+1}^{n} A_j X_1 C_{L_q}^{j-p-1} J_q C_{L_p}^{p} = \left(\sum_{j=p+1}^{n} A_j X_1 C_{L_q}^{j-p-1} Y_1 \right) X_2 C_{L_p}^{p}$$

$$= B_p[-B_0, -B_1, \cdots, -B_{p-1}],$$

$$\sum_{j=p+2}^{n} A_j X_1 C_{L_q}^{j-p-2} J_q C_{L_p}^{p+1} = \left(\sum_{j=p+2}^{n} A_j X_1 C_{L_q}^{j-p-2} Y_1 \right) X_2 C_{L_p}^{p+1} = 0,$$

$$\vdots$$

$$A_n X_1 J_q C_{L_p}^{n-1} = (A_n X_1 Y_1) X_2 C_{L_p}^{n-1} = 0.$$

From these expressions and taking into account that $B_p = A_n$, we conclude that (2.9) is true. □

Proof of Theorem 1. If $L(\lambda)$ can be factorized into the form (2.1), (2.2), and $A_n = I$, we know from [6] that

$$Q^{-1} C_L Q = E, \tag{2.11}$$

where Q is the unit lower block-triangular matrix given by

$$Q = \operatorname{col}\,[XE^j]_{j=0}^{n-1} = \begin{bmatrix} I & 0 \\ Q_1 & Q_2 \end{bmatrix} \begin{matrix} mq \\ mp \end{matrix} \qquad \begin{matrix} mq & mp \end{matrix} \tag{2.12}$$

with

$$X = [X_1, \ 0\ ], \qquad X_1 = [I, 0, \cdots, 0], \tag{2.13}$$

$$\begin{matrix} mq & mp \end{matrix}$$

$$Q_1 = \operatorname{col}\,[X_1 C_{L_q}^{j}]_{j=q}^{n-1} \tag{2.14}$$

and Q_2 is the unit lower block-triangular matrix defined by

$$Q_2 = \begin{bmatrix} I & & & & & \\ C_{q-1} & I & & & & \\ \cdot & C_{q-1} & \cdot & & & \\ \cdot & \cdot & \cdot & \cdot & & \\ C_0 & \cdot & \cdot & \cdot & & \\ 0 & C_0 & \cdot & \cdot & \cdot & \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & I \\ 0 & 0 & \cdot & \cdot & C_{q-1} & I \end{bmatrix}^{-1} . \tag{2.15}$$

The notation $\operatorname{col}\,[UV^j]_{j=0}^{n-1}$ is used to represent matrix $[U^T, (UV)^T, \cdots, (UV^{n-1})^T]^T$.

We will prove that in the nonmonic case we obtain the equivalence between C_L and E by substituting for Q^{-1}, in (2.11), another unit lower block-triangular matrix P, which we construct from the coefficients of the right divisor $L_q(\lambda)$ and matrix A_n.

From (1.5) and (2.12) we obtain that

$$C_L Q = \operatorname{col}\left[XE, XE^2, \cdots, XE^{n-1}, -\sum_{j=0}^{n-1} A_j XE^j \right], \tag{2.16}$$

and from (2.6) and (2.13)

$$
\begin{aligned}
X &= [I,0,0,\cdots,0\,|\,0,0,\cdots,0],\\
XE &= [0,I,0,\cdots,0\,|\,0,0,\cdots,0],\\
&\vdots\\
XE^{q-1} &= [0,0,0,\cdots,I\,|\,0,0,\cdots,0],\\
XE^{q} &= [-C_0,-C_1,\cdots,-C_{q-1}\,|\,I,0,\cdots,0].
\end{aligned}
\tag{2.17}
$$

We thus have the following expressions:

$$
\begin{aligned}
&\sum_{j=0}^{q-1} C_jXE^j + XE^q = [0,0,0,\cdots,0\,|\,I,0,0,\cdots,0],\\
&\sum_{j=0}^{q-1} C_jXE^{j+1} + XE^{q+1} = [0,0,0,\cdots,0\,|\,0,I,0,\cdots,0],\\
&\qquad\vdots\\
&\sum_{j=0}^{q-1} C_jXE^{j+p-1} + XE^{n-1} = [0,0,0,\cdots,0\,|\,0,0,0,\cdots,I],\\
&\sum_{j=0}^{q-1} C_jXE^{j+p} + XE^{n} = [0,0,0,\cdots,0\,|\,-B_0,-B_1,\cdots,-B_{p-1}].
\end{aligned}
\tag{2.18}
$$

If $L(\lambda)$ can be factorized into the form (2.1), (2.2) then, by Proposition 1,

$$
\sum_{j=0}^{n} A_jXE^j = [0,0,0,\cdots,0\,|\,(I-A_n)B_0,(I-A_n)B_1,\cdots,(I-A_n)B_{p-1}].
\tag{2.19}
$$

From (2.19) and the last expression of (2.18), we obtain that

$$
\sum_{j=0}^{q-1} A_nC_jXE^{j+p} - \sum_{j=0}^{n-1} A_jXE^j = [0,0,\cdots,0-B_0,-B_1,\cdots,-B_{p-1}].
\tag{2.20}
$$

Consider now the unit lower block-triangular matrix defined by

$$
P = \begin{bmatrix} I & 0 \\ P_1 & P_2 \end{bmatrix} \begin{matrix} mq \\ mp \end{matrix},
\qquad
\begin{matrix} mq & mp \end{matrix}
\tag{2.21}
$$

where

$$
[P_1,P_2] = \left[\begin{array}{cccc|cccccc}
C_0 & \cdot & \cdot & C_{q-1} & I & 0 & 0 & \cdot & 0 & 0 & 0\\
0 & C_0 & \cdot & \cdot & C_{q-1} & I & 0 & \cdot & 0 & 0 & 0\\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot\\
0 & 0 & \cdot & 0 & 0 & C_0 & \cdot & \cdot & C_{q-1} & I & 0\\
0 & 0 & \cdot & 0 & 0 & 0 & A_nC_0 & \cdot & \cdot & A_nC_{q-1} & I
\end{array}\right].
\tag{2.22}
$$

We then note that matrix P is related to matrix Q, given by (2.12)–(2.15), by means of

$$P = \operatorname{diag}(I, \cdots, I, A_n)Q^{-1} + \operatorname{diag}(0, \cdots, 0, I - A_n)$$

and $P = Q^{-1}$ when $A_n = I$.

Premultiplying (2.16) by P and taking into account the expressions (2.17), (2.18), and (2.20), we obtain that

$$PC_LQ = \begin{bmatrix} C_{L_q} & J_q \\ 0 & C_{L_p} \end{bmatrix}.$$

Furthermore, from (2.21), (2.22), and (2.12), we have

$$P \operatorname{diag}(I, \cdots, I, A_n)Q = \begin{bmatrix} I & & & & & & \\ & \ddots & & & & & \\ C_0 & \cdots & C_{q-1} & I & & & \\ & \ddots & \ddots & \ddots & \ddots & & \\ & & & C_0 & \cdots & C_{q-1} & I \\ & & & & A_nC_0 & \cdots & A_nC_{q-1} \end{bmatrix} \begin{bmatrix} X \\ XE \\ \vdots \\ XE^{n-2} \\ A_nXE^{n-1} \end{bmatrix}$$

$$= \begin{bmatrix} X \\ XE \\ \vdots \\ XE^{q-1} \\ \sum_{j=0}^{q-1} C_jXE^j + XE^q \\ \vdots \\ \sum_{j=0}^{q-1} C_jXE^{j+p-2} + XE^{n-2} \\ \sum_{j=0}^{q-1} A_nC_jXE^{j+p-1} + A_nXE^{n-1} \end{bmatrix}.$$

From (2.17), (2.18) we then conclude that

$$P \operatorname{diag}(I, \cdots, I, A_n)Q = \operatorname{diag}(I, \cdots, I, A_n)$$

and the pencils $C_L(\lambda)$ and $E(\lambda)$ are strictly equivalent. □

We now prove the necessary and sufficient conditions for the factorization problems stated in the Introduction.

3. Necessary and sufficient conditions for nonlinear factorization.

THEOREM 2. *Let* $L(\lambda) = \sum_{j=0}^{n} A_j\lambda^j$ *be a matrix polynomial* (*not necessarily regular*) *with* $A_i \in C^{m\times m}$, $i = 0, 1, \cdots, n$. $L(\lambda)$ *is factorizable in the form*

$$L(\lambda) = L_1(\lambda)L_2(\lambda)\cdots L_k(\lambda), \tag{3.1}$$

where $L_i(\lambda)$ *is a matrix polynomial of degree* n_i, $i = 1, \cdots, k$, $\sum_{i=1}^{k} n_i = n$ *and* $L_2(\lambda), \cdots,$

$L_k(\lambda)$ are monic if and only if there exist unit lower block-triangular matrices $P, Q \in C^{mn\times mn}$ such that

$$PC_L(\lambda)Q = E(\lambda) \tag{3.2}$$

with

$$C_L(\lambda) = \operatorname{diag}(I, \cdots, I, A_n)\lambda - C_L,$$

and

$$E(\lambda) = \operatorname{diag}(I, \cdots, I, A_n)\lambda - \begin{bmatrix} C_{L_k} & J_k & & & \\ & C_{L_{k-1}} & J_{k-1} & & \\ & & \cdot & \cdot & \\ & & & C_{L_2} & J_2 \\ & & & & C_{L_1} \end{bmatrix},$$

where

$$J_i = \begin{bmatrix} 0 & 0 \cdots 0 \\ \cdot & \cdot \quad \cdot \\ \cdot & \cdot \quad \cdot \\ 0 & 0 \quad 0 \\ I & 0 \cdots 0 \end{bmatrix} \in C^{mn_i \times mn_{i-1}}, \qquad i = 2, \cdots, k \tag{3.3}$$

and C_L, C_{L_i} are the companion matrices of $L(\lambda)$, $L_i(\lambda)$, $i = 1, 2, \cdots, k$.

Proof. We will prove the necessity by induction on $k \geqq 2$. The result was proved for $k = 2$ in Theorem 1. Let us assume that the result is true for $k - 1$; then we will prove that this assertion also holds for k.

We consider the nonmonic matrix polynomial

$$L_p(\lambda) = L_1(\lambda)L_2(\lambda)\cdots L_{k-1}(\lambda) = \sum_{j=0}^{p} B_j\lambda^j, \qquad p = \sum_{i=1}^{k-1} n_i. \tag{3.4}$$

By Theorem 1, and from (3.1), (3.4), there exist unit lower block-triangular matrices $R_1, S_1 \in C^{mn\times mn}$ such that

$$\begin{aligned} C_L &= R_1 \begin{bmatrix} C_{L_k} & J_k \\ 0 & C_{L_p} \end{bmatrix} S_1, \\ \operatorname{diag}(I, \cdots, I, A_n) &= R_1 \operatorname{diag}(I, \cdots, I, A_n) S_1, \end{aligned} \tag{3.5}$$

where J_k is an $mn_k \times mp$ block matrix with the identity in the southwest corner and zeros elsewhere.

From (3.1), (3.4) we note that $B_p = A_n$ and, according to the induction hypothesis, there exist unit lower block-triangular matrices $R_2, S_2 \in C^{mp\times mp}$ such that

$$\begin{aligned} C_{L_p} &= R_2 \begin{bmatrix} C_{L_{k-1}} & J_{k-1} & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & C_{L_2} & J_2 \\ & & & & C_{L_1} \end{bmatrix} S_2, \\ \operatorname{diag}(I, \cdots, I, A_n) &= R_2 \operatorname{diag}(I, \cdots, I, A_n) S_2. \end{aligned} \tag{3.6}$$

From (3.5) and (3.6) we conclude that

$$C_L = R \begin{bmatrix} C_{L_k} & J_k & & & \\ & C_{L_{k-1}} & J_{k-1} & & \\ & & \cdot & \cdot & \\ & & & C_{L_2} & J_2 \\ & & & & C_{L_1} \end{bmatrix} S, \tag{3.7}$$

$$\operatorname{diag}(I, \cdots, I, A_n) = R \operatorname{diag}(I, \cdots, I, A_n) S,$$

where R and S are unit lower block-triangular matrices given by

$$R = R_1 \operatorname{diag}(I, R_2),$$

$$S = \operatorname{diag}(I, S_2) S_1.$$

The necessity now follows from (3.7).

We prove the sufficiency using the same idea that appears in [1] for the linear factorization problem given by (1.4). Let P and Q be unit lower block-triangular matrices such that

$$PC_LQ = \begin{bmatrix} C_{L_k} & J_k & & & \\ & C_{L_{k-1}} & J_{k-1} & & \\ & & \cdot & \cdot & \\ & & & C_{L_2} & J_2 \\ & & & & C_{L_1} \end{bmatrix}, \tag{3.8}$$

$$P \operatorname{diag}(I, \cdots, I, A_n) Q = \operatorname{diag}(I, \cdots, I, A_n).$$

From (3.1), let $H(\lambda)$ be the nonmonic matrix polynomial given by

$$L_1(\lambda) L_2(\lambda) \cdots L_k(\lambda) = \sum_{j=0}^{n-1} H_j \lambda^j + A_n \lambda^n = H(\lambda). \tag{3.9}$$

Then, as has been shown in the first part of the proof, there exist unit lower block-triangular matrices R, S such that

$$C_H = R \begin{bmatrix} C_{L_k} & J_k & & & \\ & C_{L_{k-1}} & J_{k-1} & & \\ & & \cdot & \cdot & \\ & & & C_{L_2} & J_2 \\ & & & & C_{L_1} \end{bmatrix} S, \tag{3.10}$$

$$\operatorname{diag}(I, \cdots, I, A_n) = R \operatorname{diag}(I, \cdots, I, A_n) S.$$

From (3.8) and (3.10) we obtain that

$$\begin{aligned} RPC_LQS &= C_H, \\ RP \operatorname{diag}(I, \cdots, I, A_n) QS &= \operatorname{diag}(I, \cdots, I, A_n), \end{aligned} \tag{3.11}$$

where RP and QS are unit lower block-triangular matrices. Set $RP = U = [U_{ij}]$ and $(QS)^{-1} = V = [V_{ij}]$, $i, j = 1, \cdots, n$, where

$$U_{ij} = V_{ij} = \begin{cases} I & \text{if } i = j, \\ 0 & \text{if } i < j. \end{cases} \tag{3.12}$$

From (3.11) we get

$$UC_L = C_H V,$$
$$U \operatorname{diag}(I, \cdots, I, A_n) = \operatorname{diag}(I, \cdots, I, A_n)V, \tag{3.13}$$

and by means of algebraic manipulations we obtain that

$$U_{ij} = V_{ij} = 0, \qquad j < i. \tag{3.14}$$

From (3.12)–(3.14) we deduce that $A_i = H_i$, $i = 0, 1, \cdots, n-1$. Then by (3.9), the sufficiency is proved. □

COROLLARY 1. *Let $L(\lambda) = \sum_{j=0}^{n} A_j \lambda^j$ be a matrix polynomial (not necessarily regular) with $A_i \in C^{m\times m}$ $i = 0, 1, \cdots, n$, and let $L_i(\lambda)$ be a nonmonic matrix polynomial of degree n_i, $i = 1, 2, \cdots, k$, $\sum_{i=1}^{k} n_i = n$. Then $L(\lambda)$ is factorizable in the form*

$$L(\lambda) = L_1(\lambda)L_2(\lambda)\cdots L_k(\lambda), \tag{3.15}$$

$$L_i(0) = I, \qquad i = 2, \cdots, k \tag{3.16}$$

if and only if there exist unit lower block-triangular matrices $P, Q \in C^{mn\times mn}$ such that

$$PR_L(\lambda)Q = F(\lambda)$$

with

$$R_L(\lambda) = \operatorname{diag}(I, \cdots, I, A_0)\lambda - R_L,$$

$$F(\lambda) = \operatorname{diag}(I, \cdots, I, A_0)\lambda - \begin{bmatrix} R_{L_k} & J_k & & & \\ & R_{L_{k-1}} & J_{k-1} & & \\ & & \ddots & \ddots & \\ & & & R_{L_2} & J_2 \\ & & & & R_{L_1} \end{bmatrix},$$

where R_L, R_{L_i} are the comonic companion matrices of $L(\lambda)$, $L_i(\lambda)$, $i = 1, 2, \cdots, k$, and matrix J_i is defined by (3.3).

Proof. We consider matrix polynomials given by

$$N(\lambda) = \lambda^n L(\lambda^{-1}),$$
$$N_i(\lambda) = \lambda^{n_i} L_i(\lambda^{-1}), \qquad i = 1, 2, \cdots, k.$$

From (3.16), $N_2(\lambda), \cdots, N_k(\lambda)$ are monic and $L(\lambda)$ can be factorized into the form (3.15), (3.16) if and only if

$$N(\lambda) = N_1(\lambda)N_2(\lambda)\cdots N_k(\lambda).$$

Applying Theorem 1, the result is obtained, because the companion matrices of $N(\lambda)$ and $N_i(\lambda)$ are, respectively, equal to the comonic companion matrices of $L(\lambda)$, $L_i(\lambda)$, $i = 1, 2, \cdots, k$. □

COROLLARY 2. *Let $L(\lambda) = \sum_{j=0}^{n} A_j \lambda^j$ be a matrix polynomial (not necessarily regular) with $A_i \in C^{m\times m}$, $i = 0, 1, \cdots, n$ and let $L_i(\lambda)$ be a nonmonic matrix polynomial of degree n_i, $i = 1, 2, \cdots, k$, $\sum_{i=1}^{k} n_i = n$ such that*

$$\det L_i(\lambda) \not\equiv 0, \qquad i = 2, \cdots, k. \tag{3.17}$$

Let $\alpha \in C$ be such that $L_i(\alpha)$ is nonsingular for each $i = 2, \cdots, k$. We consider matrix

polynomials given by

$$M(\lambda) = L(\lambda+\alpha),$$

$$M_1(\lambda) = L_1(\lambda+\alpha)\left[\prod_{j=2}^{k} L_j(\alpha)\right],$$

$$M_i(\lambda) = \left[\prod_{j=i}^{k} L_j(\alpha)\right]^{-1} L_i(\lambda+\alpha)\left[\prod_{j=i+1}^{k} L_j(\alpha)\right], \qquad i = 2, \cdots, k-1,$$

$$M_k(\lambda) = L_k(\alpha)^{-1} L_k(\lambda+\alpha).$$

Then

$$L(\lambda) = L_1(\lambda)L_2(\lambda)\cdots L_k(\lambda) \tag{3.18}$$

if and only if there exist unit lower block-triangular matrices $P, Q \in C^{mn\times mn}$ *such that*

$$PR_M(\lambda)Q = G(\lambda)$$

with

$$R_M(\lambda) = \operatorname{diag}(I, \cdots, I, M(0))\lambda - R_M,$$

$$G(\lambda) = \operatorname{diag}(I, \cdots, I, M(0))\lambda - \begin{bmatrix} R_{M_k} & J_k & & & \\ & R_{M_{k-1}} & J_{k-1} & & \\ & & \ddots & \ddots & \\ & & & R_{M_2} & J_2 \\ & & & & R_{M_1} \end{bmatrix},$$

where J_i *is defined by* (3.3) *and* R_M, R_{M_i} *are the comonic companion matrices of* $M(\lambda)$, $M_i(\lambda)$, $i = 1, 2, \cdots, k$, *respectively.*

Proof. We note that the existence of $\alpha \in C$ such that $L_i(\alpha)$ is nonsingular, $i = 2, \cdots, k$, is a consequence of (3.17). Factorization (3.18) is equivalent to

$$M(\lambda) = M_1(\lambda)M_2(\lambda)\cdots M_k(\lambda),$$

where $M_i(0) = I$, $i = 2, \cdots, k$. Then, the result follows from Corollary 1. □

REFERENCES

[1] M. A. BEITIA AND I. ZABALLA, *Factorization of the matrix polynomial* $A(\lambda) = A_0\lambda^l + A_1\lambda^{l-1} + \cdots + A_l$, Internat. Conference on Linear Algebra and Applications, Valencia, September, 1987; Linear Algebra Appl., 121 (1989), pp. 423–432.

[2] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Matrix Polynomials*, Academic Press, New York, 1982.

[3] ———, *Invariant Subspaces of Matrices with Applications*, John Wiley, New York, 1986.

[4] V. HERNÁNDEZ AND F. INCERTIS, *A block bidiagonal form for block companion matrices*, Linear Algebra Appl., 75 (1986), pp. 241–256.

[5] V. KABAK, A. MARKUS, AND V. MEREUTSA, *On a connection between spectral properties of a polynomial operator bundle and its divisors* in Spectral Properties of Operators, Stiinca Kishinev, 1977, pp. 29–57. (In Russian.)

[6] J. MAROULAS, *A theorem on the factorization of matrix polynomials*, Linear Algebra Appl., 84 (1986), pp. 311–322.

[7] V. MEREUTSA, *Factorization of a polynomial operator pencil*, Mat. Issled., (1982), pp. 100–108. (In Russian.)

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 577–580, July 1991

015

INVERSION OF COVARIANCE MATRICES: EXPLICIT FORMULAE*

CZESŁAW STĘPNIAK†

Abstract. The inverse of the covariance matrix in a linear experiment with nonbalanced two-way hierarchical classification of the random effects is presented.

Key words. patterned matrix, covariance matrix, inversion, unbalanced model, hierarchical classification

AMS(MOS) subject classifications. primary 15A09; secondary 62J99

1. Introduction. Many statistical procedures refer to linear models, for which the best linear unbiased estimation and the minimum norm quadratic unbiased estimation require inversion of the covariance matrix (cf. Rao (1973)). The problem often reduces, in fact, to inversion of a *patterned matrix*, where the pattern is induced by the configuration of the random effects.

In the classical (i.e., balanced) case, the random effects are subject to hierarchical or cross classification with equal frequencies in subclasses. Then the covariance matrix may be easily expressed in a canonical form and there is no problem with its inversion. A key to recognition of the pattern in the unbalanced case can be found, for instance, in Stepniak (1983). We refer to Graybill (1969) as a good introduction to inverting patterned matrices. Some special computing techniques can be also taken from Householder (1957), Greenberg and Sarhan (1959), Dwyer (1964), and Searle (1966).

This note deals with the inversion of the covariance matrix in the unbalanced hierarchical classification. So far the problem has been solved explicitly for the one-way case only. Moreover, a stage-by-stage iterative computational procedure was given by LaMotte (1972). We present an explicit form of the inversion of the covariance matrix in a linear experiment with the unbalanced two-way classification of the random effects.

2. Definitions and initial reduction. The covariance matrix in a linear experiment with hierarchical classification of the random effects can be formally defined as follows.

Let n, q, and $k_1, \cdots, k_q$ be arbitrary positive integers such that $q < n$ and $k_1 < k_2 < \cdots < k_q < n$. Moreover, let $n_{11}, \cdots, n_{1k_1}; n_{21}, \cdots, n_{2k_2}; \cdots; n_{q1}, \cdots, n_{qk_q}$ be positive integers such that

$$\sum_{j=1}^{k_i} n_{ij} = n, \qquad i = 1, \cdots, q,$$

and all the elements of the matrix $V_i - V_{i+1}$, where

$$V_i = \operatorname{diag}(J_{n_{i1}}, \cdots, J_{n_{ik_i}}), \tag{1}$$

are nonnegative for $i = 1, \cdots, q-1$, and J_{mn} is the $m \times n$ matrix consisting of ones; when $m = n$, we write J_m instead of J_{mm}.

Any matrix Σ of the form

$$\Sigma = \gamma_0 I_n + \sum_{i=1}^{q} \gamma_i V_i,$$

* Received by the editors May 22, 1989; accepted for publication (in revised form) June 26, 1990.

† Institute of Applied Mathematics, Agricultural University of Lublin, Akademicka 13, PL-20-934 Lublin, Poland.

where V_i is defined by (1), while γ_0 is positive and $\gamma_1, \cdots, \gamma_q$ are nonnegative scalars, is said to be the covariance matrix in a linear experiment with q-way hierarchical classification of the random effects (cf. Stępniak (1983, § 3)). The classification is said to be *balanced* if the matrices $V_1, \cdots, V_q$ may be presented in the form

$$V_i = I_{k_i} \otimes J_{r_i}, \qquad i = 1, \cdots, q,$$

where $\otimes$ denotes the Kronecker product of matrices; otherwise it is said to be *unbalanced*.

The simplest example of hierarchical classification is one-way classification. For this case

$$\Sigma = \gamma_0 I_n + \gamma_1 \operatorname{diag}(J_{n_1}, \cdots, J_{n_k})$$

and

$$\Sigma^{-1} = \gamma_0^{-1}(I_n - \rho A),$$

where

$$A = \operatorname{diag}[(1+\rho n_1)^{-1} J_{n_1}, \cdots, (1+\rho n_k)^{-1} J_{n_k}]$$

and $\rho = \gamma_1/\gamma_0$.

In this paper we focus on two-way hierarchical classification. For this case the covariance matrix may be presented in the form

$$\Sigma = \gamma_0 \operatorname{diag}(B_1, \cdots, B_k),$$

where $B_1, \cdots, B_k$ are matrices of type

$$B = I_m + \rho \operatorname{diag}(J_{m_1}, \cdots, J_{m_r}) + \lambda J_m, \tag{2}$$

while $m, m_1, \cdots, m_r$ are positive integers satisfying $m = \sum_{i=1}^{r} m_i$, and ρ and λ are nonnegative scalars defined by $\rho = \gamma_1/\gamma_0$ and $\lambda = \gamma_2/\gamma_0$.

In this way the problem of inversion of the covariance matrix in a linear experiment with a two-way hierarchical classification of the random effects reduces to inversion of the matrix (2).

3. Inversion of the matrix B. We are seeking the inversion of the matrix

$$B = I_m + \rho \operatorname{diag}(J_{m_1}, \cdots, J_{m_r}) + \lambda J_m$$

in the form

$$I - X, \tag{3}$$

where

$$X = \begin{bmatrix} x_{11}J_{m_1m_1} & x_{12}J_{m_1m_2} & \cdots & x_{1r}J_{m_1m_r} \\ x_{21}J_{m_2m_1} & x_{22}J_{m_2m_2} & \cdots & x_{2r}J_{m_2m_r} \\ \cdots & \cdots & \cdots & \cdots \\ x_{r1}J_{m_rm_1} & x_{r2}J_{m_rm_1} & \cdots & x_{rr}J_{m_rm_r} \end{bmatrix}. \tag{4}$$

The matrix equation

$$B(I - X) = I \tag{5}$$

may be rewritten as a system of linear equations

$$x_{1i}+\rho m_1 x_{1i}+\lambda\sum_{j=1}^{r} m_j x_{ji}-\lambda=0,$$

$$x_{2i}+\rho m_2 x_{2i}+\lambda\sum_{j=1}^{r} m_j x_{ji}-\lambda=0,$$

$$\cdots$$

$$x_{ii}+\rho m_i x_{ii}-\rho+\lambda\sum_{j=1}^{r} m_j x_{ji}-\lambda=0,$$

$$\cdots$$

$$x_{ri}+\rho m_r x_{ri}+\lambda\sum_{j=1}^{r} m_j x_{ji}-\lambda=0,$$

$i = 1, \cdots, r$, or equivalently, as

$$\begin{aligned}
&(1+\rho m_1+\lambda m_1)x_{1i}+\lambda m_2 x_{2i}+\cdots+\lambda m_r x_{ri}=\lambda,\\
&\lambda m_1 x_{1i}+(1+\rho m_2+\lambda m_2)x_{2i}+\cdots+\lambda m_r x_{ri}=\lambda,\\
&\cdots\\
&\lambda m_1 x_{1i}+\cdots+(1+\rho m_i+\lambda m_i)x_{ii}+\cdots+\lambda m_r x_{ri}=\lambda+\rho,\\
&\cdots\\
&\lambda m_1 x_{1i}+\lambda m_2 x_{2i}+\cdots+(1+\rho m_r+\lambda m_r)x_{ri}=\lambda
\end{aligned}\tag{6}$$

for $i = 1, \cdots, r$.

Now let us write the system (6) in matrix form:

$$Wx_i=b_i,\qquad i=1,\cdots,r,$$

where $x_i = (x_{1i}, \cdots, x_{ri})'$ and b_i is the vector of size $r \times 1$ with the ith component equal to $\rho + \lambda$, and λ otherwise. Then the solution of (5) may be presented in the form (4) with

$$x_{ij}=\frac{|W_{ij}|}{|W|},\qquad i=1,\cdots,r,\qquad j=1,\cdots,r,$$

where W_{ij} is the matrix obtained from W by replacing its ith column by b_j.

By routine algebra we get

$$|W|=\prod_{i=1}^{r}(1+\rho m_i)+\lambda\sum_{i=1}^{r}\prod_{j\neq i} m_i(1+\rho m_j)$$

and

$$|W_{ij}|=\begin{cases}(\rho+\lambda)\displaystyle\prod_{k\neq i}(1+\rho m_k)+\rho\lambda\sum_{k\neq i} m_k\prod_{r\neq i,k}(1+\rho m_r) & \text{for } i=j,\\[2ex] \lambda\displaystyle\prod_{k\neq i,j}(1+\rho m_k) & \text{for } i\neq j.\end{cases}$$

Thus, in consequence, the inverse of the matrix B may be presented in the form (3),

where X is given by (4) with

$$x_{ij} = \begin{cases} \dfrac{\dfrac{\lambda}{(1+\rho m_i)^2}}{1+\lambda \sum_{k=1}^{r} \dfrac{m_k}{1+\rho m_k}} + \dfrac{\rho}{1+\rho m_i} & \text{for } i=j, \\ \dfrac{\dfrac{\lambda}{(1+\rho m_i)(1+\rho m_j)}}{1+\lambda \sum_{k=1}^{r} \dfrac{m_k}{1+\rho m_k}} & \text{for } i \neq j. \end{cases}$$

Acknowledgment. It is a pleasure for me to thank the Associate Editor Ingram Olkin for calling my attention to some references and for his help in the preparation of the final version of the work.

REFERENCES

P. S. DWYER (1964), *Matrix inversion with the square root method*, Technometrics, 7, pp. 197–213.

F. A. GRAYBILL (1969), *Introduction to Matrices with Applications in Statistics*, Wadsworth, Belmont, CA.

B. G. GREENBERG AND A. E. SARHAN (1959), *Matrix inversion, its interest and application in analysis of data*, J. Amer. Statist. Assoc., 54, pp. 755–766.

A. S. HOUSEHOLDER (1957), *A survey of some closed methods for inverting matrices*, J. Soc. Indust. Appl. Math., 5, pp. 155–169.

L. R. LAMOTTE (1972), *Notes on the covariance matrix of a random, nested* ANOVA *model*, Ann. Math. Statist., 43, pp. 659–662.

C. R. RAO (1973), *Linear Statistical Inference and Its Applications*, Second Edition, John Wiley, New York.

S. R. SEARLE (1966), *Matrix Analysis for the Biological Scientists (Including Applications in Statistics)*, John Wiley, New York.

C. STĘPNIAK (1983), *Optimal allocation of units in experimental designs with hierarchical and cross classification*, Ann. Inst. Statist. Math. A, 35, pp. 461–473.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 581–591, July 1991

016

THE EQUATION $AXB + CYD = E$ OVER A PRINCIPAL IDEAL DOMAIN*

A. BÜLENT ÖZGÜLER†

Abstract. This paper considers the equation $AXB + CYD = E$ in matrices over a principal ideal domain. Under the assumption that $[A : C]$ is left invertible and $[B' : D']'$ is right invertible in the domain, it is shown that this equation is solvable if and only if both $AX + YD = E$ and $XB + CY = E$ are solvable. The set of all solutions to the equation are shown to be in bijective correspondence with the set of solutions to the latter two equations modulo an equivalence relation.

Key words. linear matrix equations, equations over a ring, multivariable control systems, algebraic control

AMS(MOS) subject classifications. 15A06, 15A33, 93C35, 93C05, 93B25

1. Introduction.

We are concerned with the solvability of a linear matrix equation of the type

$$AXB + CYD = E \tag{1}$$

over an arbitrary principal ideal domain (PID) **D**. The set of solutions (X, Y) of this equation is also examined. The special cases of this equation that are of particular significance are those of type $AXB = E$ and $AX + YD = E$. These figure as the main solvability conditions of various algebraic control problems. The general equation (1) itself comes up in the analysis of the regulator problems (see, e.g., [2] and [10]).

Being linear, these equations are easily analyzed using Kronecker products and the theory of linear vector equation $Ax = b$ over a PID. This well-known approach can also be used to yield a description of the set of all solutions to (1) in terms of a particular solution. Although this may be completely satisfactory from a purely mathematical point of view (it encompasses a verifiable solvability condition and a constructive procedure to obtain solutions, etc.), it has a major drawback from a system theoretic viewpoint. The drawback is that the structure of the original matrices composing the problem data are lost in the solvability condition and in the description of the set of solutions. (The extensive literature on linear matrix equations in mathematical journals show that such a concern is not actually peculiar to system theorists.) In what follows, we state and derive alternative solvability conditions in which the structure of the matrices A, B, C, D, E is preserved and which are still suitable for obtaining a description of the set of all solutions to the equation in terms of a particular solution.

Among the many papers which deal with (1) *in the special case where* **D** *is a field*, [1] and [11] are noteworthy. In [1], a necessary and sufficient condition for (1) to be consistent (i.e., a solvability condition) is given in terms of *g-inverses* of the matrices A, B, C, D. In [11], for the special case $B = I$, $C = I$, it is shown that (1) is solvable if and only if

$$\begin{bmatrix} A & E \\ 0 & D \end{bmatrix}, \quad \begin{bmatrix} A & 0 \\ 0 & D \end{bmatrix}$$

are equivalent over the field **D** (which amounts to the equality of ranks of these two

* Received by the editors September 18, 1989; accepted for publication (in revised form) July 25, 1990.
† Electrical and Electronics Engineering, Bilkent University, Bilkent 06533, Ankara, Turkey (ozguler@trbilun.bitnet).

matrices). This result of [11] extends to the general case of **D** being a PID, as we emphasize below in Lemma 2.

Some of the terms that we use in the subsequent sections are now defined. A matrix M over **D** (i.e., with entries in **D**) is called *left unimodular* (or *left invertible over* **D**) if and only if $MN = I$ for some matrix N over **D**. It is *right unimodular* (or *right invertible over* **D**) if and only if $LM = I$ for some matrix L over **D**. It is *unimodular* if and only if it is both left and right unimodular. Two matrices A, B are *left coprime* if and only if the matrix $[A : B]$ is left unimodular and *right coprime* if and only if $[A' : B']'$ is right unimodular, where "prime" denotes "transpose." Let a matrix M over **D** have full row rank over the field of fractions **F** of **D**. Then, there exists a square matrix L and a left unimodular matrix $\bar{M}$, both over **D**, such that $M = L\bar{M}$. Such a matrix L is called *a greatest left factor of* M and it is denoted by $L = \mathbf{glf}(M)$. A $\mathbf{glf}(M)$ is unique up to right multiplications by unimodular matrices. For a full column rank matrix M over **D**, *a greatest right factor of* M, $\mathbf{grf}(M)$, is the transpose of a greatest left factor of M'.

2. The equations $AXB = E$ and $AX + YD = E$. Let $A \in \mathbf{D}^{a\times k}$, $B \in \mathbf{D}^{l\times b}$, and $E \in \mathbf{D}^{a\times b}$. Let $U \in \mathbf{D}^{a\times a}$ be a unimodular matrix such that

$$UA = \begin{bmatrix} \bar{A} \\ 0 \end{bmatrix}$$

for some full row rank matrix $\bar{A} \in \mathbf{D}^{r_A\times k}$, where r_A is the rank of A over the field of fractions **F** of **D**. Also let $V \in \mathbf{D}^{b\times b}$ be a unimodular matrix such that

$$BV = [\bar{B} \quad 0]$$

for some full column rank matrix $\bar{B} \in \mathbf{D}^{l\times r_B}$, where r_B is the rank of B over **F**. Define

$$UEV =: \begin{bmatrix} \bar{E} & \bar{E}_{12} \\ \bar{E}_{21} & \bar{E}_{22} \end{bmatrix},$$

where the matrix on the right-hand side is partitioned so that $\bar{E} \in \mathbf{D}^{r_A\times r_B}$. Furthermore, let

$$L_A := \mathbf{glf}(\bar{A}), \qquad R_B := \mathbf{grf}(\bar{B}),$$

so that

$$\bar{A} = L_A\hat{A}, \qquad \bar{B} = \hat{B}R_B,$$

for some left invertible $\hat{A}$ and right invertible $\hat{B}$ over **D**.

LEMMA 1. *Equation*

$$AXB = E \tag{2}$$

has a solution over **D** *if and only if*

(i) $\bar{E}_{12} = 0, \quad \bar{E}_{21} = 0, \quad \bar{E}_{22} = 0,$

(ii) $L_A^{-1}\bar{E}R_B^{-1} \in \mathbf{D}^{r_A\times r_B}$.

Proof. See, e.g., [9] for the proof. □

Let $A \in \mathbf{D}^{a\times k}$, $D \in \mathbf{D}^{n\times b}$, and $E \in \mathbf{D}^{a\times b}$. The solvability of

$$AX + YD = E \tag{3}$$

has been examined by Roth [11] and the following is a fundamental result. (The original proof by Roth is for the special case where **D** is the ring of polynomials. However, the same proof applies to arbitrary PIDs with almost no change.)

LEMMA 2. *Equation* (3) *is solvable over* **D** *if and only if*

$$\begin{bmatrix} A & E \\ 0 & D \end{bmatrix}, \quad \begin{bmatrix} A & 0 \\ 0 & D \end{bmatrix}$$

are equivalent over **D**.

Still, a special case of (1) is obtained by letting $E = I$ in (3). It is interesting to note that the analysis of (3) can be reduced to the analysis of an equation of the type $AX + YD = I$ by redefining the matrices A and D. The following is an unpublished result of Fuhrmann [4].

LEMMA 3. *Equation* (3) *has a solution if and only if*

$$\begin{bmatrix} A & 0 \\ 0 & I \end{bmatrix}\hat{X} + \hat{Y}\begin{bmatrix} I & E \\ 0 & D \end{bmatrix} = I \tag{4}$$

is solvable, or equivalently, if and only if

$$\begin{bmatrix} A & E \\ 0 & I \end{bmatrix}\tilde{X} + \tilde{Y}\begin{bmatrix} I & 0 \\ 0 & D \end{bmatrix} = I \tag{5}$$

is solvable.

Proof. Given a solution (X, Y) of (3), we let

$$\hat{Y} := \begin{bmatrix} I & -Y \\ 0 & 0 \end{bmatrix}, \qquad \hat{X} := \begin{bmatrix} 0 & -X \\ 0 & I \end{bmatrix},$$

which clearly satisfies (4). Conversely, given a solution $(\hat{X}, \hat{Y})$ of (4), partition $\hat{X}$ and $\hat{Y}$ compatibly and define $X := -\hat{X}_{12} - \hat{X}_{11}E$, $Y := -\hat{Y}_{12}$. Now, (3) is satisfied by X and Y. The fact that "(4) is solvable if and only if (5) is" can be established by simple unimodular transformations on the equations. □

3. The equation $AXB + CYD = E$. Let $A \in \mathbf{D}^{a\times k}$, $B \in \mathbf{D}^{l\times b}$, $C \in \mathbf{D}^{a\times m}$, $D \in \mathbf{D}^{m\times b}$, and $E \in \mathbf{D}^{a\times b}$. The solvability of the general equation (1) over **D** is now considered. This equation may be rewritten as

$$[A \quad C]\begin{bmatrix} X & 0 \\ 0 & Y \end{bmatrix}\begin{bmatrix} B \\ D \end{bmatrix} = E \tag{6}$$

and hence, it is of the type (2), where a solution of decentralized structure is sought over **D**. Let U and V be unimodular matrices over **D** such that

$$U[A \quad C] = \begin{bmatrix} \bar{A} & \bar{C} \\ 0 & 0 \end{bmatrix},$$

$$\begin{bmatrix} B \\ D \end{bmatrix}V = \begin{bmatrix} \bar{B} & 0 \\ \bar{D} & 0 \end{bmatrix},$$

where $[\bar{A} : \bar{C}]$ is of full row rank and $[\bar{B}' : \bar{D}']'$ is of full column rank. Define

$$\begin{bmatrix} \bar{E} & \bar{E}_{12} \\ \bar{E}_{21} & \bar{E}_{22} \end{bmatrix} := UEV,$$

where the partition is such that $\bar{E}$ is of size rank $[A : C] \times$ rank $[B' : D']'$. Further-

more, let

$$[\bar{A} \quad \bar{C}] = L[\hat{A} \quad \hat{C}],$$

$$\begin{bmatrix} \bar{B} \\ \bar{D} \end{bmatrix} = \begin{bmatrix} \hat{B} \\ \hat{D} \end{bmatrix} R,$$

for a left invertible $[\hat{A} : \hat{C}]$ and right invertible $[\hat{B}' : \hat{D}']'$ over **D**. Thus,

$$L = \mathbf{glf}\,[\bar{A} \quad \bar{C}], \qquad R = \mathbf{grf} \begin{bmatrix} \bar{B} \\ \bar{D} \end{bmatrix}.$$

Define

$$\hat{E} := L^{-1}\bar{E}R^{-1},$$

which is a matrix over the field of fractions of **D**. The following is a direct consequence of Lemma 1.

PROPOSITION 4. *Equation* (1) *is solvable if and only if all three of the following conditions hold*:

(i) $\bar{E}_{12} = 0, \quad \bar{E}_{21} = 0, \quad \bar{E}_{22} = 0.$

(ii) $\hat{E}$ *is over* **D**.

(iii) *Equation*

$$\hat{A}X\hat{B} + \hat{C}Y\hat{D} = \hat{E} \tag{7}$$

is solvable over **D**.

Furthermore, if the conditions (i) *and* (ii) *hold, then* (X, Y) *is a solution of* (1) *if and only if it is a solution of* (7).

Using this result, we can without loss of generality consider the solvability of (1) under the assumptions

$$[A \quad C] \text{ is left unimodular,} \tag{8}$$

$$\begin{bmatrix} B \\ D \end{bmatrix} \text{ is right unimodular.} \tag{9}$$

By these assumptions, there exist matrices $\bar{M}, \bar{N}, \bar{L}, \bar{K}$ over **D** such that

$$\begin{bmatrix} A & C \\ \bar{N} & -\bar{M} \end{bmatrix}, \qquad \begin{bmatrix} \bar{K} & D \\ \bar{L} & -B \end{bmatrix}$$

are unimodular matrices. Showing their inverses explicitly, in compatibly partitioned form, we have

$$\begin{bmatrix} A & C \\ \bar{N} & -\bar{M} \end{bmatrix} \begin{bmatrix} M & \bar{C} \\ N & -\bar{A} \end{bmatrix} = I, \qquad \begin{bmatrix} \bar{B} & \bar{D} \\ L & -K \end{bmatrix} \begin{bmatrix} \bar{K} & D \\ \bar{L} & -B \end{bmatrix} = I. \tag{10}$$

Note. In the case where $[A : C]$ and/or $[B' : D']$ is square and hence already unimodular, we can let the following expressions $\bar{C}, \bar{A}, \bar{M}, \bar{N}$ and/or $\bar{B}, \bar{D}, \bar{K}, \bar{L}$ be zero and all subsequent claims remain valid.

We can now give the following alternative conditions for the solvability of (1).

THEOREM 5. *Let* (8) *and* (9) *hold. Then, the following are equivalent statements*:

(i) *Equation* (1) *is solvable.*

(ii) *The equation*

$$A\bar{C}\bar{X} + \bar{Y}\bar{D}B = EKB - AME \tag{11}$$

is solvable.

(iii) *Both of the equations*

$$AX_1 + Y_1D = E, \qquad X_2B + CY_2 = E \tag{12}$$

are solvable.

(iv) *The equivalence over* **D**

$$\begin{bmatrix} C & E & 0 & 0 \\ 0 & B & 0 & 0 \\ 0 & 0 & A & E \\ 0 & 0 & 0 & D \end{bmatrix} \equiv \begin{bmatrix} C & 0 & 0 & 0 \\ 0 & B & 0 & 0 \\ 0 & 0 & A & 0 \\ 0 & 0 & 0 & D \end{bmatrix}$$

holds.

(v) *The equation*

$$\begin{bmatrix} C & 0 \\ 0 & A \end{bmatrix}\hat{X} + \hat{Y}\begin{bmatrix} B & 0 \\ 0 & D \end{bmatrix} = \begin{bmatrix} E & 0 \\ 0 & E \end{bmatrix} \tag{13}$$

is solvable.

Proof. We establish the following chain of implications: (iii) $\Rightarrow$ (i) $\Rightarrow$ (ii) $\Rightarrow$ (iv) $\Rightarrow$ (v) $\Rightarrow$ (iii). [(iii) $\Rightarrow$ (i)]. Given X_1, Y_1, X_2, Y_2 satisfying (12) let $X := X_1K + M(Y_1\bar{K} + X_2\bar{L})\bar{D}$ and $Y := Y_2L + N(Y_1\bar{K} + X_2\bar{L})\bar{B}$. Using (10), it is straightforward to verify that (X, Y) is a solution to (1).

[(i) $\Rightarrow$ (ii)]. Given X and Y satisfying (1), let

$$\bar{X} := \bar{N}(XB - ME) - \bar{M}(YD - NE), \qquad \bar{Y} := (CY - EL)\bar{K} - (AX - EK)\bar{L}.$$

Using (10), it can be seen that these matrices satisfy (11).

[(ii) $\Rightarrow$ (iv)]. If (11) is solvable over **D**, by Lemma 2, we have

$$\begin{bmatrix} A\bar{C} & EKB - AME \\ 0 & \bar{D}B \end{bmatrix} \equiv \begin{bmatrix} A\bar{C} & 0 \\ 0 & \bar{D}B \end{bmatrix}.$$

Extending each matrix by identity matrices of size $a + b$, we also have

$$\begin{bmatrix} A\bar{C} & EKB - AME & 0 & 0 \\ 0 & \bar{D}B & 0 & 0 \\ 0 & 0 & I_a & 0 \\ 0 & 0 & 0 & I_b \end{bmatrix} \equiv \begin{bmatrix} A\bar{C} & 0 & 0 & 0 \\ 0 & \bar{D}B & 0 & 0 \\ 0 & 0 & I_a & 0 \\ 0 & 0 & 0 & I_b \end{bmatrix}.$$

By simple unimodular operations, which also employ the unimodular matrices appearing in (10), it is not difficult to see that

$$\begin{bmatrix} A\bar{C} & EKB - AME & 0 & 0 \\ 0 & \bar{D}B & 0 & 0 \\ 0 & 0 & I_a & 0 \\ 0 & 0 & 0 & I_b \end{bmatrix} \equiv \begin{bmatrix} C & E & 0 & 0 \\ 0 & D & 0 & 0 \\ 0 & 0 & A & E \\ 0 & 0 & 0 & D \end{bmatrix},$$

$$\begin{bmatrix} A\bar{C} & 0 & 0 & 0 \\ 0 & \bar{D}B & 0 & 0 \\ 0 & 0 & I_a & 0 \\ 0 & 0 & 0 & I_b \end{bmatrix} \equiv \begin{bmatrix} C & 0 & 0 & 0 \\ 0 & B & 0 & 0 \\ 0 & 0 & A & 0 \\ 0 & 0 & 0 & D \end{bmatrix}.$$

This yields the equivalence claimed in (iv).

[(iv) $\Rightarrow$ (v)]. By permutation of the rows and columns of the first matrix, it follows that

$$\begin{bmatrix} C & E & 0 & 0 \\ 0 & B & 0 & 0 \\ 0 & 0 & A & E \\ 0 & 0 & 0 & D \end{bmatrix} \equiv \begin{bmatrix} C & 0 & E & 0 \\ 0 & A & 0 & E \\ 0 & 0 & B & 0 \\ 0 & 0 & 0 & D \end{bmatrix}.$$

Similarly,

$$\begin{bmatrix} C & 0 & 0 & 0 \\ 0 & B & 0 & 0 \\ 0 & 0 & A & 0 \\ 0 & 0 & 0 & D \end{bmatrix} \equiv \begin{bmatrix} C & 0 & 0 & 0 \\ 0 & A & 0 & 0 \\ 0 & 0 & B & 0 \\ 0 & 0 & 0 & D \end{bmatrix}.$$

Hence the matrices on the right-hand sides are equivalent. By Lemma 2, the result follows.

[(v) $\Rightarrow$ (iii)]. This is immediate on partitioning the matrices $\hat{X}$, $\hat{Y}$ conformably in (13). □

Remarks. (1) Given two equivalent block-diagonal matrices, it is by no means true that their corresponding diagonal blocks are equivalent. For this reason, one point may look peculiar in Theorem 5. Comparing the conditions (iii) and (iv) and interpreting (12) via Lemma 2, it is clear that (iv) implies the equivalence of the corresponding diagonal blocks of the matrices appearing in (iv). This quite surprising fact is a consequence of the special structure of the block entries together with our coprimeness assumptions among the matrices A, C and B, D.

(2) Let **D** be a field. Then, equivalence over **D** is simply rank equality and the condition (iii) of Theorem 5 together with Proposition 4 easily yields that (1) is solvable if and only if all the rank equalities below hold:

$$\operatorname{rank}[A \quad C \quad E] = \operatorname{rank}[A \quad C \quad 0],$$

$$\operatorname{rank}\begin{bmatrix} E \\ D \\ B \end{bmatrix} = \operatorname{rank}\begin{bmatrix} 0 \\ D \\ B \end{bmatrix},$$

$$\operatorname{rank}\begin{bmatrix} A & E \\ 0 & D \end{bmatrix} = \operatorname{rank}\begin{bmatrix} A & 0 \\ 0 & D \end{bmatrix},$$

$$\operatorname{rank}\begin{bmatrix} C & E \\ 0 & B \end{bmatrix} = \operatorname{rank}\begin{bmatrix} C & 0 \\ 0 & B \end{bmatrix}.$$

This is an alternative to the condition given in [1] and it is more directly related to the problem data.

4. The set of all solutions. We now relate the set of solutions of the equation (1) and the set of solutions of the uncoupled equations (12).

Let us denote

$$\mathscr{S} := \{(X, Y) : AXB + CYD = E\},$$

$$\mathscr{S}_{12} := \{(X_1, Y_1, X_2, Y_2) : AX_1 + Y_1D = E, X_2B + CY_2 = E\}.$$

These are the solution sets of (1) and (12), respectively, over **D**. On $\mathscr{S}$, we define an equivalence relation by

$$(X, Y) \cong (\bar{X}, \bar{Y})$$

if and only if

$$\bar{X} = X + \bar{C}\Theta\bar{D}, \qquad \bar{Y} = Y - \bar{A}\Theta\bar{B} \tag{14}$$

for some Θ over **D**, where $\bar{C}$, $\bar{D}$, $\bar{A}$, $\bar{B}$ are the matrices in (10). Similarly, on $\mathscr{S}_{12}$, we define an equivalence relation by

$$(X_1, Y_1, X_2, Y_2) \cong (\bar{X}_1, \bar{Y}_1, \bar{X}_2, \bar{Y}_2)$$

if and only if

$$\bar{X}_1 = X_1 + \Theta_1 D, \qquad \bar{Y}_1 = Y_1 - A\Theta_1, \qquad \bar{X}_2 = X_2 + C\Theta_2, \qquad \bar{Y}_{21} = Y_{21} - \Theta_2 B, \tag{15}$$

for some Θ_1 and Θ_2 over **D**. Let $[\mathscr{S}] = \{[X, Y]\}$ and $[\mathscr{S}_{12}] = \{[X_1, Y_1, X_2, Y_2]\}$ denote the sets of equivalence classes induced by these equivalence relations.

THEOREM 6. *The map* ψ: $[\mathscr{S}_{12}] \to [\mathscr{S}]$ *defined by*

$$[X_1, Y_1, X_2, Y_2] \mapsto [X_1K + M(Y_1\bar{K} + X_2\bar{L})\bar{D},\ Y_2L + N(Y_1\bar{K} + X_2\bar{L})\bar{B}],$$

is a bijection.

Proof. We show that the map is well defined, onto, and one-to-one.

[Well-defined]. If $(\bar{X}_1, \bar{Y}_1, \bar{X}_2, \bar{Y}_2) \in [X_1, Y_1, X_2, Y_2]$, then there exist Θ_1, Θ_2 over **D** such that (15) holds. Defining

$$\Theta := \bar{N}\Theta_1\bar{K} + \bar{M}\Theta_2\bar{L},$$

it is easy to verify, using (10), that

$$\bar{X}_1K + M(\bar{Y}_1\bar{K} + \bar{X}_2\bar{L})\bar{D} = X_1K + M(Y_1\bar{K} + X_2\bar{L})\bar{D} + \bar{C}\Theta\bar{D},$$

$$\bar{Y}_2L + N(\bar{Y}_1\bar{K} + \bar{X}_2\bar{L})\bar{B} = Y_2L + N(Y_1\bar{K} + X_2\bar{L})\bar{B} - \bar{A}\Theta\bar{B}.$$

Consequently, $\psi([\bar{X}_1, \bar{Y}_1, \bar{X}_2, \bar{Y}_2]) = \psi([X_1, Y_1, X_2, Y_2])$ and the map is well defined.

[Onto]. Given $[X, Y] \in [\mathscr{S}]$, consider

$$X_1 := XB, \qquad Y_1 := CY, \qquad X_2 := AX, \qquad Y_2 := YD.$$

Then, an easy computation employing (10) yields

$$\psi([X_1, Y_1, X_2, Y_2]) = [X + \bar{C}\Theta\bar{D},\ Y - \bar{A}\Theta\bar{B}], \qquad \Theta := \bar{M}Y\bar{K} - \bar{N}X\bar{L}.$$

Consequently, $\psi([X_1, Y_1, X_2, Y_2]) = [X, Y]$ and the map is onto.

[One-to-one]. Let $\psi([X_1, Y_1, X_2, Y_2]) = [X, Y]$ and $\psi([\bar{X}_1, \bar{Y}_1, \bar{X}_2, \bar{Y}_2]) = [\bar{X}, \bar{Y}]$ and suppose $[X, Y] = [\bar{X}, \bar{Y}]$. Then, there exists Θ over **D** such that

$$\bar{X} = \bar{X}_1K + M(\bar{Y}_1\bar{K} + \bar{X}_2\bar{L})\bar{D} = X + \bar{C}\Theta\bar{D} = X_1K + M(Y_1\bar{K} + X_2\bar{L})\bar{D} + \bar{C}\Theta\bar{D},$$

$$\bar{Y} = \bar{Y}_2L + N(\bar{Y}_1\bar{K} + \bar{X}_2\bar{L})\bar{B} = Y - \bar{A}\Theta\bar{B} = Y_2L + N(Y_1\bar{K} + X_2\bar{L})\bar{B} - \bar{A}\Theta\bar{B}.$$

Now, define

$$\Theta_1 := \bar{C}\Theta\bar{B} - MZ\bar{B} + (\bar{X}_1 - X_1)L, \qquad \Theta_2 := \bar{A}\Theta\bar{D} + NZ\bar{D} - (\bar{Y}_2 - Y_2)K,$$

where

$$Z := (\bar{Y}_1 - Y_1)\bar{K} + (\bar{X}_2 - X_2)\bar{L}.$$

It is straightforward to verify, using (10) and the equalities $AX_1 + Y_1D = E$, $X_2B + CY_2 = E$, $A\bar{X}_1 + \bar{Y}_1D = E$, $\bar{X}_2B + C\bar{Y}_2 = E$, that (15) holds. Hence, $(\bar{X}_1, \bar{Y}_1, \bar{X}_2, \bar{Y}_2) \in [X_1, Y_1, X_2, Y_2]$ and the map ϕ is one-to-one. □

The significance of Theorem 6 above is that once a characterization of the set of equivalence classes of solutions to $AX + YD = E$ is known, we can obtain a characterization of $\mathscr{S}$ via the map ψ. It is also possible to establish a similar bijective correspondence between $[\mathscr{S}]$ and a set of equivalence classes of solutions to (11) (or to (13)). A realization theoretic approach that yields the set of all solutions to $AX + YD = E$ or its dual $XB + CY = E$ is contained in [3] for the case where **D** is the polynomial ring, and in [8] for the case where **D** is the ring of stable or proper stable rational functions. Another related result is Theorem 4.6 of [7] which yields, indirectly through "skew-complements," a polynomial parameterization of all solutions to $AX + YD = I$. Although the latter result is again for the case where **D** is the ring of polynomials, its modification for the rings of stable and proper stable rational functions is straightforward. In order to clarify the relation between the parameterization of the set of (equivalence classes of) skew-complements and the set of (equivalence classes) of solutions to $AX + YD = I$, we first recall the following definitions from [12] and [17].

DEFINITION. Let $A \in \mathbf{D}^{a\times k}$, $D \in \mathbf{D}^{m\times a}$. Suppose that $k + m - a \geqq 0$. The pair (A, D) is called **skew-prime** over **D** if and only if there exist $\hat{A} \in \mathbf{D}^{m\times(k+m-a)}$, $\hat{D} \in \mathbf{D}^{(k+m-a)\times k}$ such that

$$DA = \hat{A}\hat{D},$$

where $(D, \hat{A})$ is left and $(A, \hat{D})$ is right coprime over **D**. In case one exists, the pair $(\hat{A}, \hat{D})$ is called a **skew-complement** of (A, D).

By a result of [12], it is known that (A, D) is skew-prime if and only if there exists a solution (X, Y) to the equation $AX + YD = I$. The set of skew-complements and the set of solutions to this equation are in bijective correspondence modulo suitable equivalence relations. We now establish this bijection explicitly. We assume below that $AX + YD = I$ is solvable so that $k + m - a \geqq 0$ necessarily holds. In the case where $k + m - a = 0$, it can be shown that $[\mathscr{X}]$ defined below consists of a single element and that $DA = 0$ so that $(0, 0)$ is the unique skew-complement of (A, D). Hence, assume below that $k + m - a > 0$.

On the set of skew-complements $\mathscr{T}$ of (A, D), we define an equivalence relation by

$$(\hat{A}, \hat{D}) \cong (\tilde{A}, \tilde{D}) \quad \text{iff } \sim\hat{A} = \tilde{A}U, \qquad \hat{D} = U^{-1}\tilde{D},$$

for some unimodular U over **D**. Let $[\mathscr{T}]$ denote the set of equivalence classes induced by this equivalence relation. Also let

$$\mathscr{X} := \{(X, Y) : AX + YD = I\}$$

denote the set of solutions to $AX + YD = I$ and define an equivalence relation on $\mathscr{X}$ by

$$(X, Y) \cong (\tilde{X}, \tilde{Y}) \quad \text{iff } X = \tilde{X} + \Theta D, \qquad Y = \tilde{Y} - A\Theta,$$

for some Θ over **D**. With abuse of notation, we let $[\mathscr{X}]$ denote the set of equivalence classes induced by this equivalence relation.

THEOREM 7. *The sets* $[\mathscr{T}]$ *and* $[\mathscr{X}]$ *are in bijective correspondence.*

Proof. Consider a map defined as

$$\phi : [\mathscr{X}] \to [\mathscr{T}] : [X, Y] \mapsto [\hat{A}, \hat{D}],$$

where $\hat{A}$, $\hat{D}$ are such that

$$\begin{bmatrix} Y & -A \\ \hat{Y} & \hat{D} \end{bmatrix} \begin{bmatrix} D & \hat{A} \\ -X & \hat{X} \end{bmatrix} = I \tag{16}$$

for some $\hat{X}$, $\bar{Y}$ over **D**. The proof consists of showing that ϕ is well defined, one-to-one, and onto.

[**Well-defined**]. We first show that given $[X, Y] \in [\mathscr{X}]$, there exist matrices $\hat{A}$, $\hat{D}$, $\hat{X}$, $\hat{Y}$ such that (16) holds. Since $(X, Y) \in \mathscr{X}$, it holds that $AX + YD = I$. Let $\bar{Y}$, $\bar{D}$ be such that

$$\begin{bmatrix} Y & -A \\ \bar{Y} & \bar{D} \end{bmatrix}$$

is a completion of the left unimodular $[Y : -A]$ to a unimodular matrix over **D**. Partitioning its inverse compatibly, we have the equality

$$\begin{bmatrix} Y & -A \\ \bar{Y} & \bar{D} \end{bmatrix} \begin{bmatrix} K & \hat{A} \\ -L & \hat{X} \end{bmatrix} = I.$$

Comparing $AX + YD = I$ and $AL + YK = I$ and noting that the columns of $[\hat{A}' : \hat{X}']'$ are a basis for the kernel of $[Y : -A]$, it follows that $D = K - \hat{A}\Theta$ and $X = L + \hat{X}\Theta$ for some Θ over **D**. Letting $\hat{Y} := \bar{Y} + \Theta Y$, $\hat{D} := \bar{D} - \Theta A$, we obtain (16). We next show that "if $[X, Y] = [\tilde{X}, \tilde{Y}]$, then $\phi([X, Y]) = \phi([\tilde{X}, \tilde{Y}])$." Let $[X, Y] = [\tilde{X}, \tilde{Y}]$ so that $\tilde{X} = X - \Theta D$, $\tilde{Y} = Y + A\Theta$ for some Θ over **D**. If we denote $[\hat{A}, \hat{D}] := \phi([X, Y])$ and $[\tilde{A}, \tilde{D}] := \phi([\tilde{X}, \tilde{Y}])$, then by the definition of ϕ, there exist matrices $\hat{X}$, $\hat{Y}$, M, N such that (16) and

$$\begin{bmatrix} Y + A\Theta & -A \\ N & \tilde{D} \end{bmatrix} \begin{bmatrix} D & \tilde{A} \\ -(X - \Theta D) & M \end{bmatrix} = I$$

hold. By simple row and column operations, it also holds that

$$\begin{bmatrix} Y & -A \\ \hat{N} & \tilde{D} \end{bmatrix} \begin{bmatrix} D & \tilde{A} \\ -X & \hat{M} \end{bmatrix} = I, \tag{17}$$

where $\hat{M} := M - \Theta\tilde{A}$, $\hat{N} := N + \tilde{D}\Theta$. Comparing the equalities $Y\tilde{A} = A\hat{M}$, $Y\hat{A} = A\hat{X}$ obtained from (16) and (17), it follows that $\hat{A} = \tilde{A}U$, $\hat{X} = \hat{M}U$ for some matrix U over **D**. By right coprimeness of $(\hat{A}, \hat{X})$, the matrix U is actually unimodular. Similarly, the equalities $\hat{N}D = \tilde{D}X$, $\hat{Y}D = \hat{D}X$ yield that $\hat{D} = V\tilde{D}$, $\hat{Y} = V\hat{N}$ for a unimodular V over **D**. Employing these and comparing various equalities implied by (16) and (17), it is easy to obtain the following four equalities: $\hat{A}\hat{Y} = \hat{A}(VU)^{-1}\hat{Y}$, $\hat{X}\hat{D} = \hat{X}(VU)^{-1}\hat{D}$, $\hat{A}\hat{D} = \hat{A}(VU)^{-1}\hat{D}$, $\hat{X}\hat{Y} = \hat{X}(VU)^{-1}\hat{Y}$. By left unimodularity of $[\hat{Y} : \hat{D}]$ and by right unimodularity of $[\hat{A}' : \hat{X}']'$, these four equalities imply that $VU = I$. Hence, $\hat{D} = U^{-1}\tilde{D}$, $\hat{A} = \tilde{A}U$, which prove that $[\hat{A}, \hat{D}] = [\tilde{A}, \tilde{D}]$.

[**One-to-one**]. We need to establish that "if $\phi([X, Y]) = \phi([\tilde{X}, \tilde{Y}])$, then $[X, Y] = [\tilde{X}, \tilde{Y}]$." Let us denote, as before, $[\hat{A}, \hat{D}] := \phi([X, Y])$ and $[\tilde{A}, \tilde{D}] := \phi([\tilde{X}, \tilde{Y}])$. If $[\hat{A}, \hat{D}] = [\tilde{A}, \tilde{D}]$, then $\hat{A} = \tilde{A}U$, $\hat{D} = U^{-1}\tilde{D}$ for some unimodular U. By definition of ϕ, there exist matrices $\hat{X}$, $\hat{Y}$, M, N over **D** such that (16) and

$$\begin{bmatrix} \tilde{Y} & -A \\ N & \tilde{D} \end{bmatrix} \begin{bmatrix} D & \tilde{A} \\ -\tilde{X} & M \end{bmatrix} = I$$

hold. Substituting $\tilde{D} = U\hat{D}$, $\tilde{A} = \hat{A}U^{-1}$ in this equality and performing simple unimodular transformations, we obtain

$$\begin{bmatrix} \tilde{Y} & -A \\ \hat{N} & \hat{D} \end{bmatrix} \begin{bmatrix} D & \hat{A} \\ -\tilde{X} & \hat{M} \end{bmatrix} = I, \tag{18}$$

where $\hat{M} := MU$, $\hat{N} := U^{-1}N$. Comparing various equalities following from (16) and (18), it is easy to see that $\tilde{Y} = Y - A\Theta$, $\hat{N} = \hat{Y} + \hat{D}\Theta$, $\tilde{X} = X + \Psi D$, $\hat{M} = \hat{X} - \Psi\hat{A}$ for some Θ and Ψ over $\mathbf{D}$. The following four equalities are also obtained by these last equalities and by comparison of various equalities from (16) and (18): $\hat{D}\Theta D = \hat{D}\Psi D$, $A\Theta D = A\Psi D$, $A\Theta\hat{A} = A\Psi\hat{A}$, $\hat{D}\Theta\hat{A} = \hat{D}\Psi\hat{A}$. By left unimodularity of $[D : \hat{A}]$ and by right unimodularity of $[A' : \hat{D}']'$, it now follows that $\Theta = \Psi$. Hence, $\tilde{X} = X + \Theta D$, $\tilde{Y} = Y - A\Theta$, implying that $[X, Y] = [\tilde{X}, \tilde{Y}]$.

[Onto]. Let $[\hat{A}, \hat{D}] \in [\mathscr{T}]$, so that $DA = \hat{A}\hat{D}$ with $(D, \hat{A})$ left and $(A, \hat{D})$ right coprime. Thus there exist matrices X, Y, $\hat{X}$, $\hat{Y}$ satisfying

$$\begin{bmatrix} D & \hat{A} \\ -X & \hat{X} \end{bmatrix}\begin{bmatrix} Y & -A \\ \hat{Y} & \hat{D} \end{bmatrix} = I.$$

This last equality is identical with (16), and hence, $\phi([X, Y]) = [\hat{A}, \hat{D}]$. □

The results of Theorems 6 and 7 above and Theorem 4.6 of [7] yield a parameterization of all solutions to (1) in the cases where $\mathbf{D}$ is one of the rings of polynomials, stable rational functions, or proper stable rational functions.

5. Comments. There are at least two directions for further research. These concern the following equations over $\mathbf{D}$:

$$\sum_{i=1}^{N} A_i X_i B_i = C, \tag{19}$$

$$\sum_{i=1}^{N} A_i X B_i = C. \tag{20}$$

Considering (19), we note that the cases $N = 1$ and $N = 2$ are the objects of Lemma 1 and Theorem 5, respectively. The simplicity of these results gives some hope of tackling the general case (19) via the equivalence over $\mathbf{D}$ of suitable matrices; although presently no such result is available for $N > 2$. Existing results concerning (20) have been summarized in the excellent survey by Hautus in [6] together with some improvements. However, even in the case $N = 2$, no result that respects the structure of the matrices A_i, B_i, and C is currently available except in a special case where $B_i = q_i(B)$ for some matrix B over $\mathbf{D}$ and polynomials $q_i(x)$ with coefficients in $\mathbf{D}$. This is not quite surprising in view of our characterization results above. The solvability of $AXB + CXD = E$ is equivalent to the existence of an element of the form (X, X) in the solution set $\mathscr{S}$ of (1). Checking the existence of such an element seems to be a formidable task as the set of equivalence classes $[\mathscr{S}_{12}]$, and hence $[\mathscr{S}]$, exhibits a rich and complicated structure (see [7]). There *is*, however, a number of papers containing structural results on the equation $AXB + CXD = E$ in the special case where $\mathbf{D}$ is a field. See, e.g., [11], [5], and the references listed in [5].

Acknowledgments. I am indebted to the referees for pointing out the relevance of [1], [5], and [6] to the work reported here, and especially to one of them, who made me realize my ignorance of the vast literature in the field case.

REFERENCES

[1] J. K. BAKSALARY AND R. KALA, *The matrix equation $AXB + CYD = E$*, Linear Algebra Appl., 30 (1980), pp. 141–147.

[2] L. CHENG AND J. B. PEARSON, JR., *Synthesis of linear multivariable regulators*, IEEE Trans. Automat. Control, 26 (1981), pp. 194–202.

[3] E. EMRE AND L. M. SILVERMAN, *The equation $XR + QY = \Phi$: A characterization of solutions*, SIAM J. Control Optim., 19 (1981), pp. 33–38.

[4] P. A. FUHRMANN, Unpublished notes.

[5] V. HERNÁNDEZ AND M. GASSÓ, *Explicit solution of the matrix equation $AXB - CXD = E$*, Linear Algebra Appl., 121 (1989), pp. 333–344.

[6] M. L. J. HAUTUS, *On the solvability of linear matrix equations*, Memorandum 1982-07, Department of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven, the Netherlands, April 1982.

[7] P. P. KHARGONEKAR, T. T. GEORGIOU, AND A. B. ÖZGÜLER, *Skew-prime polynomial matrices: The polynomial model approach*, Linear Algebra Appl., 50 (1983), pp. 403–435.

[8] P. P. KHARGONEKAR AND A. B. ÖZGÜLER, *System theoretic and algebraic aspects of the rings of stable and stable proper rational functions*, Linear Algebra Appl., 66 (1985), pp. 123–167.

[9] A. B. ÖZGÜLER AND V. ELDEM, *Disturbance decoupling problems via dynamic output feedback*, IEEE Trans. Automat. Control, 30 (1985), pp. 756–766.

[10] L. PERNEBO, *An algebraic theory for design of controllers for linear multivariable systems—Part* II: *Feedback realizations and feedback design*, IEEE Trans. Automat. Control, 26 (1981), pp. 183–193.

[11] W. E. ROTH, *The equations $AX - YB = C$ and $AX - XB = C$ in matrices*, Proc. Amer. Math. Soc., 3 (1952), pp. 392–396.

[12] W. A. WOLOVICH, *Skew-prime polynomial matrices*, IEEE Trans. Automat. Control, 23 (1978), pp. 880–887.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 592–596, July 1991

017

PERTURBATION THEORY OF A NONLINEAR GAME OF VON NEUMANN*

EZIO MARCHI†, JORGE A. OVIEDO†, AND JOEL E. COHEN‡

Abstract. Von Neumann and others considered a two-person zero-sum game with nonlinear payoff function x^TAy/x^TBy, where A and B are $m \times n$ matrices, x^T is the row m-vector strategy of the maximizing player (player 1), and y is the column n-vector strategy of the minimizing player (player 2). This game is defined to be completely mixed if every solution (or optimal strategy) (x, y) is such that all elements of x and all elements of y are positive. In such a game, it is supposed that the matrices A and B are infinitesimally perturbed by matrices of perturbations, i.e., multiple elements of each matrix are perturbed simultaneously. The effect of such perturbations on the solution and value of the game is calculated.

Key words. zero-sum game, two-person game, nonlinear game, perturbation theory, von Neumann model, economic model, stochastic game

AMS(MOS) subject classifications. primary 90D05; secondary 90A16

1. Introduction. This paper develops the perturbation theory of a finite, two-person, zero-sum game with a nonlinear payoff function proposed by von Neumann [13] in a model of economic growth. Subsequent development of the model has been synthesized by Morgenstern and Thompson [11]. The same payoff function appears in a special case of a stochastic game proposed by Shapley [12]. Because the game has more than economic interpretations, we shall not emphasize the economic view of the game nor restrict our assumptions to those that might be plausible in an economic application.

Von Neumann's game has m pure strategies for player 1, the maximizing player, and n pure strategies for player 2, the minimizing player, where $1 \leqq m, n < \infty$. The strategy of player 1 is specified by a row m-vector x^T, where x_i is the probability that player 1 chooses pure strategy i, for $i = 1, \cdots, m$. The strategy of player 2 is specified by a column n-vector y, where y_j is the probability that player 2 chooses pure strategy j, for $j = 1, \cdots, n$. The payoff function of the game, that is, the amount of money player 2 must pay player 1 if player 1 has strategy x^T and player 2 has strategy y, is x^TAy/x^TBy, where A and B are real $m \times n$ matrices. This payoff function is defined (though possibly equal to $+\infty$) provided its numerator and denominator are not simultaneously equal to zero; additional conditions will be provided to assure that the payoff function is always defined. As there does not appear to be a standard name for this game, we shall call it a *rational* game specified by (A, B), because the payoff function is a ratio of linear forms.

Marchi [8] extended and generalized the equilibrium points of a rational game to an n-person game with a rational payoff function. Marchi [7] and Marchi, Tarazaga, and Elorza [9] applied such results to expanding economies.

The perturbation theory of a game describes how small variations in the parameters of the payoff function affect the solution and value of the game. The perturbation theory

* Received by the editors March 29, 1989; accepted for publication (in revised form) December 15, 1989. This work was supported in part by National Science Foundation grants BSR 84-07461 and BSR 87-05047, and by the hospitality of Mr. and Mrs. William T. Golden.

† Instituto de Matemática Aplicada, Universidad Nacional de San Luis, 5700 San Luis, República Argentina (banyc!atina!imasl!oviedo@uunet.uu.net).

‡ Rockefeller University, 1230 York Avenue, Box 20, New York, New York 10021 (cohen@rockvax.rockefeller.edu). This work began during this author's visit to Argentina in 1987, arranged through the Sistema Para el Apoyo a la Investigación y Desarrollo de la Ecología en la República Argentina, and supported by CONICET, Argentina.

of games in general, and of rational games in particular, is of practical interest for both estimation and control. The value of a rational game has an economic interpretation as the asymptotic rate of change (growth or decrease) of an economy. In economic applications of the game, the matrices A (the output matrix) and B (the input matrix) must be estimated from data. The first derivative of the value with respect to the elements of A and B indicates the value's sensitivity to errors in the values of these elements, and therefore indicates which elements should be measured with greatest precision. Kuhn and Tucker [6, p. viii] recognized the importance of perturbation theory for control in their introduction to the work of Mills [10]: "This study promises practical application whenever these parameters [the matrix elements] can be controlled or altered since it indicates which changes will have a beneficial effect on the value."

To our knowledge, the perturbation theory of rational games has not been studied before, except in the linear special case when $B = J_{m,n}$, where $J_{m,n}$ is the $m \times n$ matrix with every element equal to one. In this case, a rational game reduces to an ordinary two-person zero-sum matrix game. Previous studies of the perturbation theory of zero-sum matrix games are reviewed by Cohen [1] and Cohen, Marchi, and Oviedo [3].

We now establish notation and state some results which are mostly standard or readily proved.

Let $P_n = \{x \in R^n\colon x_i \geqq 0,\ i = 1, 2, \cdots, n, \text{ and } \sum_{i=1}^n x_i = 1\}$ and $P_n^+ = \{x \in P_n\colon x_i > 0,\ i = 1, \cdots, n\}$. Vectors are assumed to be column vectors, and the superscript T denotes transpose.

Given two real $m \times n$ matrices A and B we say that a pair of vectors $(\bar{x}^T, \bar{y}) \in P_m \times P_n$ is a *solution of the matricial problem* if

$$(\bar{x}^T A\bar{y})(x^T B\bar{y}) - (x^T A\bar{y})(\bar{x}^T B\bar{y}) \geqq 0 \quad \forall x \in P_m,$$

$$(\bar{x}^T Ay)(\bar{x}^T B\bar{y}) - (\bar{x}^T A\bar{y})(\bar{x}^T By) \geqq 0 \quad \forall y \in P_n.$$

To prove the existence of a solution to the matricial problem, we recall a result of Marchi [8] which is a special case of a theorem of Karamardian [5].

LEMMA. *Consider a real continuous function $\phi\colon \Sigma \times \Sigma \to \Re$ defined on the Cartesian square of Σ, a nonempty, compact, convex set in a Euclidean space. If $\phi(\cdot, \tau)$ is concave for any $\tau \in \Sigma$, then there exists a point $\bar{\sigma} \in \Sigma$ such that*

$$\phi(\bar{\sigma}, \bar{\sigma}) = \max_{\sigma \in \Sigma} \phi(\sigma, \bar{\sigma}).$$

THEOREM 1. *For any real $m \times n$ matrices A and B, a matricial problem has a solution.*

The theorem is easily proved by using the lemma. Essentially this theorem was known to von Neumann [13]. Under the conditions given by von Neumann [13], namely,

$$a_{ij} \geqq 0, \qquad b_{ij} \geqq 0, \qquad a_{ij} + b_{ij} > 0, \qquad i, j = 1, \cdots, n,$$

the solution of the matricial problem determines a solution or saddle point of the rational zero-sum game with payoff function $x^T Ay / x^T By$. In what follows, instead of von Neumann's assumptions we assume $B > 0$, i.e., every element b_{ij} of B is positive real. Shapley [12] considers the same assumption. Then a solution of the matricial problem satisfies

$$\frac{x^T A\bar{y}}{x^T B\bar{y}} \leqq \frac{\bar{x}^T A\bar{y}}{\bar{x}^T B\bar{y}} \leqq \frac{\bar{x}^T Ay}{\bar{x}^T By} \quad \forall x \in P_m, \quad \forall y \in P_n,$$

which is a saddle point of the rational game. Any saddle point determines the value v of

the game $v = (\bar{x}^T A\bar{y})/(\bar{x}^T B\bar{y})$, and furthermore

$$v = \max_{x \in P_m} \min_{y \in P_n} (x^T Ay)/(x^T By) = \min_{y \in P_n} \max_{x \in P_m} (x^T Ay)/(x^T By).$$

Solutions are interchangeable. That is, if $(\bar{x}, \bar{y})$ and $(\tilde{x}, \tilde{y})$ are two saddle points of a rational game, then $(\bar{x}, \tilde{y})$ and $(\tilde{x}, \bar{y})$ are also saddle points. The proof is identical to the proof for ordinary zero-sum matrix games. Optimal strategies for both players may be defined as in matrix games. The set of optimal strategies of each player is a nonempty convex polyhedron.

A rational game is defined to be *completely mixed* (abbreviated "cm") when, for every solution (x, y), $x > 0$ and $y > 0$. When $B = J_{m,n}$, this definition reduces to Kaplansky's [4] definition of a completely mixed matrix game. Completely mixed rational games exist. For example, let $m = n$, $B = J_{n,n}$ and let A be a diagonal matrix with positive elements on the main diagonal. This rational game is an ordinary zero-sum matrix game, and Kaplansky [4] proved that it is cm.

In a rational game (A, B) with $B > 0$ (as we assume throughout), if (x, y) is a solution, then $(Ay)_i/(By)_i < v$ implies $x_i = 0$ and $(x^T A)_j/(x^T B)_j > v$ implies $y_j = 0$. Therefore, in a cm rational game specified by (A, B) with $B > 0$, if (x, y) is a solution, then $(Ay)_i/(By)_i = v$, for all $i = 1, \cdots, m$, and $(x^T A)_j/(x^T B)_j = v$, for all $j = 1, \cdots, n$. Equivalently, $Ay = vBy$ and $x^T A = vx^T B$. If, in addition, $A \neq 0$ and $A \geqq 0$ (i.e., each element a_{ij} of A is nonnegative real), then $v > 0$. The proofs of these remarks are straightforward and are omitted.

Let $\rho(A)$ denote the spectral radius (maximum modulus of the eigenvalues) of an $n \times n$ matrix A. Under certain conditions, there is a direct connection between the value of a cm rational game specified by (A, B) and the spectral radius of $A^{-1}B$.

PROPOSITION 1. *In a* cm *rational game specified by* (A, B) *with* $m = n$ *and* $B > 0$, *if* A *is nonsingular and* $A^{-1}B > 0$, *then* $v = 1/\rho(A^{-1}B) > 0$ *and, for every solution* (x, y), y *is unique and* $x^T B$ *is unique. These are the right and left positive eigenvectors of* $A^{-1}B$ *corresponding to the eigenvalue* $1/v$.

Proof. By Perron's theorem for positive matrices applied to $A^{-1}B$, $A^{-1}B$ has a unique positive right eigenvector in P_n^+. But from previous remarks, $y = vA^{-1}By$. As $y > 0$, $A^{-1}B > 0$, evidently $v > 0$ and $v^{-1}y = A^{-1}By$, so y must be that unique right eigenvector corresponding to the positive eigenvalue v^{-1} and there can exist no other $\eta \in P_n^+$ such that $A\eta = vB\eta$. Similarly, $x^T A(A^{-1}B) = vx^T B(A^{-1}B)$ or $(x^T B)v^{-1} = (x^T B)(A^{-1}B)$. □

This proposition has slightly stronger assumptions and arrives at slightly stronger conclusions than Theorem 5(a) of Cohen and Friedland [2].

While $x^T B$ is unique, under the assumptions of Proposition 1, it is clear that x^T need not be unique.

2. Perturbation theory. Let A, B, G, H be fixed $n \times n$ real matrices, $B > 0$, and for each real number α, define

$$L = L(\alpha) = A + \alpha G, \qquad M = M(\alpha) = B + \alpha H.$$

It is clear that if $B > 0$ and A is nonsingular and $A^{-1}B > 0$, then there exists a real number $r > 0$ such that, for all real α with $|\alpha| < r$, (i) $M(\alpha) > 0$, (ii) $L(\alpha)$ is nonsingular, (iii) $[L(\alpha)]^{-1}M(\alpha) > 0$, and (iv) $\rho([L(\alpha)]^{-1}M(\alpha))$ is analytic in α.

Define a rational game specified by (A, B) to be nonsingular if A and B are both $n \times n$ and both nonsingular.

PROPOSITION 2. *Suppose a nonsingular rational game specified by* (A, B) *is* cm, $B > 0$, *and* $A^{-1}B > 0$. *Then there exists a real number* r *such that if* $|\alpha| < r$, *the rational*

game specified by $(L(\alpha), M(\alpha))$ *is nonsingular and* cm, *the value* $v(\alpha)$ *of that game is given by* $v(\alpha) = 1/\rho([L(\alpha)]^{-1}M(\alpha))$, *and the solution* $(x(\alpha), y(\alpha))$ *of that game is unique.*

In other words, for sufficiently small perturbations (measured by α), under the common assumptions of Propositions 1 and 2, the conclusions of Proposition 1 about the unperturbed rational game specified by (A, B) carry over to the perturbed rational game specified by $(L(\alpha), M(\alpha))$.

Proof. If A and B are nonsingular, then so are sufficiently small perturbations of A and B. Thus, the rational game specified by $(L(\alpha), M(\alpha))$ is nonsingular for small enough values of α. By Proposition 1, the game specified by (A, B) has solutions (x, y) such that y and $z^T = x^TB$ are unique. Because B^{-1} exists, $z^TB^{-1} = x^T$ is also unique. As (x, y) is the solution of a cm rational game, $x > 0$ and $x^T(AB^{-1}) = vx^T$, i.e., x^T is the left eigenvector of AB^{-1} corresponding to the eigenvalue v. Sufficiently small perturbations of $A = L(0)$ and $B = M(0)$ to $L(\alpha)$ and $M(\alpha)$, respectively, will result in a sufficiently small perturbation of v to $v(\alpha)$ such that the corresponding left eigenvector $x^T(\alpha)$ of $L(\alpha)[M(\alpha)]^{-1}$ remains positive and the corresponding right eigenvector $y(\alpha)$ of $[L(\alpha)]^{-1}M(\alpha)$ remains positive. That $(x(\alpha), y(\alpha))$ is unique is guaranteed because $y(\alpha)$ and $z^T(\alpha) = x^T(\alpha)M(\alpha)$ are unique by the Perron theorem and therefore $x^T(\alpha)$ is unique by the nonsingularity of $M(\alpha)$. Thus for small enough α, every solution of $(L(\alpha), M(\alpha))$ is positive, i.e., $(L(\alpha), M(\alpha))$ is cm. Proposition 1 then guarantees that $v(\alpha) = 1/\rho([L(\alpha)]^{-1}M(\alpha))$. □

THEOREM 2. *In a nonsingular* cm *rational game specified by* (A, B) *with* $B > 0$ *and* $A^{-1}B > 0$, *let* A *be perturbed to* $A + \alpha G$ *and* B *be perturbed to* $B + \alpha H$. *Then there exists* $r > 0$ *such that, for* $|\alpha| < r$, $dv(\alpha)/d\alpha$ *exists. Moreover, evaluated at* $\alpha = 0$, *the derivative is*

$$\frac{dv(0)}{d\alpha} = \frac{x^T(G - vH)y}{x^TBy},$$

where (x, y) *and* v *are the solution and value of the original game specified by* (A, B).

Proof. The existence of the derivative follows from Proposition 2 and preceding remarks. Now use the chain rule. If $(x(\alpha), y(\alpha))$ and $v(\alpha)$ are the solution and value of the nonsingular cm rational game specified by $(L(\alpha), M(\alpha))$, then

$$\begin{aligned}
\frac{dv(\alpha)}{d\alpha} &= \frac{d}{d\alpha}\left(\frac{x^T(\alpha)L(\alpha)y(\alpha)}{x^T(\alpha)M(\alpha)y(\alpha)}\right) \\
&= \left\{x^T(\alpha)M(\alpha)y(\alpha)\left[\frac{dx^T(\alpha)}{d\alpha}L(\alpha)y(\alpha) + x^T(\alpha)Gy(\alpha) + x^T(\alpha)L(\alpha)\frac{dy(\alpha)}{d\alpha}\right]\right. \\
&\quad -(x^T(\alpha)L(\alpha)y(\alpha))\left[\frac{dx^T(\alpha)}{d\alpha}M(\alpha)y(\alpha) + x^T(\alpha)Hy(\alpha)\right. \\
&\qquad \left.\left. + x^T(\alpha)M(\alpha)\frac{dy(\alpha)}{d\alpha}\right]\right\} \Big/ (x^T(\alpha)M(\alpha)y(\alpha))^2 \\
&= \left[\frac{dx^T(\alpha)}{d\alpha}(L(\alpha) - v(\alpha)M(\alpha))y(\alpha)\right. \\
&\quad + x^T(\alpha)(G - v(\alpha)H)y(\alpha) \\
&\quad \left. + x^T(\alpha)(L(\alpha) - v(\alpha)M(\alpha))\frac{dy(\alpha)}{d\alpha}\right] \Big/ (x^T(\alpha)M(\alpha)y(\alpha)).
\end{aligned}$$

Because the game specified by $(L(\alpha), M(\alpha))$ is cm and $M(\alpha) > 0$, it follows that $(L(\alpha) - v(\alpha)M(\alpha))y(\alpha) = 0$ and $x^T(\alpha)(L(\alpha) - v(\alpha)M(\alpha)) = 0$. Then taking $\alpha = 0$ gives the claimed formula. □

Under the assumptions of Theorem 2, the derivatives $d^2v(0)/d\alpha^2$, $dx^T(0)/d\alpha$ and $dy(0)/d\alpha$ exist and satisfy

$$\frac{d^2v(0)}{d\alpha^2} = \frac{dx^T(0)}{d\alpha}\left(G - vH - \frac{dv(0)}{d\alpha}B\right)\frac{y}{x^TBy} + \frac{x^T}{x^TBy}\left(G - vH - \frac{dv(0)}{d\alpha}B\right)\frac{dy(0)}{d\alpha} - 2\frac{x^THy}{x^TBy}\frac{dv(0)}{d\alpha},$$

$$\frac{dx^T(0)}{d\alpha}(A - vB) = \left(\frac{dv(0)}{d\alpha}\right)x^TB - x^T(G - vH),$$

$$(A - vB)\frac{dy(0)}{d\alpha} = \left(\frac{dv(0)}{d\alpha}\right)By - (G - vH)y.$$

These formulas follow from applying the chain rule to, respectively, the formula for $dv(0)/d\alpha$ and the identities

$$x^T(\alpha)[L(\alpha) - v(\alpha)M(\alpha)] = 0, \qquad [L(\alpha) - v(\alpha)M(\alpha)]y^T(\alpha) = 0.$$

It is not difficult to verify that when $B = J_{n,n}$ and $H = 0$, the preceding formulas reduce to those found for ordinary zero-sum matrix games by Mills [10] and Cohen [1].

A task for the future is to derive perturbation results similar to the preceding under weaker or different conditions from those assumed in Theorem 2.

Acknowledgments. Several careful referees improved this paper.

REFERENCES

[1] J. E. COHEN, *Perturbation theory of completely mixed matrix games*, Linear Algebra Appl., 79 (1986), pp. 153–162.

[2] J. E. COHEN AND S. FRIEDLAND, *The game-theoretic value and the spectral radius of a nonnegative matrix*, Proc. Amer. Math. Soc., 93 (1985), pp. 205–211.

[3] J. E. COHEN, E. MARCHI, AND J. A. OVIEDO, *Perturbation theory of completely mixed bimatrix games*, Linear Algebra Appl., 114/115 (1989), pp. 169–180.

[4] I. KAPLANSKY, *A contribution to von Neumann's theory of games*, Ann. of Math., 46 (1945), pp. 474–479.

[5] S. KARAMARDIAN, *Generalized complementarity problem*. J. Optim. Theory Appl., 8 (1971), pp. 161–168.

[6] H. W. KUHN AND A. W. TUCKER, EDS., *Preface*, Linear Inequalities and Related Systems, Princeton University Press, Princeton, NJ, 1956.

[7] E. MARCHI, *El modelo de crecimiento de von Neumann para un número arbitrario de países*, Rev. Un. Mat. Argentina, 29 (1979), pp. 85–95.

[8] ———, *Equilibrium points of rational N-person games*, J. Math. Anal. Appl., 54 (1976), pp. 1–4.

[9] E. MARCHI, P. TARAZAGA, AND E. ELORZA, *Further topics in von Neumann growth model*, Portugal. Math., 42 (1983/4), pp. 255–264.

[10] H. D. MILLS, *Marginal values of matrix games and linear programs*, in Linear Inequalities and Related Systems, H. W. Kuhn and A. W. Tucker, eds., Princeton University Press, Princeton, NJ, 1956, pp. 183–193.

[11] O. MORGENSTERN AND G. L. THOMPSON, *Mathematical Theory of Expanding and Contracting Economies*, D. C. Heath, Lexington, MA, 1976.

[12] L. S. SHAPLEY, *Stochastic games*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), pp. 1095–1100.

[13] J. VON NEUMANN, *Über ein ökonomisches Gleichungssystem und eine Verallgemeinerung des Brouwerschen Fixpunktsatzes*, Ergebnisse eines Mathematischen Kolloquiums, 8 (1935/6), pp. 73–83. Translated by G. Morgenstern in Review of Economic Studies, 13 (1945/6), pp. 1–9. Reprint of English translation in Collected Works 6, pp. 29–37, Macmillan, New York, 1963.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 3, pp. 597–609, July 1991

018

EIGENVALUES OF THE LAPLACIAN THROUGH BOUNDARY INTEGRAL EQUATIONS*

YA YAN LU†‡ AND SHING-TUNG YAU†

Abstract. A numerical method for the eigenvalue problem of the Laplacian in two-dimensional domains is developed in this paper. This method requires $O(N)$ operations for calculating one eigenvalue in each iteration step, where N is the number of boundary points in the discretization. It is based on the boundary integral formulation which reduces the computation of the eigenvalues to the zeros of the function $\mu_1(\lambda)$ defined as the smallest eigenvalue of a related matrix. Iteration methods such as the Lanczos method are used to compute $\mu_1(\lambda)$, which requires the multiplication of an $N \times N$ matrix with a vector. The multipole expansion techniques developed for the potential problems by Rokhlin [*J. Comput. Phys.*, 60 (1985), pp. 187–207] are applied and extended here, and the number of operations is reduced to $O(N)$ for this multiplication. The zeros of $\mu_1(\lambda)$ are found by the method of quadratic interpolation. A method for finding the kth eigenvalue with the value of k prespecified is also presented. It is based on continuously tracing the eigenvalue while the domain is deforming to (or from) the unit disk. Only five values of μ_1 are required in this tracing process.

Key words. eigenvalue problem, Laplace operator, boundary integral equation

AMS(MOS) subject classifications. 65N25, 65R20

1. Introduction.

The eigenvalue problem for the Laplacian arises in many physical problems, such as vibration theory, acoustic scattering, and harbor oscillations. It has been extensively studied both theoretically and numerically [7]. In this paper, we develop a numerical method based on the boundary integral formulation for any two-dimensional domain with a smooth boundary.

Traditionally, numerical methods, such as finite difference and finite element methods, are based on the discretization of the interior of the domain. The number of nodes in the whole domain $\tilde{N}$ is proportional to the square of the number of nodes N on the boundary. The resulting matrix problem involves $\tilde{N} \times \tilde{N}$ sparse matrices. For large $\tilde{N}$, iteration methods, such as the Lanczos method, are usually used. A few extreme eigenvalues can be obtained in a relatively small number of iterations. However, if more eigenvalues or a higher eigenvalue are desired, not only could the number of iterations be large, but the reorthogonalization process in each iteration also becomes expensive. In each iteration step, a multiplication of the sparse $\tilde{N} \times \tilde{N}$ matrix with a vector is necessary, which requires $O(\tilde{N})$ operations.

The method in this paper is based on the boundary integral formulation of the eigenvalue problem for the Laplacian. When the boundary is discretized to N nodes, the integral equation is approximately reduced to a matrix problem with an $N \times N$ matrix A. The entries of A are functions of λ and the matrix becomes singular when λ is an eigenvalue of the Laplacian. The approach suggested by Hutchinson [6] is to evaluate the determinant of A at different values of λ and locate the value of λ where the determinant is zero. Since A is not sparse, the calculation of its determinant requires $O(N^3)$ operations. Therefore, this method is not efficient compared with the domain methods.

In this paper, instead of computing the determinant of A, we compute the smallest eigenvalue $\mu_1(\lambda)$ of the matrix $A^T A$. When λ is an eigenvalue of the Laplacian, $\mu_1(\lambda)$

* Received by the editors December 13, 1989; accepted for publication (in revised form) May 15, 1990.

† Department of Mathematics, Harvard University, Cambridge, Massachusetts 02138.

‡ The research of this author was supported by National Science Foundation/Defense Advanced Research Projects Agency grant DMS 87-11394.

becomes zero when the matrix A becomes singular. Iteration methods such as the Lanczos method [8], [3] can be used to compute the eigenvalues of the symmetric matrix A^TA. The smallest eigenvalue $\mu_1(\lambda)$ emerges in just a few iterations. In each iteration, a matrix vector multiplication involving the matrix A^TA is necessary. Furthermore, the number of operations for the matrix vector action can be reduced to $O(N)$ when we carefully apply and extend the multipole expansion techniques developed by Rokhlin [9] for the potential problem (see also [4]).

Rokhlin [9] studied the boundary value problem for the Laplace equation. When the boundary integral formulation is used, the discretized problem is a matrix problem with an $N \times N$ nonsparse matrix. Generalized conjugate gradient [5] methods can be used with a matrix vector multiplication in each step. Using careful expansions at many locations, the structure of the matrix is fully utilized and the number of operations required for the matrix vector multiplication are reduced to $O(N)$. The efficiency of this method is most evident when we compute the solution at only a few points in the interior of the domain.

The case that we will study in this paper is more difficult because the kernel in the integral formulation of the Helmholtz equation is more complicated. It involves Bessel functions and the parameter λ. We overcome this difficulty by using the polynomial approximation of the Frobenius expansions for small arguments of the Bessel functions, and Taylor expansions for large arguments. However, we should point out that the $O(N)$ operations for this matrix vector multiplication involves a constant which depends on the accuracy that we need to have (and it is usually quite large), and one should use a direct matrix vector multiplication when N is less than several thousands.

When the kth eigenvalue of the Laplacian is desired with a prespecified value of k (which could be large), we present a method based on tracing of the kth eigenvalue while the domain is continuously deforming from the unit disk to the original domain that we are studying. For such a tracing to be practical, we use a five point scheme, which keeps the minimum information for where the kth eigenvalue is located and for whether or not there are any other eigenvalues nearby.

2. Boundary integral formulation. We study the eigenvalue problem of the Laplace operator in a two-dimensional domain Ω,

$$\Delta\phi + \lambda\phi = 0 \quad \text{in} \quad \Omega,$$

with Dirichlet boundary condition $\phi = 0$ or Neumann boundary condition $\partial\phi/\partial\nu = 0$ on $\partial\Omega$.

This problem can be reduced to a one-dimensional nonlinear eigenvalue problem for an integral equation on the boundary. For the two-dimensional Helmholtz equation, the related Green's function satisfying the equation

$$\nabla^2 g + \lambda g = \delta(x - x')$$

and the Sommerfeld radiation condition at infinity is taken as

$$g(x, x'; \lambda) = \frac{1}{4} Y_0(\sqrt{\lambda}|x - x'|) + \frac{i}{4} J_0(\sqrt{\lambda}|x - x'|),$$

where Y_0 and J_0 are zeroth order Bessel functions.

For the Neumann boundary condition, we apply Green's formula, to arrive at the equations

$$\phi(x') - \int_{\partial\Omega} \phi(x) \frac{\partial g(x,x';\lambda)}{\partial \nu(x)} ds(x) = 0 \quad \text{for} \quad x' \in \Omega, \tag{1}$$

$$\frac{1}{2}\phi(x') - \int_{\partial\Omega} \phi(x) \frac{\partial g(x,x';\lambda)}{\partial \nu(x)} ds(x) = 0 \quad \text{for} \quad x' \in \partial\Omega, \tag{2}$$

where $\partial\Omega$ is assumed to be smooth and $\nu(x)$ is the normal vector of $\partial\Omega$ at x. Similarly, for a Dirichlet boundary condition, we have the following two equations:

$$\phi(x') + \int_{\partial\Omega} g(x,x';\lambda) \frac{\partial \phi(x)}{\partial \nu(x)} ds(x) = 0 \quad \text{for} \quad x' \in \Omega, \tag{3}$$

$$\frac{1}{2}\frac{\partial \phi(x')}{\partial \nu(x')} + \int_{\partial\Omega} \frac{\partial g(x,x';\lambda)}{\partial \nu(x')} \frac{\partial \phi(x)}{\partial \nu(x)} ds(x) = 0 \quad \text{for} \quad x' \in \partial\Omega. \tag{4}$$

Note that in the above boundary integral equations, the kernels $\partial g(x,x';\lambda)/\partial\nu(x)$ and $\partial g(x,x';\lambda)/\partial\nu(x')$ are smooth for all values of $x, x' \in \partial\Omega$. In fact, the kernel of the Neumann problem is

$$\begin{aligned} K(x,x',\lambda) &= \frac{\partial g(x,x';\lambda)}{\partial \nu(x)} = \nu(x).\nabla g(x,x';\lambda) \\ &= \nu(x).\nabla \left(\frac{1}{4} Y_0(\sqrt{\lambda}|x-x'|) + \frac{i}{4} J_0(\sqrt{\lambda}|x-x'|) \right) \\ &= -\frac{\sqrt{\lambda}}{4} (Y_1(\sqrt{\lambda}|x-x'|) + iJ_1(\sqrt{\lambda}|x-x'|)) \nu(x).\frac{x-x'}{|x-x'|}. \end{aligned}$$

Although $Y_1 \to \infty$ as $x \to x'$, the angle between $\nu(x)$ and $x - x'$ tends to $\pi/2$ in this limit. Using the asymptotic formula for $Y_1(x)$ as $x \to 0$, we can prove that as $x' \to x$,

$$K(x,x',\lambda) = -\frac{1}{4\pi}\kappa(x)$$

where $\kappa(x)$ is the curvature of $\partial\Omega$ at x. The kernel for the Dirichlet problem is

$$\frac{\partial g(x,x';\lambda)}{\partial \nu(x')} = \frac{\partial g(x',x;\lambda)}{\partial \nu(x')} = K(x',x,\lambda)$$

since the Green's function $g(x,x';\lambda)$ is symmetric with respect to x and x'.

The boundary integral equations (2) and (4) are eigenvalue problems in λ, since $\phi(x)$ and $\partial\phi(x)/\partial\nu$ can be nontrivial solutions only when λ is an eigenvalue of the original problem. However, the difficulty related to this problem is that λ is involved transcendentally in the kernel $K(x,x',\lambda)$. Assuming that the boundary is given by $x_1, x_2, \cdots, x_N$ and the corresponding arclength for the node x_i is ds_i, (2) and (4) are approximated by

$$\frac{1}{2}\phi(x_i) - \sum_{j=1}^{N} K(x_j,x_i,\lambda)\phi(x_j)ds_j = 0, \tag{5}$$

$$\frac{1}{2}\frac{\partial \phi}{\partial \nu}(x_i) + \sum_{j=1}^{N} K(x_i,x_j,\lambda)\frac{\partial \phi}{\partial \nu}(x_j)ds_j = 0, \tag{6}$$

respectively. For simplicity, the arclengths for different nodes are assumed to be the same, i.e., $ds_1 = ds_2 = \cdots = ds_n = ds$. The above conditions are further reduced to

$$[I - 2ds(K(x_j, x_i, \lambda))] \begin{pmatrix} \phi(x_1) \\ \phi(x_2) \\ \vdots \\ \phi(x_N) \end{pmatrix} = 0,$$

$$[I + 2ds(K(x_i, x_j, \lambda))] \begin{pmatrix} \phi_\nu(x_1) \\ \phi_\nu(x_2) \\ \vdots \\ \phi_\nu(x_N) \end{pmatrix} = 0,$$

where I is the $N \times N$ unit matrix and $(K(x_i, x_j, \lambda))$ denotes the $N \times N$ matrix whose (i, j) entry is $K(x_i, x_j, \lambda)$. The condition that λ be an eigenvalue is approximated by the existence of a nontrivial solution to the above equations, i.e.,

$$\det[I - 2ds(K(x_j, x_i, \lambda))] = 0, \tag{7}$$

$$\det[I + 2ds(K(x_i, x_j, \lambda))] = 0, \tag{8}$$

for Dirichlet and Neumann boundary conditions, respectively.

One can try to locate the eigenvalues by finding the zeros of the above determinants as functions of λ. However, the evaluation of the determinants above requires $O(N^3)$ operations, which makes this approach inefficient. Nevertheless, the condition that the coefficient matrix be singular is equivalent to the existence of a zero eigenvalue. Therefore, we are led to consider the smallest eigenvalue of

$$A_D = [I + 2ds(K(x_i, x_j, \lambda))][I + 2ds(K(x_i, x_j, \lambda))^H], \tag{9}$$

$$A_N = [I - 2ds(K(x_j, x_i, \lambda))][I - 2ds(K(x_j, x_i, \lambda))^H], \tag{10}$$

for Dirichlet and Neumann boundary conditions, respectively. With the Lanczos method, the smallest eigenvalue of A_D or A_N can be obtained in relatively few iterations. The multiplication of matrix A_D or A_N with a vector is necessary in each iteration step, which essentially involves the evaluation of

$$(K(x_i, x_j, \lambda)) \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{pmatrix} \quad \text{and} \quad (K(x_i, x_j, \lambda))^H \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{pmatrix}$$

for arbitrary constants $c_1, c_2, \cdots, c_N$. The direct calculation of the above matrix vector multiplication requires $O(N^2)$ operations. Since the number of iterations is much less than N in the Lanczos method, this method is already an improvement over that of Hutchinson [6]. Furthermore, we will demonstrate that it is possible to use the idea of multipole expansions to reduce the number of operations required in this multiplication to only $O(N)$. We first illustrate the basic idea of the multipole expansion method in a simple integral equation which is of interest by itself.

3. Integral equations. Consider the following integral equation:

$$y(x) + \int_0^1 K(x - x')y(x')dx' = f(x),$$

where the kernel K can be written in the special form $K(x-x')$. When the interval $[0,1]$ is discretized to n nodes, $x_1, x_2, \cdots, x_n$, the integral equation is approximated by a set of linear equations

$$y(x_i) + \sum_{j=1}^{n} w_j K(x_i - x_j) y(x_j) = f(x_i)$$

for $i = 1, 2, \cdots, n$. If we choose equispaced nodes, then $x_i - x_j = (i-j)h$, and the resulting matrix is a Toeplitz matrix. The direct method for solving this linear system needs $O(n^2)$ operations because of the special structure.

An iterative method is also possible, which takes $O(n^2)$ operations in each step, as it is necessary to evaluate a matrix vector multiplication which is equivalent to evaluating

$$g_i = \sum_{j=1}^{n} c_j K(x_i - x_j)$$

for some constants $c_1, c_2, \cdots, c_n$. However, when the kernel satisfies a certain condition, the idea of multipole expansion can be applied to this problem, which results in a method with $O(n)$ operations in each iteration step.

Consider the case where the function $K(t)$ for $-1 < t < 1$ satisfies the following uniform convergence radius condition. For any t_0 in this region, the Taylor expansion

$$K(t) = K(t_0) + K'(t_0)(t-t_0) + \frac{K''(t_0)}{2}(t-t_0)^2 + \cdots$$

has a radius of convergence larger than a certain constant r_0. We can choose $r_0 > 0$, such that whenever $|t - t_0| < r_0$, the value of the M term truncation of the above Taylor expansion gives rise to accuracy ϵ. This latter condition is related to the uniform boundedness of the Mth derivative. Then we can cut the interval of $[0,1]$ into several pieces of length $2r_0$, say $I_1, I_2, I_3, \cdots, I_P$. The value P is approximately $1/(2r_0)$. In these subintervals, we have the center points $C_1, C_2, \cdots, C_P$. Then

$$\begin{aligned} g_i &= \sum_{p=1}^{P} \sum_{x_j \in I_p} c_j K(x_i - C_p + C_p - x_j) \\ &\simeq \sum_{p=1}^{P} \sum_{m=0}^{M-1} \frac{K^{(m)}(x_i - C_p)}{m!} \sum_{x_j \in I_p} c_j (C_p - x_j)^m. \end{aligned}$$

Therefore, we first calculate

$$\sum_{x_j \in I_p} c_j (C_p - x_j)^m$$

for $p = 1, 2, \cdots, P$ and $m = 0, 1, \cdots, M-1$. All these calculations can be done in $O(n)$ operations when P, M are $O(1)$ constants. After this step, the values of g_i can be obtained in $O(1)$ operations. The total number of operations for computing all $g_1, g_2, \cdots, g_n$ is still $O(n)$. This method can be efficient when there exist a large r_0 and a small M such that $M/(2r_0)$ is much less than n. The number of operations is proportional to $O(M/(2r_0)n)$.

Of course, this procedure is based on the existence of a uniform radius of convergence for the Taylor expansion of the kernel K. In the case of the potential problem studied by Rokhlin and the problem in this paper, the radius of such a Taylor expansion may depend on the ratio of $|t-t_0|/|t_0|$. When t_0 is close to zero, the radius of convergence becomes small. If such a singularity exists, we need to treat the problem more carefully.

4. Matrix vector multiplication.

4.1. Basic notations. In this section, we develop an $O(N)$ algorithm for the multiplication of the related matrix with a vector. For simplicity, we take only the real part of the kernel into account. In this case, the values of λ such that the matrix is singular correspond to the eigenvalues of the Laplacian for the domain Ω and those for the exterior of Ω. We introduce complex notation for the boundary points $z_1, z_2, \cdots, z_N$. The boundary $\partial\Omega$ is assumed to be smooth with total length L, and these boundary points are equispaced with distance $ds = L/N$ between any two nearby points. Without loss of generality, we assume $N = 2^n$ for an integer n and introduce the following partition for these N boundary points:

$$\{z_1, z_2, \cdots, z_N\} = \bigcup_{k=1}^{2^{n-m}} A_k^{(m)},$$

where $A_1^{(m)}$ is the first $M = 2^m$ boundary points, $A_2^{(m)}$ is the second M points, etc. In general,

$$A_k^{(m)} = \{z_{(k-1)M+1}, z_{(k-1)M+2}, \cdots, z_{kM}\}.$$

We also define the center of $A_k^{(m)}$ as

$$c_k^{(m)} = \tfrac{1}{2}[z_{(k-1)M+M/2} + z_{(k-1)M+M/2+1}].$$

We define $W_k^{(m)}$ to be the union of all those $A_l^{(m)}$ which are at least Mds away from the center of $A_k^{(m)}$, i.e.,

$$W_k^{(m)} = \{z_j \in A_l^{(m)} \text{ for some } l, \text{ such that } \operatorname{dist}(A_l^{(m)}, c_k^{(m)}) \geq Mds\}$$

where the distance between a point and a set is defined as the minimum of the distance between the point and any point in the set.

Since only the real parts of the kernels $K(x_i, x_j, \lambda)$ are used, the matrix vector multiplication steps for both the Dirichlet and the Neumann problems require the evaluations of

$$G_i^1 = -\sum_{j\neq i} \frac{\sqrt{\lambda}c_j}{4} Y_1(\sqrt{\lambda}|x_i - x_j|)\nu(x_i) \cdot \frac{x_i - x_j}{|x_i - x_j|}, \qquad i = 1, 2, \cdots, N$$

and

$$G_i^2 = -\sum_{j\neq i} \frac{\sqrt{\lambda}c_j}{4} Y_1(\sqrt{\lambda}|x_i - x_j|)\nu(x_j) \cdot \frac{x_j - x_i}{|x_i - x_j|}, \qquad i = 1, 2, \cdots, N$$

for any given real numbers $c_1, c_2, \cdots, c_N$.

For G_i^1, using complex notation for the boundary points, we first compute

$$g_i = \sum_{j \neq i} \frac{c_j}{z_i - z_j} \sqrt{\lambda}|z_i - z_j| Y_1(\sqrt{\lambda}|z_i - z_j|), \qquad i = 1, 2, \cdots, N, \tag{11}$$

then take the dot product of $\nu(x_i)$ with the vector $(Re(g_i), -Im(g_i))$. The values of G_i^1 are thus obtained. For G_i^2, the two components of the vectors $c_j \nu(x_j)$ are used to define the real and imaginary parts of the complex numbers $\tilde{c}_j$; then we compute g_i as in (11) with c_j replaced by $\tilde{c}_j$. G_i^2 is the real part of g_i. Therefore, the essential part of the matrix vector multiplication is reduced to computing $g_1, g_2, \cdots, g_N$ defined in (11) for an arbitary set of complex numbers $c_1, c_2, \cdots, c_N$.

4.2. Near field expansion. We first study the expansion for

$$\varphi_k^{(m,l)}(z) = \sum_{z_j \in A_l^{(m)}} \frac{c_j}{z - z_j} \sqrt{\lambda}|z - z_j| Y_1(\sqrt{\lambda}|z - z_j|),$$

where $z \in A_k^{(m)}$ and $A_l^{(m)}$ is contained in $W_k^{(m)}$ such that $\sqrt{\lambda}|z - z_j| \leq 3$ for all $z_j \in A_l^{(m)}$ and $z \in A_k^{(m)}$. We seek the following expansion:

$$\varphi_k^{(m,l)}(z) = \sum_{t,s \geq 0} a_{tsk}^{(m,l)} (z - c_k^{(m)})^t (\overline{z} - \overline{c}_k^{(m)})^s.$$

For $0 < x < 3$, we have the approximate expansion

$$xY_1(x) \simeq \sum_{i=1}^{7} \left[\alpha_i x^{2(i-1)} + \beta_i x^{2i} \ln \frac{x}{2} \right]$$

valid to 10^{-8} in this range. The coefficients α_i, β_i are given in the following table:

i	α_i	β_i
1	-0.6366198	0.3183099
2	2.4578789×10^{-2}	$-3.9788724 \times 10^{-2}$
3	2.6768777×10^{-2}	1.6578500×10^{-3}
4	$-1.8058748 \times 10^{-3}$	$-3.4531940 \times 10^{-5}$
5	4.7613944×10^{-5}	4.3015641×10^{-7}
6	$-6.7905638 \times 10^{-7}$	$-3.4242205 \times 10^{-9}$
7	5.2447967×10^{-9}	$1.3284848 \times 10^{-11}$

After some calculation, we end up with

$$\varphi_k^{(m,l)}(z) = \sum_{t,s \geq 0} e_{tsk}^{(m)} \sum_{z_j \in A_l^{(m)}} c_j \theta^{t-1} \overline{\theta}^{s-1},$$

where $\theta = [z_j - c_l^{(m)} - (z - c_k^{(m)})]/w$, $w = c_k^{(m)} - c_l^{(m)}$, and

$$e_{tsk}^{(m)} = \sum_{i=0}^{7} \left[\alpha'_{its} + \beta'_{its} \ln \frac{\sqrt{\lambda}|w|}{2} \right] \lambda^i w^{i-1} \overline{w}^i.$$

The coefficients $\alpha'_{its}, \beta'_{its}$ are given by the following formula:

$$\alpha'_{0t1} = \alpha_0,$$

$$\alpha'_{its} = (-1)^{t+s}\binom{i-1}{t-1}\binom{i}{s-1}\alpha_i - \frac{1}{2}\binom{i}{s-1}\sum_{p=1}^{\min(i,t-1)}\beta_i\frac{(-1)^{p+s}}{t-p}\binom{i-1}{p-1}$$
$$-\frac{1}{2}\binom{i-1}{t-1}\sum_{q=1}^{\min(i+1,s-1)}\beta_i\frac{(-1)^{q+t}}{s-q}\binom{i}{q-1} \quad \text{for} \quad i \geq 1,$$

$$\beta'_{its} = (-1)^{t+s}\beta_i\binom{i-1}{t-1}\binom{i}{s-1} \quad \text{for} \quad i \geq 1.$$

The coefficients $a^{(m,l)}_{tsk}$ are obtained when the binomial expansion of $\theta, \overline{\theta}$ are used.

4.3. Far field expansion. In the case when $\sqrt{\lambda}|z-z_j| > 2$, we use direct Taylor expansion for

$$f(z, z_j) = \frac{1}{z - z_j}\sqrt{\lambda}|z - z_j|Y_1(\sqrt{\lambda}|z - z_j|).$$

Introducing the function

$$F(u, v) = \sqrt{\frac{v}{u}}Y_1(\sqrt{uv}),$$

we find that the derivatives of F are given by

$$\frac{\partial^{t+s}F}{\partial u^t \partial v^s} = \sum_{i=0}^{t+s}\gamma_i^{(t,s)}v^{(i+1-2s)/2}u^{(i-1-2t)/2}Y_1^{(i)}(\sqrt{uv}),$$

where the coefficients satisfy the following recurrence relations:

$$\gamma_0^{(0,0)} = 1,$$

$$\gamma_i^{(t+1,s)} = \begin{cases} -\frac{1+2t}{2}\gamma_0^{(t,s)} & i = 0, \\ -\frac{1+2t-i}{2}\gamma_i^{(t,s)} + \frac{1}{2}\gamma_{i-1}^{(t,s)} & 0 < i < t+s+1, \\ \frac{1}{2}\gamma_{t+s}^{(t,s)} & i = t+s+1, \end{cases}$$

$$\gamma_i^{(t,s+1)} = \begin{cases} \frac{1-2s}{2}\gamma_0^{(t,s)} & i = 0, \\ \frac{1-2s+i}{2}\gamma_i^{(t,s)} + \frac{1}{2}\gamma_{i-1}^{(t,s)} & 0 < i < t+s+1, \\ \frac{1}{2}\gamma_{t+s}^{(t,s)} & i = t+s+1. \end{cases}$$

The function $f(z, z_j)$ is then expanded as follows:

$$f(z, z_j) = \sum_{t,s\geq 0}\frac{\sqrt{\lambda}^{t+s+1}}{t!s!}\frac{\partial^{t+s}F}{\partial u^t \partial v^s}\Big|_{u=\sqrt{\lambda}(c_k^{(m)}-z_j), v=\overline{u}}(z - c_k^{(m)})^t(\overline{z} - \overline{c}_k^{(m)})^s.$$

It is clear that $a^{(m)}_{tsk}$, defined before, is the summation of the coefficients in the above equation over all z_j in $A_l^{(m)}$.

4.4. Multiplication steps. After we have obtained the expansion coefficients for $\varphi_k^{(m,l)}$ for the near field and the far field, we are ready to consider the general steps for the evaluation of $g_1, g_2, \cdots, g_N$.

We first define m_0 to be the largest integer satisfying the condition

$$m_0 \le n - \log(\sqrt{\lambda} L).$$

For $m = m_0$, we calculate the expansion for

$$\varphi_k^{(m_0)}(z) = \sum_{z_j \in W_k^{(m_0)}} \frac{c_j}{z - z_j} \sqrt{\lambda} |z - z_j| Y_1(\sqrt{\lambda}|z - z_j|)$$

with $z \in A_k^{(m_0)}$. That is,

$$\varphi_k^{(m_0)}(z) = \sum_{t,s \ge 0} a_{tsk}^{(m_0)} (z - c_k^{(m_0)})^t (\overline{z} - \overline{c}_k^{(m_0)})^s.$$

The contribution $a_{tsk}^{(m_0,l)}$ to the coefficients $a_{tsk}^{(m_0)}$ by $A_l^{(m_0)}$ is calculated through the formula in the previous two subsections. Depending on whether $A_l^{(m_0)}$ is a near field or far field of $A_k^{(m_0)}$, we use the different approaches studied before. The fact that m_0 satisfies the above inequality guarantees that $A_l^{(m_0)}$ can always be put into one of the two cases.

When we have the level m expansions for $\varphi_k^{(m)}(z)$, the level $m-1$ expansion can be obtained combining the level m results and supplementing a few level $m-1$ near field expansions. More precisely, we have

$$W_{2k-1}^{(m-1)} = \bigcup_{l=l_1,l_2,\cdots} A_l^{(m-1)} \bigcup W_k^{(m)}, \tag{12}$$

$$W_{2k}^{(m-1)} = \bigcup_{l=l_1',l_2',\cdots} A_l^{(m-1)} \bigcup W_k^{(m)}. \tag{13}$$

The values of $l_1, l_1', l_2, l_2', \cdots$ reflect the geometry of the boundary and must be checked each time. In most cases, we only need the following: $l_1 = 2k-3, l_2 = 2k+1, l_3 = 2k+2, l_1' = 2k-3, l_2' = 2k-2, l_3' = 2k+2$. Therefore, we have,

$$\begin{aligned}
\varphi_{2k-1}^{(m-1)}(z) &= \varphi_k^{(m)}(z) + \sum_{l=l_1,l_2,\cdots} \varphi_k^{(m-1,l)}(z) \\
&= \sum_{t,s \ge 0} a_{tsk}^{(m)} (z - c_k^{(m)})^t (\overline{z} - \overline{c}_k^{(m)})^s \\
&\quad + \sum_{l=l_1,l_2,\cdots} \sum_{t,s \ge 0} a_{ts,2k-1}^{(m-1,l)} (z - c_{2k-1}^{(m-1)})^t (\overline{z} - \overline{c}_{2k-1}^{(m-1)})^s.
\end{aligned}$$

The first term in the above equation can be put in the new form, that is,

$$\sum_{t,s \ge 0} \left\{ \sum_{p \ge t, q \ge s} \binom{p}{t} \binom{q}{s} (c_{2k-1}^{(m-1)} - c_k^{(m)})^{p-t} (\overline{c}_{2k-1}^{(m-1)} - \overline{c}_k^{(m)})^{q-s} \right\} \times (z - c_{2k-1}^{(m-1)})^t (\overline{z} - \overline{c}_{2k-1}^{(m-1)})^s.$$

After we reach the $m = 1$ level, the problem is almost done, since $W_k^{(1)}$ is almost all these N boundary points. We simply evaluate it at $z = z_{2k-1}, z_{2k}$ and add a few nearby terms which are not included in $W_k^{(1)}$. The values of $g_1, g_2, \cdots, g_N$ are thus obtained.

5. Zeros of $\mu_1(\lambda)$. In the previous section, we presented the details for the multiplication of the related $N \times N$ matrix $(K(x_i, x_j, \lambda))$ with an arbitrary vector. The whole process can be carried out in about kN operations, where k is a constant related to the finite accuracy that we need to attain. Usually, for a few digits precision, the constant k is estimated to be around several thousand. Therefore, when the number of points on the boundary for a discretization is not too large, we should use the direct matrix vector multiplication. The method of the previous section for an $O(N)$ multiplication is only useful when we are interested in some very large eigenvalues of the domain.

This matrix vector multiplication is used in the Lanczos method (or other iteration methods) to determine the smallest eigenvalue $\mu_1(\lambda)$ of A_D or A_N as a function of λ, depending on the type of the boundary conditions. The eigenvalues of the Laplacian correspond to the zeros of $\mu_1(\lambda)$. Since the matrix A_D and A_N are the products of $I \pm 2ds(K(x_i, x_j, \lambda))$ with their complex transposes, the zeros of $\mu_1(\lambda)$, which are the same as the values of λ such that $\det(A_D)$, $\det(A_N)$ are zero, must be at least double zeros. In fact, if λ_i is an eigenvalue of multiplicity m, it must be a zero of $\mu_1(\lambda)$ of multiplicity $2m$. One possible way [10] to calculate a zero of $\mu_1(\lambda)$, say λ_i, is based on Newton's method with a transformation that changes $\mu_1(\lambda)$ to a new function $T(\lambda)$ such that λ_i is a single zero of it. The trouble with this method is that we need more than one evaluation of the function μ_1 in each step, and these values are expensive to obtain. Therefore, we turn to the quadratic interpolation method, in which, for three known points at $\lambda_1, \lambda_2, \lambda_3$, we seek λ_* such that $(\lambda_*, \mu_1(\lambda_*))$ and $(\lambda_j, \mu_1(\lambda_j))$, $j = 1, 2, 3$, are on the same parabola with a minimum located at λ_*. The new value of λ_* is then used with the previous calculated values for a further iteration. In principle, the final point we come to should have a zero μ_1 value, and the generated series converges to a zero of μ_1 quadratically.

A typical graph for $\mu_1(\lambda)$ is shown in Fig. 1. When only the real part of the matrix is used, both the eigenvalues of the domain Ω and those of the exterior of Ω are zeros of $\mu_1(\lambda)$. The convergence of the zeros of $\mu_1(\lambda)$ towards the eigenvalues of the Laplacian, when N tends to infinity, is illustrated in the following table for the first eigenvalue of the unit disk.

N	λ_1
10	5.8
20	5.79
45	5.784
90	5.7833
180	5.78319

The exact value for λ_1 is $5.7831859629\cdots$. It should be noted that for a fixed value of N, the error is larger for larger eigenvalues, as expected.

6. The kth eigenvalue. The method that we presented in the previous sections allows us to find the value of an eigenvalue near some initial starting point that we provide. Very often, in many practical applications, we would like to know which

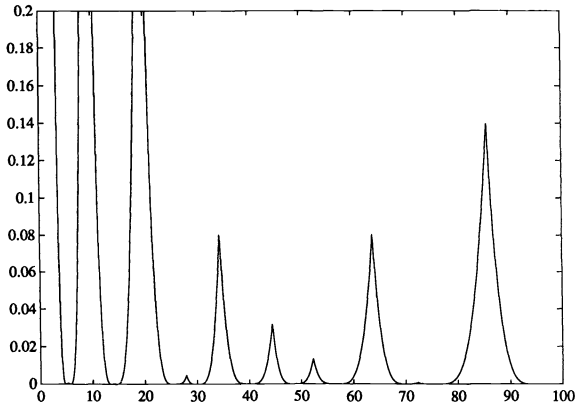


FIG. 1. $\mu_1(\lambda)$ *versus* λ *for the unit disk. Zeros of the function* μ_1 *correspond to the eigenvalues of the Laplacian for the unit disk or the domain outside the unit disk subject to Dirichlet boundary conditions.*

eigenvalue it is according to the ordering from smaller ones to larger ones:

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k \leq \cdots.$$

In other words, we know that λ_* is an eigenvalue of the Laplacian in a two-dimensional domain with an appropriate boundary condition, and we want to find the value of k, such that $\lambda_k = \lambda_*$. Furthermore, it is also very common that we are given the value of k at the beginning and are asked to find the kth eigenvalue λ_k. We present a method for these problems based on a five point tracing of the eigenvalues with a continuous deformation of the domain to the unit circle.

We consider a set of domains Ω_t continuously varying with the parameter t, such that $\Omega_t|_{t=0} = \Omega$ (the original domain that we are concerned with), and $\Omega_t|_{t=1}$ is the unit disk. Such a continuous deformation can usually be constructed by a linear superposition of weights t and $1-t$ for representations of the domain and the unit circle with their parameters scaled the same. Suppose that we start from the kth eigenvalue of the unit circle; we follow the development of this eigenvalue as t decreases from 1 to 0. The principle guiding us is that we must keep track of the kth eigenvalue. Therefore, when there is a crossing between other branches of eigenvalues with the one that we are tracing, we must make an appropriate switch in order to keep track of the kth eigenvalue. Such is the case when two eigenvalues get close to each other and become a double eigenvalue, then bifurcate to two eigenvalues again as t varies. There is necessarily a switch of the branches for the kth eigenvalue, when there is such a crossing.

The main issue of the above tracing proposal is its efficiency. Apparently, it is very time-consuming to find the values of λ_k and its nearby eigenvalues for all values of t discretized for certain accuracy. The five point tracing algorithm is the scheme that makes this problem accessible in reality. It is based on the fact that the trace of the kth eigenvalue and its crossing with other branches can be realized by the values of $\mu_1(\lambda)$ at five points for each t. Consider a grid for λ with constant grid size $\delta\lambda$. For five nearby points on the grid, $s^n_{-2}, s^n_{-1}, s^n_0, s^n_1, s^n_2$ at $t = t^n$, if s^n_0 is the closest to λ_k at $t = t^n$, and there is no other eigenvalue nearby, we have

$$\mu_1(s^n_{-2}) > \mu_1(s^n_{-1}) > \mu_1(s^n_0) < \mu_1(s^n_1) < \mu_1(s^n_2). \tag{14}$$

At the next time step, $t = t^{n+1}$, we first evaluate μ_1 at the three points, s^n_{-1}, s^n_0, s^n_1. Then we choose s^{n+1}_0 as the s^n_j among the above three having the smallest μ_1 value. s^{n+1}_0 is the closest point on the grid to the kth eigenvalue in the next time step. The other four points are defined as

$$s^{n+1}_j = s^{n+1}_0 + j\delta\lambda \quad \text{for} \quad j = \pm 1, \pm 2.$$

If there is no other eigenvalue except λ_k in the region defined by the five points, an inequality similar to (14) is valid again. If some eigenvalue is nearby, we can detect this from the μ_1 values on these five points. For example, if λ_{k-1} appears in this region, then condition (14) becomes

$$\mu_1(s^{n+1}_{-2}) < \mu_1(s^{n+1}_{-1}) > \mu_1(s^{n+1}_0) < \mu_1(s^{n+1}_1) < \mu_1(s^{n+1}_2).$$

We follow the development of this pattern as t varies, until a later time, say $t = t^m$, when the two minima merge together. This is a different situation, since it represents a double eigenvalue near s^m_0. Therefore, we really need to keep track of the nearby two points, which have the smallest and second smallest μ_1 value. This fact and the side from which the other eigenvalue has come are remembered, and we anticipate that there is a bifurcation when t is further varied. In order to capture this bifurcation, it is necessary to keep track of six points denoted by $s^m_j, j = \pm 0, \pm 1, \pm 2$, where s_{-0}, s_{+0} are two distinct points at distance $\delta\lambda$ from each other. The following relation holds:

$$s^m_{-2} > s^m_{-1} > s^m_{-0} \approx s^m_{+0} < s^m_1 < s^m_2.$$

To follow the bifurcation of the eigenvalues, we evaluate μ_1 for $t = t^{m+1}$ at the four points $s^m_{-1}, s^m_{-0}, s^m_{+0}, s^m_1$. If the smallest and second smallest of the four values we obtain are located at two nearby points, then we pick $s^{m+1}_{-0}, s^{m+1}_{+0}$ as these two points. If the smallest and second smallest among these four points are separated, then we pick s^{m+1}_0 as one of these two points and switch back to the five point tracing scheme. The choice of s^{m+1}_0 reflects the kind of merging we have. If the merging is between λ_{k-1} and λ_k, then we need to select s^{m+1}_0 with a larger value, so that we are still on the branch for λ_k. Similarly, if it is the λ_{k+1} that merges with λ_k, then we select the point with a smaller s^m_j value as s^{m+1}_0. In any event, if such a separation occurs, we switch back to the five point tracing and continue our process. The value $\delta t = |t^m - t^{m+1}|$ can be fixed or varied at each step; the only necessary condition is that the ratio $\delta t/\delta\lambda$ should be smaller than the absolute value of the slope (or slopes at a crossing) of the curve (curves) of t versus λ_k (and $\lambda_{k\pm 1}$). For the starting point at $t = 1$, i.e., the initial unit disk, we may often start from a double root, so that a six point tracing will be necessary to begin with.

7. Discussion. We conclude this paper with a number of comments.

First, the discretization of the integral equation is based on the trapezoidal rule. This numerical quadrature formula is preferred here not only for its simplicity but also for its high order of accuracy for the periodic functions [2].

Second, the boundary value problem or the scattering problem related to the Helmholtz equation can also be handled with this $O(N)$ algorithm of matrix vector multiplication. A different kind of generalized conjugate gradient method, which utilizes the above matrix vector multiplication in each of its iteration steps, can be used as the main process.

Finally, it is evident that the method used in this paper can be generalized to other eigenvalue or boundary value problems [1] where boundary integral formulations, particularly with smooth kernels, are available.

REFERENCES

[1] C. A. BREBBIA, J. C. F. TELLES, AND L. C. WROBEL, *Boundary Element Techniques*, Springer-Verlag, Berlin, New York, 1984.

[2] L. M. DELVES AND J. L. MOHAMAD, *Computational Methods for Integral Equations*, Cambridge Univeristy Press, New York, 1985.

[3] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1984.

[4] W. HACKBUSCH AND Z. P. NOWAK, *On the fast matrix multiplication in the boundary element method by panel clustering*, Numer. Math., 54 (1989) pp. 463–491.

[5] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 409–436.

[6] R. HUTCHINSON, *Determination of membrane vibration characteristics by integral equation method*, in Recent Advances in Boundary Element Methods, C. A. Brebbia, ed., Pentech Press, London, 1978, pp. 301–315.

[7] J. R. KUTTLER AND V. G. SIGILLITO, *Eigenvalues of the Laplacian in two dimensions*, SIAM Rev., 26 (1984), pp. 163–194.

[8] C. LANCZOS, *An iteration method for the solutions of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Bur. Standards, 45 (1950), pp. 255–282.

[9] V. ROKHLIN, *Rapid solution of integral equations of classical potential theory*, J. Comput. Phys., 60 (1985), pp. 187–207.

[10] T. J. YPMA, *Finding a multiple zero by transformations and Newton-like methods*, SIAM Rev., 25 (1983), pp. 365–378.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 4, pp. 611–624, October 1991

ITERATIVE DESCENT ALGORITHMS FOR A ROW SUFFICIENT LINEAR COMPLEMENTARITY PROBLEM*

JONG-SHI PANG†

Abstract. The class of row sufficient linear complementarity problems was introduced in a recent paper by Cottle, Pang, and Venkateswaran [*Linear Algebra Appl.*, 114/115 (1989), pp. 231–249]. In the present paper, two iterative descent algorithms for solving such a linear complementarity problem are developed. One of the algorithms is based on a symmetric variational inequality formulation of the problem, and the other algorithm is an interior-point method which requires a strict feasibility assumption on the problem. Convergence of both algorithms is established. As a by-product of the investigation, a certain property of a column sufficient matrix is uncovered which leads to a constructive way of determining the solvability of a column sufficient linear complementarity problem.

Key words. linear complementarity problem, matrix splitting, sufficient matrices, interior-point method

AMS(MOS) subject classification. 90C33

1. Introduction. In a recent paper [8], Cottle, Pang, and Venkateswaran introduced the classes of row and column sufficient matrices and discussed their connection with the linear complementarity problem (LCP). These new matrix classes are generalizations of the positive semidefinite and the P-matrices whose fundamental role in the study of the linear complementarity problem is well recognized. It was shown that the matrix sufficiency properties provide interesting characterizations for the solutions of the corresponding linear complementarity problems. The authors of the cited paper also pointed out that Lemke's almost complementary pivoting algorithm [16] can be successfully employed to process a linear complementarity problem of the row sufficient type. In a subsequent paper [6], Cottle established the same conclusion for the Cottle–Dantzig principal pivoting algorithm [7]. As a reference of a solution method for solving a general linear complementarity problem, we mention [22], which describes a global optimization-based algorithm that does not depend on any special property of the defining matrix of the problem.

The motivation of the present paper is twofold. First of all, it is well known that a symmetric positive semidefinite linear complementarity problem can be solved by a host of efficient iterative algorithms, such as the family of successive overrelaxation methods [18], [21]; a paper by Cheng [5] demonstrates that the gradient projection algorithm in nonlinear programming can be applied to solve the asymmetric linear complementarity problem with a P-matrix. Since the class of row sufficient matrices extends the symmetric positive semidefinite and the (asymmetric) P-matrices, it is natural to ask whether the whole class of row sufficient linear complementarity problems can be processed by some iterative scheme(s). One goal of this paper is to develop a large family of such iterative algorithms as a generalization of Cheng's work. The resulting algorithms are related to the projected gradient methods for linearly constrained optimization problems specialized to a quadratic program [4].

A second motivation of this paper stems from the recent interest in interior-point methods for solving mathematical programs. Many papers have discussed these methods in the context of the linear complementarity problem (see [13], [14], [15], [26], [27],

* Received by the editors September 8, 1989; accepted for publication (in revised form) February 28, 1990. This work was based on research supported by National Science Foundation grant ECS-8717968.

† Department of Mathematical Sciences, The Whiting School of Engineering, The Johns Hopkins University, Baltimore, Maryland 21218 (MSC_WJP@JHUVMS.BITNET).

[28] and the references therein). In all these papers (with the exception of [28]), the linear complementarity problem is assumed to be of either the positive semidefinite or the P type. The paper [28] attempts to generalize the results to a broader class of problems and discusses the role of a certain quantity in the computational complexity of these interior-point methods. A second objective of the present paper is to develop a special version of the interior-point method for solving a row sufficient linear complementarity problem. We establish the (infinite) convergence of the method; unlike the cited references, the complexity issue is not treated here. As a by-product of this part of the investigation, we uncover a certain property of a column sufficient matrix which leads to a constructive way of determining the solvability of a column sufficient linear complementarity problem.

The organization of the remaining sections of this paper is as follows. In the next section, we review the notion of matrix sufficiency and its characterization in terms of certain solution properties of the linear complementarity problem; § 3 describes the family of iterative splitting algorithms for solving a row sufficient linear complementarity problem. The fourth section treats the interior-point method; in the fifth section, we discuss the solvability of a column sufficient linear complementarity problem. Finally, some concluding remarks are made in the sixth and last section.

2. Review. We begin with the definition [8].

DEFINITION 1. A matrix $M \in R^{n\times n}$ is row sufficient if the implication holds:

$$[\max_{1\leqq i\leqq n} x_i(M^Tx)_i \leqq 0] \Rightarrow [x_i(M^Tx)_i = 0 \quad \text{for all } i = 1, \cdots, n].$$

The matrix M is column sufficient if M^T is row sufficient; M is sufficient if it is both row and column sufficient.

A matrix $M \in R^{n\times n}$ is a P-*matrix* (P_0-*matrix*) if all its principal minors are positive (nonnegative). In [8], it was pointed out how row (and column) sufficient matrices are related to the classes of P- and P_0-matrices. In particular, a row (or column) sufficient matrix must be a P_0-matrix, and a P-matrix must be sufficient. Throughout this paper, we use the same notation to denote matrices of a particular type as well as the class to which they belong. For example, we speak of the class P of P-matrices.

We define the linear complementarity problem. Given a vector $q \in R^n$ and a matrix $M \in R^{n\times n}$, this problem, which is denoted by LCP(q, M), is to find a vector $x \in R^n$ such that

$$x \geqq 0, \qquad w = q + Mx \geqq 0, \qquad x^Tw = 0.$$

The feasible region of LCP(q, M) is denoted by F(q, M); we have

$$F(q,M) = \{x \in R^n : x \geqq 0, q + Mx \geqq 0\}.$$

A vector $x \in F(q, M)$ is said to be *strictly feasible* if $x > 0$ and $q + Mx > 0$.

Associated with the LCP(q, M) is the *natural* quadratic program denoted by NQP(q, M)

$$\begin{aligned} &\text{minimize} \quad x^T(q+Mx) \\ &\text{subject to} \quad x \in F(q,M). \end{aligned}$$

The stationary point problem associated with NQP(q, M) is the variational inequality problem of finding a vector $x \in F(q, M)$ such that

$$(y-x)^T(q+Nx) \geqq 0 \quad \text{for all } y \in F(q,M)$$

where

$$N = M + M^T \tag{1}$$

is twice the symmetric part of M. Note that N is symmetric. The latter variational problem is denoted by VI(q, M). Let

$$\theta(x) = x^T(q + Mx)$$

denote the objective function of the NQP(q, M).

In [8], a characterization of row sufficiency was obtained in terms of the relationship between the Karush–Kuhn–Tucker points of NQP(q, M) and the solutions of the LCP(q, M). In what follows, we rephrase this characterization in terms of the problem VI(q, M).

THEOREM 1. *Let $M \in R^{n \times n}$ be given. Then, M is row sufficient if and only if for all $q \in R^n$, the (possibly empty) solution sets of the two problems* LCP(q, M) *and* VI(q, M) *coincide.*

The above characterization does not assert that a row sufficient LCP is always solvable. As pointed out in [8], a row sufficient matrix must belong to the class Q_0, that is, for any vector q, the LCP(q, M) is solvable if and only if it is feasible. In particular, if LCP(q, M) has a strictly feasible solution, then it has a complementary solution.

In the case of a column sufficient matrix, the following characterization was obtained [8].

THEOREM 2. *Let $M \in R^{n \times n}$ be given. Then, M is column sufficient if and only if for all $q \in R^n$, the (possibly empty) solution set of* LCP(q, M) *is convex.*

Unlike a row sufficient matrix, a column sufficient matrix does not necessarily belong to the class Q_0; indeed, for the vector q and matrix M below,

$$q = \begin{bmatrix} -2 \\ -1 \\ -2 \end{bmatrix}, \qquad M = \begin{bmatrix} 1 & 1 & 0 \\ -5 & 1 & 0 \\ -1 & 1 & 0 \end{bmatrix},$$

it is not difficult to verify that M is column sufficient, and that the LCP(q, M) has a feasible solution $x = (0, 2, 0)^T$ but has no complementary solution. Since a column sufficient matrix must belong to the class P_0, the results in [1], [2] provide necessary and sufficient conditions for a column sufficient matrix to be in Q_0 and in Q (Q is the class of matrices M for which the LCP(q, M) has a solution for all q); nevertheless, these results fail to be applicable to determine the solvability of LCP(q, M) for a specific q.

3. A symmetric VI-based splitting algorithm. Based on Theorem 1, we introduce a matrix-splitting algorithm for solving the LCP(q, M) with M being a row sufficient matrix. Let

$$N = B + C$$

be a given *splitting* of the (symmetric) matrix N defined in (1). For our purpose here, the matrix B is taken to be symmetric positive definite.

ALGORITHM I. Let $x^0 \in \mathrm{F}(q, M)$ be given. In general, given a feasible vector $x^k \in \mathrm{F}(q, M)$, let $x^{k+1/2}$ be the (unique) solution of the quadratic program

$$\begin{aligned} &\text{minimize} && x^T(q + Cx^k) + \tfrac{1}{2} x^T B x \\ &\text{subject to} && x \in \mathrm{F}(q, M). \end{aligned} \tag{2}$$

Set $d^k = x^{k+1/2} - x^k$. Define the step size τ_k as follows: if $(d^k)^T M d^k$ is nonpositive, set

$\tau_k = 1$; otherwise, let τ_k be a nonnegative number such that

$$\theta(x^k + \tau_k d^k) = \min \{\theta(x^k + \tau d^k) : x^k + \tau d^k \in \mathrm{F}(q, M), \tau \geqq 0\}.$$

Set $x^{k+1} = x^k + \tau_k d^k$ and test x^{k+1} for termination. Repeat the process if x^{k+1} fails the termination test.

Several remarks should be made, the foremost of which is the fact that the algorithm requires the feasibility of the LCP(q, M) and maintains this property throughout. As a consequence of this feasibility condition, each subproblem (2) indeed has a unique optimal solution $x^{k+1/2}$ by the symmetry and positive definiteness of the matrix B. This solution $x^{k+1/2}$ can be computed by many efficient iterative methods such as those described in [17]. In this context, Algorithm I becomes a hybrid iterative scheme for solving the LCP(q, M), which comprises two levels of iterations: inner and outer. Each outer iteration corresponds to an update of the subproblem (2) to be solved by the inner iterations; the step-size determination is performed at the outer iteration.

By the minimum principle, the solution $x^{k+1/2}$ of (2) satisfies the variational inequality

$$(y - x^{k+1/2})^T (q + Cx^k + Bx^{k+1/2}) \geqq 0, \quad \text{for all } y \in \mathrm{F}(q, M). \tag{3}$$

Substituting $y = x^k \in \mathrm{F}(q, M)$, we derive

$$(x^k - x^{k+1/2})^T (q + Cx^k + Bx^{k+1/2}) \geqq 0,$$

which implies

$$(d^k)^T (q + Nx^k) \leqq -(x^{k+1/2} - x^k)^T B (x^{k+1/2} - x^k) \leqq 0, \tag{4}$$

where the last inequality follows because B is positive definite. By noticing

$$\nabla\theta(x^k) = q + Nx^k$$

where $\theta(x)$ is the objective function of NQP(q, M), we conclude that

$$(d^k)^T \nabla\theta(x^k) \leqq 0; \tag{5}$$

moreover, equality holds only if $x^{k+1/2} = x^k$, in which case x^k must be a solution of VI(q, M), and thus of LCP(q, M), by Theorem 1. Summarizing this discussion, we have proven the following important descent result.

LEMMA 1. *Suppose that x^k is not a solution of* LCP(q, M). *Then the vector d^k generated by Algorithm* I *is a direction of descent for the objective function of* NQP(q, M).

In the discussion below, we assume that strict inequality holds in (5). The determination of the step size τ_k can be better understood by considering the one-dimensional function

$$\delta(t) = \theta(x^k + td^k), \qquad t > 0.$$

We may write

$$\theta(x^k + td^k) = \theta(x^k) + t(d^k)^T (q + Nx^k) + \frac{t^2}{2} (d^k)^T M d^k.$$

Suppose that $(d^k)^T M d^k \leqq 0$. Then, with the step size $t = 1$, we obtain

$$\theta(x^k + d^k) \leqq \theta(x^k) + (d^k)^T (q + Nx^k);$$

thus, the next iterate $x^{k+1} = x^k + d^k$ satisfies

$$\theta(x^{k+1}) - \theta(x^k) \leqq \sigma\tau_k (d^k)^T (q + Nx^k) \tag{6}$$

for any $\sigma \in (0, 1)$ by the descent condition (5).

Suppose now that $(d^k)^T M d^k > 0$. Then the one-dimensional function $\delta(t)$ is strictly convex in t and its unconstrained global minimum is attained at the value

$$t_k = -\frac{(d^k)^T (q + Nx^k)}{(d^k)^T M d^k}.$$

Note that $t_k \geqq 0$ by (5). If the vector $x^k + t_k d^k$ is in the feasible region $\mathrm{F}(q, M)$, then the step size τ_k is equal to t_k; in this case, it is not difficult to show that the inequality (6) must hold as an equation with $\sigma = \frac{1}{2}$. On the other hand, if $x^k + t_k d^k$ lies outside of the feasible region $\mathrm{F}(q, M)$, then we must have

$$1 \leqq \tau_k < t_k;$$

in this case, one can easily show that the inequality (6) also holds with $\sigma = \frac{1}{2}$. Moreover, the actual computation of the desired step size τ_k is not difficult at all because of the quadratic nature of the one-dimensional function $\delta(t)$ and the polyhedrality of the feasible region $\mathrm{F}(q, M)$.

Summarizing the above discussion, we have established the following descent property of the sequence $\{x^k\}$ produced by Algorithm I.

PROPOSITION 1. *Let x^{k+1} be generated as in Algorithm* I. *Then*,

$$\theta(x^{k+1}) - \theta(x^k) \leqq \frac{\tau_k}{2} (d^k)^T (q + Nx^k) \leqq 0.$$

Before completing the convergence proof, we make some further comments on Algorithm I when the $\mathrm{LCP}(q, M)$ is derived from a primal-dual pair of linear programs. To be specific, let us consider the linear program

$$\begin{aligned} &\text{minimize} && c^T z \\ &\text{subject to} && Az \geqq b, \qquad z \geqq 0 \end{aligned}$$

and its dual

$$\begin{aligned} &\text{maximize} && b^T y \\ &\text{subject to} && A^T y \leqq c, \qquad y \geqq 0. \end{aligned}$$

The corresponding $\mathrm{LCP}(q, M)$ in this case is defined by

$$q = \begin{bmatrix} c \\ -b \end{bmatrix}, \qquad M = \begin{bmatrix} 0 & -A^T \\ A & 0 \end{bmatrix};$$

the vector x (in the LCP) is composed of the primal and dual variables z and y; the objective function $\theta(x) = c^T z - b^T y$ is equal to the gap between the primal and dual objective values at an arbitrary primal-dual pair of vectors z and y. The matrix M is skew-symmetric, hence N is identically equal to zero. Consequently, the step size τ_k in Algorithm I is equal to unity at each iteration.

Consider the choice of an identity matrix for B. Then C is equal to the negative identity matrix. With this choice, the problem (2) decomposes into two independent

subproblems: one in the primal variables z,

$$\begin{aligned} &\text{minimize} && c^T z + \tfrac{1}{2}(z-z^k)^T(z-z^k) \\ &\text{subject to} && Az \geqq b, \qquad z \geqq 0 \end{aligned}$$

(where z^k is the current (primal) feasible iterate), and the other in the dual variables y,

$$\begin{aligned} &\text{minimize} && -b^T y + \tfrac{1}{2}(y-y^k)^T(y-y^k) \\ &\text{subject to} && A^T y \leqq c, \qquad y \geqq 0 \end{aligned}$$

(where y^k is the current (dual) feasible iterate). It is easy to see that these subproblems are precisely those arising from the *proximal point algorithm* [23], [24] specialized to solve a primal-dual pair of linear programs [19], [9]. Consequently, we may conclude that in the context of linear programming, Algorithm I, with a diagonal choice of B, reduces to the well-known proximal point algorithm.

The following result establishes the convergence of Algorithm I.

THEOREM 3. *Let M be a row sufficient matrix and B a symmetric positive definite matrix. Suppose that the* LCP(q, M) *is feasible. Then, every accumulation point of the sequence $\{x^k\}$ produced by Algorithm* I *is a solution of* LCP(q, M).

Proof. Let $\tilde{x}$ be the limit of a subsequence $\{x^k : k \in \kappa\}$. The sequence $\{\theta(x^k)\}$ is nonincreasing by Proposition 1. Since the subsequence $\{\theta(x^k) : k \in \kappa\}$ converges, the entire sequence $\{\theta(x^k)\}$ must be bounded below and hence must converge. Thus, $\{\theta(x^{k+1}) - \theta(x^k)\} \to 0$. By Proposition 1 again, it follows that

$$\lim_{k \to \infty} \tau_k (d^k)^T (q + Nx^k) = 0.$$

There are two cases:

(i) $\liminf_{k \to \infty, k \in \kappa} \tau_k > 0$;

(ii) $\liminf_{k \to \infty, k \in \kappa} \tau_k = 0$.

In case (i), we must have

$$\lim_{k \to \infty, k \in \kappa} (d^k)^T (q + Nx^k) = 0, \tag{7}$$

which in view of the inequality (4) and the positive definiteness of B, implies

$$\lim_{k \to \infty, k \in \kappa} (x^{k+1/2} - x^k) = 0.$$

Consequently, it follows that $\lim_{k \to \infty, k \in \kappa} x^{k+1/2} = \tilde{x}$. Passing to the limit as $k \to \infty$, $k \in \kappa$ in the expression (3), we conclude that $\tilde{x}$ is a solution of the VI(q, M). Consequently, $\tilde{x}$ solves LCP(q, M) by Theorem 1.

Consider case (ii). By the inequality (4) and the Cauchy–Schwartz inequality, we obtain

$$\alpha \| d^k \|_2^2 \leqq (d^k)^T B d^k \leqq -(d^k)^T (q + Nx^k) \leqq \| d^k \|_2 \| q + Nx^k \|_2,$$

where $\alpha > 0$ is the smallest eigenvalue of the (symmetric positive definite) matrix B. Cancelling one term $\| d^k \|_2$, we conclude that the sequence of directions $\{d^k : k \in \kappa\}$ is bounded. Without loss of generality, we may assume that $\lim_{k \to \infty, k \in \kappa} \tau_k = 0$. By the preceding analysis of the step-size determination, we deduce that for all $k \in \kappa$ sufficiently large,

$$\tau_k = -\frac{(d^k)^T (q + Nx^k)}{(d^k)^T M d^k}.$$

Hence, it follows that the limit condition (7) again holds. The remaining proof of case (i) therefore applies. This completes the proof of the theorem. □

Theorem 3 does not assert the existence of an accumulation point of the sequence $\{x^k\}$. In order for this to hold, it suffices that the level set

$$\{x \in \mathrm{F}(q,M) : \theta(x) \leqq \theta(x^0)\}$$

be bounded. In turn, the latter condition holds if the matrix M belongs to the class R_0, i.e., if the homogeneous LCP$(0, M)$ has zero as the unique solution (this is true because if $\{u^k\}$ were any unbounded sequence belonging to the above level set, each limit point of the normalized sequence $\{u^k/\|u^k\|\}$ would be a nonzero solution of the LCP$(0, M)$). In particular, if M is a P-matrix, then the sequence $\{x^k\}$ produced by Algorithm I must be bounded and every accumulation point must be a solution of LCP(q, M). Since the LCP(q, M) has a unique solution by the P-property of M, it follows that the sequence $\{x^k\}$ must converge to that solution. Summarizing this discussion, we have proven the following consequences of Theorem 3.

COROLLARY 1. *Let $M \in R^{n\times n}$ and $q \in R^n$ be given.*

(a) *If M is row sufficient and belongs to the class* R_0, *then for any initial vector $x^0 \in \mathrm{F}(q, M)$, the sequence $\{x^k\}$ produced by Algorithm* I *is bounded*; *moreover, every accumulation point solves the* LCP(q, M).

(b) *If M is a* P*-matrix, then for any initial vector $x^0 \in \mathrm{F}(q, M)$, the sequence $\{x^k\}$ produced by Algorithm* I *converges to the unique solution of* LCP(q, M).

It should be pointed out that under the assumptions in part (a) of the corollary, the solvability (and hence the feasibility) of the LCP(q, M) are guaranteed by the results in [8], [2]; the main conclusion of this part of the result is the boundedness of the sequence produced by Algorithm I.

4. An interior-point method. In this section, we describe an interior-point method for solving a row sufficient LCP. The basic assumption required for this method is the existence of a strictly feasible vector for the problem. Recall that this assumption can always be satisfied in the case of a positive semidefinite matrix. This is accomplished by considering a modified LCP, as suggested in [14]. Indeed, given an $n \times n$ LCP(q, M) with M being positive semidefinite, one can always associate with it an augmented $(n+1)\times(n+1)$ LCP$(\tilde{q}, \tilde{M})$, where

$$\tilde{q} = \begin{bmatrix} q \\ \sigma \end{bmatrix}, \qquad \tilde{M} = \begin{bmatrix} M & e \\ -e^T & 0 \end{bmatrix},$$

with e being the n-vector of all ones and $\sigma > 0$ being an appropriate constant; this augmented LCP satisfies the following properties:

(i) The matrix $\tilde{M}$ remains positive semidefinite;

(ii) A strictly feasible vector $\tilde{x}$ is readily available for the augmented problem;

(iii) If the given LCP(q, M) has a solution, then the last component in any solution $\tilde{x}$ of LCP$(\tilde{q}, \tilde{M})$ must be zero.

When M is merely row sufficient, it is generally not true that the above augmented matrix $\tilde{M}$ will remain row sufficient. An example is the following matrix

$$M = \begin{bmatrix} 0 & 5 \\ -1 & 1 \end{bmatrix},$$

which can be easily verified to be row sufficient. The augmented matrix $\tilde{M}$ is not row sufficient because the defining property of row sufficiency fails with the vector

$(0, -3, 0)^T$. Consequently, the augmentation suggested in [14] fails to be applicable for a row sufficient LCP.

As an alternative, consider the $2n \times 2n$ LCP(q', M'), where

$$q' = \begin{bmatrix} q \\ \sigma e \end{bmatrix}, \qquad M' = \begin{bmatrix} M & I \\ -I & 0 \end{bmatrix},$$

and $\sigma > 0$ is an appropriately chosen scalar. In what follows, the magnitude of σ is not explicitly specified because we are not concerned with complexity issues here; we simply refer to σ as a sufficiently large positive quantity. The reader can consult [14] if he is interested in how σ ought to be defined in compliance with the size of the data q and M. (See also the proof of Proposition 2, below, and the next section.)

The matrix M' is row sufficient if M is so; to verify this, suppose $x' \in R^{2n}$ is such that

$$(x')_i((M')^T x')_i \leqq 0 \quad \text{for all } i = 1, \cdots, 2n.$$

Write $x' = (u, v)$ with u and v both n-vectors. Then, we have for all $i = 1, \cdots, n$,

$$u_i(M^T u)_i - u_i v_i \leqq 0, \qquad v_i u_i \leqq 0, \tag{8}$$

which imply

$$u_i(M^T u)_i \leqq 0. \tag{9}$$

By the row sufficiency of M, it follows that equality holds in (9) for all $i = 1, \cdots, n$; hence the same is true in (8). This establishes the row sufficiency of M'.

The augmented LCP(q', M') obviously has a strictly feasible vector provided that σ is large enough. To see this, choose an arbitrary positive vector $u \in R^n$; then pick $\sigma > \max_i u_i$ and $v \in R^n_+$ such that $q + Mu + v > 0$. This pair of vectors u and v provides a desired strictly feasible vector for LCP(q', M'). With the availability of such a strictly feasible vector, one could apply the interior-point method (to be described later) to the augmented LCP(q', M'); if the computed solution (u, v) has $v = 0$, then u is a solution of the original LCP(q, M). In particular, if the LCP(q', M') has the property that all of its solutions have the v-component equal to zero, then one can safely apply the interior-point method to it and recover a solution to the original problem LCP(q, M) easily. It turns out that the required property of M for this statement to hold is *column* (and not row) sufficiency. We state this result more precisely in the following proposition and shall return to discuss more of its implications in the next section.

PROPOSITION 2. *If M is column sufficient and if* LCP(q, M) *has a solution, then for all σ sufficiently large, any solution (u, v) of* LCP(q', M') *must have $v = 0$, and therefore u is a solution of the* LCP(q, M).

Proof. To prove the assertion, let σ be greater than the largest component in all basic (complementary) solutions of LCP(q, M) (there are only finitely many such solutions). Suppose that (u, v) is an arbitrary solution of the augmented LCP(q', M'). Write $y = q + Mu + v$ and $z = \sigma e - u$. Let x be a basic solution of the original LCP(q, M). Then, for $i = 1, \cdots, n$,

$$0 \geqq (u-x)_i(y-w)_i = (u-x)_i(M(u-x))_i + (u-x)_i v_i.$$

If $v_i > 0$ for some i, then $u_i = \sigma$ by complementarity; thus, $u_i > x_i$ by the choice of σ. Consequently, each product $(u - x)_i v_i$ is nonnegative. It follows that for each i,

$$0 \geqq (u-x)_i(M(u-x))_i.$$

By the column sufficiency of M, we deduce

$$0 = (u-x)_i (M(u-x))_i$$

for all i. Thus, $v = 0$, as desired. □

Summarizing the above discussion, we conclude that if M is a sufficient matrix, then by considering an augmented LCP if necessary, we can assume with no loss of generality that the LCP(q, M) has a strictly feasible vector. At this time, we do not fully understand how essential the column sufficiency property is for Proposition 2 to be valid; the question of whether a (possible different) modified LCP can be derived which satisfies all the desirable properties when M is row sufficient remains unanswered.

From this point on, we consider an $n \times n$ LCP(q, M) where M is row sufficient and for which there exists a vector $x > 0$ satisfying $q + Mx > 0$. Fix a scalar $\rho > n$ and consider the real-valued function $\phi : R^n_{++} \times R^n_{++} \to R$ defined by

$$\phi(x, w) = \rho \log (x^T w) - \sum_{i=1}^{n} \log (x_i \omega_i),$$

where "log" denotes the natural logarithm. This is the "merit" function for this class of descent methods; it is well defined whenever x and w are both positive. The function $\phi(x, w)$ first appeared in a paper by Todd and Ye [25] dealing with linear programming, and was used in [13], which treats the LCP. The following result lists several useful properties of this function. (A word about notation: if d is a vector, we denote by diag (d) the diagonal matrix whose diagonal entries are the components of d.)

PROPOSITION 3. *Let x and w be two positive n-vectors. Let X = diag (x) and W = diag (w). Then,*

$$\phi(x, w) \geqq (\rho - n) \log (x^T w) \tag{10}$$

$$(\nabla_x \phi(x, w))_i (\nabla_w \phi(x, w))_i = x_i w_i \left(\frac{\rho}{x^T w} - \frac{1}{x_i w_i} \right)^2 \quad \text{for all } i \tag{11}$$

$$(\nabla_x \phi(x, w))^T \nabla_w \phi(x, w) > 0. \tag{12}$$

Proof. The verification of (10) and (11) is fairly straightforward. We now prove (12). Suppose that $(\nabla_x \phi(x, w))^T \nabla_w \phi(z, w) = 0$. According to (11), it follows that for each i,

$$\frac{\rho}{x^T w} - \frac{1}{x_i w_i} = 0.$$

This implies

$$\rho x^T w = n x^T w,$$

which is a contradiction because $\rho > n$. □

The next result is an immediate consequence of the expressions (11) and (12) in the above proposition and provides an important justification for the descent step of the interior-point method.

COROLLARY 2. *If M is row sufficient, then for $x, w > 0$,*

$$\nabla_x \phi(x, w) + M^T \nabla_w \phi(x, w) \neq 0.$$

Proof. Suppose the contrary. By (11), it follows that for all i,

$$(\nabla_w \phi(x, w))_i (M^T \nabla_w \phi(x, w))_i = -(\nabla_w \phi(x, w))_i (\nabla_x \phi(x, w))_i \leqq 0.$$

Hence, by the row sufficiency of M, we deduce

$$(\nabla_w \phi(x, w))^T \nabla_x \phi(x, w) = 0,$$

which contradicts (12). □

Remark. By using a sign-reversing characterization of a P_0-matrix [11], we can show that the above corollary remains valid when M is a P_0-matrix. This can be proved by a slight modification of the argument used above.

We now describe the interior-point method for solving the LCP(q, M) when M is row sufficient. In the algorithm below, the scalar $\beta \in (0, 1)$ controls the step size in each descent iteration and ensures the strict feasibility of the iterates obtained; $\gamma \in (0, 1)$ is the usual backtracking factor required in an Armijo-type line search step; and $\sigma \in (0, \frac{1}{2})$ determines the amount of sufficient decrease in the line search (we refer to [3] for more discussion of line-search procedures of this type).

ALGORITHM II. Let $\beta, \gamma \in (0, 1)$ and $\sigma \in (0, \frac{1}{2})$ be given. Let x^0 be a strictly feasible point of LCP(q, M) and let $w^0 = q + Mx^0$. In general, given the pair $(x^k, w^k) > 0$, let

$$\nabla_x \phi_k = \nabla_x \phi(x^k, w^k), \qquad \nabla_w \phi_k = \nabla_w \phi(x^k, w^k),$$

and

$$X^k = \operatorname{diag}(x^k), \qquad W^k = \operatorname{diag}(w^k).$$

Solve the problem below to obtain the search direction (d_x^k, d_w^k):

$$\begin{aligned} &\text{minimize} && (\nabla_x \phi_k)^T d_x + (\nabla_w \phi_k)^T d_w \\ &\text{subject to} && d_w = M d_x, \qquad \|(X^k)^{-1} d_x\|_2^2 + \|(W^k)^{-1} d_w\|_2^2 \leqq \beta^2. \end{aligned}$$

Let m_k be the smallest nonnegative integer m such that

$$\phi(x^k + \gamma^m d_x^k, w^k + \gamma^m d_w^k) - \phi(x^k, w^k) \leqq \sigma \gamma^m [(\nabla_x \phi_k)^T d_x^k + (\nabla_w \phi_k)^T d_w^k]$$

and set

$$(x^{k+1}, w^{k+1}) = (x^k, w^k) + \gamma^{m_k}(d_x^k, d_w^k).$$

If (x^{k+1}, w^{k+1}) satisfies a prescribed termination rule, stop; otherwise, repeat the general step with k replaced by $k + 1$.

The search direction (d_x^k, d_w^k) admits an explicit expression. Indeed, let

$$p^k = \nabla_x \phi_k + M^T \nabla_w \phi_k,$$

$$M^k = (X^k)^{-2} + M^T (W^k)^{-2} M;$$

the vector p^k is nonzero by Corollary 2 and the matrix M^k is clearly symmetric and positive definite; hence, the scalar

$$\lambda_k = \frac{\sqrt{(p^k)^T (M^k)^{-1} p^k}}{\beta}$$

is positive. We have

$$d_x^k = -\frac{1}{\lambda_k} (M^k)^{-1} p^k, \qquad d_w^k = M d_x^k.$$

Since

$$(\nabla_x \phi_k)^T d_x^k + (\nabla_w \phi_k)^T d_w^k = -\lambda_k \beta^2 < 0, \tag{13}$$

it follows that (d_x^k, d_w^k) indeed provides a descent direction for the function $\phi(x, w)$. Moreover, it is obvious that for any $\tau \in [0, 1)$, the vector pair

$$(x^k(\tau), w^k(\tau)) = (x^k, w^k) + \tau(d_x^k, d_w^k)$$

remains positive; in particular, so does (x^{k+1}, w^{k+1}), defined in Algorithm II.

In view of (13), the sequence $\{x^k\}$ satisfies the inequality

$$\phi(x^{k+1}, w^{k+1}) - \phi(x^k, w^k) \leqq -\sigma\beta^2\gamma^{m_k}\lambda_k < 0, \tag{14}$$

which implies that the sequence $\{\phi(x^k, w^k)\}$ is decreasing. Thus, by (10), the sequence

$$\{(x^k)^T w^k\}$$

is bounded. Since $x^k \in \mathrm{F}(q, M)$ for each k, the discussion in the paragraph preceding Corollary 1 of § 3 shows that the sequence $\{x^k\}$ is bounded if, for example, M belongs to the class R_0. The following theorem establishes the convergence of Algorithm II.

THEOREM 4. *Let M be a row sufficient matrix. Suppose that the* LCP(q, M) *has a strictly feasible solution. Then every accumulation point of the sequence $\{x^k\}$ produced by Algorithm* II *is a solution of* LCP(q, M).

Proof. Let $\tilde{x}$ be the limit of a subsequence $\{x^k : k \in \kappa\}$ and let $\tilde{w} = q + M\tilde{x}$. Clearly, the pair $(\tilde{x}, \tilde{w})$ is nonnegative. Since $\phi(\tilde{x}, \tilde{w}) < \infty$, it follows that either $\tilde{x}^T\tilde{w} = 0$ or $(\tilde{x}, \tilde{w}) > 0$. In the former case, the theorem is proven. Therefore we assume that the latter holds. Let $\tilde{p}$ and $\tilde{M}$ denote the limits of the sequences $\{p^k : k \in \kappa\}$ and $\{M^k : k \in \kappa\}$, respectively. The matrix $\tilde{M}$ remains positive definite; moreover, the sequence of scalars $\{\lambda_k : k \in \kappa\}$ converges to

$$\tilde{\lambda} = \frac{\sqrt{\tilde{p}^T\tilde{M}^{-1}\tilde{p}}}{\beta},$$

which is positive, and the sequence of directions $\{(d_x^k, d_w^k) : k \in \kappa\}$ converges to $(\tilde{d}_x, \tilde{d}_w)$, where

$$\tilde{d}_x = -\frac{1}{\tilde{\lambda}}\tilde{M}^{-1}\tilde{p}, \qquad \tilde{d}_w = M\tilde{d}_x.$$

Since the sequence $\{\phi(x^{k+1}, w^{k+1}) - \phi(x^k, w^k)\}$ converges to zero, the inequality (14) implies

$$\lim_{k \to \infty, k \in \kappa} \gamma^{m_k} = 0$$

or equivalently,

$$\lim_{k \to \infty, k \in \kappa} m_k = \infty.$$

Hence, both sequences $\{(x^{k+1}, w^{k+1}) : k \in \kappa\}$ and $\{(x^k + \tau_k d_x^k, w^k + \tau_k d_w^k) : k \in \kappa\}$, where $\tau_k = \gamma^{m_k - 1}$, converge to $(\tilde{x}, \tilde{w})$. By the definition m_k, we have

$$\frac{\phi(x^k + \tau_k d_x^k, w^k + \tau_k d_w^k) - \phi(x^k, w^k)}{\tau_k} > -\sigma\beta^2\lambda_k;$$

on the other hand, (14) implies

$$\frac{\phi(x^{k+1}, w^{k+1}) - \phi(x^k, w^k)}{\gamma^{m_k}} \leqq -\sigma\beta^2\lambda_k.$$

Passing to the limit $\{k \to \infty, k \in \kappa\}$ in the last two inequalities and noting that ϕ is F-differentiable at $(\tilde{x}, \tilde{w})$, we deduce

$$\nabla_x \phi(\tilde{x}, \tilde{w})^T \tilde{d}_x + \nabla_w \phi(\tilde{x}, \tilde{w})^T \tilde{d}_w = -\sigma \tilde{\lambda} \beta^2.$$

On the other hand, passing to the same limit in (13), we obtain

$$\nabla_x \phi(\tilde{x}, \tilde{w})^T \tilde{d}_x + \nabla_w \phi(\tilde{x}, \tilde{w})^T \tilde{d}_w = -\tilde{\lambda} \beta^2,$$

which is a contradiction. This establishes the theorem. □

Kojima and his colleagues [12] have developed a version of the interior-point method, which can be proven subsequentially convergent for the class of linear complementarity problems with a P_0-matrix; our results (Corollary 2 and Theorem 4) are also valid for such a matrix. It appears that there are considerable overlaps between the two sets of results.

5. More on the augmented problem. In this section, we expand on the discussion of the augmented LCP introduced in § 4. Let the $n \times n$ LCP(q, M) be given. For our purpose here, we consider the $2n \times 2n$ LCP(q', M') where

$$q' = \begin{bmatrix} q \\ \sigma e \end{bmatrix}, \qquad M' = \begin{bmatrix} M & I \\ -I & tI \end{bmatrix}, \tag{15}$$

where σ and t are two positive scalars. Similar to the proof of Proposition 2, choose σ such that for all $\alpha \subseteq \{1, \cdots, n\}$ with $M_{\alpha\alpha}$ nonsingular,

$$\sigma > \max\,(0, -((M_{\alpha\alpha})^{-1} q_\alpha)_i) \quad \text{for all } i \in \alpha;$$

the scalar $t > 0$ is arbitrary.

We recall [10] that a matrix $M \in R^{n \times n}$ is *semimonotone* (or equivalently, belongs to class E_0) if for every $0 \neq x \geqq 0$, there exists an index i such that $x_i > 0$ and $(Mx)_i \geqq 0$. Clearly, a column sufficient matrix must be semimonotone, and so must a P_0-matrix. The following proposition identifies two important properties of the augmented matrix M' in (15) when $M \in E_0$.

PROPOSITION 4. *Let $t > 0$ be arbitrary. If M is semimonotone, then $M' \in E_0 \cap R_0$.*

Proof. The proof is very similar to the argument in § 4 preceding Proposition 2. For completeness, we give a detailed proof. Let $x' = (u, v) \in R^{2n}_+$ be a given nonzero vector. Without loss of generality, we may assume that $u \neq 0$. By the semimonotonicity of M, there exists an index $i \in \{1, \cdots, n\}$ such that $u_i > 0$ and $(Mu)_i \geqq 0$. For such an index i, we have $(Mu + v)_i \geqq 0$; this establishes the semimonotonicity of M'.

To show $M' \in R_0$, suppose that $x' = (u, v)$ is a nonzero solution of the $2n \times 2n$ homogeneous LCP$(0, M')$. If $u = 0$, the fact that $t > 0$ yields $v = 0$, which contradicts the nonzero assumption of (u, v). Suppose that $u \neq 0$. As above, let i be an index such that $u_i > 0$ and $(Mu)_i \geqq 0$. By complementarity, we derive $(Mu + v)_i = 0$, which implies that $v_i = (Mu)_i = 0$. On the other hand, we also have

$$0 \leqq -u_i + tv_i = -u_i < 0,$$

which is a contradiction. This completes the proof. □

Remarks. (1) If M' is semimonotone, then so is M because the semimonotone property is inherited by principal submatrices.

(2) The fact that $t > 0$ is essential for the augmented matrix M' to belong to R_0; indeed, trivial examples can be constructed to show that the matrix M' is not in R_0 when $t = 0$.

According to the results in [10], [20], if a matrix belongs to $E_0 \cap R_0$, the corresponding LCP can be solved by Lemke's algorithm. Now, suppose that M is column sufficient. By the same proof, one can show that Proposition 2 remains valid for the matrix M' and vector q' in (15) and for σ chosen as above. Solving the augmented LCP(q', M') by Lemke's algorithm, one computes a solution (u, v). If $v = 0$, then a solution to the original LCP(q, M) is obtained; otherwise, one concludes that the original LCP(q, M) does not have a solution. In this fashion, one can successfully determine whether or not the LCP(q, M) is solvable, and compute a solution if it exists. This procedure requires only that M be column sufficient; no other assumption is required.

6. Concluding remarks. In the last two sections, we have presented two algorithms for solving an LCP with a row sufficient matrix. These algorithms are infinite and descent in nature and therefore differ from the existing methods for solving this class of LCPs, which are finite pivotal schemes [2], [6].

The first algorithm is based on a symmetric variational inequality formulation of the problem and requires solving subproblems which are strictly convex quadratic programs. The implementation of this algorithm requires a feasible vector of the LCP to be used as a starting iterate. There is considerable flexibility in using the algorithm, since the splitting $N = B + C$ is quite arbitrary. The reader is referred to [17] for an in-depth discussion on the family of matrix-splitting-based iterative methods for solving related LCPs and quadratic programs.

The second algorithm is based on the interior-point concept and is actually applicable to a broader class of LCPs than those of the row sufficient type. This algorithm requires solving systems of linear equations defined by symmetric positive definite matrices. An important drawback of Algorithm II is that a strictly feasible vector must be available in order to initiate it.

It is not easy to evaluate the practical performance of the two algorithms without actual implementation. The above discussion briefly outlines some potential advantages and disadvantages of the algorithms. Numerical results on the algorithms are not available at this time.

Acknowledgment. The author is grateful to Professor R. W. Cottle, who brought to his attention a talk which Professor M. Kojima gave at Stanford University on the interior-point method. This led to the subsequent contact with Professor Kojima and to the private communication, [12].

REFERENCES

[1] M. AGANAGIĆ AND R. W. COTTLE, *A note on Q-matrices*, Math. Programming, 16 (1979), pp. 374–377.

[2] ———, *A constructive characterization of Q_0-matrices with nonnegative principal minors*, Math. Programming, 37 (1987), pp. 223–231.

[3] D. M. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[4] P. H. CALAMAI AND J. J. MORÉ, *Projected gradient methods for linearly constrained problems*, Math. Programming, 39 (1987), pp. 93–116.

[5] Y. C. CHENG, *On the gradient-projection method for solving the nonsymmetric linear complementarity problem*, J. Optim. Theory Appl., 43 (1984), pp. 527–541.

[6] R. W. COTTLE, *The principal pivoting method revisited*, Math. Programming, Ser. B, 48 (1990), pp. 369–386.

[7] R. W. COTTLE AND G. B. DANTZIG, *Complementary pivot theory of mathematical programming*, in Mathematics for the Decision Sciences, Part 1, G. B. Dantzig and A. F. Veinott, Jr., eds., American Mathematical Society, Providence, RI, 1968, pp. 115–136.

[8] R. W. COTTLE, J. S. PANG, AND V. VENKATESWARAN, *Sufficient matrices and the linear complementarity problem*, Linear Algebra Appl., 114/115 (1989), pp. 231–249.
[9] R. DE LEONE AND O. L. MANGASARIAN, *Serial and parallel solution of large scale linear programs by augmented Lagrangian successive overrelaxation*, in Optimization, Parallel Processing and Applications, A. Kurzhanski, K. Neuman, and D. Pallasche, eds., Lecture Notes in Economics and Mathematical Systems 304, Springer-Verlag, Berlin, New York, 1988, pp. 103–124.
[10] B. C. EAVES, *The linear complementarity problem*, Management Sci., 17 (1971), pp. 621–634.
[11] M. FIEDLER AND V. PTÁK, *On matrices with non-positive off-diagonal elements and positive principal minors*, Czechoslovak Math. J., 12 (1962), pp. 382–400.
[12] M. KOJIMA, Private communication, August, 1989.
[13] M. KOJIMA, N. MEGIDDO, AND Y. YE, *An interior point potential reduction algorithm for the linear complementarity problem*, Res. Report RJ 6486, IBM Almaden Research Center, San José, CA, 1988.
[14] M. KOJIMA, S. MIZUNO, AND A. YOSHISE, *A polynomial-time algorithm for a class of linear complementarity problems*, Math. Programming, 44 (1989), pp. 1–26.
[15] ———, *An $O(\sqrt{n}L)$ iteration potential reduction algorithm for linear complementarity problems*, Res. Report B-217, Department of Information Sciences, Tokyo Institute of Technology, Tokyo, Japan, 1988.
[16] C. E. LEMKE, *Bimatrix equilibrium points and mathematical programming*, Management Sci., 11 (1965), pp. 681–689.
[17] Y. Y. LIN AND J. S. PANG, *Iterative methods for large convex quadratic programs: A survey*, SIAM J. Control Optim., 25 (1987), pp. 383–411.
[18] O. L. MANGASARIAN, *Solution of symmetric linear complementarity problems by iterative methods*, J. Optim. Theory Appl., 22 (1977), pp. 465–485.
[19] ———, *Iterative solution of linear programs*, SIAM J. Numer. Anal., 18 (1981), pp. 606–614.
[20] J. S. PANG, *On Q-matrices*, Math. Programming, 17 (1979), pp. 243–247.
[21] ———, *More results on the convergence of iterative methods for the symmetric linear complementarity problem*, J. Optim. Theory Appl., 49 (1986), pp. 107–134.
[22] P. M. PARDALOS AND J. B. ROSEN, *Global optimization approach to the linear complementarity problem*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 341–353.
[23] R. T. ROCKAFELLAR, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.
[24] ———, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.
[25] M. J. TODD AND Y. YE, *A centered projective algorithm for linear programming*, Math. Oper. Res., 15 (1990), pp. 508–529.
[26] P. TSENG, *Complexity analysis of a linear complementarity algorithm based on a Lyapunov function*, Tech. Report LIDS-P-1883, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1989 (revised version).
[27] Y. YE, *A further result on the potential reduction algorithm for the* P*-matrix linear complementarity problem*, manuscript, Department of Management Sciences, The University of Iowa, Iowa City, IA, 1988.
[28] Y. YE AND P. PARDALOS, *A class of linear complementarity problems solvable in polynomial time*, manuscript, Department of Management Sciences, The University of Iowa, Iowa City, IA, 1989.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 4, pp. 625–644, October 1991

A BLACK BOX GENERALIZED CONJUGATE GRADIENT SOLVER WITH INNER ITERATIONS AND VARIABLE-STEP PRECONDITIONING*

O. AXELSSON† AND P. S. VASSILEVSKI‡

Abstract. The generalized conjugate gradient method proposed by Axelsson is studied in the case when a variable-step preconditioning is used. This can be the case when the preconditioned system is solved approximately by an auxiliary (inner) conjugate gradient method, for instance, and the thus-obtained quasi residuals are used to construct the next search vector in the outer generalized cg-iteration method.

A monotone convergence of the method is proved and a rough convergence rate estimate is derived, provided the variable-step preconditioner (generally, a nonlinear mapping) satisfies a continuity and a coercivity assumption.

These assumptions are verified for application of the method for two-level grids and indefinite problems. This variable-step preconditioning involves, for the two-level case, the solution of the coarse grid problem and problems for the nodes on the rest of the grid—both by auxiliary (inner) iterative methods. For the indefinite problems that are considered, the special block structure of the matrix is utilized—also in an outer-inner iterative method.

For both the outer and inner iterations, parameter-free preconditioned generalized conjugate gradient methods are advocated. For indefinite problems the method used offers an alternative to the well-known Uzawa algorithm.

Key words. generalized conjugate gradient method, variable-step preconditioning, two-level method, two-grid method, indefinite problems

AMS(MOS) subject classifications. 65F10, 65N20, 65N30

1. Introduction.

We consider the solution of the system of linear equations,

$$A\mathbf{x} = \mathbf{b} \tag{1.1}$$

by a GCG (generalized conjugate gradient) method, in the form proposed by Axelsson [1] and further developed in [2]. In general, A may be a nonsymmetric and/or indefinite matrix. A may even be a rectangular matrix, if only its column rank is complete.

The GCG method from [1] consists of the following steps.

Given a set of search directions $\{\mathbf{d}^{(s)}\}_{s=0}^{k-1}$ orthogonal with respect to $(\cdot, \cdot)_1$, one computes a new approximation $\mathbf{x}^{(k)}$, such that the quadratic functional

$$f(\mathbf{x}) = \tfrac{1}{2}(\mathbf{r}, \mathbf{r})_0 \tag{1.2}$$

is minimized over the shifted space

$$\mathbf{x}^{(0)} + \text{span}\,\{\mathbf{d}^{(s)}\}_{s=0}^{k-1},$$

where $\mathbf{x}^0$ is an initial approximation, $(\cdot, \cdot)_0$ and $(\cdot, \cdot)_1$ are inner products, and $\mathbf{r} = A\mathbf{x} - \mathbf{b}$ is the residual. Since the column rank of A is complete, there exists a unique minimizer of (1.2) on any space of vectors of dimension m, the column rank of A.

* Received by the editors July 10, 1989; accepted for publication (in revised form) May 18, 1990.

† Faculty of Mathematics and Informatics, Catholic University, Toernooiveld, 6525 ED Nijmegen, the Netherlands.

‡ Institute of Mathematics and Center of Informatics and Computer Technology, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria. The research of this author was performed while the author was visiting the Catholic University, Nijmegen, the Netherlands. This research was partly supported by Bulgarian Committee of Science grant 55, 26-3-1987 and by Stichting Mathematisch Centrum, Amsterdam.

Determining in this way the next approximation

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \sum_{s=1}^{k} \alpha_{k-s}^{(k-1)} \mathbf{d}^{(k-s)},$$

and the next residual

$$\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} + \sum_{s=1}^{k} \alpha_{k-s}^{(k-1)} A\mathbf{d}^{(k-s)},$$

in order to accelerate the convergence one uses a preconditioning step, i.e., one computes by some procedure corresponding to a matrix B, called preconditioner to A, the vector or pseudoresidual,

$$\tilde{\mathbf{r}}^{(k)} = B\mathbf{r}^{(k)}. \tag{1.3}$$

Then the next search vector is defined by

$$\mathbf{d}^{(k)} = -\tilde{\mathbf{r}}^{(k)} + \sum_{s=1}^{k} \beta_{k-s}^{(k-1)} \mathbf{d}^{(k-s)}.$$

The coefficients $\beta_{k-s}^{(k-1)}$ are determined from the orthogonality conditions

$$(\mathbf{d}^{(k)}, \mathbf{d}^{(k-s)})_1 = 0, \qquad s = 1, 2, \cdots, k.$$

As is readily seen, this approach is quite general and can be used for an arbitrary mapping B,

$$\mathbf{r} \to B[\mathbf{r}]. \tag{1.4}$$

In practice B is chosen to approximate the inverse of A, if this exists, or at any rate such that BA is sufficiently close to the identity operator. In the general case when B is a nonlinear mapping, we shall assume a certain coercivity and boundedness condition that generalizes this matrix property.

In the literature, various iterative methods with inner-outer iterations have been considered, e.g., by Axelsson [3]; Golub and Overton [12]; Bank, Welfert, and Yserentant [10]; and Verfürth [15]. However, as an outer iterative method, they used a stationary iterative method, i.e., not a conjugate gradient method.

The algebraic multilevel method considered by Axelsson and Vassilevski [5], [6] can also be seen as an inner-outer iterative method. Here the inner iterations on a given discretization level correspond to the approximate solution of the coarse-grid problem in the two-level grid context of the method, by a Chebyshev iterative method and to two problems for the nodes not lying on the coarse grid, also solved by an iterative method. However, this method is parameter-dependent, i.e., certain parameters required in the iterative method must be estimated.

In this paper we analyse the GCG method in the general case of variable-step (i.e., generally a nonlinear mapping) preconditioner $B[\cdot]$ under the following assumptions:

(i) coercivity, i.e., there exists a positive constant δ_1, such that

$$(AB[\mathbf{v}], \mathbf{v})_0 \geqq \delta_1 (\mathbf{v}, \mathbf{v})_0, \quad \text{all } \mathbf{v}, \tag{1.5}$$

(ii) continuity, i.e., there exists a positive constant δ_2, such that

$$\|AB[\mathbf{v}]\|_0 \leqq \delta_2 \|\mathbf{v}\|_0, \quad \text{all } \mathbf{v}.$$

Under these assumptions we prove in § 2 that the GCG method converges monotonically and at least with a rate given by the inequality,

$$\|\mathbf{r}^{(k)}\|_0 \leqq \sqrt{1-(\delta_1/\delta_2)^2}\,\|\mathbf{r}^{(k-1)}\|_0.$$

These results are based on already-proven similar results, say, in the linear case (i.e., a fixed matrix as a preconditioner) in Axelsson [1].

In § 3 we express these conditions as algebraic conditions and verify them in § 4 in the context of the two-level nonsymmetric preconditioning method, studied in the symmetric case in Axelsson and Gustafsson [4], with corresponding variable-step preconditioners. By the theory presented in §§ 2 and 3, we thereby give a mathematical justification of the numerical experiments presented in Axelsson and Gustafsson [4], when the preconditioned conjugate gradient (PCG) method is used as an inner iterative method, to solve the systems of equations corresponding to the nodes on the finer level, not lying on the coarse grid.

In § 4 we also demonstrate the algebraic conditions for indefinite matrices on a common block form. For indefinite problems, our method offers an alternative to the Uzawa algorithm, used, for instance, in Verfürth [15] and Langer and Queck [13].

2. The generalized conjugate gradient method with variable-step preconditioning. Following an earlier presentation of Axelsson [1], the GCG method with variable-step preconditioning is defined as follows.

Let $\mathbf{x}^{(0)}$ be an initial approximation, $\mathbf{r}^{(0)} = A\mathbf{x}^{(0)} - \mathbf{b}$ the initial residual, $\tilde{\mathbf{r}}^{(0)} = B[\mathbf{r}^{(0)}]$ a corresponding pseudoresidual, and $\mathbf{d}^{(0)} = -\tilde{\mathbf{r}}^{(0)}$ an initial search vector. For $k = 1, 2, \cdots$, let

$$\{\mathbf{d}^{(s)}\}_{s=0}^{k-1}$$

be $(\cdot, \cdot)_1$-orthogonal search vectors. Then the next approximation

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \sum_{s=1}^{k} \alpha_{k-s}^{(k-1)} \mathbf{d}^{(k-s)}$$

is determined from

$$\frac{\partial}{\partial \alpha_s^{(k-1)}} \varphi = 0, \qquad s = 0, 1, \cdots, k-1, \tag{2.1}$$

where

$$\varphi = \varphi(\alpha_0^{(k-1)}, \cdots, \alpha_{k-1}^{(k-1)}) = \tfrac{1}{2}(\mathbf{r}^{(k)}, \mathbf{r}^{(k)})_0,$$

and

$$\begin{aligned} \mathbf{r}^{(k)} &= A\mathbf{x}^{(k)} - \mathbf{b} \\ &= \mathbf{r}^{(k-1)} + \sum_{s=1}^{k} \alpha_{k-s}^{(k-1)} A\mathbf{d}^{(k-s)}. \end{aligned} \tag{2.2}$$

We have the following lemma.

LEMMA 2.1. (a) $(\mathbf{r}^{(k)}, A\mathbf{d}^{(s)})_0 = 0$, $s = 0, 1, \cdots, k-1$;

(b) $\Lambda^{(k)}\alpha^{(k)} = \gamma^{(k)}$, *where* $\Lambda^{(k)}$ *is the matrix with entries*

$$\Lambda_{i,j}^{(k)} = (A\mathbf{d}^{(k-l)}, A\mathbf{d}^{(k-j)})_0, \qquad 1 \leqq l \leqq k, 1 \leqq j \leqq k,$$

$$(\alpha^{(k)})_j = \alpha_{k-j}^{(k-1)}, \qquad j = 1, \cdots, k,$$

and

$$(\gamma^{(k)})_1 = -(\mathbf{r}^{(k-1)}, A\mathbf{d}^{(k-1)})_0, \qquad (\gamma^{(k)})_j = 0, \qquad j = 2, 3, \cdots, k.$$

Proof. (See [1]. As it is short, we present it here also.) Equations (2.1) and (2.2) give

$$0 = \frac{\partial}{\partial \alpha_s^{(k-1)}} \varphi = (\mathbf{r}^{(k)}, A\mathbf{d}^{(s)})_0 = 0, \qquad s = 0, 1, \cdots, k-1,$$

or

$$\sum_{j=1}^{k} \alpha_{k-j}^{(k-1)} (A\mathbf{d}^{(k-j)}, A\mathbf{d}^{(k-1)})_0 = -(\mathbf{r}^{(k-1)}, A\mathbf{d}^{(k-l)})_0, \qquad l = 1, \cdots, k,$$

which proves part (a) and also part (b), using an induction hypothesis. □

The inner products $(\cdot, \cdot)_0$ and $(\cdot, \cdot)_1$ can be chosen independently of each other. For any pair of inner products, $\Lambda^{(k)}$ is nonsingular. However, for practical reasons we shall here consider two special cases.

Case 1. $(\mathbf{u}, \mathbf{v})_1 = (A\mathbf{u}, A\mathbf{v})_0$.

Case 2. $(\mathbf{u}, \mathbf{v})_1 = (\mathbf{u}, \mathbf{v})_0$.

LEMMA 2.2. (a) *In Case* 1, $(\mathbf{u}, \mathbf{v})_1 = (A\mathbf{u}, A\mathbf{v})_0$, *we have that* $\Lambda^{(k)}$ *is diagonal and*

$$\begin{aligned} &\alpha_s^{(k-1)} = 0, \qquad s = 0, 1, \cdots, k-2, \\ &\alpha_{k-1}^{(k-1)} = -(\mathbf{r}^{(k-1)}, A\mathbf{d}^{(k-1)})_0 / (\mathbf{d}^{(k-1)}, \mathbf{d}^{(k-1)})_1. \end{aligned} \tag{2.3}$$

(b) *In Case* 2, $(\mathbf{u}, \mathbf{v})_1 = (\mathbf{u}, \mathbf{v})_0$, *we have that* $\Lambda^{(k)}$ *equals* $\Lambda^{(k-1)}$ *augmented with a row and a column.*

(c) $\Lambda^{(k)}$ *is symmetric and positive definite.*

Proof (see [1]). (a) In Case 1, the $(\cdot, \cdot)_1$ orthogonality of $\mathbf{d}^{(s)}$, $s = 0, 1, \cdots, k-1$, shows that

$$(A\mathbf{d}^{(k-j)}, A\mathbf{d}^{(k-l)})_0 = (\mathbf{d}^{(k-j)}, \mathbf{d}^{(k-l)})_1 = 0, \qquad j \neq l, \qquad j, l = 1, 2, \cdots, k.$$

Part (a) then follows from part (b) of Lemma 2.1. For any pair of inner products the orthogonality of $\{\mathbf{d}^{(s)}\}$ implies in particular that this vector set is linearly independent. Since A has complete column rank, the set $\{A\mathbf{d}^{(s)}\}_{s=0}^{k-1}$ is also linearly independent, so $\Lambda^{(k)}$ is nonsingular. The last parts of the statement follow by construction of $\Lambda^{(k)}$ and by the linear independence of the set $\{\mathbf{d}^{(s)}\}$. □

At the kth step of the GCG method, the new search direction is defined by

$$\mathbf{d}^{(k)} = -\tilde{\mathbf{r}}^{(k)} + \sum_{s=1}^{k} \beta_{k-s}^{(k-1)} \mathbf{d}^{(k-s)}, \tag{2.4}$$

where

$$\tilde{\mathbf{r}}^{(k)} = B[\mathbf{r}^{(k)}] \tag{2.5}$$

and the parameters $\{\beta_{k-s}^{(k-1)}\}_{s=1}^{k}$ are determined from the orthogonality conditions

$$(\mathbf{d}^{(k)}, \mathbf{d}^{(j)})_1 = 0, \qquad j = 0, 1, \cdots, k-1,$$

i.e.,

$$\beta_j^{(k-1)} = \frac{(\tilde{\mathbf{r}}^{(k)}, \mathbf{d}^{(j)})_1}{(\mathbf{d}^{(j)}, \mathbf{d}^{(j)})_1}, \qquad j = 0, 1, \cdots, k-1. \tag{2.6}$$

LEMMA 2.3. (a) $(\mathbf{r}^{(k-1)}, A\mathbf{d}^{(k-1)})_0 = -(\mathbf{r}^{(k-1)}, AB[\mathbf{r}^{(k-1)}])_0$,
(b) $\alpha_{k-1}^{(k-1)} = (\mathbf{r}^{(k-1)}, AB[\mathbf{r}^{(k-1)}])_0 \det(\Lambda^{(k-1)})/\det(\Lambda^{(k)})$,
(c) $\alpha_{k-1}^{(k-1)} > 0$ *if and only if* $(\mathbf{r}^{(k-1)}, AB[\mathbf{r}^{(k-1)}])_0 > 0$.
(Note that in Case 1, the expression in (b) can be further simplified, as shown in Lemma 2.2(a).)

Proof (see [1]). Equation (2.4) shows that

$$(\mathbf{r}^{(k-1)}, A\mathbf{d}^{(k-1)})_0 = -(\mathbf{r}^{(k-1)}, A\tilde{\mathbf{r}}^{(k-1)})_0 + \sum_{s=2}^{k} \beta_{k-s}^{(k-2)} (\mathbf{r}^{(k-1)}, A\mathbf{d}^{(k-s)})_0.$$

Part (a) follows now from Lemma 2.1(a) and (2.5). Part (b) follows from Lemma 2.2 and Cramer's rule. Since $\Lambda^{(k-1)}$ is a Gramian-type matrix, its determinant is positive, so part (c) follows directly from part (b). □

Lemmata 2.1, 2.2, and 2.3 show now that one GCG step of the algorithms in Cases 1 and 2, respectively, takes the following forms (in practice, we frequently let $(\mathbf{u}, \mathbf{v})_0 = \mathbf{u}^t\mathbf{v}$).

ALGORITHM 1.
Compute $A\mathbf{d}^{(k-1)}$
Compute $(A\mathbf{d}^{(k-1)}, A\mathbf{d}^{(k-1)})_0$
Compute $(\mathbf{r}^{(k-1)}, A\mathbf{d}^{(k-1)})_0$
$\alpha_{k-1}^{(k-1)} = -(\mathbf{r}^{(k-1)}, A\mathbf{d}^{(k-1)})_0/(A\mathbf{d}^{(k-1)}, A\mathbf{d}^{(k-1)})_0$
$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \alpha_{k-1}^{(k-1)}\mathbf{d}^{(k-1)}$
$\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} + \alpha_{k-1}^{(k-1)} A\mathbf{d}^{(k-1)}$
Compute $\tilde{\mathbf{r}}^{(k)} = B[\mathbf{r}^{(k)}]$
Compute $A\tilde{\mathbf{r}}^{(k)}$
Compute $\beta_j^{(k-1)} = (A\tilde{\mathbf{r}}^{(k)}, A\mathbf{d}^{(j)})_0/(A\mathbf{d}^{(j)}, A\mathbf{d}^{(j)})_0$, $\quad j = 0, \cdots, k-1$
$\mathbf{d}^{(k)} = -\tilde{\mathbf{r}}^{(k)} + \sum_{s=1}^{k} \beta_{k-s}^{(k-1)}\mathbf{d}^{(k-s)}$

ALGORITHM 2.
Compute $A\mathbf{d}^{(k-1)}$
Compute $(A\mathbf{d}^{(k-1)}, A\mathbf{d}^{(k-j)})_0$, $\quad j = 1, \cdots, k$
Compute $(\mathbf{r}^{(k-1)}, A\mathbf{d}^{(k-1)})_0$
Solve $\Lambda^{(k)}\alpha^{(k)} = \gamma^{(k)}$
$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \sum_{s=1}^{k} \alpha_{k-s}^{(k-1)}\mathbf{d}^{(k-s)}$
$\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)} + \sum_{s=1}^{k} \alpha_{k-s}^{(k-1)} A\mathbf{d}^{(k-s)}$ (or $\mathbf{r}^{(k)} = A\mathbf{x}^{(k)} - \mathbf{b}$)
Compute $\tilde{\mathbf{r}}^{(k)} = B[\mathbf{r}^{(k)}]$
Compute $\beta_j^{(k-1)} = (\tilde{\mathbf{r}}^{(k)}, \mathbf{d}^{(j)})_0/(\mathbf{d}^{(j)}, \mathbf{d}^{(j)})_0$, $\quad j = 0, \cdots, k-1$
$\mathbf{d}^{(k)} = -\tilde{\mathbf{r}}^{(k)} + \sum_{s=1}^{k} \beta_{k-s}^{(k-1)}\mathbf{d}^{(k-s)}$

Note that in Algorithm 1 we need two multiplications with the matrix A, while in Algorithm 2 only one such multiplication is required. On the other hand, Algorithm 2 requires $k-1$ more inner products and $2(k-1)$ more vector updates per iteration step. Hence, if the number of iterations (k) is sufficiently small or if the cost of a matrix multiplication with A is sufficiently big, Algorithm 2 can be more efficient than Algorithm 1.

We now estimate the rate of convergence of the algorithms.

THEOREM 2.1. *Let the preconditioner* $B[\cdot]$ *satisfy the assumptions* (i) *and* (ii), *i.e.*,
(i) $(\mathbf{v}, AB[\mathbf{v}])_0 \geqq \delta_1(\mathbf{v}, \mathbf{v})_0$, *all* $\mathbf{v}$;
(ii) $\|AB[\mathbf{v}]\|_0 \leqq \delta_2\|\mathbf{v}\|_0$, *all* $\mathbf{v}$, *for some positive constants* δ_1, δ_2. *Then the variable-step* GCG *method converges monotonically and the following convergence rate estimate is valid*:

$$\|\mathbf{r}^{(k)}\|_0 \leqq \sqrt{1-(\delta_1/\delta_2)^2}\,\|\mathbf{r}^{(k-1)}\|_0, \qquad k = 1, 2, \cdots.$$

Proof (see Theorem 2.2 in [1]). Lemma 2.1 shows that

$$(\mathbf{r}^{(k)},\mathbf{r}^{(k)})_0=(\mathbf{r}^{(k)},\mathbf{r}^{(k-1)}+\sum_{j=1}^{k}\alpha_{k-j}^{(k-1)}A\mathbf{d}^{(k-j)})_0=(\mathbf{r}^{(k)},\mathbf{r}^{(k-1)})_0$$

$$=(\mathbf{r}^{(k-1)}+\sum_{j=1}^{k}\alpha_{k-j}^{(k-1)}A\mathbf{d}^{(k-j)},\mathbf{r}^{(k-1)})_0$$

$$=(\mathbf{r}^{(k-1)},\mathbf{r}^{(k-1)})_0+\alpha_{k-1}^{(k-1)}(\mathbf{r}^{(k-1)},A\mathbf{d}^{(k-1)})_0.$$

Hence Lemma 2.3 shows that

$$(\mathbf{r}^{(k)},\mathbf{r}^{(k)})_0=(\mathbf{r}^{(k-1)},\mathbf{r}^{(k-1)})_0-(\mathbf{r}^{(k-1)},AB[\mathbf{r}^{(k-1)}])_0^2\det(\Lambda^{(k-1)})/\det(\Lambda^{(k)}).$$

It is shown in [1] that

$$\det(\Lambda^{(k)})/\det(\Lambda^{(k-1)})=\min_{\mathbf{g}\in W_{k-2}}\|AB[\mathbf{r}^{(k-1)}]-\mathbf{g}\|_0^2 \tag{2.7}$$

where W_{k-2} is the vectorspace spanned by $\{A\mathbf{d}^{(s)}\}_{s=0}^{k-2}$. Simply letting $\mathbf{g}=0$ in (2.7), we get the upperbound

$$(\mathbf{r}^{(k)},\mathbf{r}^{(k)})_0\leqq(\mathbf{r}^{(k-1)},\mathbf{r}^{(k-1)})_0-((\mathbf{r}^{(k-1)},AB[\mathbf{r}^{(k-1)}])_0^2/\|AB[\mathbf{r}^{(k-1)}]\|_0)^2.$$

Assumptions (i) and (ii) now show Theorem 2.1. □

Remark 2.1. Note that Theorem 2.1 also holds for a variable-step preconditioner, i.e., the preconditioner can change from one step to the next. In fact, it is readily seen that the rate of convergence estimate in Theorem 2.1 can be derived even for the steepest descent algorithm where $\alpha_s^{(k-1)}=0$, $s=0,1,\cdots,k-2$, and $\beta_s^{(k-1)}=0$, $s=0,1,\cdots,k-1$.

Remark 2.2. If $\mathbf{d}^{(k)}=0$, for some k, then it follows, by (2.4), that $\tilde{\mathbf{r}}^{(k)}$ is a linear combination of $\{\mathbf{d}^{(k-s)}\}_{s=1}^{k}$. Then

$$(\mathbf{r}^{(k)},A\tilde{\mathbf{r}}^{(k)})_0=\sum_{s=1}^{k}\beta_{k-s}^{(k-1)}(\mathbf{r}^{(k)},A\mathbf{d}^{(k-s)})_0=0,$$

by Lemma 2.1. By the coercivity assumption (i), we then have,

$$0=(\mathbf{r}^{(k)},A\tilde{\mathbf{r}}^{(k)})_0=(\mathbf{r}^{(k)},AB[\mathbf{r}^{(k)}])_0\geqq\delta_1\|\mathbf{r}^{(k)}\|_0^2.$$

Hence $\mathbf{r}^{(k)}=0$, i.e., the problem has been solved.
Thus we proved the following result.

THEOREM 2.2. *If the preconditioner* $B[\cdot]$ *satisfies the coercivity assumption* (i), *then the* (*variable-step*) *preconditioned* GCG *method with this preconditioner cannot fail.*

Note finally that even if A is indefinite, for instance, there can still exist a mapping $B[\cdot]$ for which the coercivity and boundedness assumptions hold.

Remark 2.3. When applying the GCG method there is a simple way to automatically determine if the preconditioner is sufficiently accurate. We simply check the sign of $(\mathbf{r}^{(k-1)}, AB[\mathbf{r}^{(k-1)}])_0$, which by Lemma 2.3(a) equals $-(\mathbf{r}^{(k-1)}, A\mathbf{d}^{(k-1)})_0$. Equivalently, we can check the sign of $\alpha_{k-1}^{(k-1)}$. If this is negative, we restart the algorithm without updating the approximation at the last step and compute in the following iterations a more accurate preconditioner B by making the inner iteration parameters ε_1, ε_2 (see § 3) smaller, or by simply performing more inner iterations. This corresponds to one form of a variable-step preconditioning and makes the algorithm a "black box" solver.

3. Verification of coercivity and continuity assumptions. We consider here two important types of problems where matrices on a two-by-two block form naturally arise. Hence, consider a matrix A, partitioned as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}. \tag{3.1}$$

A can be symmetric or nonsymmetric, but in the case when A is indefinite we assume here that A is symmetric.

We want to solve

$$A\mathbf{x} = \mathbf{b}$$

or

$$\begin{aligned} A_{11}\mathbf{x}_1 + A_{12}\mathbf{x}_2 &= \mathbf{b}_1 \\ A_{21}\mathbf{x}_1 + A_{22}\mathbf{x}_2 &= \mathbf{b}_2. \end{aligned} \tag{3.2}$$

We shall assume that A_{11} is invertible (in fact even with a positive-definite symmetric part, $\frac{1}{2}(A_{11} + A_{11}^T)$). However, in the case where $A_{22} = 0$ and A_{11} is singular or indefinite (a case occurring frequently in constrained optimization problems), we consider the equivalent system (that is, with the same solution),

$$\begin{aligned} \left(A_{11} + \frac{1}{\varepsilon} A_{12} A_{12}^T\right)\mathbf{x}_1 + A_{12}\mathbf{x}_2 &= \mathbf{b}_1 + \frac{1}{\varepsilon} A_{12}\mathbf{b}_2, \\ A_{12}^T \mathbf{x}_1 &= \mathbf{b}_2, \end{aligned}$$

and we assume then that, for some $\varepsilon > 0$, $A_{11} + (1/\varepsilon)A_{12}^T A_{12}$ has a positive-definite symmetric part. Hence, we must assume that A_{11} is positive definite on the nullspace of A_{12}^T. Therefore, we might as well assume that A_{11} is positive definite from the onset.

We assume also that

$$S = A_{22} - A_{21}A_{11}^{-1}A_{12}$$

is definite (positive or negative) and that A_{22} is positive definite if S is positive definite and A_{22} is negative semidefinite if S is negative definite.

This means that A_{22} is definite (positive or negative) on the nullspace of A_{12}.

Next we consider the following exact block-form of the inverse of A (see, for an earlier derivation, Banachiewicz [8]), which is readily derived by inverting the block matrix factorization of A,

$$A = \begin{bmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} I & A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix}$$

and hence

$$\begin{aligned} A^{-1} &= \begin{bmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{bmatrix} \\ &= \begin{bmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11}^{-1} & 0 \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{bmatrix}. \end{aligned}$$

Note that the application of this form to compute $A^{-1}\mathbf{v}$ for any block vector

$$\mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix},$$

involves the following steps.

ALGORITHM 3.

(1) $\mathbf{w}_1 = A_{11}^{-1}\mathbf{v}_1$;
(2) $\mathbf{w}_2 = -A_{21}\mathbf{w}_1 + \mathbf{v}_2$;
(3) $\mathbf{x}_2 = S^{-1}\mathbf{w}_2$;
(4) $\mathbf{y}_1 = A_{12}\mathbf{x}_2$;
(5) $\mathbf{z}_1 = A_{11}^{-1}\mathbf{y}_1$;
(6) $\mathbf{x}_1 = \mathbf{w}_1 - \mathbf{z}_1$.

Then

$$A^{-1}\mathbf{v} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}.$$

The preconditioner, approximating A^{-1}, is now defined as follows. Every occurrence of the inverse of A_{11} (i.e., steps (1) and (5) above) is replaced by an (inner) iterative method to solve the corresponding systems with A_{11}, i.e., $A_{11}\mathbf{w}_1 = \mathbf{v}_1$ and $A_{11}\mathbf{z}_1 = \mathbf{y}_1$, above. Likewise, the occurrence of S^{-1} in (3) is replaced by an (inner) iterative method, i.e., to solve $S\mathbf{x}_2 = \mathbf{w}_2$ approximately. In all cases we iterate until the iteration error is sufficiently small.

In order to define preconditioner B on each (outer iteration) step we need (in general, nonlinear) mappings

$$\mathbf{v}_1 \to B_{11}[\mathbf{v}_1], \qquad \mathbf{v}_2 \to C[\mathbf{v}_2]$$

such that

$$\|A_{11}B_{11}[\mathbf{v}_1] - \mathbf{v}_1\|_0 \leqq \varepsilon_1 \|\mathbf{v}_1\|_0, \quad \text{all } \mathbf{v}_1 \tag{3.3a}$$

$$\|SC[\mathbf{v}_2] - \mathbf{v}_2\|_0 \leqq \varepsilon_2 \|\mathbf{v}_2\|_0, \quad \text{all } \mathbf{v}_2, \tag{3.3b}$$

and ε_1, ε_2 are sufficiently small positive numbers.

The application of the (variable-step) preconditioner $B = B[\,\cdot\,]$ involves, therefore, the following steps.

ALGORITHM 4.

(1) $\mathbf{w}_1 = B_{11}[\mathbf{v}_1]$;
(2) $\mathbf{w}_2 = -A_{21}\mathbf{w}_1 + \mathbf{v}_2$;
(3) $\mathbf{x}_2 = C[\mathbf{w}_2]$;
(4) $\mathbf{y}_1 = A_{12}\mathbf{x}_2$;
(5) $\mathbf{z}_1 = B_{11}[\mathbf{y}_1]$;
(6) $\mathbf{x}_1 = \mathbf{w}_1 - \mathbf{z}_1$.

Then

$$B[\mathbf{v}] = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}.$$

We shall now estimate the deviation of $AB[\mathbf{v}]$ from $\mathbf{v}$. Note first, then, that Algorithm 4 shows that

$$
\begin{aligned}
AB[\mathbf{v}] &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} A_{11}(\mathbf{w}_1 - \mathbf{z}_1) + A_{12}C[\mathbf{w}_2] \\ A_{21}(\mathbf{w}_1 - \mathbf{z}_1) + A_{22}C[\mathbf{w}_2] \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} + \begin{bmatrix} A_{11}(\mathbf{w}_1 - B_{11}[\mathbf{y}_1]) - \mathbf{v}_1 + A_{12}C[\mathbf{w}_2] \\ A_{21}(\mathbf{w}_1 - B_{11}[\mathbf{y}_1]) + A_{22}C[\mathbf{w}_2] - \mathbf{v}_2 \end{bmatrix} \qquad (3.4) \\
&= \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} + \begin{bmatrix} (A_{11}\mathbf{w}_1 - \mathbf{v}_1) - (A_{11}\mathbf{z}_1 - \mathbf{y}_1) \\ A_{22}C[\mathbf{w}_2] - \mathbf{w}_2 - A_{21}B_{11}[\mathbf{y}_1] \end{bmatrix}.
\end{aligned}
$$

Therefore, since $C[\mathbf{w}_2] = \mathbf{x}_2$ and $B_{11}[\mathbf{y}_1] = \mathbf{z}_1$, the application of the preconditioner on each (outer iteration) step can be realized as Algorithm 4′.

ALGORITHM 4′. Given sufficiently small positive numbers $\varepsilon_1, \varepsilon_2$, we iterate in steps (1) and (5) with some method until the iterations $\mathbf{w}_1, \mathbf{z}_1$ satisfy

$$
\|A_{11}\mathbf{w}_1 - \mathbf{v}_1\|_0 \leqq \varepsilon_1 \|\mathbf{v}_1\|_0, \qquad \|A_{11}\mathbf{z}_1 - \mathbf{y}_1\|_0 \leqq \varepsilon_1 \|\mathbf{y}_1\|_0, \tag{3.5a}
$$

and in step (3) until the iteration $\mathbf{x}_2$ satisfies

$$
\|A_{22}\mathbf{x}_2 - A_{21}\mathbf{z}_1 - \mathbf{w}_2\|_0 \leqq \varepsilon_2 \|\mathbf{w}_2\|_0 \tag{3.5b}
$$

where

$$
\mathbf{w}_2 = \mathbf{v}_2 - A_{21}\mathbf{w}_1.
$$

Since (3.5b) involves the computation of $\mathbf{z}_1$ in step (5), the test (3.5b) is actually performed after step (5). This means that we may have to repeat steps (4) and (5) if (3.5b) fails to be satisfied. Hence, in practice, it can be advisable to choose a certain (fixed) number of iterations in step (3) and test on the sign of $\alpha_{k-1}^{(k-1)}$, instead, as was already mentioned in Remark 2.2. If the sign test is violated, we repeat Algorithm 4 with smaller values of $\varepsilon_1, \varepsilon_2$.

To continue the estimate of $AB[\mathbf{v}] - \mathbf{v}$ in (3.4), note first that by (3.5a),

$$
\|A_{11}B_{11}[\mathbf{v}_1] - \mathbf{v}_1\|_0 = \|A_{11}\mathbf{w}_1 - \mathbf{v}_1\|_0 \leqq \varepsilon_1 \|\mathbf{v}_1\|_0, \tag{3.6}
$$

and hence that

$$
\begin{aligned}
\|\mathbf{w}_2\|_0 &\leqq \|\mathbf{v}_2\|_0 + \|A_{21}A_{11}^{-1}A_{11}\mathbf{w}_0\|_0 \leqq \|\mathbf{v}_2\|_0 + \|A_{21}A_{11}^{-1}\|_0 \|A_{11}B_{11}[\mathbf{v}_1]\|_0 \\
&\leqq \|\mathbf{v}_2\|_0 + \|A_{21}A_{11}^{-1}\|_0 (1 + \varepsilon_1)\|\mathbf{v}_1\|_0. \qquad (3.7)
\end{aligned}
$$

Next note that

$$
\begin{aligned}
\|SC[\mathbf{w}_2] - \mathbf{w}_2\|_0 &= \|S\mathbf{x}_2 - \mathbf{w}_0\|_0 \\
&= \|A_{22}\mathbf{x}_2 - A_{21}A_{11}^{-1}A_{12}\mathbf{x}_2 - \mathbf{w}_2\|_0 \\
&= \|A_{22}\mathbf{x}_2 - A_{21}B_{11}[\mathbf{y}_1] - \mathbf{w}_2 + A_{21}(B_{11}[\mathbf{y}_1] - A_{11}^{-1}A_{12}\mathbf{x}_2)\|_0 \\
&\quad \times \|A_{22}\mathbf{x}_2 - A_{21}\mathbf{z}_1 - \mathbf{w}_2 + A_{21}A_{11}^{-1}[A_{11}\mathbf{z}_1 - \mathbf{y}_1]\|_0 \\
&\leqq \varepsilon_2 \|\mathbf{w}_2\|_0 + \|A_{21}A_{11}^{-1}\|_0 \varepsilon_1 \|\mathbf{y}_1\|_0 \quad \text{(by (3.5b), (3.5a))}.
\end{aligned}
$$

Further,

$$
\|\mathbf{y}_1\|_0 = \|A_{12}\mathbf{x}_2\|_0 = \|A_{12}S^{-1}SC[\mathbf{w}_2]\| \leqq \|A_{12}S^{-1}\|_0 \|SC[\mathbf{w}_2]\|_0. \tag{3.8}
$$

Hence

$$\|SC[\mathbf{w}_2]-\mathbf{w}_2\|_0 \leqq \varepsilon_2\|\mathbf{w}_2\|_0+\varepsilon_1\|A_{21}A_{11}^{-1}\|_0\|A_{12}S^{-1}\|_0[\|SC[\mathbf{w}_2]-\mathbf{w}_2\|_0+\|\mathbf{w}_2\|_0],$$

so, if ε_1 is sufficiently small,

$$\|SC[\mathbf{w}_2]-\mathbf{w}_2\|_0 \leqq (\varepsilon_2+\varepsilon_1\|A_{21}A_{11}^{-1}\|_0\|A_{12}S^{-1}\|_0)\|\mathbf{w}_2\|/(1-\varepsilon_1\|A_{21}A_{11}^{-1}\|_0\|A_{12}S^{-1}\|_0)$$

and

(3.9)

$$\|SC[\mathbf{w}_2]\|_0 \leqq \|SC[\mathbf{w}_2]-\mathbf{w}_2\|_0+\|\mathbf{w}_2\|_0 \leqq (1+\varepsilon_2)\|\mathbf{w}_2\|_0/(1-\varepsilon_1\|A_{21}A_{11}^{-1}\|_0\|A_{12}S^{-1}\|_0).$$

Therefore, by (3.8) and (3.9)

(3.10)

$$\|A_{11}\mathbf{z}_1-\mathbf{y}_1\|_0 \leqq \varepsilon_1\|\mathbf{y}_1\|_0 \leqq \varepsilon_1\|A_{12}S^{-1}\|_0(1+\varepsilon_2)\|\mathbf{w}_2\|_0/(1-\varepsilon_1\|A_{21}A_{11}^{-1}\|_0\|A_{12}S^{-1}\|_0).$$

Finally (3.4), (3.5a), and (3.10) show that

$$\begin{aligned}\|AB[\mathbf{v}]-\mathbf{v}\|_0^2 &\leqq (\|A_{11}\mathbf{w}_1-\mathbf{v}_1\|_0+\|A_{11}\mathbf{z}_1-\mathbf{y}_1\|_0)^2+\|A_{22}\mathbf{x}_2-A_{21}\mathbf{z}_1-\mathbf{w}_2\|_0^2\\ &\leqq [\varepsilon_1\|\mathbf{v}_1\|_0+\varepsilon_1\|A_{12}S^{-1}\|_0(1+\varepsilon_2)\|\mathbf{w}_2\|_0/(1-\varepsilon_1\|A_{21}A_{11}^{-1}\|_0\|A_{12}S^{-1}\|_0)]^2\\ &\quad+\varepsilon_2^2\|\mathbf{w}_2\|_0^2,\end{aligned}$$

so by (3.7)

$$\|AB[\mathbf{v}]-\mathbf{v}\|_0^2 \leqq (\varepsilon_1+\varepsilon_2)^2C_1(\sigma_1,\sigma_2)\|\mathbf{v}\|_0^2, \tag{3.11}$$

where $\|\mathbf{v}\|_0^2 = \|\mathbf{v}_1\|_0^2 + \|\mathbf{v}_2\|_0^2$, $C_1 = C_1(\sigma_1, \sigma_2)$,

$$\begin{aligned}C_1(\sigma_1,\sigma_2) &= [1+\sigma_1\sigma_2(1+\varepsilon_2)(1+\varepsilon_1)/(1-\varepsilon_1\sigma_1\sigma_2)]^2+2\sigma_2(1+\varepsilon_2)^2\\ &\quad+[1+\sigma_1(1+\varepsilon_2)/(1-\varepsilon_1\sigma_1\sigma_2)]^2\end{aligned}$$

and

$$\sigma_1 = \|A_{12}S^{-1}\|_0, \qquad \sigma_2 = \|A_{21}A_{11}^{-1}\|_0.$$

We summarize the result in the following theorem.

THEOREM 3.1. *Let the norms* $\|\cdot\|_0$ *in the vectorspaces for* $\mathbf{v}_1$, $\mathbf{v}_2$, *respectively, be such that*

$$\sigma_1 = \|A_{12}S^{-1}\|_0, \qquad \sigma_2 = \|A_{21}A_{11}^{-1}\|_0$$

are bounded uniformly with respect to the problem parameter. Then for ε_1, ε_2 *sufficiently small, the mapping* $B[\cdot]$ *defined by Algorithm* 4, *with* $B_{11}[\cdot]$ *and* $C[\cdot]$ *satisfying* (3.3a,b), *is coercive and bounded; that is,*

$$(\mathbf{v}, AB[\mathbf{v}])_0 \geqq [1-C_1^{1/2}(\varepsilon_1+\varepsilon_2)]\|\mathbf{v}\|_0^2, \quad \textit{all } \mathbf{v},$$

where $C_1 = C_1(\sigma_1, \sigma_2)$ *is a function of* σ_1, σ_2, *bounded for all bounded values of* σ_1, σ_2, *and*

$$\|AB[\mathbf{v}]\|_0 \leqq [1+C_1^{1/2}(\varepsilon_1+\varepsilon_2)]\|\mathbf{v}\|_0, \quad \textit{all } \mathbf{v},$$

respectively.

Proof. Equation (3.10) shows that

$$|\,\|AB[\mathbf{v}]\|_0-\|\mathbf{v}\|_0| \leqq \|AB[\mathbf{v}]-\mathbf{v}\|_0 \leqq C_1^{1/2}(\varepsilon_1+\varepsilon_2)\|\mathbf{v}\|_0.$$

Hence

$$[1-C_1^{1/2}(\varepsilon_1+\varepsilon_2)]\|\mathbf{v}\|_0 \leqq \|AB[\mathbf{v}]\|_0 \leqq [1+C_1^{1/2}(\varepsilon_1+\varepsilon_2)]\|\mathbf{v}\|_0. \tag{3.12}$$

Further, (3.11) shows that

$$\|\mathbf{v}\|_0^2 - 2(\mathbf{v}, AB[\mathbf{v}])_0 + \|AB[\mathbf{v}]\|_0^2 \leqq C_1(\varepsilon_1+\varepsilon_2)^2\|\mathbf{v}\|_0^2.$$

Hence

$$2(\mathbf{v}, AB[\mathbf{v}])_0 \geqq [1-C_1(\varepsilon_1+\varepsilon_2)^2]\|\mathbf{v}\|_0^2 + \|AB[\mathbf{v}]\|_0^2,$$

and this, together with the left-hand side part of (3.12), show that

$$2(\mathbf{v}, AB[\mathbf{v}])_0 \geqq [2-2C_1^{1/2}(\varepsilon_1+\varepsilon_2)]\|\mathbf{v}\|_0^2. \qquad \square$$

In the next section we make the corresponding choice of norms $\|\mathbf{v}_1\|_0$, $\|\mathbf{v}_2\|_0$ in two particular and important applications: the two-level multilevel method for nonselfadjoint elliptic problems and a mixed finite element discretization of the Stokes problem or of the second-order elliptic equation.

4. Applications for the two-level multigrid method for nonselfadjoint elliptic problems and for problems arising in mixed finite element solution of elliptic equations.

PROBLEM 4.1. *Consider the following boundary value problem,*

$$-\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right) + \mathbf{v}\cdot\nabla u + bu = f(x), \qquad x \in \Omega \subset \mathbb{R}^2,$$

$u = 0$ *on* $\Gamma = \partial\Omega$.

Here Ω is a polygonal domain and the matrix $[a_{ij}(x)]_{i,j=1}^2$ is assumed to be symmetric and uniformly positive definite on $x \in \bar{\Omega}$.

The form

$$a(u,w) = \int_\Omega \sum a_{ij}(x)\frac{\partial u}{\partial x_i}\frac{\partial w}{\partial x_j}\,dx + \int_\Omega [(\mathbf{v}\cdot\nabla u)w + buw]\,dx$$

is assumed to be H^1-coercive, that is,

$$\begin{aligned} a(u,u) &= \int_\Omega \sum a_{ij}(x)\frac{\partial u}{\partial x_i}\frac{\partial u}{\partial x_j}\,dx + \int_\Omega [b - \tfrac{1}{2}\operatorname{div}\mathbf{v}]u^2\,dx \\ &\geqq c_0|u|_{1,\Omega}^2 + b_0|u|_{0,\Omega}^2. \end{aligned}$$

for some $c_0 > 0$, $b_0 \geqq 0$.

To satisfy this, it suffices to have

$$b(x) - \tfrac{1}{2}\operatorname{div}\mathbf{v}(x) \geqq b_0 \geqq 0.$$

The standard variational formulation of this problem is:

Find $u \in H_0^1(\Omega)$ such that

$$a(u,\phi) = (f,\phi), \qquad \text{all } \phi \in H_0^1(\Omega).$$

Since the form $a(\cdot,\cdot)$ is H^1-coercive, this problem has a unique solution $u \in H_0^1(\Omega)$ for any right-hand side function $f \in L_2(\Omega)$.

Also, we shall need the bilinear form $\mathring{a}(\cdot, \cdot)$, the symmetric part of $a(\cdot, \cdot)$, defined by

$$\mathring{a}(u,\phi)=\int_\Omega \sum a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial \phi}{\partial x_j}\,dx+\int_\Omega [b-\tfrac{1}{2}\operatorname{div} \mathbf{v}]u\phi\,dx. \tag{4.1}$$

Consider now a finite element space V split up into two spaces V_1, V_2 such that

$$V=V_1+V_2, \qquad V_1\cap V_2=\{0\}.$$

The following are examples of such partitionings.

Example 4.1. Let τ_2 be a triangulation of Ω, consisting of a set of nonoverlapping triangles. Let

$$V_2=\operatorname{span}\{\phi_i^{(2)}\}_{i=1}^{n_2},$$

where n_2 is the number of vertices in τ_2 not lying on Γ_D and where $\phi_i^{(2)}$ is piecewise linear on the triangles in τ_2,

$$\phi_i^{(2)}(x_j^{(2)})=\delta_{i,j},$$

and $x_j^{(2)}$ runs over all vertices of the triangles in τ_2. By a refining procedure, e.g., by bisection or by pairwise connecting the center points of the edges of the triangles (see Fig. 4.1), we get a finer triangulation τ_1. Then V_1 is defined by

$$V_1=\operatorname{span}\{\phi_i^{(1)}\}_{i=1}^{n_1},$$

where $\phi_i^{(1)}$ forms a nodal basis in V_1 and are piecewise linear on the triangles in τ_2 and vanish on the vertices of the triangles in τ_1 (except on the ith).

FIG. 4.1

Example 4.2. Let τ_2 be a triangulation of Ω, as in Example 4.1; let V_2 be defined in the same way; and let V_1 be the set of continuous functions, which are piecewise polynomials of degree p in each triangle, vanishing on the vertices of the triangles in τ_2 and spanning the complete monomials up to degree p, except 1, x_1, and x_2. Here p is a fixed integer greater than 1.

Example 4.3. Assume that Ω can be divided into a set of rectangular elements. To the vertices of the elements we associate piecewise bilinear functions, which span the space V_2. V_1 is spanned by the corresponding serendipity piecewise polynomials of degree less than or equal to p, which vanish on the vertices of the elements (see Fig. 4.2).

In all these examples we have two disjoint sets of nodes N_2, N_1, such that

$$V_1=\{\phi\in V \text{ and } \phi(x_j^{(2)})=0, x_j^{(2)}\in N_2\}.$$

Using this block ordering of the nodes, namely first ordering nodes in N_1 and then in N_2, we get the following two-by-two block form of the stiffness matrix A:

$$A=\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \tag{4.2}$$

where

$$A_{11} = \{a(\phi_j^{(1)}, \phi_i^{(1)})\}_{x_i, x_j \in N_1},$$

$$A_{21} = \{a(\phi_j^{(1)}, \phi_i^{(2)})\}_{x_j \in N_1, x_i \in N_2}, \qquad A_{12} = \{a(\phi_j^{(2)}, \phi_i^{(1)})\}_{x_i \in N_1, x_j \in N_2},$$

and

$$A_{22} = \{a(\phi_j^{(2)}, \phi_i^{(2)})\}_{x_i, x_j \in N_2}.$$

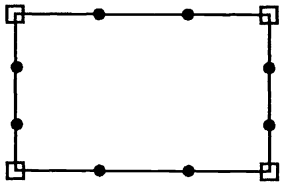


FIG. 4.2

The symmetric part of A, $\mathring{A} = \frac{1}{2}(A + A^T)$, is obtained from the bilinear form $\mathring{a}(\cdot, \cdot)$ defined by (4.1); that is,

$$\mathring{A} = \begin{bmatrix} \mathring{A}_{11} & \mathring{A}_{12} \\ \mathring{A}_{21} & \mathring{A}_{22} \end{bmatrix} \tag{4.3}$$

with

$$\mathring{A}_{11} = \{\mathring{a}(\phi_j^{(1)}, \phi_i^{(1)})\}_{x_i, x_j \in N_1},$$

$$\mathring{A}_{12}^T = \mathring{A}_{21} = \{\mathring{a}(\phi_j^{(1)}, \phi_i^{(2)})\}_{x_j \in N_1, x_i \in N_2},$$

and

$$\mathring{A}_{22} = \{\mathring{a}(\phi_j^{(2)}, \phi_i^{(2)})\}_{x_i, x_j \in N_2}.$$

The following strengthened Cauchy inequality, proved in Bank and Dupont [9] and Axelsson and Gustafsson [4], will be used later:

There exists a constant $\gamma \in (0, 1)$, independent of the mesh parameter (but dependent on p), such that

$$\mathbf{v}_1^t \mathring{A}_{12} \mathbf{v}_2 \leqq \gamma \{\mathbf{v}^t \mathring{A}_{11} \mathbf{v}\}^{1/2} \{\mathbf{v}_2^t \mathring{A}_{22} \mathbf{v}_2\}^{1/2} \quad \text{for all } \mathbf{v}_1, \mathbf{v}_2. \tag{4.4}$$

We shall also use the following relations, valid for any s.p.d. (symmetric, positive-definite) stiffness matrix partitioned into the block form (4.3).

LEMMA 4.1. *Let* $\mathring{S} = \mathring{A}_{22} - \mathring{A}_{21}\mathring{A}_{11}^{-1}\mathring{A}_{12}$. *Then*

(a) *The condition number of* $\mathring{A}_{11}$ *is bounded above by a number independent of the mesh parameter*;

(b) $1 - \gamma^2 \leqq \mathbf{v}_2^t \mathring{S} \mathbf{v}_2 / \mathbf{v}_2^t \mathring{A}_{22} \mathbf{v}_2 \leqq 1$, *for all* $\mathbf{v}_2$, *where* γ *is the constant in* (4.4).

These results have been proved in Axelsson and Gustafsson [4].

Since the bilinear form $a(\cdot, \cdot)$ is bounded on $H_0^1 \times H_0^1$, one can easily verify the following estimate.

LEMMA 4.2. *There exists a constant* $\gamma_2 \geqq 1$, *such that*

$$\mathbf{v}^t A \mathbf{w} \leqq \gamma_2 (\mathbf{v}^t \mathring{A} \mathbf{v})^{1/2} (\mathbf{w}^t \mathring{A} \mathbf{w})^{1/2} \quad \textit{for all } \mathbf{v}, \mathbf{w}.$$

COROLLARY 4.1. *Consider the Schur complements*

$$S = A_{22} - A_{21}A_{11}^{-1}A_{12}, \qquad \mathring{S} = \mathring{A}_{22} - \mathring{A}_{21}\mathring{A}_{11}^{-1}\mathring{A}_{12},$$

of A and $\mathring{A}$, partitioned into block forms (4.2) *and* (4.3), *respectively. Then the following inequality is valid*:

$$\mathbf{v}_2^t S \mathbf{w}_2 \leqq \gamma_2^2 \{\mathbf{v}_2^t \mathring{S} \mathbf{v}_2\}^{1/2} \{\mathbf{w}_2^t \mathring{S} \mathbf{w}_2\}^{1/2} \quad \textit{for all } \mathbf{v}, \mathbf{w},$$

and

$$\mathbf{v}_2^t S \mathbf{v}_2 \geqq \mathbf{v}_2^t \mathring{S} \mathbf{v}_2 \quad \textit{for all } \mathbf{v}_2.$$

Proof. (See also Ewing, Lazarov, Pasciak, and Vassilevski [11] and Axelsson and Vassilevski [7]. Since the proof is short, we present it here for completeness.)

Given $\mathbf{v}_2$, $\mathbf{w}_2$, choose $\mathbf{v}_1$ arbitrary and $\mathbf{w}_1$, so that

$$A\mathbf{w} = \begin{bmatrix} 0 \\ S\mathbf{w}_2 \end{bmatrix},$$

that is,

$$A_{11}\mathbf{w}_1 + A_{12}\mathbf{w}_2 = 0.$$

Then, by Lemma 4.2, we have

$$\begin{aligned} \mathbf{v}_2^t S \mathbf{w}_2 = \mathbf{v}^t A \mathbf{w} &\leqq \gamma_2 (\mathbf{v}^t \mathring{A} \mathbf{v})^{1/2} (\mathbf{w}^t \mathring{A} \mathbf{w})^{1/2} \\ &= \gamma_2 (\mathbf{v}^t \mathring{A} \mathbf{v})^{1/2} (\mathbf{w}_2^t S \mathbf{w}_2)^{1/2}. \end{aligned} \tag{4.5}$$

If we choose $\mathbf{v}_2 = \mathbf{w}_2$ above, then

$$\mathbf{w}_2^t S \mathbf{w}_2 \leqq \gamma_2^2 \mathbf{v}^t \mathring{A} \mathbf{v}$$

and hence

$$\{\mathbf{w}_2^t S \mathbf{w}_2\}^{1/2} \leqq \gamma_2 \left\{ \inf_{\mathbf{v}_1} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{w}_2 \end{bmatrix}^t \mathring{A} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{w}_2 \end{bmatrix} \right\}^{1/2} = \gamma_2 (\mathbf{w}_2^t \mathring{S} \mathbf{w}_2)^{1/2}.$$

Inserting the last inequality into (4.5), we get

$$\begin{aligned} \mathbf{v}_2^t S \mathbf{w}_2 &\leqq \gamma_2^2 \left\{ \inf_{\mathbf{v}_1} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix}^t \mathring{A} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} \right\}^{1/2} \{\mathbf{w}_2^t \mathring{S} \mathbf{w}_2\}^{1/2} \\ &= \gamma_2^2 \{\mathbf{v}_2^t \mathring{S} \mathbf{v}_2\}^{1/2} \{\mathbf{w}_2^t \mathring{S} \mathbf{w}_2\}^{1/2}. \end{aligned}$$

The last inequality follows from

$$\mathbf{v}^t A \mathbf{v} = \mathbf{v}^t \mathring{A} \mathbf{v} \geqq \mathbf{v}_2^t \mathring{S} \mathbf{v}_2 \quad \text{for all } \mathbf{v}_1,$$

and hence for $\mathbf{v}_1$, such that $A_{11}\mathbf{v}_1 + A_{12}\mathbf{v}_2 = 0$, that is,

$$A\mathbf{v} = \begin{bmatrix} 0 \\ S\mathbf{v}_2 \end{bmatrix},$$

we get

$$\mathbf{v}^t A \mathbf{v} = \mathbf{v}_2^t S \mathbf{v}_2 \geqq \mathbf{v}_2^t \mathring{S} \mathbf{v}_2. \qquad \square$$

COROLLARY 4.2. *Assume that the eigenvalues of $\mathring{A}_{11}$ are contained in the interval $[\alpha_1, \alpha_2]$, which is defined in Lemma* 4.1, *and hence where $\alpha_1 > 0$. Then the eigenvalues*

of A_{11} are contained in the fixed segment of a disc in the right-half complex plane

$$\{z\in C;\operatorname{Re} z\geqq\alpha_1,\ |z|\leqq\gamma_2\alpha_2\}.$$

Proof. By Lemma 4.2, we have

$$\begin{aligned}\mathbf{v}_1^t A_{11}\mathbf{w}_1 &\leqq \gamma_2\{\mathbf{v}_1^t\mathring{A}_{11}\mathbf{v}_1\}^{1/2}\{\mathbf{w}_1^t\mathring{A}_{11}\mathbf{w}_1\}^{1/2}\\ &\leqq \gamma_2\alpha_2\|\mathbf{v}_1\|\,\|\mathbf{w}_1\|.\end{aligned}$$

Hence the eigenvalues of A_{11} satisfy

$$|\lambda(A_{11})|\leqq\gamma_2\alpha_2$$

and

$$\operatorname{Re}\lambda(A_{11})\geqq\lambda_{\min}[\tfrac{1}{2}(A_{11}+A_{11}^T)]\geqq\alpha_1.$$ □

Corollary 4.2 shows that A_{11} is well conditioned and hence, we can solve the system with A_{11} (occurring in Algorithm 3) using inner iterations with a generalized conjugate gradient method in a number of iterations independent of the mesh parameter and to any desired relative accuracy (see Axelsson [1], for instance). In practice, one will use a preconditioned form of the generalized conjugate gradient method, with a preconditioner, such as the diagonal part, or an incomplete factorization of A_{11}. Further, Lemma 4.1(b) and Corollary 4.1 show that we can solve the system with S occurring in Algorithm 3 with $\mathring{A}_{22}$ (a coarse-grid symmetric and positive-definite stiffness matrix) as a spectrally equivalent preconditioner in a preconditioned generalized conjugate gradient method. Since $\mathring{A}_{22}$ corresponds to a stiffness matrix on the coarse mesh (τ_2) we can expect to be able to solve systems with $\mathring{A}_{22}$ with much less cost than A. In the symmetric case, this was discussed in Axelsson and Gustafsson [4]. Alternatively, we can solve $\mathring{A}_{22}$ using a recursive factorization with two-by-two block matrix splittings. This has been analysed in Axelsson and Vassilevski [5], [6] and shall now be discussed further in the present context.

In order to apply the theory in § 3, we need to define the norms $\|\mathbf{v}_1\|_0$ and $\|\mathbf{v}_2\|_0$ and estimate the corresponding numbers

$$\sigma_1=\|A_{12}S_2^{-1}\|_0,\qquad \sigma_2=\|A_{21}A_{11}^{-1}\|_0.$$

We choose here

$$\|\mathbf{v}_1\|_0=\{\mathbf{v}_1^t\mathring{A}_{11}^{-1}\mathbf{v}_1\}^{1/2},\qquad \|\mathbf{v}_2\|_0=\{\mathbf{v}_2^t\mathring{S}^{-1}\mathbf{v}_2\}^{1/2}. \tag{4.6}$$

For practical purposes, however, one must choose

$$\|\mathbf{v}_1\|_0=\{\mathbf{v}_1^t\mathring{B}_{11}\mathbf{v}_1\}^{1/2},\qquad \|\mathbf{v}_2\|_0=\{\mathbf{v}_2^t\mathring{A}_{22}^{-1}\mathbf{v}_2\}^{1/2}, \tag{4.6′}$$

with $\mathring{B}_{11}$ s.p.d. and spectrally equivalent to $\mathring{A}_{11}^{-1}$ (such as $\mathring{B}_{11}^{-1}=\operatorname{diag}(\mathring{A}_{11})$), which give uniformly equivalent norms to the previous ones. To simplify the presentation, we consider here only the choice (4.6).

By the definition of σ_1, we have

$$\begin{aligned}\sigma_1^2&=\sup_{\mathbf{v}_2}\left(\frac{\|A_{12}S^{-1}\mathbf{v}_2\|_0}{\|\mathbf{v}_2\|_0}\right)^2\\ &=\sup_{\mathbf{v}_2}\frac{(A_{12}S^{-1}\mathbf{v}_2)^T\mathring{A}_{11}^{-1}A_{12}S^{-1}\mathbf{v}_2}{\mathbf{v}_2^t\mathring{S}^{-1}\mathbf{v}_2}\\ &=\sup_{\mathbf{v}_2}\frac{\mathbf{v}_2^tS^{-T}A_{12}^T\mathring{A}_{11}^{-1}A_{12}S^{-1}\mathbf{v}_2}{\mathbf{v}_2^t\mathring{S}^{-1}\mathbf{v}_2}\end{aligned}$$

$$= \sup_{\mathbf{v}_2} \frac{\mathbf{v}_2^t [\hat{S}^{1/2} S^{-T} \hat{S}^{1/2}][\hat{S}^{-1/2} A_{12}^T \mathring{A}_{11}^{-1/2}] R_1 R_2 \mathbf{v}_2}{\mathbf{v}_2^t \mathbf{v}_2}$$

$$= \sup_{\mathbf{v}_2} \frac{\mathbf{v}_2^t (R_1 R_2)^T R_1 R_2 \mathbf{v}_2}{\mathbf{v}_2^t \mathbf{v}_2} \leqq \|R_1\|^2 \|R_2\|^2,$$

where $R_1 = \mathring{A}_{11}^{-1/2} A_{12} \hat{S}^{-1/2}$, $R_2 = \hat{S}^{1/2} S^{-1} \hat{S}^{1/2}$. Hence

$$\sigma_1 = \|R_1\| \|R_2\|.$$

Note that

$$\|R_1\| = \sup_{\mathbf{v}_1, \mathbf{v}_2} \frac{|\mathbf{v}_1^t A_{12} \mathbf{v}_2|}{\{\mathbf{v}_1^t \mathring{A}_{11} \mathbf{v}_1\}^{1/2} \{\mathbf{v}_2^t \hat{S} \mathbf{v}_2\}^{1/2}},$$

so by Lemma 4.1,

$$\|R_1\| \leqq \gamma_2 \{\mathbf{v}_2^t \mathring{A}_{22} \mathbf{v}_2 / \mathbf{v}_2^t \hat{S} \mathbf{v}_2\}^{1/2} \leqq \gamma_2 / (1 - \gamma^2)^{1/2}.$$

Further, Corollary 4.1 shows that

$$\|R_2\| \leqq 1. \tag{4.7}$$

Hence

$$\sigma_1 \leqq \gamma_2 / (1 - \gamma^2)^{1/2}.$$

Also,

$$\sigma_2^2 = \|A_{21} A_{11}^{-1}\|_0^2$$

$$= \sup_{\mathbf{v}_1} \left(\frac{\|A_{21} A_{11}^{-1} \mathbf{v}_1\|_0}{\|\mathbf{v}_1\|_0} \right)^2$$

$$= \sup_{\mathbf{v}_1} \frac{(A_{21} A_{11}^{-1} \mathbf{v}_1)^T \hat{S}^{-1} A_{21} A_{11}^{-1} \mathbf{v}_1}{\mathbf{v}_1^t \mathring{A}_{11}^{-1} \mathbf{v}_1}$$

$$= \sup_{\mathbf{v}_1} \frac{\mathbf{v}_1^t G_2^T G_2 \mathbf{v}_1}{\mathbf{v}_1^t \mathbf{v}_1} = \|G_2\|^2$$

where

$$G_2 = \hat{S}^{-1/2} A_{21} A_{11}^{-1} \mathring{A}_{11}^{1/2}.$$

Hence

$$\sigma_2 = \sup_{\mathbf{v}_2, \mathbf{w}_1} \frac{\mathbf{v}_2^t \hat{S}^{-1/2} A_{21} A_{11}^{-1} \mathring{A}_{11}^{1/2} \mathbf{w}_1}{\|\mathbf{v}_2\| \|\mathbf{w}_1\|} \leqq \gamma_2 \sup_{\mathbf{v}_2, \mathbf{w}_1} \frac{\{\mathbf{v}_2^t \hat{S}^{-1/2} \mathring{A}_{22} \hat{S}^{-1/2} \mathbf{v}_2\}^{1/2} \{\mathbf{w}_1^t G_1^T G_1 \mathbf{w}_1\}^{1/2}}{\|\mathbf{v}_2\| \|\mathbf{w}_1\|},$$

where $G_1 = \mathring{A}_{11}^{1/2} A_{11}^{-1} \mathring{A}_{11}^{1/2}$, so by Lemma 4.2,

$$\sigma_2 \leqq \gamma_2 (1 - \gamma^2)^{-1/2} \|G_1\|,$$

and by the construction of $\mathring{A}_{11}$, $\|G_1\| \leqq 1$, so

$$\sigma_2 \leqq \gamma_2 / (1 - \gamma^2)^{1/2}. \tag{4.8}$$

We summarize the result in the following theorem.

THEOREM 4.1. *Let the norms* $\|\mathbf{v}_1\|_0$, $\|\mathbf{v}_2\|_0$ *be defined by* (4.6) (*or by* (4.6′)). *Then for* ε_1, ε_2 *sufficiently small, the mapping* $B[\cdot]$ *defined by Algorithm* 4, *with* $B_{11}[\cdot]$ *and* $C[\cdot]$ *satisfying* (3.3), *is an optimal order variable-step preconditioning.* □

PROBLEM 4.2. *Consider the following saddle point problem*

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}$$

with A_{11} symmetric, positive definite, $A_{21} = A_{12}^T$ and A_{22} negative semidefinite. In certain applications, such as mixed finite element discretizations of second-order elliptic problems, A_{22} is, in fact, zero.

In this case the Schur complement, $-S$,

$$-S = A_{22} - A_{21} A_{11}^{-1} A_{12}$$

is negative definite.

In order to apply the two-by-two block variable step (with inner iterations) preconditioner B, we need to specify $B_{11}[\mathbf{w}_1]$, a mapping which approximates the solution of

$$A_{11}\mathbf{v}_1 = \mathbf{w}_1,$$

and a mapping $C[\mathbf{w}_2]$, which approximates the solution of

$$S\mathbf{v}_2 = \mathbf{w}_2.$$

Note that, hence, $-C[\cdot]$ will approximate $-S^{-1}$.

If A is derived by a finite element approximation of the Stokes problem, then $B_{11}[\mathbf{w}_1]$ can, for example, be the approximation of $A_{11}^{-1}\mathbf{w}_1$, which one gets when applying ν steps of an optimal order preconditioned conjugate gradient method, for instance, for the solution of the Poisson equation based on a multigrid or on the algebraic multilevel preconditioner in Axelsson and Vassilevski [5], [6]. Let us denote the corresponding optimal (symmetric and positive-definite) preconditioning matrix in this inner iterative method by $\mathring{A}_{11}$.

In the case when A is derived by a mixed finite element discretization of second-order elliptic equations, $\mathring{A}_{11}$ will be, say, a lumped mass matrix, or more generally a (modified) incomplete factorization of A_{11}, since cond $(A_{11}) = O(1)$.

Since the actions of the Schur complement S are not generally available, first use an approximation of S by

$$\hat{S} = A_{22} - A_{21}\hat{A}_{11}^{-1}A_{12} \tag{4.9}$$

where $\hat{A}_{11}^{-1}$ is generally a more accurate approximation to A_{11}^{-1} than is $\mathring{A}_{11}^{-1}$. $\hat{A}_{11}^{-1}$ is obtained by a fixed number of steps of an optimal stationary (inner) iterative method.

Finally, let D be an optimal order preconditioner to S. For the Stokes problem, D can, for example, be the unity matrix (see Langer and Queck [13], Verfürth [15]). For the mixed finite element approximation of second-order elliptic problems, the (best) choice of D is not clear, as it can depend on the discretization used. However, for the lowest order Raviart–Thomas finite element spaces discretization (see, for example, [14]), D can be chosen as a multigrid step applied to the corresponding equation obtained after elimination of the velocity.

Then $C[\mathbf{w}_2]$ corresponds to the approximations obtained by a fixed step preconditioned conjugate gradient method with D as a preconditioner applied to solve the system $\hat{S}\mathbf{v}_2 = \mathbf{w}_2$ (see (4.9)).

The corresponding norms are chosen as

$$\|\mathbf{v}_1\|_0 = \{\mathbf{v}_1^T \mathring{A}_{11}^{-1}\mathbf{v}_1\}^{1/2}, \qquad \|\mathbf{v}_2\|_0 = \{\mathbf{v}_2^T D^{-1}\mathbf{v}_2\}^{1/2}. \tag{4.10'}$$

However, in order to somewhat simplify the analysis, we shall assume that

$$\|\mathbf{v}_1\|_0 = \{\mathbf{v}_1^T A_{11}^{-1}\mathbf{v}_1\}^{1/2}, \qquad \|\mathbf{v}_2\|_0 = \{\mathbf{v}_2^T S^{-1}\mathbf{v}_2\}^{1/2}. \tag{4.10}$$

Finally, it remains to estimate

$$\sigma_1 = \|A_{12}S^{-1}\|_0, \qquad \sigma_2 = \|A_{21}A_{11}^{-1}\|_0.$$

We have

$$\begin{aligned}
\sigma_1^2 &= \sup_{\mathbf{v}_2}\left(\frac{\|A_{12}S^{-1}\mathbf{v}_2\|_0}{\|\mathbf{v}_2\|_0}\right)^2 \\
&= \sup_{\mathbf{v}_2}\frac{(A_{12}S^{-1}\mathbf{v}_2)^T A_{11}^{-1}(A_{12}S^{-1}\mathbf{v}_2)}{\mathbf{v}_2^t S^{-1}\mathbf{v}_2} \\
&\leqq \sup_{\mathbf{v}_2}\frac{\mathbf{v}_2^t S^{-1/2}A_{21}A_{11}^{-1}A_{12}S^{-1/2}\mathbf{v}_2}{\mathbf{v}_2^t\mathbf{v}_2} \\
&\leqq \sup_{\mathbf{v}_2}\frac{\mathbf{v}_2^t S^{-1/2}(A_{21}A_{11}^{-1}A_{12}-A_{22})S^{-1/2}\mathbf{v}_2}{\mathbf{v}_2^t\mathbf{v}_2} \\
&\leqq 1.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\sigma_2^2 &= \sup_{\mathbf{v}_1}\left(\frac{\|A_{21}A_{11}^{-1}\mathbf{v}_1\|_0}{\|\mathbf{v}_1\|_0}\right)^2 \\
&= \sup_{\mathbf{v}_1}\frac{\mathbf{v}_1^t A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1}\mathbf{v}_1}{\mathbf{v}_1^t A_{11}^{-1}\mathbf{v}_1} \\
&= \sup_{\mathbf{v}_1}\frac{\mathbf{v}_1^t A_{11}^{-1/2}A_{12}S^{-1}A_{21}A_{11}^{-1/2}\mathbf{v}_1}{\mathbf{v}_1^t\mathbf{v}_1}.
\end{aligned}$$

We have

$$\begin{aligned}
\mathbf{w}_1^t A_{12}\mathbf{w}_2 &= (A_{11}^{1/2}\mathbf{w}_1)^T(A_{11}^{-1/2}A_{12}\mathbf{w}_2) \\
&\leqq \|A_{11}^{1/2}\mathbf{w}_1\|(\mathbf{w}_2^t A_{21}A_{11}^{-1}A_{12}\mathbf{w}_2)^{1/2} \\
&\leqq \|A_{11}^{1/2}\mathbf{w}_1\|\,\|S^{1/2}\mathbf{w}_2\|.
\end{aligned}$$

Hence, with $\mathbf{v}_1 = A_{11}^{1/2}\mathbf{w}_1$, $\mathbf{v}_2 = S^{1/2}\mathbf{w}_2$, we find

$$\mathbf{v}_1^t A_{11}^{-1/2}A_{12}S^{-1/2}\mathbf{v}_2 \leqq \|\mathbf{v}_1\|\,\|\mathbf{v}_2\|,$$

that is,

$$\sigma_2 = \|S^{-1/2}A_{21}A_{11}^{-1/2}\| \leqq 1.$$

Here the norm is the standard Euclidean norm, $\|\mathbf{v}\| = \{\sum v_i^2\}^{1/2}$.

Remark 4.1. Note that

$$\frac{|(A_{21}\mathbf{x},\mathbf{y})|}{(A_{11}\mathbf{x},\mathbf{x})^{1/2}(S\mathbf{y},\mathbf{y})^{1/2}} = \frac{|(S^{-1/2}A_{21}A_{11}^{-1/2}\mathbf{x},\mathbf{y})|}{\|\mathbf{x}\|\,\|\mathbf{y}\|} \leqq \sigma_2 \leqq 1 \quad \text{for all } \mathbf{x},\mathbf{y}.$$

This shows the upper bound in the well-known Babuska–Brezzi condition. Note also that

$$\sup_{\mathbf{x}\neq\mathbf{0}} \frac{(A_{21}\mathbf{x},\mathbf{y})}{(A_{11}\mathbf{x},\mathbf{x})^{1/2}(S\mathbf{y},\mathbf{y})^{1/2}} = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{(\mathbf{x},A_{11}^{-1/2}A_{12}S^{-1/2}\mathbf{y})}{\|\mathbf{x}\|\|\mathbf{y}\|} = \frac{\|A_{11}^{-1/2}A_{12}S^{-1/2}\mathbf{y}\|}{\|\mathbf{y}\|}.$$

Hence

$$\inf_{\mathbf{y}\neq 0}\sup_{\mathbf{x}\neq 0} \frac{(A_{21}\mathbf{x},\mathbf{y})}{(A_{11}\mathbf{x},\mathbf{x})^{1/2}(S\mathbf{y},\mathbf{y})^{1/2}} = \mu_1,$$

where μ_1^2 is the smallest eigenvalue of the matrix

$$S^{-1/2}A_{21}A_{11}^{-1}A_{12}S^{-1/2} = (A_{11}^{-1/2}A_{12}S^{-1/2})^T(A_{11}^{-1/2}A_{12}S^{-1/2}).$$

This is positive (i.e., the Babuska–Brezzi inf-sup condition is satisfied) if and only if A_{12} has a complete column rank. The algebraic formulation of the Babuska–Brezzi condition can be found earlier in Bank, Welfert, and Yserentant [10].

The result for Problem 4.2 is summarized in the following theorem.

THEOREM 4.2. *Let the norms* $\|\mathbf{v}_1\|_0$, $\|\mathbf{v}_2\|_0$ *be defined by* (4.10′). *Then the mapping* $B[\cdot]$ *defined by Algorithm* 4, *with* ε_1, ε_2 *sufficiently small and* $B_{11}[\cdot]$, $C[\cdot]$ *defined accordingly as above for Problem* 4.2, *gives an optimal variable-step preconditioner.*

5. Conclusion. We have derived a general framework for a parameter-free variable step preconditioned generalized conjugate gradient method, which is applied for solving two-by-two block matrix problems arising, for example, in two-level nonsymmetric problems, as well as for indefinite saddle-point problems. For them, the general coercivity and boundedness properties of the variable-step preconditioner have been verified. The method can be implemented as a black box solver for any problem satisfying the coercivity and boundedness assumptions.

REFERENCES

[1] O. AXELSSON, *A generalized conjugate gradient, least square method*, Numer. Math., 51 (1987), pp. 209–227.

[2] ———, *A restarted version of a generalized preconditioned conjugate gradient method*, Comm. Appl. Numer. Meth., 4 (1988), pp. 521–530.

[3] ———, *Notes on the numerical solution of the biharmonic equation*, IMA J. Appl. Math., 11 (1973), pp. 213–226.

[4] O. AXELSSON AND I. GUSTAFSSON, *Preconditioning and two-level multigrid methods of arbitrary degree of approximation*, Math. Comp., 40 (1983), pp. 219–242.

[5] O. AXELSSON AND P. S. VASSILEVSKI, *Algebraic multilevel preconditioning methods*, I, Numer. Math., 56 (1989), pp. 157–177.

[6] ———, *Algebraic multilevel preconditioning methods*, II, SIAM J. Numer. Anal., 27 (1990), pp. 1569–1590.

[7] ———, *A survey of multilevel preconditioned iterative methods*, BIT, 29 (1989), pp. 769–793.

[8] T. BANACHIEWICZ, *Zur Berechnung der Determinanten, wie auch die Inversen und zur darauf basierten Auflösung der Systeme lineairen Gleichungen*, Acta Astronom. Ser. C, 3 (1937), pp. 41–67.

[9] R. E. BANK AND T. DUPONT, *Analysis of a two-level scheme for solving finite element equations*, Report CNA-159, Center for Numerical Analysis, University of Texas, Austin, TX, 1980.

[10] R. E. BANK, B. D. WELFERT, AND H. YSERENTANT, *A class of iterative methods for solving saddle point problems*, Numer. Math., 56 (1990), 645–666.

[11] R. E. EWING, R. D. LAZAROV, J. E. PASCIAK, AND P. S. VASSILEVSKI, *Finite element methods for parabolic problems with time steps variable in space*, Report #1989-05, Institute for Scientific Computation, University of Wyoming, Laramie, WY, 1989.

[12] G. H. GOLUB AND M. L. OVERTON, *The convergence of inexact Chebyshev and Richardson iterative methods for solving linear systems*, Numer. Math., 53 (1988), pp. 571–593.

[13] U. LANGER AND W. QUECK, *On the convergence factor of Uzawa's algorithm*, J. Comput. Appl. Math., 15 (1986), pp. 191–202.

[14] R. A. RAVIART AND J. M. THOMAS, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of the FEM, Lecture Notes in Mathematics 606, Springer-Verlag, Berlin, New York, 1977, pp. 292–315.

[15] R. VERFÜRTH, *A combined conjugate gradient-multigrid algorithm for the numerical solution of the Stokes problem*, IMA J. Numer. Anal., 4 (1984), pp. 441–455.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 4, pp. 645–653, October 1991

A CLASS OF ARBITRARILY ILL CONDITIONED FLOATING-POINT MATRICES*

SIEGFRIED M. RUMP†

Abstract. Let $\mathbb{F}$ be a floating-point number system with basis $\beta \geqq 2$ and an exponent range consisting of at least the exponents 1 and 2. A class of arbitrarily ill conditioned matrices is described, the coefficients of which are elements of $\mathbb{F}$. Due to the very rapidly increasing sensitivity of those matrices, they might be regarded as "almost" ill posed problems.

The condition of those matrices and their sensitivity with respect to inversion is given by means of a closed formula. The condition is rapidly increasing with the dimension. For example, in the double precision of the IEEE 754 floating-point standard (base 2, 53 bits in the mantissa including implicit 1), matrices with $2n$ rows and columns are given with a condition number of approximately $4 \cdot 10^{32n}$.

Key words. condition number, sensitivity, ill conditioned, linear systems, floating-point number systems

AMS(MOS) subject classifications. 15A12, 65F05, 65G05

0. Introduction. It is a trivial fact that there are arbitrarily ill conditioned *real* matrices. In this paper we concentrate on matrices that are exactly representable in some floating-point number system $\mathbb{F}$. There is no restriction to the basis and only a trivial technical assumption on the exponent range of $\mathbb{F}$. For fixed $\mathbb{F}$ there are finitely many square matrices with n rows and a maximum condition number less than ∞ for given n.

The well-known schemes for constructing ill-conditioned matrices suffer from the fact that for given $\mathbb{F}$ only a few matrices are exactly representable in $\mathbb{F}$, say up to $n_{\max}$ rows. For $n > n_{\max}$ rows the entries are getting "too big." For example, let

$$(Z_n)_{ij} := \frac{\binom{n+i-1}{i-1} \cdot n \cdot \binom{n-1}{n-j}}{i+j-1},$$

as proposed by Zielke. For single precision in the IEEE 754 floating-point format (base 2 with 24 bit in the mantissa including implicit 1), we have (using infinity norm)

$$n_{\max}(Z_n) = 10 \quad \text{with } \|Z_{10}\| \cdot \|Z_{10}^{-1}\| \approx 2 \cdot 10^{14}.$$

From Pascal's triangle we get

$$(P_n)_{ij} := \binom{i+j-1}{i-1}$$

with

$$n_{\max}(P_n) = 15 \quad \text{with } \|P_{15}\| \cdot \|P_{15}^{-1}\| \approx 1 \cdot 10^{16}.$$

The classical example for ill-conditioned matrices is Hilbert matrices, the ijth component of which is $1/(i + j - 1)$. In order to make them exactly representable in a binary floating-point format, we may use their inverses, or we may multiply the entire matrix by $lcm(1, 2, \cdots, 2n - 1)$. We call the latter matrix H_n^*. Then

$$n_{\max}(H_n^{-1}) = 7 \quad \text{with } \|H_7\| \cdot \|H_7^{-1}\| \approx 5 \cdot 10^8$$

* Received by the editors August 31, 1989; accepted for publication (in revised form) March 13, 1990.
† Informatik III, Technische Universität, 2100 Hamburg 90, Federal Republic of Germany.

and

$$n_{\max}(H_n^*) = 10 \quad \text{with } \|H_{10}^*\| \cdot \|H_{10}^{*-1}\| \approx 2 \cdot 10^{13}.$$

The second method is obviously much more effective with respect to generating exactly representable ill-conditioned matrices. The class of matrices to be described in the following has no restriction in the dimension. In the single precision IEEE 754 floating-point number system, there are 10×10-matrices with condition number $1.1 \cdot 10^{78}$.

1. The class of matrices. Let $\mathbb{F}$ be a floating-point number system with base β, i.e., $\mathbb{F}$ consists of real numbers of the form

$$\chi = \pm 0 \cdot m_1 m_2 \cdots m_\lambda \cdot \beta^e \tag{1.1}$$

with

$$0 \leqq m_i < \beta \quad \text{for } 1 \leqq i \leqq \lambda \quad \text{and} \quad e_{\min} \leqq e \leqq e_{\max}. \tag{1.2}$$

We do not require numbers in the gradual underflow range and assume

$$m_1 \neq 0 \quad \text{if } \chi \neq 0. \tag{1.3}$$

Let $\mathbb{F}$ consist at least of all real numbers $\chi \in \mathbb{R}$ with a representation satisfying (1.1)–(1.3) and assume $e_{\min} \leqq 1$, $2 \leqq e_{\max}$.

Consider Pell's equation (see [1])

$$P^2 - k \cdot Q^2 = 1 \tag{1.4}$$

for positive integers P, Q, and k. If β is a square, let k be the smallest prime factor of β, otherwise set $k = \beta$. Then (1.4) has infinitely many solutions (P, Q) (see [1]).

Let P, Q be numbers satisfying Pell's equation (1.4) for some k and let

$$P = \sum_{\nu=0}^{n} p_i \cdot \sigma^i \quad \text{and} \quad Q = \sum_{\nu=0}^{n} q_i \cdot \sigma^i \tag{1.5}$$

with $p_n \neq 0$ or $q_n \neq 0$ for some $\sigma \in \mathbb{N}$ and $|p_i|, |q_i| < \sigma$, $i = 0 \cdots n$. Furthermore, in this section we assume that $0 \leqq p_i, q_i < \sigma$ for $i = 0 \cdots n$.

In practical applications a typical choice for σ is β^λ. However, in this section we are interested in minimum requirements for the floating-point number system $\mathbb{F}$. Therefore we set $\sigma = k$.

For $\sigma = k$ the numbers p_i, q_i are of $\mathbb{F}$ if $e_{\min} \leqq 1 \leqq e_{\max}$ and so is $k \cdot q_i$ because $k \cdot q_i < k^2 \leqq \beta^2$. To store the number 1 requires 1 to be an admissible exponent; to store $k \cdot q_i$ requires 1 or 2 to be admissible exponents. Therefore

$$p_i, k \cdot q_i, 1, \sigma \in \mathbb{F} \quad \text{if } e_{\min} \leqq 1 \quad \text{and} \quad 2 \leqq e_{\max}$$

and the matrix

$$c_n := \begin{bmatrix} p_n & p_{n-1} & \cdots & p_1 & p_0 & kq_n & kq_{n-1} & & kq_1 & kq_0 \\ q_n & q_{n-1} & \cdots & q_1 & q_0 & p_n & p_{n-1} & & p_1 & p_0 \\ 1 & -\sigma & & & & & & & & \\ & 1 & -\sigma & & & & & & & \\ & & & \cdots & & & & & & \\ & & & & 1 & -\sigma & & & & \\ & & & & & & 1 & -\sigma & & \\ & & & & & & & 1 & -\sigma & \\ & & & & & & & & \cdots & \\ & & & & & & & & 1 & -\sigma \end{bmatrix} \tag{1.6}$$

consists only of components that are exactly representable in $\mathbb{F}$. Since (1.4) has infinitely many solutions, the class of matrices C_n defined by (1.6) consists of elements with an arbitrarily large number of rows.

2. Properties of the matrices. In this section some properties of the matrices defined by (1.6) will be studied. Here, no restrictions on k or σ with respect to β are necessary; our only assumptions are (1.5) and (1.4). In the following, especially, the assumption $0 \leqq p_i, q_i < \sigma$ for $i = 0 \cdots n$ is not necessary.

Throughout this paper we use componentwise ordering of matrices, i.e., $A \leqq B : \langle = \rangle$ $a_{ij} \leqq b_{ij}$ and the componentwise absolute value $|A| = (|A_{ij}|)$, which is again a matrix.

The condition number $\|C_n\| \cdot \|C_n^{-1}\|$ for the ∞-norm will be calculated along with the sensitivity of C_n. Rohn, in [3], gave a nice definition of the sensitivity of a matrix C with respect to inversion: Let B be a matrix of relative distance less than or equal to α to C, i.e., $|B - C| \leqq \alpha \cdot |C|$, then

$$s_{ij}^{\alpha}(C) := \max \left\{ \frac{|B_{ij}^{-1} - C_{ij}^{-1}|}{|C_{ij}^{-1}|}; |B - C| \leqq \alpha \cdot |C| \right\},$$

provided $C_{ij}^{-1} \neq 0$ and

$$s_{ij}(C) := \lim_{\alpha \to 0_+} \frac{s_{ij}^{\alpha}(C)}{\alpha}.$$

In [3], Rohn proves an explicit formula for the sensitivity matrix $S = (s_{ij}(C))$:

$$s_{ij}(C) = \frac{(|C^{-1}| \cdot |C| \cdot |C^{-1}|)_{ij}}{|C^{-1}|_{ij}} \quad \text{for } C_{ij}^{-1} \neq 0. \tag{2.1}$$

LEMMA 1. $\det(C_0) = 1$, $\|C_0\|_{\infty} \|C_0^{-1}\|_{\infty} = (P + kQ)^2$, *and* $s_{ij}(C_0) = 4P^2 - 3$ *for* $i = j$ *and* $s_{ij}(C_0) = 4P^2 - 1$ *for* $i \neq j$.

Proof. For $n = 0$, (1.6) writes

$$C_0 = \begin{pmatrix} P & kQ \\ Q & P \end{pmatrix} \quad \text{with } C_0^{-1} = \begin{pmatrix} P & -kQ \\ -Q & P \end{pmatrix},$$

as follows from (1.4). Then the first two statements are obvious; for the third, a short computation yields

$$(s_{ij}(C_0)) = \begin{pmatrix} \zeta & \eta \\ \eta & \zeta \end{pmatrix} \quad \text{with } \zeta = P^2 + 3kQ^2, \qquad \eta = 3P^2 + kQ^2. \qquad \square$$

In the following we will show that for $n > 0$ the condition and sensitivity of C_n increase compared to those of C_0.

For the rest of the paper we frequently use

$$C := C_n \in \mathbb{R}^{(2n+2) \times (2n+2)} \quad \text{with components } c_{ij},\ 0 \leqq i, j \leqq 2n+1. \tag{2.2}$$

The indices of matrices start with 0 with the exception of A and B, to be defined later. Those are $(n + 1) \times n$-matrices with row indices starting with σ and column indices starting with 1.

LEMMA 2. *The matrices* C_n *are not singular*: $\det(C_n) = (-1)^n$.

Proof. Define

$$s := (\sigma^n, \sigma^{n-1}, \cdots, \sigma, 1)^t \in \mathbb{R}^{n+1} \tag{2.3}$$

and

$$x := \begin{pmatrix} P\cdot s \\ -Q\cdot s \end{pmatrix} = \begin{pmatrix} P\cdot\sigma^n \\ \vdots \\ P\cdot 1 \\ -Q\cdot\sigma^n \\ \vdots \\ -Q\cdot 1 \end{pmatrix} \in \mathbb{R}^{2n+2}. \tag{2.4}$$

Then

$$(p_n, \cdots, p_0)\cdot s = P \quad \text{and} \quad (q_n, \cdots, q_0)\cdot s = Q, \tag{2.5}$$

and using (2.2),

$$\sum_{\nu=0}^{2n+1} c_{0\nu}\cdot x_\nu = P^2 - kQ^2 = 1,$$

$$\sum_{\nu=0}^{2n+1} c_{1\nu}\cdot x_\nu = PQ - QP = 0 = \sum_{\nu=0}^{2n+1} c_{i\nu}\cdot x_\nu \quad \text{for } i \geqq 2.$$

This means that x is the first column of C^{-1} and, especially,

$$(C^{-1})_{2n+1,0} = -Q. \tag{2.6}$$

Therefore $-Q = -\det(\bar{C})/\det(C)$ with

$$\bar{C} := \begin{pmatrix} q_n \cdots q_0 & p_n \cdots p_1 \\ \Sigma & 0 \\ 0 & \Sigma^* \end{pmatrix},$$

and

$$\Sigma := \begin{pmatrix} 1 & -\sigma & & & \\ & 1 & -\sigma & & \\ & & \cdots & & \\ & & & 1 & -\sigma \end{pmatrix}, \qquad \Sigma^* := \begin{pmatrix} 1 & -\sigma & & & \\ & 1 & -\sigma & & \\ & & \cdots & & \\ & & & & 1 \end{pmatrix}.$$

But $\det(\bar{C}) = \det(\bar{\bar{C}})$ with

$$\bar{\bar{C}} := \begin{pmatrix} q_n \cdots q_0 \\ \Sigma \end{pmatrix}$$

and $\bar{\bar{C}}\cdot s = Q\cdot e$ with $e = (1, 0, \cdots, 0)^t$. This implies that

$$(\bar{\bar{C}}^{-1})_{00} = \sigma^n/Q = \det(\hat{C})/\det(\bar{\bar{C}})$$

with

$$\hat{C} := \begin{pmatrix} -\sigma & & & \\ 1 & -\sigma & & \\ & \cdots & & \\ & & 1 & -\sigma \end{pmatrix} \in \mathbb{R}^{n\times n}, \qquad \det(\hat{C}) = -(1)^n\cdot\sigma^n.$$

Therefore

$$\det(C) = \frac{\det(\bar{C})}{Q} = \frac{\det(\bar{\bar{C}})}{Q} = \frac{\det(\hat{C})\cdot Q}{\sigma^n\cdot Q} = (-1)^n. \qquad \square$$

Next we calculate the inverse of $C = C_n$ explicitly. The first column is already given by (2.4), the second is given by

$$y := \begin{pmatrix} -k\cdot Q\cdot s \\ P\cdot s \end{pmatrix} \in \mathbb{R}^{2n+2}, \qquad C\cdot y = (0,1,0,\cdots,0)^t. \tag{2.7}$$

Formulas (2.4) and (2.7) imply, especially, that $-Q$ and P are the first two elements of the last row of C^{-1}. Let

$$(-QP\alpha_n \cdots \alpha_1 \beta_n \cdots \beta_1) \in \mathbb{R}^{2n+2} \tag{2.8}$$

be the last row of C^{-1}. Then multiplication with the first $n+1$ columns of C yields

$$\begin{aligned} &-Q\cdot p_n + p\cdot q_n + \alpha_n = 0, \\ &-Q\cdot p_{n-1} + P\cdot q_{n-1} - \sigma\cdot\alpha_n + \alpha_{n-1} = 0, \\ &\qquad\cdots \\ &-Q\cdot p_1 + P\cdot q_1 - \sigma\cdot\alpha_2 + \alpha_1 = 0, \\ &-Q\cdot p_0 + P\cdot q_0 - \sigma\cdot\alpha_1 = 0. \end{aligned} \tag{2.9}$$

Setting $\alpha_0 = \alpha_{n+1} = 0$ by definition gives

$$-Q\cdot p_i + P\cdot q_i - \sigma\cdot\alpha_{i+1} + \alpha_i = 0 \quad \text{for } i = 0\cdots n \tag{2.10}$$

and by successively adding the equations in (2.9), multiplied by σ, yields

$$\alpha_i = Q\cdot\sum_{\nu=i}^{n} p_\nu\cdot\sigma^{\nu-i} - P\cdot\sum_{\nu=i}^{n} q_\nu\cdot\sigma^{\nu-i} \quad \text{for } i = 1\cdots n. \tag{2.11}$$

By treating the last $n+1$ columns of C in the same way, we obtain

$$\begin{aligned} &-k\cdot Q\cdot q_i + P\cdot p_i - \sigma\cdot\beta_{i+1} + \beta_i = 0 \quad \text{for } i = 1\cdots n, \\ &-k\cdot Q\cdot q_0 + P\cdot p_0 - \sigma\cdot\beta_1 = 1, \end{aligned} \tag{2.12}$$

setting $\beta_0 = \beta_{m+1} = 0$ by definition, and

$$\beta_i = P\cdot\sum_{\nu=i}^{n} p_\nu\cdot\sigma^{\nu-i} - k\cdot Q\cdot\sum_{\nu=i}^{n} q_\nu\cdot\sigma^{\nu-i} \quad \text{for } i = 1\cdots n. \tag{2.13}$$

According to our assumption (1.5), $p_n \neq 0$ or $q_n \neq 0$ and

$$\sum_{\nu=i}^{n} p_\nu\cdot\sigma^{\nu-i} < \sigma^n \leqq P \quad \text{or} \quad \sum_{\nu=i}^{n} q_\nu\cdot\sigma^{\nu-i} < Q \quad \text{for } i \geqq 1.$$

Moreover, gcd $(P, kQ) = 1$ such that (2.11) and (2.13) imply

$$\alpha_i \neq 0 \quad \text{and} \quad \beta_i \neq 0 \quad \text{for } i = 1\cdots n. \tag{2.14}$$

Let $\iota_i \in \mathbb{R}^{n+1,n+1}$ be a matrix with 1 in the ith upper diagonal and 0 elsewhere such that

$$\iota_i\cdot s = (\sigma^{n-i}, \cdots, \sigma, 1, 0, \cdots, 0)^t \in \mathbb{R}^{n+1}, \tag{2.15}$$

using s from (2.3). Then we are ready to describe C^{-1} as follows.

LEMMA 3. *The inverse of* $C = C_n$ *defined by* (1.6) *is given by*

$$\left[\begin{array}{c|c|c|c} P\cdot S & -k\cdot Q\cdot S & B & k\cdot A \\ \hline -Q\cdot S & P\cdot S & A & B \end{array}\right] \begin{matrix} 0 \\ n+1 \\ n+2 \\ 2n+1 \end{matrix} \tag{2.16}$$

$$\begin{matrix} 0 & 1 & 2 & n+1 & n+2 & 2n+1 \end{matrix}$$

with

$$A := (\alpha_n s, \cdots, \alpha_1 s) \in \mathbb{R}^{n+1,n},$$

and

$$B := ((\beta_n I + \iota_n)\cdot s, \cdots, (\beta_1 I + \iota_1)\cdot s) \in \mathbb{R}^{n+1,n}.$$

Proof. For the matrices $A = (a_{ij})$ and $B = (b_{ij})$, we have

$$a_{ij} = \alpha_{n-j+1}\cdot \sigma^{n-i}, \quad \text{and}$$

$$b_{ij} = \begin{cases} \beta_{n-j+1}\cdot \sigma^{n-i}, & j \leqq i, \\ \beta_{n-j+1}\cdot \sigma^{n-i} + \sigma^{j-i+1} & j \geqq i+1 \end{cases} \tag{2.17}$$

for $i = 0\cdots n$, $j = 1\cdots n$ (the row indices start with 0, the column indices with 1). Denote the matrix defined by (2.16) by Γ. Then for $0 \leqq i, j \leqq n$, we have

$$(\Gamma\cdot C)_{ij} = P\cdot s_i\cdot p_{n-j} - k\cdot Q\cdot s_i\cdot q_{n-j} + b_{i,j+1} - \sigma\cdot b_{ij}$$

where the third summand cancels for $j = n$, the fourth for $j = 0$. Using $\beta_0 = \beta_{n+1} = 0$ and (2.17) yields

$$(\Gamma\cdot C)_{ij} = \begin{cases} t(i,j) & \text{for } j < i, \\ t(i,j) + \sigma^{j-i} & \text{for } j = i, \\ t(i,j) + \sigma^{j-i} + \sigma^{j-i-1} & \text{for } j > i \end{cases}$$

using the abbreviation

$$t(i,j) := \sigma^{n-i}\cdot(P\cdot p_{n-j} - k\cdot Q\cdot q_{n-j} + \beta_{n-j} - \sigma\cdot\beta_{n-j+1}).$$

Therefore, for $0 \leqq i, j \leqq n$,

$$(\Gamma\cdot C)_{ij} = \sigma^{n-i}\cdot(P\cdot p_{n-j} - k\cdot Q\cdot q_{n-j} + \beta_{n-j} - \sigma\cdot\beta_{n-j+1}) + \delta_{ij} \tag{2.18}$$

using Kronecker's delta. Since later on we will need $|C^{-1}|\cdot|C|$, we write down the explicit formulae for the other components of $\Gamma\cdot C$. For $0 \leqq i \leqq n$, $n+1 \leqq j \leqq 2n+1$ derives

$$(\Gamma\cdot C)_{ij} = \sigma^{n-i}\cdot k\cdot(P\cdot q_{n-j} - Q\cdot p_{n-j} + \alpha_{n-j} - \sigma\cdot\alpha_{n-j+1}); \tag{2.19}$$

for $n+1 \leqq i \leqq 2n+1$, $0 \leqq j \leqq n$ derives

$$(\Gamma\cdot C)_{ij} = \sigma^{n-i}\cdot(-Q\cdot p_{n-j} + P\cdot q_{n-j} + \alpha_{n-j} - \sigma\cdot\alpha_{n-j+1}); \tag{2.20}$$

and for $n+1 \leqq i, j \leqq 2n+1$ derives

$$(\Gamma\cdot C)_{ij} = \sigma^{n-i}\cdot(-k\cdot Q\cdot q_{n-j} + P\cdot p_{n-j} + \beta_{n-j} - \sigma\cdot\beta_{n-j+1}) + \delta_{ij}. \tag{2.21}$$

The identities (2.10) and (2.12) prove $(\Gamma\cdot C)_{ij} = \delta_{ij}$. □

For the condition of C using the ∞-norm and $\alpha_i \neq 0$,

$$\|C_n\|_\infty \cdot \|C_n^{-1}\|_\infty > \left\{ \sum_{\nu=0}^{n} (p_\nu + k \cdot q_\nu) \right\} \cdot \{ \sigma^n \cdot (P + k \cdot Q) \}$$

$$= \left\{ \sum_{\nu=0}^{n} (\sigma^n p_\nu + k \sigma^n q_\nu) \right\} \cdot (P + kQ) \geqq (P + k \cdot Q)^2. \tag{2.22}$$

We calculate the sensitivity $s_{ij}(C)$ according to (2.1) for $0 \leqq i \leqq n$, $j = 0$. By (2.18) we have

$$(|C^{-1}| \cdot |C|)_{i\nu} \geqq \sigma^{n-i} \cdot (P \cdot |p_{n-\nu}| + k \cdot Q \cdot |q_{n-\nu}| + |\beta_{n-\nu}| + \sigma \cdot |\beta_{n-\nu+1}|),$$

for $0 \leqq \nu \leqq n$ and by (2.19) we have

$$(|C^{-1}| \cdot |C|)_{i\nu} \geqq \sigma^{n-i} \cdot k \cdot (P \cdot |q_{n-\nu}| + Q \cdot |p_{n-\nu}| + |\alpha_{n-\nu}| + \sigma \cdot |\alpha_{n-\nu+1}|),$$

for $n + 1 \leqq \nu \leqq 2n + 1$.

Using $\alpha_\nu, \beta_\nu \neq 0$ we get, for $0 \leqq i \leqq n$,

$$(|C^{-1}| \cdot |C| \cdot |C^{-1}|)_{i0}$$

$$= \sum_{\nu=0}^{n} (|C^{-1}| \cdot |C|)_{i\nu} \cdot |C^{-1}|_{\nu 0} + \sum_{\nu=n+1}^{2n+1} (|C^{-1}| \cdot |C|)_{i\nu} \cdot |C^{-1}|_{\nu 0}$$

$$\geqq \sigma^{n-i} \cdot \sum_{\nu=0}^{n} \{ (P \cdot |p_{n-\nu}| + k \cdot Q \cdot |q_{n-\nu}|) \cdot P \cdot \sigma^{n-\nu} + k \cdot (P \cdot |q_{n-\nu}| + Q \cdot |p_{n-\nu}|) \cdot Q \cdot \sigma^{n-\nu} \}$$

$$+ \sigma^{n-i} \cdot \left\{ \sum_{\nu=0}^{n} (|\beta_{n-\nu}| + \sigma \cdot |\beta_{n-\nu+1}|) \cdot P \cdot \sigma^{n-\nu} \right.$$

$$\left. + \sum_{\nu=0}^{n} (|\alpha_{n-\nu}| + \sigma \cdot |\alpha_{n-\nu+1}|) \cdot kQ\sigma^{n-\nu} \right\}$$

$$\geqq \sigma^{n-i} \cdot P \cdot (P^2 + kQ^2 + kQ^2 + kQ^2) + \sigma^{n-i} \cdot P \cdot 4$$

$$= \sigma^{n-i} \cdot P \cdot (4P^2 - 3 + 4) > \sigma^{n-i} \cdot P \cdot (4P^2)$$

using $k \cdot Q \geqq P$. Together with $|C^{-1}|_{i0} = \sigma^{n-i} \cdot P \neq 0$,

$$S_{i0}(C) > 4P^2 \quad \text{for } 0 \leqq i \leqq n$$

follows. This proves the following theorem.

THEOREM 4. *The matrix C defined by* (1.6) *satisfies*

$$\|C\|_\infty \cdot \|C^{-1}\|_\infty > (P + k \cdot Q)^2$$

and there are components of C of which the sensitivity defined by (2.1) *is greater than* $4 \cdot P^2$.

3. Some examples. For given k, suitable pairs (P, Q) satisfying Pell's equation $P^2 - k \cdot Q^2 = 1$ are easily generated. Given some (P_0, Q_0) unequal, the trivial solution is $(1, 0)$, and successive solutions are

$$(P_{i+1}, Q_{i+1}) = (P_i P_0 + k Q_i Q_0, Q_i P_0 + P_i Q_0).$$

For a floating-point number system given by (1.1)–(1.3), a choice for σ is β^λ. Any expansion (1.5) of P, Q is suitable. The coefficients p_i, q_i are calculated successively.

Some bits can be saved by the following observation. If some coefficient p_i is divisible by β or by a power of β, then p_i and the following p_j, $j > i$ are expressed with a corresponding exponent. If the last digit m_λ in the mantissa of p_{i+1} is equal to $\beta - 1$, then p_i can be replaced by $p_i - \sigma$ and p_{i+1} by $p_{i+1} + 1$, the latter being divisible by β.

For example, let $P = 73942$, $\beta = 10$, $\sigma = 100$. Then expanding P yields $(p_2, p_1, p_0) = (7, 39, 42)$ and this is reduced by the method just described to $(p_1, p_0) = (74 \cdot 10^1, -58)$. This method is especially useful for base 2.

For a given number P, the corresponding coefficients p_i, $i = 0 \cdots n$ can be calculated by the following algorithm:

$e = 0$; $i = 0$;
repeat
  while $P \bmod \beta = 0$ do $\{P = P/\beta;\ e = e+1\}$;
  $q = \lfloor P/\sigma \rfloor$; $r = P - \sigma \cdot q$;
  if $(q \bmod \beta \neq \beta - 1)$ *or* $(q < \beta)$
    then $\{p_j = r \cdot \beta^e;\ P = q\}$
    else $\{p_i = (r - \sigma) \cdot \beta^e;\ P = q+1\}$;
  $i = i+1$
until $P = 0$;

For $k = 2$, successive pairs P, Q are $(3, 2)$, $(17, 12)$, $(99, 70) \cdots$. In Table 1 we display some values for p_i, q_i for single and double precision. For the individual value of n (resulting in a $2n \times 2n$-matrix C) we choose the maximum values (P, Q) being representable by $(p_{n-1}, \cdots, p_0)$ and $(q_{n-1}, \cdots, q_0)$. In the columns of Table 1, the condition number is given followed by the coefficients p_i and q_i, both in descending order. The coefficients are given by two numbers m and e such that $m \cdot 2^e$ is the actual coefficient. For example, $q_4 = 1175 \cdot 2^{22}$ for $n = 5$ (yielding a 10×10-matrix). Our algorithm yields a higher condition than the expected maximum $4 \cdot 2^{24 \cdot 2n} \approx 7 \cdot 10^{72}$, especially for this 10×10-matrix.

For double precision we choose different values for k yielding the coefficients in Table 2. These coefficients are, of course, only samples used to construct matrices of the general form (1.6). We conclude by writing the 6×6-matrix for single precision explicitly.

TABLE 1
p_i, q_i for binary format, 24 bit precision; $k \equiv 2$.

Cond	1.3E+030	2.2E+044	6.5E+060	1.1E+078	4.8E+090	1.7E+107
p_i	15248163 2	3527199 3	6929233 6	425393 14	2161033 8	8490761 10
	11171905 0	6746489 1	9763077 3	6127903 11	5075327 7	15520103 6
		−8816797 0	12608263 1	−10707825 7	8241033 6	6855055 5
q_i	84235 9		−6160127 0	7194379 1	−9934673 5	6997339 4
	−3559681 3	1247053 4		−2285085 0	−5752371 1	−11831695 3
		13508351 2	1224927 8		12291875 0	9051609 1
		−14061827 1	−5131195 6	1175 22		−11093871 0
			14870387 5	−14199789 15	47753 13	
			−7145793 4	12492253 13	−15523515 12	3001937 11
				9093109 10	−1620555 9	12103369 10
				10074835 1	14867027 6	−13213329 9
					14366575 3	−9497253 7
					−4879973 1	−3241495 4
						8507481 3
						−1367575 2

TABLE 2
p_i, q_i for binary format, 53 bit precision.

Cond k	7.0E+066 32	3.4E+097 2	2.1E+131 32	1.4E+164 2
p_i	8384758637032543 5	119071610094027 9	1838140087490775 8	1217131843483323 9
	−3529290569461695 0	−3183251058136493 3	−6618243915631817 2	5555590710757647 8
		−8183182949466111 0	−7698164339527309 1	−1048381871128883 4
q_i	5928919690858185 3		5251261634103185 0	4113071334050663 3
	−6097772977423311 1	84196342944287 9		−3228782923936605 0
		891386017353869 8	162470165079445 9	
		−1900818942150157 7	6774769086897599 6	1721284360250283 8
			4831599480133437 3	292142371452983 6
			−5900891544265983 0	−4351444206118847 4
				1403045714199203 2
				−2787903664869301 1

It is exactly storable with only 24 bits in the mantissa (and therefore in almost any floating-point number system) but matrix inversion will "fail" in almost any floating-point format available because, due to the condition number $2.2 \cdot 10^{44}$, an equivalent of approximately 44 decimal digits precision would be necessary:

$$\begin{pmatrix} 3527199\cdot 2^3 & 6746489\cdot 2^1 & -8816797\cdot 2^0 & 1247053\cdot 2^5 & 13508351\cdot 2^3 & -14061827\cdot 2^2 \\ 1247053\cdot 2^4 & 13508351\cdot 2^2 & -14061827\cdot 2^1 & 3527199\cdot 2^3 & 6746489\cdot 2^1 & -8816797\cdot 2^0 \\ 1 & -2^{24} & & & & \\ & 1 & -2^{24} & & & \\ & & & 1 & -2^{24} & \\ & & & & 1 & -2^{24} \end{pmatrix}.$$

To generate this matrix, the values $P = 7942546277405390632803$ and $Q = 5616228332641321147898$ have been used.

MATLAB [2] delivers as an estimation for the condition number of the matrix the (almost) correct answer ∞.

REFERENCES

[1] G. H. HARDY AND E. WRIGHT, *An Introduction to the Theory of Numbers*, Fifth Edition, Oxford Science Publications, Oxford, 1980, 1981, p. 442.

[2] PRO-MATLAB *User's Guide*, Vers. 32-SUN, The MathWorks, Inc., Sherborn, MA, 1987.

[3] J. ROHN, *New condition numbers for matrices and linear systems*, Computing, 41 (1989), pp. 167–169.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 4, pp. 654–660, October 1991

004

AN ALGORITHM FOR $Ax = \lambda Bx$ WITH SYMMETRIC AND POSITIVE-DEFINITE A AND B*

WANG SHOUGEN† AND ZHAO SHUQIN†

Abstract. An algorithm is given for computing the solution of the eigenvalues of $Ax = \lambda Bx$ with symmetric and positive-definite A and B. It reduces $Ax = \lambda Bx$ to the generalized singular value problem $LL^T x = \lambda(L_B L_B^T)x$ by the Cholesky decompositions $A = LL^T$ and $B = L_B L_B^T$, and then reduces the generalized singular value decomposition of L^T and L_B^T to the CS decomposition of Q by the QR decomposition $(L, L_B)^T = QR$. Finally, it reduces A and B to diagonal forms by singular value decompositions. The algorithm provided is stable and, what is more, faster than the QZ algorithm. Numerical examples are also presented.

Key words. generalized eigenvalue problem, generalized singular value problem, CS decomposition, singular value decomposition

AMS(MOS) subject classification. 65F15

1. Introduction.

Consider the generalized eigenvalue problem

$$Ax = \lambda Bx, \tag{1}$$

where A and B are $n \times n$ real symmetric and positive-definite matrices. The equations of motion for small vibrations about a position of stable equilibrium of a mechanical system operated upon by a conservative system of forces derive the generalized eigenvalue problem of the form (1) [8, p. 34].

Many efficient methods have been designed to solve the generalized eigenvalue problems. Among them, there are the stable QZ algorithm [2] and LZ algorithm [3]; the MDR algorithm [4], which can preserve the symmetry of a problem; and the Lanczos method, which is known to be well suited to the numerical solution of large sparse generalized symmetric eigenvalue problems [11]. Besides, when A and B are symmetric and B is positive definite, the problem could be reduced to a standard symmetric eigenvalue one by the Cholesky decomposition of B [1]. Last, the generalized Jacobi method has been used with some success on small A and B that are diagonally dominant [12, p. 452].

In this paper we present an algorithm for the problem (1). By means of the Cholesky decompositions of A and B, it reduces (1) to the generalized singular value problem, and then reduces the generalized singular value decomposition (GSVD) to the CS decomposition (CSD). Finally, it reduces A and B to diagonal forms by singular value decompositions (SVDs).

2. Algorithm.

Since A and B are symmetric and positive definite, there exist the Cholesky decompositions

$$A = LL^T, \qquad B = L_B L_B^T, \tag{2}$$

where L and L_B are lower triangular matrices with positive diagonal elements. The problem (1) thus becomes the generalized singular value problem [9]

$$LL^T x = \lambda(L_B L_B^T)x.$$

* Received by the editors September 26, 1988; accepted for publication (in revised form) April 6, 1990.

† Department of Mathematics, East China Normal University, Shanghai, 200062, People's Republic of China.

The problem of computing the GSVD can be reduced to the problem of computing the CSD [5], [10]. Let

$$\begin{pmatrix} L^T \\ L_B^T \end{pmatrix} = QR \tag{3}$$

be the QR decomposition of $(L, L_B)^T$, where the columns of

$$Q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}$$

are orthonormal, $Q_1, Q_2, R \in R^{n\times n}$, and R is a nonsingular upper triangular matrix. Let

$$Q_1 = U_1 C V^T, \qquad Q_2 = U_2 S V^T \tag{4}$$

be the CSD of Q, where $U_1, U_2, V \in R^{n\times n}$ are orthogonal; $S = \operatorname{diag}(s_1, \cdots, s_n)$; $C = \operatorname{diag}(c_1, \cdots, c_n)$; $s_i \geqq 0$; $c_i \geqq 0$; and $s_i^2 + c_i^2 = 1$, $(i = 1, \cdots, n)$. From (2), (3), and (4), we obtain

$$A = R^T V C^2 V^T R, \qquad B = R^T V S^2 V^T R. \tag{5}$$

Since $V^T R$ is nonsingular, the eigenvalues of the problem (1) are given by

$$\lambda_i = c_i^2 / s_i^2, \qquad (i = 1, \cdots, n). \tag{6}$$

If we set

$$P = R^{-1} V, \qquad \Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_n), \tag{7}$$

then we have, from (5),

$$AP = BP\Lambda.$$

It shows that the columns of P are the eigenvectors of (1).

If R is well conditioned with respect to inversion, then the formulas (2), (3), (4), (6), and (7) can lead to an algorithm for the computation of the eigenvalues and eigenvectors of (1).

Now we consider the computation of the eigenvalues of (1). Let $Q_1 = \bar{U}_1 \bar{C} V_1^T$ be the SVD of Q_1, where $\bar{U}_1$ and V_1 are orthogonal, and $\bar{C}$ is diagonal. We set $X = Q_2 V_1$. Let $X = \bar{U}_2 \bar{S}$ be the QR decomposition of X, where $\bar{U}_2$ is orthogonal and $\bar{S}$ is upper triangular with positive diagonal elements. Then $\bar{S}$ is diagonal [5]. Thus

$$\begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} = \begin{pmatrix} \bar{U}_1 \bar{C} \\ \bar{U}_2 \bar{S} \end{pmatrix} V_1^T$$

is the CSD of Q. The SVD $Q_1 = \bar{U}_1 \bar{C} V_1^T$ is computed first for computing the CSD (4) of Q [5], [6], where $\bar{C} = \operatorname{diag}(\bar{c}_1, \cdots, \bar{c}_n)$. In the presence of rounding error, Stewart [6] has shown that C is effectively $\bar{C}$. Likewise, if we compute the SVD $Q_2 = \bar{U}_2 \bar{S} V_2^T$, where $\bar{S} = \operatorname{diag}(\bar{s}_1, \cdots, \bar{s}_n)$, then S is effectively $\bar{S}$ in the presence of rounding error. Hence the eigenvalues of (1) are given by $\lambda_i = \bar{c}_i^2 / \bar{s}_i^2$, $(i = 1, \cdots, n)$.

From above, we obtain the following algorithm.

ALGORITHM. Let $A, B \in R^{n\times n}$ be positive definite. This algorithm produces the diagonal matrices $\bar{C}$ and $\bar{S}$ satisfying $\bar{C}^2 + \bar{S}^2 = I$, $P^T B P = \bar{S}^2$, and $P^T A P = \bar{C}^2$, where P is nonsingular.

(1) Compute the Cholesky decompositions of A and B

$$A = LL^T, \qquad B = L_B L_B^T,$$

where L and L_B are lower triangular matrices with positive diagonal elements.

(2) Compute the QR decomposition of $(L, L_B)^T$

$$\begin{pmatrix} L^T \\ L_B^T \end{pmatrix} = QR,$$

where the columns of

$$Q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}$$

are orthonormal, $Q_1, Q_2, R \in R^{n\times n}$, and R is a nonsingular upper triangular matrix.

(3) Compute the SVDs of Q_1 and Q_2

$$Q_1 = \bar{U}_1 \bar{C} V_1^T, \qquad Q_2 = \bar{U}_2 \bar{S} V_2^T,$$

where $\bar{U}_1, \bar{U}_2, V_1, V_2 \in R^{n\times n}$ are orthogonal, and $\bar{C} = \operatorname{diag}(\bar{c}_1, \cdots, \bar{c}_n)$, $\bar{S} = \operatorname{diag}(\bar{s}_1, \cdots, \bar{s}_n)$, the $\bar{c}_i$ and $\bar{s}_i$ are ordered as follows: $0 \leqq \bar{c}_1 \leqq \cdots \leqq \bar{c}_n \leqq 1$, $1 \geqq \bar{s}_1 \geqq \cdots \geqq \bar{s}_n \geqq 0$.

The QR decomposition of $F = (L, L_B)^T$ is computed in step (2) of the algorithm. Since L^T and L_B^T are upper triangular matrices, computing the QR decomposition can take advantage of the structure of the matrix F. Golub and Pereyra [13] have given a method. It results in upper triangular matrices Q_1 and Q_2 and requires $((2/3)n^3)$ flops. The SVDs of Q_1 and Q_2 are computed in step (3) of the algorithm. The reduction of a triangular matrix to bidiagonal form is given in [14]. If a modified Householder matrix [7, p. 43] is adopted, then the reduction requires n^3 flops. It requires $O(n^2)$ flops for computing the singular values from the bidiagonal form, while computing the Cholesky decomposition of a matrix of order n requires $(n^3/6)$ flops [7, p. 89]. Therefore, if we do not form $\bar{U}_1$, $\bar{U}_2$, V_1, and V_2, the total number of flops required is about $3n^3$. If only the eigenvalues are desired, then the QZ algorithm requires about $15n^3$ flops [7, p. 262].

As noted, $\bar{C}^2 + \bar{S}^2 = I$, so we may compute only the SVD of Q_1 in step (3). Then, the eigenvalues of (1) are given by $\lambda_i = \bar{c}_i^2/(1 - \bar{c}_i^2)$, $(i = 1, \cdots, n)$. Thus the algorithm requires about $2n^3$ flops. However, the cancellation of significant figures might occur in the computation of $\lambda_i = \bar{c}_i^2/(1 - \bar{c}_i^2)$.

3. Rounding-error analysis. Let $\hat{x}$ be the computed version of x. From [8, p. 232], we have

$$\hat{L}\hat{L}^T = A + E_1, \qquad \hat{L}_B \hat{L}_B^T = B + E_2, \tag{8}$$

where E_1 and E_2 are small relative to $\|A\|_2$ and $\|B\|_2$, respectively. From [8, pp. 160, 236], we know that there is an orthogonal matrix Q and a matrix (E_3^T, E_4^T) such that

$$Q\hat{R} = \begin{pmatrix} \hat{L}^T + E_3 \\ \hat{L}_B^T + E_4 \end{pmatrix}, \qquad Q = \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix}, \tag{9}$$

where E_3 and E_4 are small relative to

$$\left\| \begin{pmatrix} \hat{L}^T \\ \hat{L}_B^T \end{pmatrix} \right\|_2.$$

From [8, p. 161], we obtain

$$\hat{Q} = \begin{pmatrix} \hat{Q}_1 \\ \hat{Q}_2 \end{pmatrix} = \begin{pmatrix} Q_1 + E_5 \\ Q_2 + E_6 \end{pmatrix}, \tag{10}$$

where E_5 and E_6 are small relative to unity. Suppose the CSD (4) of $\hat{Q}$ is computed by the algorithm in [5], [6]. From [5],

$$\begin{aligned} U_1 \hat{C} V^T &= \hat{Q}_1 + E_7, \qquad U_2 \hat{S} V^T = \hat{Q}_2 + E_8, \\ U_1^T U_1 &= I, \qquad U_2^T U_2 = I, \qquad V^T V = I, \end{aligned} \tag{11}$$

where E_7 and E_8 are small relative to unity. From [6],

$$\hat{\tilde{C}} = \hat{C} + E_9, \qquad \hat{\tilde{S}} = \hat{S} + E_{10}, \tag{12}$$

where E_9 and E_{10} are small relative to unity. From (9), (10), (11), (12) we have

$$\begin{aligned} U_1 \hat{\tilde{C}} V^T &= (\hat{L}^T + E_3)\hat{R}^{-1} + E_5 + E_7 + U_1 E_9 V^T \\ &= (\hat{L}^T + E_3 + E_5 \hat{R} + E_7 \hat{R} + U_1 E_9 V^T \hat{R})\hat{R}^{-1}. \end{aligned}$$

Let $E = E_3 + E_5\hat{R} + E_7\hat{R} + U_1 E_9 V^T \hat{R}$. Then from (8),

$$V \hat{\tilde{C}}^2 V^T = \hat{R}^{-T}(A + E_1 + \hat{L}E + E^T\hat{L}^T + E^T E)\hat{R}^{-1} = \hat{R}^{-T}(A + E_c)\hat{R}^{-1},$$

where $E_c = E_1 + \hat{L}E + E^T\hat{L}^T + E^T E$ is small relative to $(\|A\|_2 + \|B\|_2)$ and symmetric. Thus

$$\hat{\tilde{C}}^2 = V^T \hat{R}^{-T}(A + E_c)\hat{R}^{-1} V.$$

Analogously, we can obtain

$$\hat{\tilde{S}}^2 = V^T \hat{R}^{-T}(B + E_s)\hat{R}^{-1} V,$$

where E_s is small relative to $(\|A\|_2 + \|B\|_2)$ and symmetric. The algorithm is stable thereby.

Let

$$\rho((\alpha, \beta), (\gamma, \delta)) = \frac{|\alpha\delta - \beta\gamma|}{\sqrt{(|\alpha|^2 + |\beta|^2)(|\gamma|^2 + |\delta|^2)}},$$

α, β, γ, and $\delta \in C$, $(\alpha, \beta) \neq (0, 0)$ and $(\gamma, \delta) \neq (0, 0)$.

Since A, B, E_c, and E_s are symmetric and A and B are positive definite, from [15] we know that if

$$\max_{\|x\|_2 = 1} \left\{ \sqrt{\frac{(x^H E_c x)^2 + (x^H E_s x)^2}{(x^H A x)^2 + (x^H B x)^2}} \right\} < 1,$$

then the generalized eigenvalue variation

$$\min_{\pi} \{ \max_{i} \rho((c_i^2, s_i^2), (\hat{\tilde{c}}^2_{\pi(i)}, \hat{\tilde{s}}^2_{\pi(i)})) \}$$

$$\leqq \max_{\|x\|_2 = 1} \{ \rho((x^H A x, x^H B x), (x^H(A + E_c)x, x^H(B + E_s)x)) \},$$

and here π runs through all permutations of $\{1, \cdots, n\}$.

TABLE 1
$n = 9$.

Our algorithm		QZ algorithm	M − W method	Correct values
$\bar{c}_i^2 + \bar{s}_i^2$	λ_i	λ_i	λ_i	λ_i
$0.100000E+01$	$0.301185E-05$	$0.298588E-05$	$-0.305790E-05$	$0.300371E-05$
$0.100000E+01$	$0.244826E-03$	$0.244887E-03$	$0.251860E-03$	$0.244832E-03$
$0.100000E+01$	$0.778315E-02$	$0.778319E-02$	$0.791056E-02$	$0.778316E-02$
$0.100000E+01$	$0.116766E+00$	$0.116766E+00$	$0.117145E+00$	$0.116766E+00$
$0.100000E+01$	$0.100000E+01$	$0.100000E+01$	$0.100073E+01$	$0.100000E+01$
$0.100000E+01$	$0.856412E+01$	$0.856413E+01$	$0.856557E+01$	$0.856413E+01$
$0.100000E+01$	$0.128483E+03$	$0.128483E+03$	$0.128484E+03$	$0.128483E+03$
$0.100000E+01$	$0.408494E+04$	$0.408496E+04$	$0.408501E+04$	$0.408489E+04$
$0.100000E+01$	$0.333536E+06$	$0.338781E+06$	$0.337572E+06$	$0.336035E+06$

4. Numerical examples. The calculations were done on an IBM-PC/AT computer in the following examples. Two programs in the STYR/MATH Users' Guide [16] have been used on the problems for comparison. The first one transforms $Ax = \lambda Bx$ to a standard symmetric eigenvalue problem and applies the QL algorithm (i.e., the Martin and Wilkinson method); the second one is the QZ algorithm.

Example 1. $B = (b_{ij})$, $A = (a_{ij})$, where $b_{ij} = \sin((i-j)(\pi/2))/(i-j)$, $a_{ij} = \pi\delta(i,j) - b_{ij}$. The calculations were done in double precision for b_{ij} and a_{ij}. For $n = 3, 4, \cdots, 15$ we have found the eigenvalues in single precision.The results, for $n = 9$ and $n = 15$, are shown in Tables 1 and 2.

As the eigenvalues computed by our algorithm and the QZ algorithm in double precision agree to six digits for $n \leqq 15$, we take them as correct results which are presented in Tables 1 and 2.

From Tables 1 and 2, it can be seen that the eigenvalues computed by our algorithm have about the same accuracy as the ones computed by the QZ algorithm. $\bar{c}_i^2 + \bar{s}_i^2 = 1$ is satisfied to working precision. In Table 1 the approximations to the smallest eigenvalues are far less accurate for the Martin–Wilkinson method (M–W method) than for the two others. In Table 2 the computed eigenvalues by the M–W method fail to agree with the correct values.

TABLE 2
$n = 15$.

Our algorithm		QZ algorithm	M − W method	Correct values
$\bar{c}_i^2 + \bar{s}_i^2$	λ_i	λ_i	λ_i	λ_i
$0.100000E+01$	$0.374126E-07$	$0.461546E-07$	$-0.342318E+00$	$0.279294E-07$
$0.100000E+01$	$0.596040E-07$	$0.668033E-07$	$-0.334077E+00$	$0.432000E-07$
$0.100000E+01$	$0.107120E-05$	$0.976377E-06$	$-0.240770E+00$	$0.105091E-05$
$0.999999E+00$	$0.393491E-04$	$0.392772E-04$	$-0.165970E+00$	$0.393447E-04$
$0.100000E+01$	$0.955453E-03$	$0.955525E-03$	$-0.119879E+00$	$0.955460E-03$
$0.100000E+01$	$0.149418E-01$	$0.149418E-01$	$-0.395875E-01$	$0.149418E-01$
$0.100000E+01$	$0.147088E+00$	$0.147088E+00$	$-0.681586E-02$	$0.147088E+00$
$0.100000E+01$	$0.100000E+01$	$0.100000E+01$	$0.709770E-02$	$0.100000E+00$
$0.100000E+01$	$0.679867E+01$	$0.679866E+01$	$0.425809E-01$	$0.679866E+01$
$0.100000E+01$	$0.669264E+02$	$0.669262E+02$	$0.112224E+00$	$0.669263E+02$
$0.100000E+01$	$0.104665E+04$	$0.104663E+04$	$0.133817E+00$	$0.104665E+04$
$0.100000E+01$	$0.254386E+05$	$0.254325E+05$	$0.282139E+00$	$0.254343E+05$
$0.100000E+01$	$0.968951E+06$	$0.980617E+06$	$0.369343E+00$	$0.977433E+06$
$0.100000E+01$	$0.501775E+08$	$+\infty$	$0.431837E+00$	$0.650511E+08$
$0.100000E+01$	$0.757954E+08$	$+\infty$	$0.484609E+00$	$0.981998E+10$

Let $Q^H AZ = T$ and $Q^H BZ = H$ be upper triangular, where Q and Z are orthogonal. Let $\hat{T}$ and $\hat{H}$ be the computed versions of T and H by the QZ algorithm, and let t_i, h_i, $\hat{t}_i$, and $\hat{h}_i$ denote the diagonal elements of T, H, $\hat{T}$, and $\hat{H}$, respectively. Analogously, we take the $\hat{\bar{c}}_i^2$ and $\hat{\bar{s}}_i^2$ computed by our algorithm in double precision as the exact c_i^2 and s_i^2, and the $\hat{t}_i$ and $\hat{h}_i$ computed by the QZ algorithm in double precision as the exact t_i and h_i. We obtain

$$\min_{\pi} \{\max_i \rho((c_i^2, s_i^2), (\hat{\bar{c}}_{\pi(i)}^2, \hat{\bar{s}}_{\pi(i)}^2))\} < 1.5E-7 \quad (n=9),$$

$$\min_{\pi} \{\max_i \rho((t_i, h_i), (\hat{t}_{\pi(i)}, \hat{h}_{\pi(i)}))\} < 3.8E-7 \quad (n=9),$$

$$\min_{\pi} \{\max_i \rho((c_i^2, s_i^2), (\hat{\bar{c}}_{\pi(i)}^2, \hat{\bar{s}}_{\pi(i)}^2))\} < 1.4E-7 \quad (n=15),$$

$$\min_{\pi} \{\max_i \rho((t_i, h_i), (\hat{t}_{\pi(i)}, \hat{h}_{\pi(i)}))\} < 1.0E+0 \quad (n=15).$$

Example 2. Let $A = U^T D_A U$, $B = U^T D_B U$, where U is a random Householder matrix and D_A and D_B are diagonal matrices with positive diagonal elements. We have constructed many problems by this method. An example of this kind is the 10×10 problem

$$U = I - 2\frac{uu^T}{\|u\|_2^2}, \qquad A = U^T D_A U, \qquad B = U^T D_B U,$$

where u, D_A, D_B are as follows:

u	D_A	D_B
0.443755E − 02	0.100000E − 03	0.200000E − 02
0.581926E + 00	0.100000E + 05	0.200000E + 04
0.428218E + 00	0.100000E − 02	0.200000E + 05
0.546248E − 01	0.100000E + 04	0.200000E − 03
0.790236E − 01	0.100000E − 01	0.200000E + 02
0.149859E + 00	0.100000E + 03	0.200000E + 00
0.674297E + 00	0.100000E + 00	0.200000E − 01
0.916294E + 00	0.100000E + 02	0.200000E + 03
0.161062E + 00	0.100000E + 01	0.200000E + 01
0.970872E + 00	0.100000E + 01	0.200000E + 01

The calculations were done in single precision. The results are shown in Table 3. Again it can be seen that the eigenvalues computed by our algorithm have about the same accuracy as the eigenvalues computed by the QZ algorithm, and $\bar{c}_i^2 + \bar{s}_i^2 = 1$ is satisfied to working precision. The approximation to the eigenvalue 0.5 is far less accurate for the M–W method than for the two others.

Let $D_A = \text{diag}(\alpha_1, \cdots, \alpha_{10})$, $D_B = \text{diag}(\beta_1, \cdots, \beta_{10})$. We obtain

$$\min_{\pi} \{\max_i \rho((\alpha_i, \beta_i), (\hat{\bar{c}}_{\pi(i)}^2, \hat{\bar{s}}_{\pi(i)}^2))\} < 3.1E-3$$

and

$$\min_{\pi} \{\max_i \rho((\alpha_i, \beta_i), (\hat{t}_{\pi(i)}, \hat{h}_{\pi(i)}))\} < 7.9E-1.$$

TABLE 3

Our algorithm		QZ algorithm	M − W method	Correct values
$\bar{c}_i^2 + \bar{s}_i^2$	λ_i	λ_i	λ_i	λ_i
$0.100000E+01$	$0.542428E-07$	$0.445096E-07$	$-0.469603E-07$	$0.500000E-07$
$0.100000E+01$	$0.500134E-03$	$0.484089E-03$	$0.499008E-03$	$0.500000E-03$
$0.100000E+01$	$0.499888E-01$	$0.497860E-01$	$0.498549E-01$	$0.500000E-01$
$0.100000E+01$	$0.500009E-01$	$0.499535E-01$	$0.500173E-01$	$0.500000E-01$
$0.100000E+01$	$0.499977E+00$	$0.499964E+00$	$0.399286E+00$	$0.500000E+00$
$0.100000E+01$	$0.500002E+00$	$0.499999E+00$	$0.500876E+00$	$0.500000E+00$
$0.100000E+01$	$0.492157E+01$	$0.498219E+01$	$0.466847E+01$	$0.500000E+01$
$0.100000E+01$	$0.499999E+01$	$0.500000E+01$	$0.500000E+01$	$0.500000E+01$
$0.100000E+01$	$0.500065E+03$	$0.500026E+03$	$0.500042E+03$	$0.500000E+03$
$0.100000E+01$	$0.523423E+07$	$+\infty$	$0.522994E+07$	$0.500000E+07$

Acknowledgments. The authors would like to thank the editor for suggesting Example 1 and references [13] and [14]. We are grateful to the referees for their help and suggestions.

REFERENCES

[1] R. S. MARTIN AND J. H. WILKINSON, *Reduction of the symmetric eigenproblem $Ax = \lambda Bx$ and related problems to standard form*, Numer. Math., 11 (1968), pp. 99–110.

[2] C. B. MOLER AND G. W. STEWART, *An algorithm for generalized matrix eigenvalue problems*, SIAM J. Numer. Anal., 10 (1973), pp. 241–256.

[3] L. KAUFMAN, *The LZ-algorithm to solve the generalized eigenvalue problem*, SIAM J. Numer. Anal., 11 (1974), pp. 997–1024.

[4] A. BUNSE-GERSTNER, *An algorithm for the symmetric generalized eigenvalue problem*, Linear Algebra Appl., 58 (1984), pp. 43–68.

[5] C. VAN LOAN, *Computing the CS and the generalized singular value decompositions*, Numer. Math., 46 (1985), pp. 479–491.

[6] G. W. STEWART, *Computing the CS decomposition of a partitioned orthonormal matrix*, Numer. Math., 40 (1982), pp. 297–306.

[7] G. H. GOLUB AND C. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[8] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, U.K., 1965.

[9] C. VAN LOAN, *Generalizing the singular value decomposition*, SIAM J. Numer. Anal., 13 (1976), pp. 76–83.

[10] G. W. STEWART, *A method for computing the generalized singular value decomposition*, in Matrix Pencils, B. Kågström and A. Ruhe, eds., Springer-Verlag, New York, 1983, pp. 207–220.

[11] T. ERICSSON AND A. RUHE, *The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems*, Math. Comp., 35 (1980), pp. 1251–1268.

[12] K. J. BATHE AND E. WILSON, *Numerical Methods in Finite Element Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[13] G. H. GOLUB AND V. PEREYRA, *The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate*, SIAM J. Numer. Anal., 10 (1973), pp. 413–432.

[14] T. F. CHAN, *An improved algorithm for computing the singular value decomposition*, ACM Trans. Math. Software, 8 (1982), pp. 72–83.

[15] J. G. SUN, *Perturbation analysis for the generalized eigenvalue and the generalized singular value problem*, in Matrix Pencils, B. Kågström and A. Ruhe, eds., Springer-Verlag, New York, 1983, pp. 221–244.

[16] STYR/MATH Users' Guide, Computing Center, Academia Sinica Publications, Beijing, People's Republic of China, 1985.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 4, pp. 661-678, October 1991

005

DIAGONALIZING THE ADAPTIVE SOR ITERATION METHOD*

JEROME DANCIS†

Abstract. The SOR iteration method is a popular method for solving the large sparse systems of linear algebraic equations which approximate many partial differential equations that arise in engineering. Often the associated SOR matrix $M^{-1}N$ is diagonalizable except at the eigenvalue $\lambda = \omega - 1$, and the noneigenvector p_* associated with the $\lambda = \omega - 1$ (i) slows down the convergence, and (ii) in the adaptive SOR method, reduces the accuracy of the calculation of the next relaxation factor ω_i. Of course, $M^{-1}N$ cannot be diagonalized, but the error vector can be pushed into the span of the eigenvectors of $M^{-1}N$, thereby eliminating the p_*-coordinate of the error vector, together with its undesirable effects. This is done with the simple polynomial acceleration associated with the polynomial $P_1(x) = (x - (\omega - 1))/(1 - (\omega - 1))$, and $P_n(x) = x^{n-1}P_1(x)$, $n = 2, 3, \cdots$.

In the adaptive SOR method, this acceleration reduces the size of the error (i) by enabling the program to update the value of ω_i sooner, and (ii) by eliminating the contribution of p_* to the error vector.

In the computer runs, using this polynomial acceleration resulted (on average) in an extra digit of accuracy over the results using the standard adaptive SOR method.

Key words. SOR iteration method, polynomial acceleration

AMS(MOS) subject classifications. 65F10, 65F50

1. Introduction. The SOR iteration method is a popular method for solving many of the large sparse systems of linear algebraic equations which approximate the partial differential equations that arise in engineering problems. The modern (1980s) approach is to use the "adaptive SOR" method described in Hageman and Young's book [2]. Under this adaptive SOR method a short increasing sequence of relaxation factors is used:

$$1 \leqq \omega_1 < \omega_2 < \cdots < \omega \approx \omega_b$$

(where ω_b is the "optimal" relaxation factor). Roughly speaking, a small number of SOR iterations are done using a relaxation factor ω_i until it is clear that (i) $\omega_i \neq \omega_b$, and (ii) certain conditions in Hageman and Young's algorithm are met; then ω_i is updated to the next relaxation factor ω_{i+1}.

When the associated SOR matrix $\mathscr{L}_\omega$ is not diagonalizable, there will be a principal vector of grade two p_* (that is, $(\mathscr{L}_\omega - (\omega - 1))^2 p_* = \mathbf{0}$). Various undesirable effects caused by these principal vectors p_* are described in [2, pp. 227, 228]. The main effects are (i) a postponing of the updating of the relaxation factor ω_i, and (ii) a slowing of the rate of convergence.

We can remove these undesirable effects by "pushing" the error vector into the span of the eigenvectors of $\mathscr{L}_\omega$. Then the principal vector will make no contribution to the error vector, and hence only the "diagonalizable part" of $\mathscr{L}_\omega$ will be acting on the error vector.

We achieve this by doing a "first-degree polynomial acceleration" (on the first SOR iteration for each relaxation factor ω_i) which we will call an "a-shift."

DEFINITION. An *a-shift* on two successive SOR iterations, say v_0 and v_1, shall mean formation of the new vector:

$$v_1^* : \leftarrow \frac{1}{1-a} v_1 - \frac{a}{1-a} v_0. \tag{1.1}$$

* Received by the editors December 28, 1987; accepted for publication (in revised form) June 8, 1990.

† Department of Mathematics, University of Maryland, College Park, Maryland 20742-4015 (jnd@hilda.umd.edu).

The details of the consequences of this a-shift (with $a = \omega_i - 1$) will be discussed later. At first glance, it might look like this move will greatly increase the size (2-norm) of the error vector. We will explain why this "should" *not* happen with the *adaptive* SOR method.

We did computer runs on the "model" problem [2, § 1.7], that is, on the 5-point rule applied to Poisson's equation on grids of size 25×25 to 45×45, which results in 625–2,025 equations. Our variation resulted in an extra digit of accuracy (on average) when compared to the standard adaptive SOR method. The results of the computer runs are presented in § 4 and in Graphs 1–8.

2. Background. It is well known (see [2]) that for a symmetric linear system $Av = w$ and a relaxation factor ω, there is the associated SOR matrix

$$\mathcal{L}_\omega = (D - \omega L)^{-1}[\omega U + (1 - \omega)D], \tag{2.1}$$

where $A = D - L - L^T$ is the usual splitting (with D the diagonal of A, and L and U the upper and lower triangular parts of A).

Young [4, p. 238] and Hageman and Kellogg [1] have shown that these associated SOR matrices are diagonalizable *except* at the eigenvalue $\omega - 1$ (for some common types of matrices), namely, Theorem 2.1.

THEOREM 2.1. *Let $\mathcal{L}_\omega$ be the associated* SOR *matrix for some symmetric block tridiagonal matrix A or a consistently ordered 2-cyclic symmetric matrix.*

(i) *If $\omega - 1$ is not an eigenvalue of $\mathcal{L}_\omega$, then $\mathcal{L}_\omega$ is diagonalizable (with complex eigenvalues);*

(ii) *If $\omega - 1$ is an eigenvalue of $\mathcal{L}_\omega$, then $\mathcal{L}_\omega$ has one or more Jordan blocks*

$$\begin{pmatrix} \omega - 1 & 1 \\ 0 & \omega - 1 \end{pmatrix};$$

all the other Jordan blocks of $\mathcal{L}_\omega$ are diagonal matrices.

2.1. The case when $\omega - 1$ is an eigenvalue of $\mathcal{L}_\omega$. In this case, there will be a unit principal vector p_* and a unit eigenvector u_* such that, with $a = \omega - 1$:

$$\mathcal{L}_\omega p_* = a p_* + u_* \quad \text{and} \quad \mathcal{L}_\omega u_* = a u_*.$$

Thus

$$J = \begin{pmatrix} a & 1 \\ 0 & a \end{pmatrix}$$

represents the restriction of $\mathcal{L}_\omega$ to the subspace Span $\{u_*, p_*\}$, with $\{u_*, p_*\}$ as the basis for its coordinate system.

Let c_* and c_p be the coefficients of u_* and p_* in the initial error vector $e^{(0)}$. Thus

$$e^{(0)} = c_* u_* + c_p p_* + \Sigma a_i u_i, \tag{2.2}$$

where the u_i are the other eigenvectors of $\mathcal{L}_\omega$ and $\mathcal{L}_\omega(u_i) = \lambda_i u_i$.

Then, the contribution to the next error vector $e^{(1)}$ by these vectors u_* and p_* is

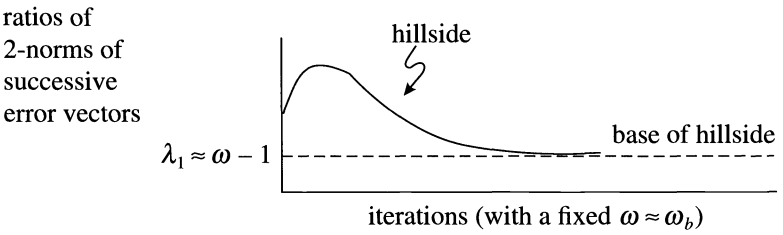
$$J\begin{pmatrix} c_* \\ c_p \end{pmatrix} = \begin{pmatrix} a & 1 \\ 0 & a \end{pmatrix}\begin{pmatrix} c_* \\ c_p \end{pmatrix} = \begin{pmatrix} ac_* + c_p \\ ac_p \end{pmatrix}.$$

Then

$$e^{(1)} = (ac_* + c_p)u_* + ac_p p_* + \Sigma a_i \lambda_i u_i. \tag{2.3}$$

That the contribution of the principal vector to the error vector results in a slowing down of the convergence has been observed in computer runs is mentioned in Chapter 7 of [4].

As Sheldon [3] has pointed out, a consequence of $\omega = \omega_b$ and the existence of the principal vector (when $\omega = \omega_b$) is that the diagrams of the ratios of the 2-norms of successive error vectors look like hillsides:



3. Removing the principal vector. In this section, we will explain how the a-shift (eq. (1.1)) "kills" the principle vector component of the error vector, thereby "pushing" the error and difference vectors into the span of the eigenvectors. The result should be the elimination or reduction of the hillsides mentioned in the preceding section.

DEFINITION. The nth *difference-vector* $r^{(n)}$ is the residual vector for the iteration equation. It is known that $r^{(n)}$ and the nth error vector are connected by the equation

$$r^{(n)} = (I - \mathcal{L}_\omega)e^{(n)}.$$

The next theorem is the main result.

THEOREM 3.1. *Suppose we are doing an* SOR *iteration procedure (with* $\mathcal{L}_\omega$ *as the associated* SOR *matrix with a fixed value for* ω*) and suppose that Theorem* 2.1 *is applicable. Suppose that a single a-shift (with* $a = \omega - 1$*) is done only after the first* SOR *iteration. Thus the iterations are*:

$$\begin{aligned} v_n &= \mathcal{L}_\omega v_{n-1} + w_0, \qquad n = 1, 3, 4, 5, \cdots, \\ v_1^* &= \frac{1}{1-a} v_1 - \frac{a}{1-a} v_0 \quad \text{where } a = \omega - 1, \\ v_2 &= \mathcal{L}_\omega v_1^* + w_0. \end{aligned} \tag{3.1}$$

Suppose that $a = \omega - 1$ *is an eigenvalue of* $\mathcal{L}_\omega$*. Then (except for* v_0 *and* v_1*) all the error and difference vectors will be in the span of the eigenvectors of* $\mathcal{L}_\omega$*.*

Remark. The matrix algebra motivation is as follows. Calculating from (3.1) we obtained

$$v_1^* = \frac{1}{1-a}(\mathcal{L}_\omega - aI)v_0 + \frac{1}{1-a} w_0.$$

Therefore, the matrix S associated with going from v_0 to v_1^* is

$$S = \frac{1}{1-a}(\mathcal{L}_\omega - aI).$$

Let

$$J = \begin{pmatrix} a & 1 \\ 0 & a \end{pmatrix}$$

again. This matrix S restricted to the subspace Span $\{u_*, p_*\}$ with $\{u_*, p_*\}$ coordinates is represented by

$$\frac{1}{1-a}(J-aI)=\frac{1}{1-a}\left[\begin{pmatrix} a & 1 \\ 0 & a \end{pmatrix}-aI\right]=\frac{1}{1-a}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Therefore the contribution by u_* and p_* to the error vector for v_1^* is

$$\frac{1}{1-a}(J-aI)\begin{pmatrix} c_* \\ c_p \end{pmatrix}=\frac{1}{1-a}\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} c_* \\ c_p \end{pmatrix}=\frac{1}{1-a}\begin{pmatrix} c_p \\ 0 \end{pmatrix}=\frac{1}{1-a}c_p u_*.$$

Observe that there *is* a u_* component but *no* p_* component. We now proceed to the proof.

Proof. Let $e_*^{(1)}$ be the error vector for v_1^*; then

$$e_*^{(1)}=v-v_1^*=\frac{1}{1-a}(v-v_1)-\frac{a}{1-a}(v-v_0)$$

$$=\frac{1}{1-a}e^{(1)}-\frac{a}{1-a}e^{(0)},$$

where $e^{(1)}$ and $e^{(0)}$ are the zeroth and first error vectors. Plugging in (2.2) and (2.3) results in

$$e_*^{(1)}=\frac{1}{1-a}\left(c_p u_*+\sum_i(\lambda_i-a)a_i u_i\right).$$

We observe that the principle vector p_* does *not* appear in this equation. Thus $e_*^{(1)}$ is in the span of the eigenvectors. The difference vector for this error vector is given by

$$r_*^{(1)}=(1-\mathscr{L}_\omega)e_*^{(1)}=c_p u_*+\sum_i\frac{1-\lambda_i}{1-a}(\lambda_i-a)a_i u_i.$$

Again, p_* does *not* appear and $r_*^{(1)}$ is a linear combination of the eigenvectors of $\mathscr{L}_\omega$.

As long as we continue to do more SOR iterations with the *same* ω and *no* more a-shifts, the error and difference vectors will be

$$e_*^{(n)}=\frac{1}{1-a}\left(a^{n-1}c_p u_*+\sum_i(\lambda_i-a)a_i^n u_i\right)$$

and

$$r_*^{(n)}=a^{n-1}c_p u_*+\sum_i\frac{1-\lambda_i}{1-a}(\lambda_i-a)a_i^n u_i.$$

Clearly, $e^{(n)}$ and $r^{(n)}$ are linear combinations of the eigenvectors (u_* and u_i) of $\mathscr{L}_\omega$.

4. The computer runs. The algorithm for our computer runs is to use a standard adaptive SOR algorithm modified by the addition of a-shifts, $a = \omega - 1$, immediately following the first SOR iteration after the updating of ω.

ALGORITHM 4.1. Use the adaptive SOR Algorithm 9-6.1 of [2] together with one modification as follows.

Let $V_{i,0}$ denote the iteration solution vector for the iteration after which ω_{i-1} is updated to ω_i

$$V_{i,1} \leftarrow \mathscr{L}_{\omega_i}(V_{i,0}) + w_{0_i},$$

$$V^*_{i,1} \leftarrow \frac{1}{2-\omega_i} V_{i,1} - \frac{\omega_i - 1}{2-\omega_i} V_{i,0},$$

$$V_{i,2} \leftarrow \mathscr{L}_{\omega_i}(V^*_{i,1}) + w_{0_i},$$

$$V_{i,n} \leftarrow \mathscr{L}_{\omega_i}(V_{i,n-1}) + w_{0_i}, \qquad n = 3, 4, \cdots$$

until the next updating of ω_i. Thus $V_{i,1}$ is obtained from $V_{i,0}$ by a single SOR iteration and $V^*_{i,1}$ is obtained from $V_{i,1}$ and $V_{i,0}$ by an a-shift with $a = \omega_i - 1$.

Our computer runs were on the popular "model" problem [2, § 1.7], which is the 5-point rule applied to Poisson's equation on (equally spaced) square grids—from a 25×25 to a 45×45 point grid, all with the dictionary ordering.

We used the problem $(M - N)v = w$ where v was chosen by a random number generator as a vector with integer entries between -999 and $+999$. Using the randomly chosen answer vector v and the computer calculated w, we then used the adaptive SOR method with initial guess $v_0 = \mathbf{0}$ to calculate approximate solutions $\{v_n\}$. Since we knew the exact value of v, the computer also calculated the 2-norms of the error vectors $(\|v - v_n\|)$.

We used these strategy parameters (see [2, pp. 228–229])

$$\text{PSP} = .01, \qquad \text{RSP} = .0001, \qquad F = .8.$$

The initial omega ω_0 was always set equal to one.

Basically we did a line-by-line translation of the 50 lines of Algorithm 9-6.1 of [2] into 50 lines of APL code. (This resulted in our APL code being as easy to read as Algorithm 9-6.1 itself.) APL does its calculations in double precision.

We did seven pairs of computer runs. The control runs used the standard adaptive SOR method [2, Algorithm 9-6.1]. The experimental computer runs used Algorithm 4.1.

The results of these computer runs are presented in the Tables 4.1, 4.2, 5.1, and 6.1 and in Graphs 1–7. The 2-norm is used throughout. The number of iteration steps for each pair of computer runs was a prechosen fixed number.

4.1. The graphs and their descriptions. For each pair of computer runs we will present a set of three graphs.

Notation.

$e^{(n)}$ is the error vector after n SOR iterations using the adaptive SOR method.

$e^{(n)}_*$ is the error vector after n SOR iterations using the adaptive SOR method together with simple a-shifts after the first iteration with each new value for omega.

↑ A solid arrow points to the step where an a-shift was performed as ω was updated.

↑ A broken arrow points to the step in which ω was updated *without* an a-shift.

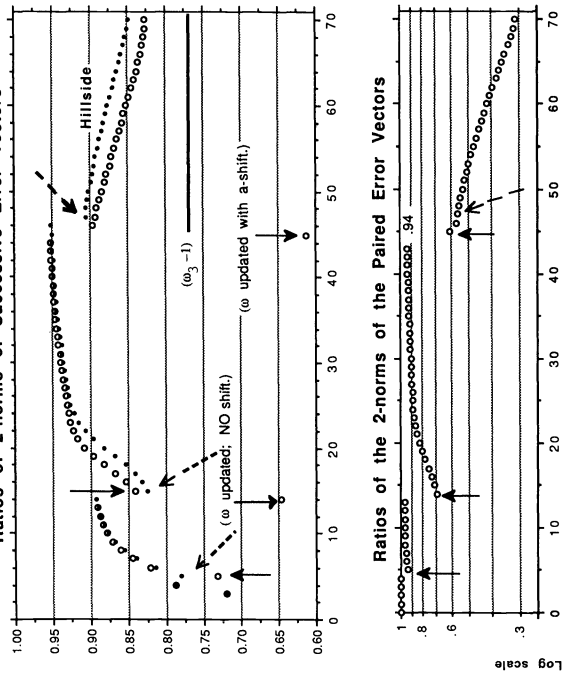
—— The thick dark line segment denotes the base of the "hillside" at "altitude" $\omega - 1$.

(i) The top graph is a graph of the ratios of the 2-norms of successive error vectors of the unshifted and shifted runs, namely,

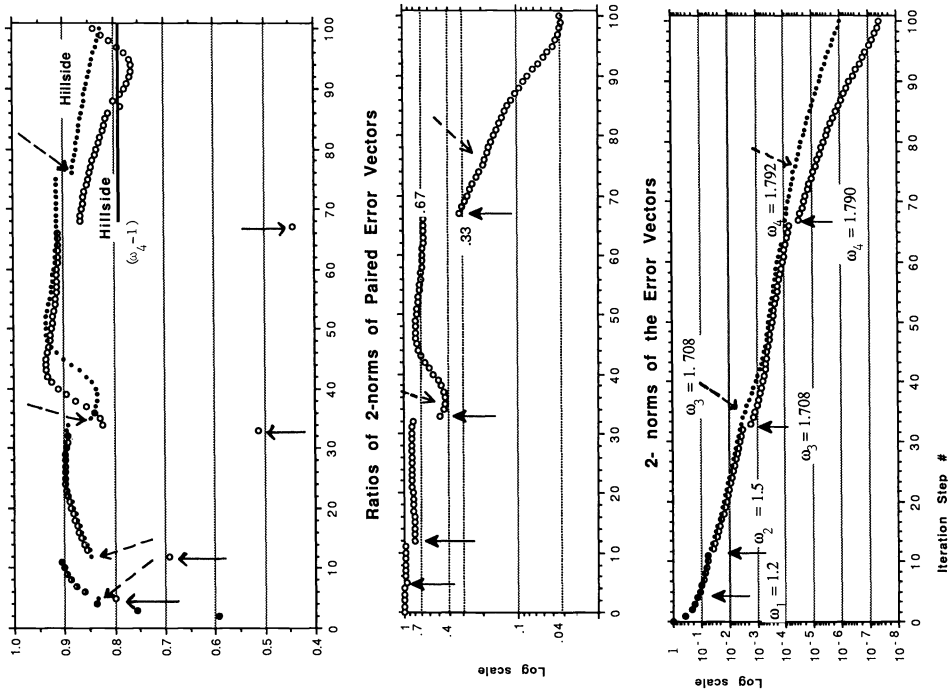
$$\|e^{(n)}\| \div \|e^{(n-1)}\| \quad \text{denoted by } \cdots,$$

$$\|e^{(n)}_*\| \div \|e^{(n-1)}_*\| \quad \text{denoted by } \circ\circ\circ.$$

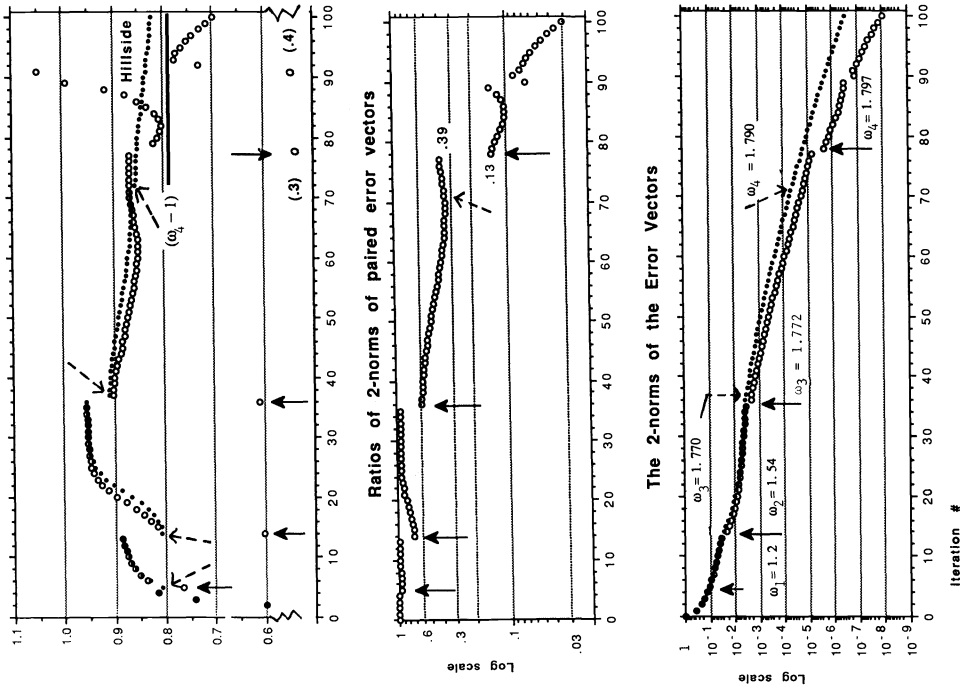
GRAPH 1
Comparisons for 625 equations.



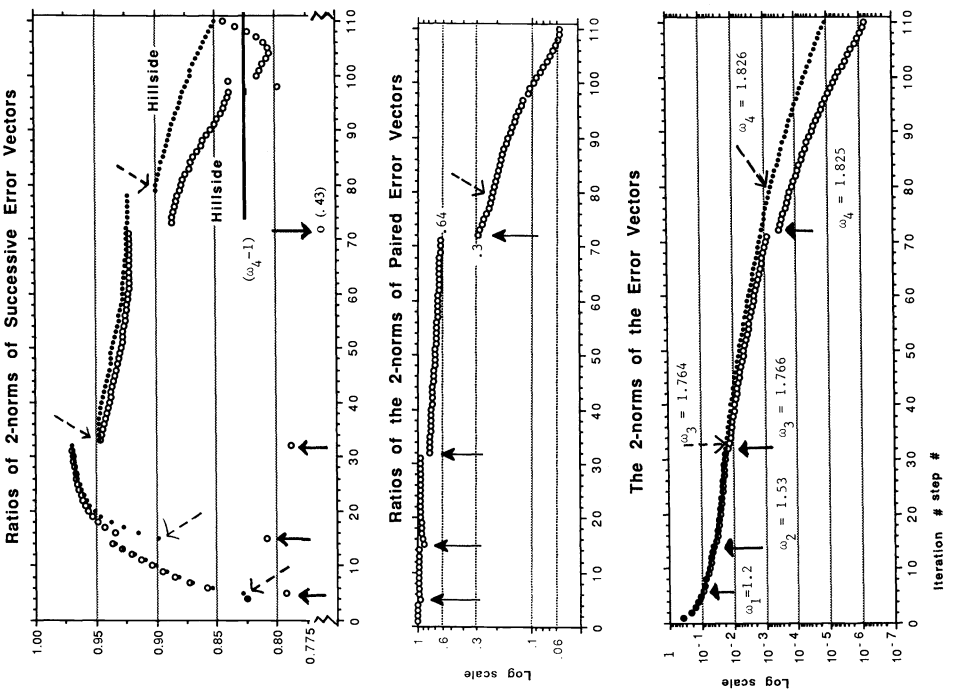
GRAPH 2
Comparisons for 676 equations—solution #1.



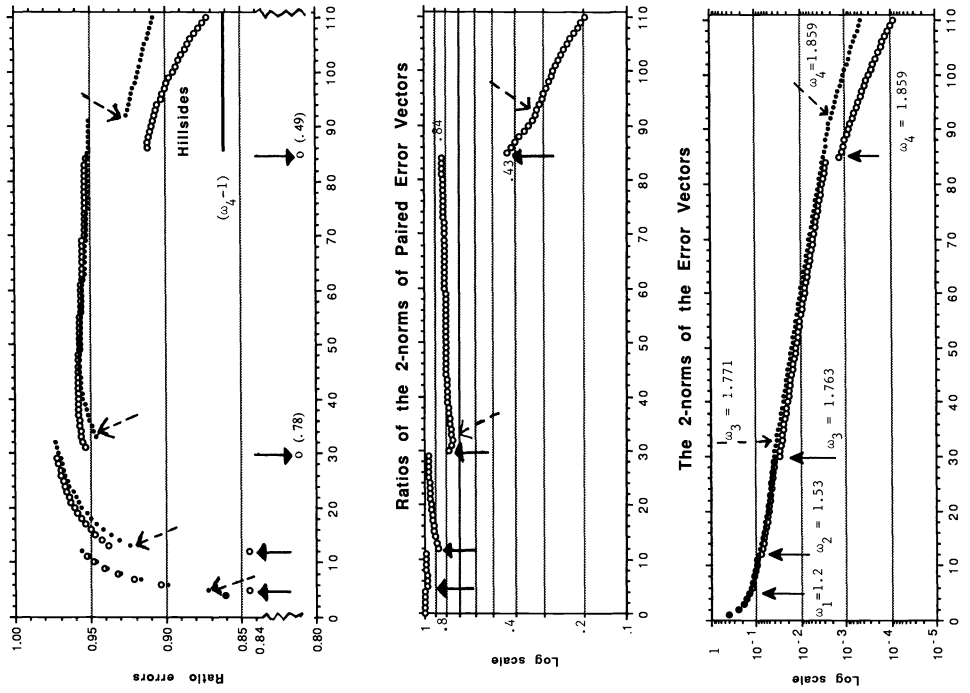
GRAPH 3
Comparisons for 676 equations—solution #2.



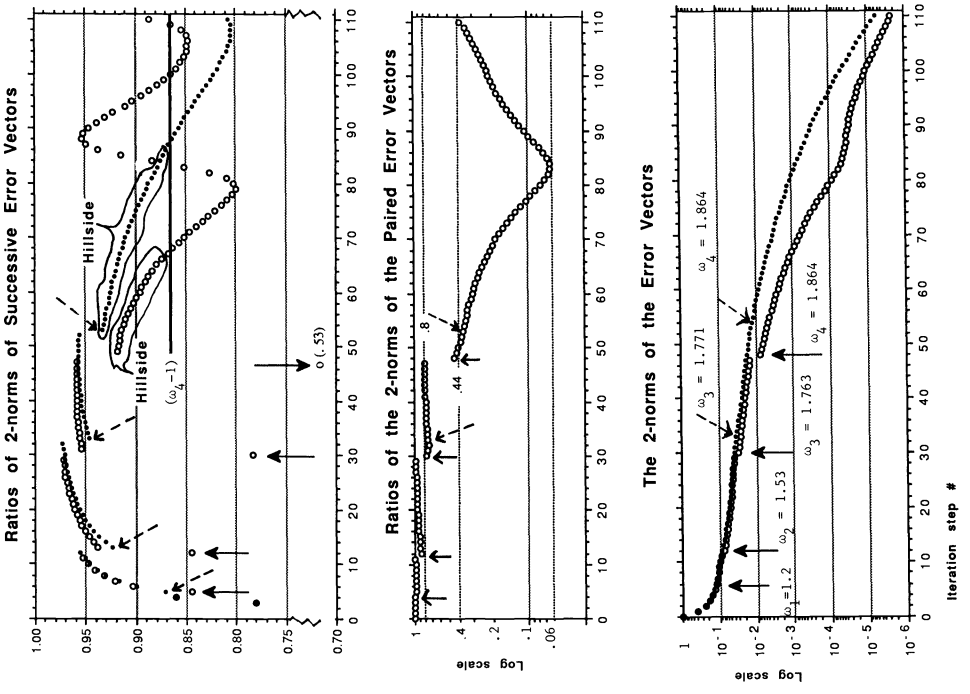
GRAPH 4
Comparisons for 1024 equations.



GRAPH 5
Comparisons for 1600 *equations with "better"* PSP = .01.

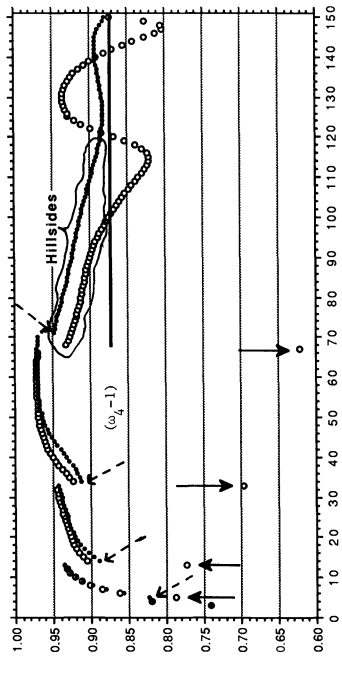


GRAPH 6
Comparisons for 1600 *equations with high* PSP = .5.



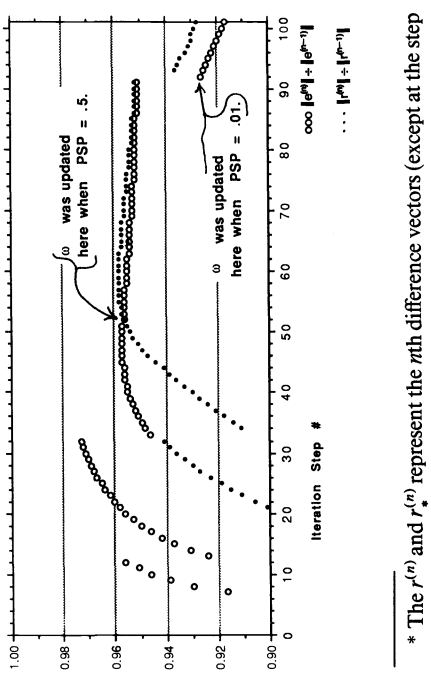
GRAPH 7

Comparisons for 2025 equations.



GRAPH 8

Ratios of 2-norms of successive error vectors and difference vectors with 1600 *equations and* PSP = .01.



* The $r^{(n)}$ and $r_*^{(n)}$ represent the nth difference vectors (except at the step where ω is updated).

These top graphs clearly show the hillsides following the update to the final ω. Both the height and the length of the hillsides of the runs with a-shifts are always less than the ones for unshifted runs. The sinusoidal behavior to the right of some hillsides is a known phenomenon due to the complex eigenvalues. Also, the error drop at each a-shift is visible (sometimes even off scale).

(ii) Each middle graph is a graph of the ratios of the paired error vectors, i.e.,

$$\|e_*^{(n)}\| \div \|e^{(n)}\| \quad \text{denoted by } \cdots .$$

This shows the cumulative improvement of the shifted runs over the unshifted runs. As such each middle graph is a consequence of the top graph.

(iii) Each bottom graph compares the 2-norms of the error vectors of the unshifted and shifted runs, namely,

$$\|e^{(n)}\| \quad \text{denoted by } \cdots ,$$
$$\|e_*^{(n)}\| \quad \text{denoted by } \circ\circ\circ .$$

The difference between the two discrete curves reflects the ratios pictured in the middle graph.

The "drop" points in the middle and bottom graphs occur at the a-shift iterations.

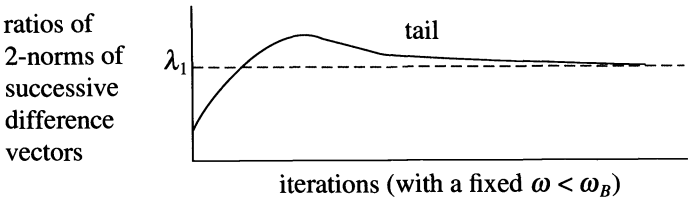
Log scales are used in the middle and bottom graphs. Every horizontal axis lists the numbers of the iteration steps.

4.2. Observed reduction of the hillside effect. In our computer runs, ω was never equal to ω_B, but for the final ω (in each run), $\omega \approx \omega_B$. This resulted in there *not* being any principal-vector-of-$\mathscr{L}_\omega$ component in the error vector when each final ω was chosen. Still, since $\omega \approx \omega_B$, the eigenvectors associated with the pair of eigenvalues near $\omega - 1$ acted in a manner similar (but reduced) to an eigenvector-principal vector pair. This is demonstrated by the low hillsides, which are clearly seen in all the graphs (of the ratios of the 2-norms of successive error vectors). In each case, both the length and the height of the hillside of each run with the a-shifts was lower than the length and height of its "control" run's hillside (without the a-shifts). Thus, doing the a-shifts, when $\omega \approx \omega_B$, seems to remove a good portion of the hillside effect caused by the "approximate eigenvector-principal vector pair." By observation, this reduction of the "hillside" effect seems to have reduced the 2-norm of the error vector by a factor of $\frac{1}{2}$ in our runs (see the lower hillside column in Table 6.1).

4.3. Contamination of the error vector. A second problem caused by principal vectors is that the contribution of this p_* vector "contaminates"[1] the error vector, thereby helping to disguise the maximum eigenvalue. The a-shifts probably reduce this "contamination" by reducing the "principal-vector"–type effect. We observed that the computer runs with a-shifts usually updated the values of ω earlier. This is tabulated in Table 4.1. (In the adaptive SOR method, all runs are updated from $\omega_0 = 1$ to ω_1 following the fourth step.)

4.4. A technical detail. In our runs we observed the following pattern for the ratios of the difference vectors when using the next-to-last ω:

[1] The term "contamination" is rigorously defined and discussed on pp. 221–224 of [2].



When ω is updated at the local maximum on the tail, then the final ω will be somewhat higher than ω_B. To avoid this, the Hageman–Young book tells us to choose a strategy parameter PSP which the computer will use in calculating a variable p^*. This number p^* is supposed to be the number of iteration steps needed to arrive past the local maximum on the tail.

My choice for these runs was PSP = .01. For comparison, I also ran a pair of computer runs on 1,600 equations with (the default value) PSP = .5. The result (using PSP = .5) was that the ω's (for both the shifted and unshifted runs) were updated to the final $\omega_4 \approx \omega_B + .006$ at the relative maximums of the tails (steps 47 and 52) instead of updating to the final $\omega_4 \approx \omega_B + .001$ further down the tail (steps 84 and 91), as occurred when PSP = .01. It turned out that the error vectors, when PSP = .5, were smaller than those when PSP = .01 in spite of the larger spectral radius of $\mathscr{L}_{\omega_4}$. This is a known phenomenon when $\omega > \omega_B$.

For our runs with 1,024 and 2,205 equations, the choice of PSP being .5 or .01 did not affect the steps at which the ω's were updated.

We present the graphs of the successive difference vectors for the runs with 1,600 equations in Graph 8.

A side effect of using these a-shifts (to remove the principal vector) is that the associated eigenvector is increased by the term

$$\frac{1}{1-a} c_p u_*$$

TABLE 4.1
The step at which ω was updated.

		ω was updated to		
Computer run		ω_2 at step #	ω_3 at step #	ω_4 at step #
625 eqs.	with shifts:	13	44	
	no shifts:	14	46	
676 eqs.	with shifts:	11	32	66
solution #1	no shifts:	11	34	75
676 eqs.	with shifts:	13	35	77
solution #2	no shifts:	13	36	71
1,024 eqs.	with shifts:	14	31	72
	no shifts:	14	32	78
1,600 eqs	with shifts:	11	29	84
PSP = .01	no shifts:	12	32	91
1,600 eqs.	with shifts:	11	29	47
PSP = .5	no shifts:	12	32	52
2,025 eqs.	with shifts:	12	32	66
	no shifts:	13	33	70

and this term might be undesirably large, especially when a is close to one (which corresponds to $a + 1 = \omega$ being close to two). This raises the possibility of the norm of the error vector increasing considerably at the time an a-shift is performed.

This did *not* occur in any of our seven computer runs. In fact, as Table 4.2 demonstrates, the opposite occurred. Not only did the norms of the error vectors decrease at the time of each a-shift (which was combined with the updating of the SOR relaxation factors ω_i), but the decrease was even greater than the corresponding decrease in the control computer runs.

We have not (as yet) investigated this effect for the *non*-adaptive SOR method.

For the red-black ordering, the principal vector may be considerably reduced by doing a single Gauss–Seidel iteration immediately following each updating of ω. This is called Sheldon's method [3], [4].

5. The eigenvalue banjos and further analysis of the consequences of the a-shift. It is well known [4, Chap. 5], [2, § 9.3] that when the relaxation factor ω is less than the optimal relaxation factor ω_b, then the set of eigenvalues of the SOR matrix $\mathcal{L}_\omega$ are on a "banjo" (see Fig. 5.1).

DEFINITION. Whenever the eigenvalues of a matrix lie on the union of a circle and a real line segment such that

(i) the center of the circle is a real number,

(ii) the circle intersects the line segment in exactly one point, and

(iii) the endpoints of the line segment are (real) eigenvalues,

then the union of this circle and line segment shall be called the *eigenvalue banjo* of the matrix.

The a-shift (defined in § 1) is the same as the polynomial acceleration associated with the first-degree polynomial

$$P_1(x) = \frac{x - a}{1 - a},$$

TABLE 4.2
Reduction of error vectors while shifting.

Computer run		Ratios of 2-norms of succesive error vectors at the updating of ω to							
		ω_1		ω_2		ω_3		ω_4	
625 eqs.	with shifts:	.73		.65		.61			
	no shifts:		.78		.83		.90		
676 eqs.	with shifts:	.80		.69		.51		.44	
solution #1	no shifts:		.83		.85		.85		.89
676 eqs.	with shifts:	.77		.60		.61		.29	
solution #2	no shifts:		.81		.81		.91		.86
1,024 eqs.	with shifts:	.79		.81		.79		.43	
	no shifts:		.83		.90		.95		.90
1,600 eqs.	with shifts:	.85		.85		.78		.49	
PSP = 0.1	no shifts:		.87		.92		.95		.93
1,600 eqs.	with shifts:	.85		.85		.78		.53	
PSP = .5	no shifts:		.87		.92		.95		.93
2,025 eqs.	with shifts:	.79		.77		.70		.62	
	no shifts:		.82		.90		.91		.95

Let $e_*^{(r_*)}$ and $e^{(r)}$ be the error vectors for the r_* and rth iteration vectors during a pair of computer runs with the $*$ indicating the run with the a-shifts. Then the ratios listed in this table are $\|e_*^{(r_*)}\|/\|e_*^{(r_*-1)}\|$ and $\|e^{(r)}\|/\|e^{(r-1)}\|$ when ω_{i-1} is updated to ω_i between steps $\#r_* - 1$ and $\#r_*$ and steps $\#r - 1$ and $\#r$, respectively.

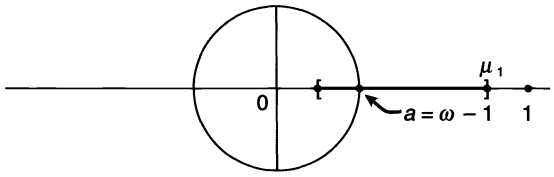


FIG. 5.1

whose unique root is at $x = a$ (also $P(1) = 1$). Since such a $P_1(x)$ is an affine function, these polynomial accelerations will shift the eigenvalues of $\mathscr{L}_\omega$ by the affine transformation $P_1(x)$ fixed at the point one. Therefore the a-shift, when $a = \omega - 1$, will shift the eigenvalue banjo of $\mathscr{L}$ to this eigenvalue banjo for $P_1(\mathscr{L}_\omega)$ (see Fig. 5.2). For example, when $a \approx .825$ (as it will be in the case of our computer run with 1,024 equations), the eigenvalue banjos are as shown in Fig. 5.3.

In our computer run, a-shifts were done at steps 5, 15, 32, and 73 (Graph 4 and Table 4.1). At step 73, $\omega = 1.825$ and so Fig. 5.3 is a correct picture of the eigenvalue banjo. We will now examine the changes in the eigenvector coordinates of the error vector during steps 33–73 in order to explain why the a-shift at step 73 reduced the 2-norm of the error vector to less than half its size instead of increasing it tenfold. Let

$$e^{(33)} = \sum_{1}^{1024} c_i u_i \tag{5.1}$$

be an error vector with u_1 and $u_{1,024}$, the eigenvectors of $\mathscr{L}_\omega$, corresponding to the eigenvalues with largest and smallest real part; the u_i are the unit eigenvectors of $\mathscr{L}_\omega$. In our example we will have

$$P_1(\mathscr{L}_{1.825}(u_{1,024})) \approx -9.4 c_{1,024} u_{1,024}. \tag{5.2}$$

Thus, this a-shift increases the coefficients of the $u_{1,024}$-part of the error vector by a factor of 9.4. On the face of it, this appears to be masochistically counterproductive. But

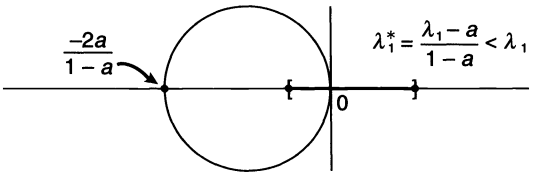


FIG. 5.2

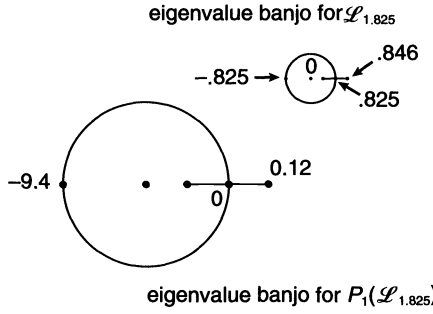


FIG. 5.3

when we look at the computer results (Graph 4 and Table 4.2), we see that the 2-norm of the error actually *drops* by more than $\frac{1}{2}$! How is this possible?

The explanation lies in the fact that the error vector was concentrated in the eigenvectors associated with the large real eigenvalues by the earlier SOR iterations which were performed using the preceding relaxation factor ω_3. We shall explain this.

Let

$$e^{(33)} = b_1 w_1 + \sum_{i=2}^{r-1} b_i w_u + \sum_{i=r}^{1,024} b_i w_i$$

be the eigenvector expansion of the error vector after the 33rd SOR iteration of our computer run (Graph 4); the w_i are the unit eigenvectors of $\mathscr{L}_{1.766}$. Here

$$\mathscr{L}_{1.766}(w_1) = \lambda_1 w_1 \approx .92 w_1,$$

and

$$\mathscr{L}_{1.766}(w_i) = \lambda_i w_i \quad \text{when } |\lambda_i| = .766$$

(and these λ_i, $i \geqq r$, are the complex eigenvalues of $\mathscr{L}_{1.766}$). (In our particular computer run with 1,024 equations, $r = 2$, but this is not important in this analysis.) The ratios of the 2-norms of the error vectors (also the difference vectors) is

$$\frac{\|e^{(32)}\|}{\|e^{(33)}\|} > .9 > .8 > .766 = \omega - 1.$$

The important consequence of this is that the contribution of the complex eigenvalues to the error vector is less than the contribution of the real eigenvalue λ_1 (and $\lambda_1 > .9$). Thus

$$\|b_1 w_1\| > \left\| \sum_{i=r}^{1,024} b_i w_i \right\|. \tag{5.3}$$

Forty iteration steps later, we see that the 72nd error vector is

$$e^{(72)} \approx \lambda_1^{40} b_1 w_1 + \sum_{i=2}^{r-1} (\lambda_i)^{40} b_i w_i + \sum_{i=r}^{1,024} \lambda_i^{40} b_i w_i,$$

$$\left|\frac{\lambda_i}{\lambda_1}\right|^{40} \approx \frac{1}{800} \quad \text{for } i = r, r+1, \cdots, 1024.$$

Combining this, we have

$$\frac{\|b_1 \lambda_1^{40} w_1\|}{\|\sum_{i=r}^{1,024} b_i \lambda_i^{40} w_i\|} > 800. \tag{5.4}$$

Thus, in the 72nd error vector, the contribution of the largest eigenvalue is more than 800 times the size of the contribution of the complex eigenvalues.

After the 72nd step, the relaxation factor was changed, which resulted in a change in the eigenvectors. Fortunately, the span of the complex eigenvectors of $\mathscr{L}_{1.766}$ is "close" to the span of the corresponding complex eigenvectors of $\mathscr{L}_{1.825}$.

Therefore, it is most likely the case that

$$|c_1| > 25 |c_{1,024}|$$

where c_1 and $c_{1,024}$ come from (5.1) and the ratio of 25 is much less than the ratio of 800

in (5.4), and hence

$$9.4\,|\,c_{1,024}\,| < .4\,|\,c_1\,|.$$

This is why increasing $c_{1,024}$ by a factor of 9.4 does not prevent the size of the error vector from being cut by a factor of .4 (while doing the a-shift).

The same type of analysis will explain why the a-shifts at the other steps when an ω_i was being updated (see Table 4.2) did not increase and instead actually decreased the size of the error in each of our seven computer runs.

We do not have proof that this type of analysis will work in all cases. Of course, when the relaxation factor is greater than the 1.872 (of our computer run with 2,025 equations), the factor, by which the coefficient ($c_{1,024}$) is multiplied, will be larger than 14. But our results certainly suggest that further experimentation with these a-shifts should be fruitful.

Table 5.1 compares the "blowup" factors to the actual reduction of the 2-norms of the error vectors as the last a-shift was done in each of our seven computer runs.

6. Summary. When the SOR iteration method is used to solve problems in which Young's theory is applicable, it is well known that the SOR iteration matrix $\mathscr{L}_\omega$ is diagonalizable except when $\omega - 1$ is an eigenvalue. In this case there is a principal vector of grade 2 associated with the eigenvalue $\omega - 1$. It is well known that this principal vector slows down the convergence process in a manner illustrated by the hillside-like figure in the graph of the ratios of successive error vectors. (See the diagram in § 2.)

We introduced an affine polynomial acceleration $P_1(\mathscr{L}_\omega)$, which we named an a-shift. In theory, these a-shifts should "kill off" the undesirable principal vector and thereby remove the undesirable hillside effect, by shifting the iteration error-vectors into

TABLE 5.1
Blowup factors at a-shifts.

Computer run	Last shift performed at step #	Blowup factor at shift	×	Partial decontamination factor when using previous ω	=	Product	<	Observed reduction at shift
625 eqs.	44	6.7	×	5×10^{-8}	=	3×10^{-7}	<	.61
676 eqs. vector #1	66	7.5	×	.0002	=	.001	<	.44
676 eqs. vector #2	77	7.9	×	.009	=	.07	<	.29
1,024 eqs.	72	9.4	×	.001	=	.01	<	.43
1,600 eqs. PSP = .01	84	12.2	×	5×10^{-6}	=	6×10^{-5}	<	.49
1,600 eqs. PSP = .5	47	12.7	×	.015	=	.19	<	.53
2,025 eqs.	66	13.6	×	.0001	=	.002	<	.62

Blowup factor = $-P_1(-a_4) = 2a_4/(1 - a_4)$, where a_4 is the last value of a.

Partial decontamination factor = $(a_3/\lambda_1)^m$, where a_3 is the next-to-last value of a; λ_1 is the largest eigenvalue of $\mathscr{L}_\omega$, for the next-to-last ω; m is the number of iteration steps using the next-to-last ω.

Product = (Blowup factor) × (Decontamination factor).

Observed reduction is the factor by which the 2-norm of the error vector was reduced by the a-shift (as in Table 4.2).

the span of the eigenvectors of $\mathscr{L}_\omega$ (see Theorem 3.1). Separately, we noted that when the red-black ordering is used, applying Sheldon's method may be more appropriate.

As is the usual situation with the adaptive SOR method, the final ω is not equal to the optimal ω_b, but is merely close to ω_b. This results in there not being any principal vector of $\mathscr{L}_\omega$. Still, when $\omega \approx \omega_b$, there is a pair of eigenvalues which are close to $\omega - 1$. The principal vector-eigenvector pair is replaced by a "nearby" eigenvector pair. As is common in numerical calculations, this results in a practical effect which is similar to the theoretical one. This is clearly demonstrated by the existence of the undesirable hillsides on the right sides of all the top graphs of Graphs 1–7.

The hillside was (essentially) eliminated by the a-shift only in the run presented in Graph 3. In each of our six other runs, the theoretical effect of eliminating the hillside was approximated by the practical result of a sizable reduction in both the height and length of the hillside. The differences in the hillsides caused by the a-shift is clearly visible on the right side in each of the top graphs of Graphs 1–7. The quantitative improvement factors that resulted from these smaller hillsides are listed in the "lower hillside" column of Table 6.1.

That each of our computer runs had its own individual features is easily observed by examining Graphs 1–7. This was even true for the two runs with 676 equations which had the same matrix, the only difference in the two problems being the two different randomly generated solution vectors. Just observe the large differences in the top graphs of Graphs 2 and 3.

Another consequence of using a-shifts is as follows. According to Table 4.1, in five of our seven computer runs, four to nine less iteration steps were performed before ω was updated to the final ω. This resulted in four to nine iteration steps being performed when a lower spectral radius was operative. This reduced the 2-norms of the error vectors by the factors listed in the "Earlier update to final ω" column of Table 6.1.

As the spectral radius of $P_1(\mathscr{L}_\omega)$ will be ten or more times the spectral radius of $\mathscr{L}_\omega$ (when $\omega \geqq 1.83$), it would be reasonable to predict that an a-shift would greatly increase the size of the error vector. In § 5, we explained why this does *not* occur with the adaptive SOR method where the a-shift is done only after the many SOR iterations (using the previous ω) have already greatly reduced the relative contribution to the error vector made by the complex eigenvalues (of the previous $\mathscr{L}_\omega$).

Instead, the 2-norms of the error vectors actually decreased at each a-shift. (The numerical results are tabulated in Table 4.2.) In fact, as the final a-shifts of our computer runs were performed, the 2-norms of the error vectors decreased by factors ranging from 0.3 to 0.64. See the "last a-shift" column of Table 6.1.

We now present Table 6.1, which tabulates the sizes of the various improvements (error-ratio drops) that resulted from using the a-shift. The column headings are now defined.

The error ratio of the 2-norms of the error vectors after Step #n is denoted by $R_n = \|e_*^{(n)}\|/\|e^{(n)}\|$, $n = 0, 1, 2, \cdots$.

Let N and N_* be the step numbers for the first step using the final ω for the unshifted run and the shifted run, respectively.

The ratio before last update is R_{N_*-1}. This ratio measures the amount of improvement that resulted from the a-shifts (not including the last one).

The error ratio drop at the last a-shift is

$$\frac{R_{N_*}}{R_{N_*-1}}.$$

TABLE 6.1
Error-ratio drops.

Computer run	Ratio before last update to final ω	×	Last a-shift	×	Earlier update to final ω	×	Lower hillside	×	After the hillside	=	Final error ratio
625 eqs.	0.94		0.64		0.94		0.55*				.31
676 eqs. solution #1	0.67		0.5		0.6		0.51		0.41		.042
676 eqs. solution #2	0.39		0.34				0.23**				.031
1,024 eqs.	0.63		0.47		0.77		0.46		0.55		.057
1,600 eqs. PSP = .01	0.89		0.52		0.76		0.59*				.2
1,600 eqs. PSP = .5	0.8		0.56		0.84		0.58		1.8		.4
2,025 eqs.	0.89		0.64		0.88		0.26		0.57		.072

* The runs, for 625 equations and for 1,600 equations with PSP = .01, ended before the successive error ratios reached the bottom of the hillside.

** For solution #2, the unshifted run updated to the last ω first. Also, the run ended before the unshifted run reached the bottom of the hillside.

The error ratio drop due to the earlier update to the final ω is

$$\frac{R_N}{R_{N_*}}.$$

Let M be the step number at which a run arrives at the bottom of the hillside (i.e., the successive error ratio is $\omega - 1$).

The error ratio drop while "going down the hillside" is

$$\frac{R_M}{R_N}.$$

Let M_2 be the last iteration step number. The error ratio drop after the hillside is (when applicable)

$$\frac{R_{M_2}}{R_M}.$$

The final error ratio is

$$R_{M_2} = R_{N_*-1} \times \frac{R_{N_*}}{R_{N_*-1}} \times \frac{R_N}{R_{N_*}} \times \frac{R_M}{R_N} \times \frac{R_{M_2}}{R_M}.$$

The costs in time involved in the calculations of the a-shifts are inconsequential since less than six of them are done in each run and each a-shift involves only a single vector addition and two scalar times vector multiplications.

In future papers, we plan to investigate the effects of a-shifts (i) while using the red-black ordering, (ii) while using the nonadaptive SOR method, and (iii) on certain worst-case–type situations for the adaptive SOR method.

7. Conclusions. In our computer runs (all with dictionary orderings) the use of a-shifts reduced the 2-norms of the error vectors by factors ranging from 0.03 to 0.4. We note that the implementation of our variation (employing a-shifts) of the adaptive SOR

method is both simple and virtually cost-free. Therefore, it seems appropriate to employ and to experiment further with these a-shifts whenever the adaptive SOR method (without a red-black ordering) is being used to solve a system of equations in which Young's theory [4] is applicable.

REFERENCES

[1] L. A. HAGEMAN AND R. B. KELLOGG, *Estimating optimum acceleration parameters for use in the successive overrelaxation and the Chebyshev polynomial methods of iteration*, WAPD-TM-592, 1966. (Available from The National Technical Information Service, Springfield, VA.)

[2] L. A. HAGEMAN AND D. YOUNG, *Applied Iterative Methods*, Academic Press, New York, 1981.

[3] J. W. SHELDON, *On the spectral norms of several iterative processes*, J. Assoc. Comput. Mach., 6 (1959), pp. 494–505.

[4] D. M. YOUNG, *Iterative Solutions of Large Linear Systems*, Academic Press, New York, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 4, pp. 679–689, October 1991

006

ORDER REDUCTIONS OF THE MARGINALS AND IDENTIFICATION OF MULTIPLE ARMA MODELS*

ANTONIE STAM† AND STEVEN C. HILLMER‡

Abstract. General conditions are derived for order reductions in the marginals of multivariate ARMA time series models. These reductions are shown to be related to the structure of the autoregressive part of the model. Particular model structures, such as block diagonal and block triangular, are analyzed as special cases because of their practical relevance for multivariate time series modeling. It is shown that the occurrence of order reductions is closely related to the issue of model identification in multiple time series.

Key words. multivariate time series, model identification, marginal models

AMS(MOS) subject classifications. 15A18, 15A21, 62H05, 62M10, 62P20, 90A20, 93E12

1. Introduction.

The general k-variate ARMA$_k(p, q)$ model can be written as

$$\Phi(B)\mathbf{z}_t = \Theta(B)\mathbf{a}_t, \tag{1.1}$$

where $\mathbf{z}_t = (z_{1t}, \cdots, z_{kt})'$ is a k-variate vector time series, $\mathbf{a}_t = (a_{1t}, \cdots, a_{kt})'$ is a k-variate white noise vector with mean $\mathbf{0}$ and covariance matrix $\Sigma_{\mathbf{a}}$, and $\Phi(B)$ and $\Theta(B)$ are polynomial matrices in B, the backshift operator, of order p and q, respectively, with roots on or inside the unit circle. $\Phi(B)$ and $\Theta(B)$ can be represented as follows:

$$\begin{cases} \Phi(B) = \mathbf{I} - \Phi_1 B - \cdots - \Phi_p B^p \\ \Theta(B) = \mathbf{I} - \Theta_1 B - \cdots - \Theta_q B^q \end{cases}. \tag{1.2}$$

A problem which has been previously studied (see [18], [3]) is that of determining the form of the models for the individual series, z_{it}, or the marginal models, in the case where the vector series is known to follow the form (1.1). Stam and Hillmer [14] explore this problem for the particular case of multivariate $AR_k(1)$ models. The purpose of this paper is to explore additional aspects related to determination of the marginals of (1.1). In particular, we extend the results of our earlier paper to the more general ARMA(p, q) case, and characterize more fully the relationship between the multivariate structure of the autoregressive operator and the marginal models in some specific situations. We also explore the problem of the identifiability of a particular parameterization related to the order reductions of the marginal models.

We illustrate by example that order reductions can occur not only when the coefficient matrices are sparse and contain many zero values, but also when these matrices are dense.

2. Preliminary results.

The typical approach to determining the form of the marginal models from the given multivariate model (1.1) is as follows. Define $|\Phi(B)|$ and Adj $[\Phi(B)]$ to be the determinant and the classical adjoint of the matrix $\Phi(B)$, respectively. Then premultiplying both sides of (1.1) by Adj $[\Phi(B)]$ gives

$$|\Phi(B)|\mathbf{z}_t = \text{Adj}\,[\Phi(B)]\Theta(B)\mathbf{a}_t. \tag{2.1}$$

It follows from (2.1) that the model for z_{it}, $i = 1, \cdots, k$ is an ARIMA $(m_1, m_2, (k-1)p + q)$, $m_1 + m_2 = kp$, where m_1 is the order of the stationary auto-

* Received by the editors February 13, 1989; accepted for publication (in revised form) June 15, 1990.
† Department of Management Sciences and Information Technology, College of Business Administration, University of Georgia, Athens, Georgia 30602 (astam@uga.bitnet).
‡ School of Business, University of Kansas, Lawrence, Kansas 66045.

regressive component and m_2 is the degree of nonstationarity (unit roots); and that the models for each of the individual series have identical autoregressive parts. It has been recognized [18], [3] that the stated orders may be reduced by cancellation of common factors in the autoregressive and moving average sides of the model, but the identification aspects of such order reductions has not been considered.

We use results developed for the first-order autoregressive case by Stam and Hillmer [14]. Therefore, it is desirable to write the model (1.1) as a higher-order $\mathrm{ARMA}_{kp}(1, q)$ model. This can be done by considering

$$(\mathbf{I}-\mathbf{\Phi}^* B)\mathbf{y}_t=\mathbf{\Theta}^*(B)\mathbf{b}_t, \tag{2.2}$$

where $\mathbf{y}_t = (\mathbf{z}_t', \cdots, \mathbf{z}_{t-p+1}')'$, $\mathbf{b}_t = (\mathbf{a}_t', \mathbf{0}', \cdots, \mathbf{0}')'$,

$$\mathbf{\Phi}^* = \begin{bmatrix} \mathbf{\Phi}_1 & \mathbf{\Phi}_2 & \cdots & \mathbf{\Phi}_{p-1} & \mathbf{\Phi}_p \\ \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ & \ddots & \ddots & \vdots & \vdots \\ \mathbf{0} & & & \mathbf{I} & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \mathbf{\Theta}^*(B) = \begin{bmatrix} \mathbf{\Theta}(B) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & & & \\ \vdots & & & \\ \mathbf{0} & & \mathbf{0} & \end{bmatrix}.$$

Based upon $\mathbf{\Phi}^*$, the λ-matrix $(\mathbf{I}\lambda - \mathbf{\Phi}^*)$ can be defined. Using the expression (2.2), the results derived by Stam and Hillmer [14] for the first-order autoregressive case will be extended to general form multiple ARIMA models.

Next, we review some useful results concerning the matrix $(\mathbf{I}\lambda - \mathbf{\Phi}^*)$, which depend on the eigenvalues and eigenvectors of $\mathbf{\Phi}^*$. It can be shown (see [13]) that the eigenvalues of $\mathbf{\Phi}^*$ are the same as the zeros of $|\mathbf{I}\lambda^p - \mathbf{\Phi}_1\lambda^{p-1} - \cdots - \mathbf{\Phi}_p|$. If $\bar{\lambda}$ is an eigenvalue of $\mathbf{\Phi}^*$, then any nonzero vector $\mathbf{x}$ satisfying $(\mathbf{I}\bar{\lambda} - \mathbf{\Phi}^*)\mathbf{x} = \mathbf{0}$ is an eigenvector of $\mathbf{\Phi}^*$. If any $\bar{\lambda}$ has algebraic multiplicity greater than its geometric multiplicity, then a nonzero generalized eigenvector $\mathbf{x}_j$ satisfying $(\mathbf{I}\bar{\lambda} - \mathbf{\Phi}^*)\mathbf{x}_j = \mathbf{x}_{j-1}$ can be found, where $\mathbf{x}_{j-1}$ is either an eigenvector or a generalized eigenvector corresponding to $\bar{\lambda}$ (see [16, p. 360]).

Let $\mathbf{A}^*(\lambda)$ be the adjoint of the matrix $(\mathbf{I}\lambda - \mathbf{\Phi}^*)$ and define

$$\mathbf{C}^*(\lambda)=\mathbf{C}_0^*+\mathbf{C}_1^*\lambda+\cdots+\mathbf{C}_l^*\lambda^l \tag{2.3}$$

as the reduced adjoint obtained from $\mathbf{A}^*(\lambda)$ by cancelling any common factors $d(\lambda)$ in $\mathbf{A}^*(\lambda)$, so that $\mathbf{C}^*(\lambda)d(\lambda) = \mathbf{A}^*(\lambda)$. In addition, the minimal polynomial $\psi(\lambda)$ corresponding to $(\mathbf{I}\lambda - \mathbf{\Phi}^*)$ has the property that $\psi(\lambda)d(\lambda) = |\mathbf{I}\lambda - \mathbf{\Phi}^*|$. The following lemma is proved in [5, p. 168].

LEMMA 1. *Let λ_j be an eigenvalue of $\mathbf{\Phi}^*$ and let v_j be the dimension of the largest Jordan block associated with λ_j. For $u = 1, \cdots, v_j$, let q_u denote the number of Jordan blocks of size $(v_j + 1 - u)$. For $q_u > 0$, let $\mathbf{p}_{uh}$, $h = 1, \cdots, q_u$ denote the eigenvector corresponding to the hth Jordan block of size $(v_j + 1 - u)$ and let $\mathbf{p}_{uh}^l$, $l = 2, \cdots, v_j + 1 - u$ denote the $(l-1)$st generalized eigenvector of that block. If $\mathbf{C}^{*(k)}(\lambda_j)$ is the kth derivative of $\mathbf{C}^*(\lambda)$ with respect to λ evaluated at $\lambda = \lambda_j$, then the column space of $\mathbf{C}^{*(n-1)}(\lambda_j)$ for $n = 1, \cdots, v_j$ is spanned by the vectors $\mathbf{p}_{uh}^l$ for $u = 1, \cdots, n$, in which $q_u > 0$, $l = 1, \cdots, n + 1 - u$, and $h = 1, \cdots, q_u$.*

Results in terms of λ_j may be expressed equivalently in terms of $B_j = 1/\lambda_j$ $(\lambda_j \neq 0)$ (see [14]). We find it convenient to do so throughout the paper.

Since in (2.2), $\mathbf{y}_t = (\mathbf{z}_t', \cdots, \mathbf{z}_{t-p+1}')'$, it is sufficient to analyze the marginal models associated with the first k rows of $\mathbf{\Phi}^*$ (see [15]). Premultiplying (2.2) by $\mathbf{A}^*(B)$ gives

$$|\mathbf{I}-\mathbf{\Phi}^* B|\,\mathbf{y}_t=\mathbf{A}^*(B)\mathbf{\Theta}^*(B)\mathbf{b}_t. \tag{2.4}$$

By writing out (2.4) according to the partitions introduced below equation (2.2), it

follows that

$$(2.5)\qquad \begin{cases} |\mathbf{I}-\Phi^* B|\mathbf{z}_t = \mathbf{A}_{11}^*(B)\Theta(B)\mathbf{a}_t \\ \quad\vdots \\ |\mathbf{I}-\Phi^* B|\mathbf{z}_{t-p} = \mathbf{A}_{p1}^*(B)\Theta(B)\mathbf{a}_t \end{cases}$$

where

$$\mathbf{A}^*(B) = \begin{bmatrix} \mathbf{A}_{11}^*(B)\cdots\mathbf{A}_{1p}^*(B) \\ \vdots \qquad\qquad \vdots \\ \mathbf{A}_{p1}^*(B)\cdots\mathbf{A}_{pp}^*(B) \end{bmatrix}$$

is partitioned in the same way as Φ^*, so that $\mathbf{A}_{ij}^*(B)$ is of dimension $k \times k$. From (2.5), $\mathbf{A}_{p1}^*(B) = B\mathbf{A}_{p-1,1}^*(B) = \cdots = B^{p-1}\mathbf{A}_{11}^*(B)$. Since $\mathbf{A}_{p1}^*(B)$ is of order at most $kp-1$, $\mathbf{A}_{11}^*(B)$ is of order at most $kp-1-(p-1) = (k-1)p$. In addition, $\Theta(B)$ is of order q. Thus the first equation in (2.5) has marginals of maximal order ARIMA $(m_1, m_2, (k-1)p+q)$ where $m_1 + m_2 = kp$.

3. Intrablock and local reductions. We are interested in the form of the marginal models corresponding to the joint model (1.1). From (2.1) we know that the model for z_{it} is of the ARIMA form with autoregressive order at most kp and moving average order at most $(k-1)p+q$. Cancellation of common factors in (2.1) may lead to reduced orders. We will consider three different types of order reduction. A reduction in every marginal model within the same subset (block) of mutually Granger causally dependent variables [7] due to the same cause (condition) will be called an *intrablock reduction*, and reduction in only some of the marginal models within a subset of Granger causally dependent variables will be called a *local reduction*. The third type of order reduction is *interblock reduction* and involves the case where the coefficient matrices can be written in block diagonal or block triangular form. We first introduce the specific structure of block diagonal and triangular matrices. Interblock reductions are further specified in § 5.

Partition the polynomial matrix $\Phi(B)$ into blocks $\Phi_{ij}(B)$, $i = 1, \cdots, m$ and $j = 1, \cdots, m$ of dimensions $k_i \times k_j$ where $\sum_{i=1}^m k_i = k$. If $\Phi(B)$ can be written such that $\Phi_{ij}(B) = \mathbf{0}$ whenever $i \neq j$, we say that $\Phi(B)$ is *block diagonal*, and if $\Phi_{ij}(B) = \mathbf{0}$ whenever $i < j$, we say that $\Phi(B)$ is *block lower triangular*. We can partition $\mathbf{z}_t$, $\mathbf{a}_t$, and $\Theta(B)$ in a way conformable to $\Phi(B)$ so that $\mathbf{z}_t = [\mathbf{z}_{1t}', \cdots, \mathbf{z}_{mt}']'$, $\mathbf{a}_t = [\mathbf{a}_{1t}', \cdots, \mathbf{a}_{mt}']'$ with $\mathbf{z}_{it}$ and $\mathbf{a}_{it}$ of dimensions $k_i \times 1$, and $\Theta(B) = [\Theta_{ij}(B)]$ with blocks Θ_{ij} of dimensions $k_i \times k_j$.

Each set of block diagonal variables $\mathbf{z}_{it}$ can be considered a "subsystem," in which each variable is Granger causally dependent on each of the other variables in the subsystem, but does not depend on any variables in other subsystems. Similarly, each set $\mathbf{z}_{it}$ of lower block triangular variables does not depend in the Granger causal sense on subsets below it in the partitioned matrix, and may depend causally on the subsets above it in the matrix [13].

In the remainder of this section, we assume that all variables in the multivariate time series model have a "feedback" relationship and are causally related in the Granger sense [6], [7], [11]. Thus, the properties derived in this section relate to intrablock and local reductions only.

3.1. Intrablock reductions. The following result characterizes the circumstances under which intrablock reductions occur.

RESULT 1. *Suppose* $\mathbf{y}_t = (\mathbf{z}_t', \cdots, \mathbf{z}_{t-p}')'$ *follows the model* (2.2) *and there is no block diagonal or block triangular structure within* Φ^*. *Suppose* Φ^* *has distinct non-*

zero eigenvalues $\lambda_1, \cdots, \lambda_m$ *and* $\lambda_0 = 0$, *and let* r_0 *be the algebraic multiplicity of* λ_0. *Let* v_j *be the size of the largest Jordan block in the Jordan canonical form corresponding to* λ_j, $j = 0, \cdots, m$. *Then the model for* z_{it} $(i = 1, \cdots, k)$ *is an* ARIMA $(m_1, m_2, r_0 + v + q - p)$ *where* $m_1 + m_2 = v = \sum_{j=1}^{m} v_j$.

The proof follows by multiplying both sides of (2.2) by $\psi^*(B)$, the minimal polynomial associated with Φ^*. This is the same approach as in Theorem 1 of Stam and Hillmer [14, p. 91]. The marginal models are represented by the first k rows of the resulting model

$$\psi^*(B)\mathbf{y}_t = \mathbf{C}^*(B)\Theta^*(B)\mathbf{b}_t. \tag{3.1}$$

Due to the zero coefficients in $\Theta^*(B)$, (3.1) can be written as (3.2), where $\mathbf{C}_{11}^*(B)$ is the upper left $k \times k$ block of $\mathbf{C}^*(B)$.

$$\psi^*(B)\mathbf{y}_t = \mathbf{C}_{11}^*(B)\Theta^*(B)\mathbf{b}_t. \tag{3.2}$$

The result follows directly from (3.2); a detailed proof can be found in [15]. Result 1 implies that the way to get intrablock reductions in the orders of the marginals is to have common eigenvalues in Φ^*. Intrablock reductions apply to each of the marginals in the block. It is also possible to have order reductions for an individual marginal within a block.

3.2. Local reductions. Using Lemma 1 and [14], [15], it can be shown that the eigenvectors and generalized eigenvectors of Φ^* play an important role in determining local reductions in the marginal models. This is summarized in the following result.

RESULT 2. *Suppose that after accounting for intrablock reductions, the multivariate model follows* (3.2). *Then for a given nonzero eigenvalue* λ_j *of* Φ^*, *there will be a local reduction in the marginal model for* z_{it} *of magnitude exactly* s_i $(s_i \leqq v_j)$, $i = 1, \cdots, k$ *if and only if the following conditions hold*:

(i) *All of the eigenvectors* $\mathbf{p}_{uh}^{l}$ *of* Φ^* *for* $u = 1, \cdots, s_i - 1$ *in which* $q_u > 0$, $l = 1, \cdots, s_i - u$ *and* $h = 1, \cdots, q_u$ *have a zero in the ith row, and*

(ii) *At least one of the eigenvectors* $\mathbf{p}_{s_i,h}^{1}, \mathbf{p}_{s_i-1,h}^{2}, \cdots, \mathbf{p}_{1,h}^{s_i}$ $h = 1, \cdots, q_u$ *does not have a zero in the ith row.*

The basic idea of this result is that local reductions will occur when there is a common factor in every element of the ith row of Adj $[\Phi(B)]$ that is also included in $|\Phi(B)|$. In this situation, the common value can be factored out of each element in the ith row of Adj $[\Phi(B)]$ and cancelled with the $|\Phi(B)|$ term for the ith marginal model. Lemma 1 provides necessary and sufficient conditions on the eigenvalues and generalized eigenvalues that will result in the above cancellation.

Results 1 and 2 provide necessary and sufficient conditions for intrablock and local reductions due to the structure of Φ^*. In the next section we investigate how intrablock and local reductions can affect the identifiability of the parameterization (2.1). In § 5 we consider the ways in which strategically located zeros in $\Phi^*(B)$ of (2.2) can lead to interblock reductions in the marginal models.

4. Identifiability. Some authors (e.g., [18], [17], [3]) have suggested that the multivariate ARMA model in (1.1) be reparameterized as in (2.1) because of the simple structure of the resulting autoregressive component in (2.1). The results derived in this paper have relevance for the identifiability of the parameterization in (2.1). In order to understand the issue, we briefly discuss the concept of identification of multiple ARMA models. A more complete discussion is given in [8]–[10] and [12, pp. 801–804].

Assuming that $\mathbf{z}_t$ conforms to the Gaussian model (1.1), identification requires unique determination of the values of p and q along with the matrices $\Phi_1, \cdots, \Phi_p$,

$\Theta_1, \cdots, \Theta_q$ from the covariances of $\mathbf{z}_t$, given a sufficiently large number of observations. If this can be done, the multiple ARMA model is identified. The model (1.1) is not generally identified without imposing additional restrictions upon the parameters. For instance, if $\mathbf{D}(B)$ is an arbitrary matrix polynomial in B, then the covariance structure of $\mathbf{z}_t$ will remain unaffected if we premultiply both sides of (1.1) by $\mathbf{D}(B)$. Thus, assuming Gaussian errors corresponding to the given covariance structure of $\mathbf{z}_t$, there will be a likelihood equivalence class of models, all of which will lead to the same likelihood function. The problem of identification is to constrain the parameters so that a unique member of this equivalence class is selected.

Hannan [10] shows that the following four conditions identify the model (1.1): (1) The leading coefficient in $\Phi(B)$ and $\Theta(B)$ is $\mathbf{I}$; (2) The zeros of $|\Phi(B)|$ and $|\Theta(B)|$ lie outside the unit circle; (3) The matrix polynomials $\Phi(B)$ and $\Theta(B)$ have no common left divisors, implying that if $\mathbf{D}(B)$ is a matrix polynomial such that $\Phi(B) = \mathbf{D}(B)\Phi_1(B)$ and $\Theta(B) = \mathbf{D}(B)\Theta_1(B)$, then $|\mathbf{D}(B)|$ is a constant independent of B; (4) The matrix $[\Phi_p \vdots \Theta_q]$ is of full rank k. Hannan has also shown that when $\Phi(B)$ is constrained, the first three conditions with an alternative fourth condition are sufficient to identify the model. In particular, if $\Phi(B)$ has a specific diagonal form $\Phi(B) = \alpha(B)\mathbf{I}$ where $\alpha(B) = 1 - \alpha_1 B - \cdots - \alpha_m B^m$, then the model is identified provided that $\alpha_m \neq 0$ and the first three conditions mentioned above are met. If $\Phi(B)$ is constrained to be lower triangular with the (i, j)th element denoted by $\phi_{ij}(B)$, the fourth condition for identifiability is that $\phi_{ij}(B)$ must not be of higher degree than $\phi_{jj}(B)$. A similar condition holds for upper triangular $\Phi(B)$.

Hannan [10] points out that issues similar to the identification of (1.1) arise in the context of control engineering where (1.1) is put in state space form. Akaike [1], [2] discusses a state space formulation of (1.1) and considers the possibility of achieving identification of the model (1.1) by making use of a state space representation and results from control theory. He introduces the idea of "block identifiability," which is based upon the condition of controllability that is common in the optimal filtering literature. Hannan [10] and Priestley [12, pp. 801–804] show that "block identifiability" is equivalent to conditions (1) through (3) above and to Θ_q being nonsingular when $q \geqq p$, Φ_p being nonsingular if $p \geqq q + 1$, and $[\Phi_p \vdots \Theta_q]$ being of full rank. Thus, block identifiability is often equivalent to the conditions derived by Hannan for the identifiability of (1.1).

Suppose that equation (1.1) is identified. Then, if one wishes to parameterize the model as in (2.1), it does not follow that this parameterization is identified. In particular, we argue that failure to recognize the types of cancellations discussed in this paper will result in a model of the form (2.1) that is not identified. For instance, consider the model (2.2), let $\psi^*(B)$ be the minimal polynomial of $\Phi^*(B)$, and let $d(B)$ be a polynomial such that $\psi^*(B)d(B) = |\Phi^*(B)|$. It follows that the adjoint $\mathbf{A}^*(B) = d(B)\mathbf{C}^*(B)$, so that the matrix $d(B)\mathbf{I}$ is a common left divisor of $|\Phi^*(B)|\mathbf{I}$ and of $\mathbf{A}^*(B)$. Thus, if there are intrablock reductions in the marginal models, it follows that $d(B) \neq 1$ and the parameterization (2.1) is not identified. Intrablock reductions are described by Result 1 above.

Another instance where (2.1) will not be identified is when $|\Phi^*(B)|$ and every element in the ith row of $\mathbf{A}^*(B)$ have a common factor $(B - B_j)^{s_i}$ $(s_i > 0)$. In this case the diagonal matrix $\mathbf{D}(B)$ with diagonal elements equal to one, except for the factor $(B - B_j)^{s_i}$ in the ith position, is a common left divisor of $|\Phi^*(B)|\mathbf{I}$ and $\mathbf{A}^*(B)$. Note that a common factor in every element of the ith row of $\mathbf{A}^*(B)$ is equivalent to a common factor in every element of the ith row of $\mathbf{C}^*(B)$ [14], [15]. Furthermore, Lemma 1 and Result 2 show that local reductions imply this situation and will in turn lead to (2.1) not being identified.

In this particular instance it is of interest to investigate whether or not the parameterization (2.1) is identified after the cancellations mentioned previously have been made. One way to establish identifiability is to verify Hannan's conditions. The first condition is trivial. The second condition is true for (2.1) since the zeros of $|\text{Adj}\,[\Phi(B)]\cdot\Theta(B)| = |\text{Adj}\,[\Phi(B)]|\cdot|\Theta(B)|$ are outside the unit circle if the zeros of $|\Phi(B)|$ and $|\Theta(B)|$ are outside the unit circle. If all the possible cancellations have been made, the third condition will be satisfied. Finally, the reduced autoregressive operator in (2.1) after cancellations is diagonal and thus lower triangular with the off-diagonal terms of the same or lower degree than the diagonal terms. Therefore the relevant fourth condition is satisfied.

Thus, if one wishes to use the parameterization (2.1), it is important to recognize that intrablock and local reductions of the types discussed in this paper will cause identifiability problems due to a common left divisor $\mathbf{D}(B) \neq \mathbf{I}$ of $|\Phi^*(B)|\mathbf{I}$. These can be diagnosed by considering the eigenvalues, eigenvectors, and generalized eigenvectors of Φ^*.

Example 1. As an example of the identification problems that can occur if the types of cancellations discussed in this paper are ignored, consider the model $(\mathbf{I} - \Phi B)\mathbf{z}_t = \mathbf{a}_t$, where

$$\Phi = \begin{bmatrix} 1.000 & 0 & .500 & .500 \\ -.214 & .571 & -.286 & -.071 \\ -.071 & -.143 & .571 & -.357 \\ .071 & .143 & .429 & 1.357 \end{bmatrix},$$

for which the Jordan form Λ and the matrix of (generalized) eigenvectors $\mathbf{P}$ are

$$\Lambda = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}, \qquad \mathbf{P} = \begin{bmatrix} 0 & 1 & -1 & 2 \\ 0 & 0 & 1 & 1 \\ -2 & 1 & 0 & 3 \\ 1 & 2 & 1 & 0 \end{bmatrix}.$$

A check of conditions (1) through (4) of [8] verifies that this $AR_4(1)$ model is identified. Transforming this model as in (2.1) yields

$$(1-B)^3(1-.5B)\mathbf{z}_t = \text{Adj}\,[\Phi(B)]\mathbf{a}_t \tag{4.1}$$

with

$$\text{Adj}\,[\Phi(B)] = \begin{bmatrix} (1-B)^2(1-.857B) & -.5B(1-B)^2 \\ -.179B(1-B)^2 & (1-B)^2(1-.75B) \\ (-.393B+.464B^2)(1-B) & .25B(1+B)(1-B) \\ -.071B(1-B)(1-.5B) & -.5B(1-B)(1-.5B) \end{bmatrix}$$

$$\begin{bmatrix} -.071B(1-B)^2 & -.143B(1-B)^2 \\ .036B(1-B)^2 & -.071B(1-B)^2 \\ (1-B)(1-1.821B+.607B^2) & (-.644B+.215B^2)(1-B) \\ .215B(1-B)(1-.5B) & (1-B)(1-.5B)(1-.572B) \end{bmatrix},$$

and it follows that

$$\mathbf{D}(B) = \begin{bmatrix} (1-B)^2 & 0 & 0 & 0 \\ 0 & (1-B)^2 & 0 & 0 \\ 0 & 0 & (1-B) & 0 \\ 0 & 0 & 0 & (1-B)(1-.5B) \end{bmatrix}$$

is a common left divisor of $(1 - B)^3(1 - .5B)\mathbf{I}$ and Adj $[\Phi(B)]$. Thus, the model in (4.1) is not identified. If $\mathbf{D}(B)$ is factored from both sides, the resulting model is

$$\Phi^*(B)\mathbf{z}_t = \Theta^*(B)\mathbf{a}_t$$

with

$$\Phi^*(B) = \begin{bmatrix} (1-B)(1-.5B) & 0 & 0 & 0 \\ 0 & (1-B)(1-.5B) & 0 & 0 \\ 0 & 0 & (1-B)^2(1-.5B) & 0 \\ 0 & 0 & 0 & (1-B)^2 \end{bmatrix}$$

and

$$\Theta^*(B) = \begin{bmatrix} (1-.857B) & -.5B & -.071B & -.143B \\ -.179B & (1-.75B) & .036B & -.071B \\ -.393B+.464B^2 & .25B(1+B) & 1-1.821B+.607B^2 & -.644B+.215B^2 \\ -.071B & -.5B & .215B & (1-.572B) \end{bmatrix}.$$

Some authors have advocated using the model (2.1) when fitting multiple ARMA models to data, because of the apparent simplicity of the autoregressive component of this model form. In particular, the form of the autoregressive operator is diagonal with identical diagonal terms in (2.1) as long as there are no cancellations. As illustrated in this example, one way to have cancellations (and as a result a nonidentified model) is to have repeated eigenvalues and/or zeros in the eigenvectors. Thus, the apparent simplicity of (2.1) may be achieved at the cost of a nonidentified model.

5. Special structures. In the previous section, only systems in which all variables have "feedback" relationships were considered, i.e., the variables were all mutually causally dependent in the Granger sense and together formed one block. A number of authors (e.g., [18], [3]) have indicated that reductions occur in some of the marginal models if the $\Phi(B)$ operator in (1.2) has a triangular or diagonal form. In this section we establish sufficiency conditions on the structure of $\Phi(B)$ for order reductions. These order reductions will be referred to as interblock reductions. Note that such reductions are in fact a special case of local reductions, because in the case of block diagonal or triangular coefficient matrices the associated matrices of eigenvectors can be written in the same block diagonal or triangular structure. Result 2, applied to the zeros in the off-diagonal blocks of the matrix of eigenvectors, explains the interblock (local) reductions. Additionally, all properties of intrablock reductions apply within a given subsystem.

5.1. Φ is block diagonal. If $\Phi(B)$ is block diagonal, it follows from (1.1) that $\mathbf{z}_{it}$ as defined above satisfies

$$\Phi_{ii}(B)\mathbf{z}_{it} = \sum_{j=1}^{m} \Theta_{ij}(B)\mathbf{a}_{jt}. \tag{5.1}$$

However, the right hand side of (5.1) can be written as a moving average model of order at most q (see [4, Chap. 4]) so that the model for $\mathbf{z}_{it}$ is given by

$$\Phi_{ii}(B)\mathbf{z}_{it} = \Theta^*_{ii}(B)\mathbf{a}^*_{it}, \tag{5.2}$$

where $\Phi_{ii}(B)$ is a polynomial matrix of order p and $\Theta^*_{ii}(B)$ a polynomial matrix of order q. From (5.2) it follows that the autoregressive part of the marginal models in the ith

block depends only upon the properties of $\Phi_{ii}(B)$ and that Results 1 and 2 can be applied to $\Phi_{ii}(B)$ when determining the marginal models for the ith block. Thus a sufficient condition for interblock reductions in the marginal models is for $\Phi(B)$ to be block diagonal.

In the specific situation where $\Phi(B) = (\mathbf{I} - \Phi B)$ and the matrix Φ is diagonal, the above arrangement implies that the autoregressive portion of the marginal model for z_{it} is $(1 - \phi_{ii}B)$ where ϕ_{ii} is the diagonal term of the ith row of Φ. Conversely, it is of interest to investigate under which conditions the knowledge that the autoregressive part of each of the marginal models is of first order will imply a diagonal form of the multivariate model. The next result addresses this issue.

RESULT 3. *Suppose* $\mathbf{z}_t$ *follows the model* $(\mathbf{I} - \Phi B)\mathbf{z}_t = \Theta(B)\mathbf{a}_t$, *and the marginal model for* z_{it} *is of the form* $(1 - \phi_{ii}B)z_{it} = \theta_i(B)a_{it}^*(i = 1, \cdots, k)$, *where* $\theta_i(B)$ *is a scalar polynomial in* B *of order at most* q, *and* a_{it}^* *is white noise. Then it is true that*:

(i) *If the reductions due to the structure of* Φ *are associated with one eigenvalue, then* Φ *is diagonal of the form* $\Phi = \phi\mathbf{I}$;

(ii) *If the reductions due to the structure of* Φ *are not intrablock reductions, and due to more than one eigenvalue, then* Φ *is diagonal, and* $\phi_{ii} \neq \phi_{jj}$ *if* $i \neq j$.

Proof. First, assume that the reductions are due to only one eigenvalue. Since all the marginals have autoregressive components of order $p = 1$, Result 1 states that the k eigenvalues of Φ must all be equal and the dimensions of the largest Jordan block must be 1×1. This implies that the Jordan form is $\Lambda = \phi\mathbf{I}$, where ϕ is the common eigenvalue. Since $\Phi = \mathbf{P}\Lambda\mathbf{P}^{-1} = \mathbf{P}[\phi\mathbf{I}]\mathbf{P}^{-1} = \phi\mathbf{I}$, it follows that Φ is diagonal and of the form desired, which proves (i).

Next, assume that there is more than one eigenvalue, and that none of the reductions are intrablock reductions. Result 2 implies that every row of the matrix $\mathbf{P}$ of eigenvectors and generalized eigenvectors has $k - 1$ zeros and thus, without loss of generality, can be taken to be diagonal. If all the eigenvalues of Φ are different, Λ is diagonal and $\Phi = \mathbf{P}\Lambda\mathbf{P}^{-1}$ is diagonal. If Φ has a repeated eigenvalue, say α, then Λ must be of the form

$$\Lambda = \begin{bmatrix} \alpha & 1 & \mathbf{0} \\ 0 & \alpha & 0 \\ \mathbf{0} & 0 & \Lambda_{11} \end{bmatrix}.$$

But as stated above, the eigenvector for α has only one nonzero element, say in row i, and the generalized eigenvector for α has only one nonzero element in some row $j \neq i$. Using Result 2 this implies that the marginal model for z_{it} has an autoregressive order greater than one, which is a contradiction. Therefore Φ cannot have any repeated roots. □

The proof of Result 3 implies that in case (i) the autoregressive components of the marginals must be *identical*, whereas in case (ii) the autoregressive coefficients must be *all different*.

The following example illustrates that one can have first-order autoregressive components in the marginal models *without* a diagonal Φ, and that the coefficient matrix can even be without blocks of zero coefficients, even if the Jordan form Λ is diagonal.

Example 2. Suppose $\mathbf{z}_t$ follows the model $(k = 4)$: $(\mathbf{I} - \Phi B)\mathbf{z}_t = \mathbf{a}_t$ with

$$\Phi = \begin{bmatrix} .875 & .175 & .025 & -.300 \\ -.350 & .650 & .350 & .000 \\ .225 & .525 & .675 & -.300 \\ -.425 & .175 & .225 & .400 \end{bmatrix}.$$

The eigenvalues of Φ are $\lambda_1 = \lambda_2 = 0.3$ and $\lambda_3 = \lambda_4 = 1$. The geometric multiplicity of

both repeated eigenvalues is two so that the Jordan form is diagonal. In addition, it can be shown that the matrix of eigenvectors (those corresponding to $\lambda = 0.3$ written on the left) is

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & -1 & 2 \\ 2 & 0 & 1 & 1 \\ -2 & 1 & 0 & 3 \\ 1 & 2 & 1 & 0 \end{bmatrix}.$$

Result 1 indicates intrablock reductions of order 2 for z_{it} ($i = 1, \cdots, 4$) and Result 2 gives an additional local reduction of order 1 in each of the four marginal models due to zeros in the leading eigenvectors. Thus, the marginal models are of the form

$$(1 - \alpha_{ii}B)z_{it} = a_{it}$$

where $\alpha_{11} = \alpha_{22} = 1$ and $\alpha_{33} = \alpha_{44} = 0.3$. Note that the intrablock reductions apply to all the marginals, but the local reductions are due to the zeros in $\mathbf{P}$ and apply to the individual series. The third and fourth marginal series are stationary $AR(1)$ processes with identical autoregressive components, while the first and second series are white noise after first differencing. It is interesting to observe that even though Φ has two eigenvalues equal to one, indicating nonstationarity, two of the marginals (z_3 and z_4) are stationary, whereas z_1 and z_2 require first differencing rather than second differencing to achieve stationarity. Also note that in this example $d(B) = (1 - B)(1 - .3B)$, so that an attempt to remove nonstationarity by factoring out $(1 - B)^2$ fails.

5.2. Φ is block triangular. Interblock reductions will also occur when $\Phi(B)$ is block triangular. We can assume, without loss of generality, that $\Phi(B)$ is lower block triangular. Partition $\Phi(B)$ and $\mathbf{z}_t$ as before. Define $\bar{\Phi}_{11}(B) = \Phi_{11}(B)$,

$$\bar{\Phi}_{jj}(B) = \begin{bmatrix} \bar{\Phi}_{j-1,j-1}(B) & \mathbf{0} \\ \Phi_{j1}(B) \cdots \Phi_{j,j-1}(B) & \Phi_{jj}(B) \end{bmatrix} \quad \text{for } j = 2, \cdots, m,$$

and $\bar{\mathbf{z}}_{jt} = (\mathbf{z}'_{1t}, \cdots, \mathbf{z}'_{mt})'$ for $j = 1, \cdots, m$. Then, because of the block lower triangular structure, it follows from (1.1) that the model for $\bar{\mathbf{z}}_{jt}$ is

$$\bar{\Phi}_{jj}(B)\bar{\mathbf{z}}_{jt} = \bar{\Theta}_{jj}(B)\bar{\mathbf{a}}_{jt}. \tag{5.3}$$

Premultiplying (5.3) by $\bar{\mathbf{A}}_{jj}(B)$, the adjoint of $\bar{\Phi}_{jj}(B)$, and using the fact that $|\bar{\Phi}_{jj}(B)| = \prod_{i=1}^{j} |\Phi_{ii}(B)|$, since $\Phi_{ii}(B)$ is block triangular, gives

$$\prod_{i=1}^{j} |\Phi_{ii}(B)| \bar{\mathbf{z}}_{jt} = \bar{\mathbf{A}}_{jj}(B)\bar{\Theta}_{jj}(B)\bar{\mathbf{a}}_{jt}.$$

Therefore, the autoregressive part of any marginal model in the jth block is $\prod_{i=1}^{j} |\Phi_{ii}(B)|$, which is a polynomial in B of order at most $p(\sum_{i=1}^{j} k_i)$, ($j = 1, \cdots, m$). This is generally of lower order than kp, the order of $|\Phi(B)| = \prod_{i=1}^{m} |\Phi_{ii}(B)|$ that would result from (2.1) and that ignores the special structure of $\Phi(B)$. Thus there will be systematic interblock reductions in some of the marginals due to the block lower triangular form. Results 1 and 2 can be applied further to determine additional intrablock reductions due to the structure of $\Phi_{ii}(B)$, $i = 1, \cdots, m$.

In the specific case in which $\mathbf{z}_t$ follows the model $(\mathbf{I} - \Phi B)\mathbf{z}_t = \Theta(B)\mathbf{a}_t$ with Φ lower triangular, the above discussion establishes that the marginal models for z_{it} have autoregressive parts of order i, and more specifically that the autoregressive part for z_{it} is equal to $\prod_{j=1}^{i} (1 - \phi_{jj}B)$. The following result establishes conditions under which the converse is true.

RESULT 4. *Suppose* $\mathbf{z}_t$ *follows the model* $(\mathbf{I} - \Phi B)\mathbf{z}_t = \Theta(B)\mathbf{a}_t$, Φ *is nonsingular, and the only order reductions in the marginal models are due to the (block) triangular structure of* Φ. *If the marginal model for each* z_{it} $(i = 1, \cdots, k)$ *has the form* $\phi_i(B)z_{it} = \theta_i(B)a_{it}$, *where* $\phi_1(B) = (1 - \phi_1 B)$ *and* $\phi_i(B) = (1 - \phi_i B)\phi_{i-1}(B)$, $i = 2, \cdots, k$, *then* Φ *is a lower triangular matrix.*

Proof. Let Λ be the Jordan form of Φ and $\mathbf{P}$ be the matrix of eigenvectors and generalized eigenvectors, so that $\Phi\mathbf{P} = \mathbf{P}\Lambda$. Note that there are no systemwide intrablock reductions because the marginal model for z_{kt} has an autoregressive component of maximal degree, namely k. From the nature of the given marginal models and using Result 2, it follows that the jth row of $\mathbf{P}$ has exactly $k - j$ zeros and that the columns with zeros also have zeros in every row above the jth row. This structure implies that $\mathbf{P}$ or a permutation of the columns of $\mathbf{P}$ is a lower triangular matrix. Let $\mathbf{S}$ be a matrix derived from the identity matrix by permutating some of the columns, so that $\mathbf{S}^2 = \mathbf{I}$. Let us define $\bar{\mathbf{P}} = \mathbf{PS}$ to be a lower triangular matrix, so that $\Phi\bar{\mathbf{P}} = \bar{\mathbf{P}}\bar{\Lambda}$, where $\bar{\Lambda} = \mathbf{S}\Lambda\mathbf{S}$. Since $\bar{\mathbf{P}}$ is lower triangular, $\bar{\mathbf{P}}^{-1}$ is lower triangular as well.

If the Jordan form of Φ is diagonal, then $\bar{\Lambda}$ is also diagonal and $\Phi = \bar{\mathbf{P}}\bar{\Lambda}\bar{\mathbf{P}}^{-1}$ is lower triangular. Thus, we need to consider the case where Λ has at least one 1 above the diagonal. In particular, we need to show that $\bar{\Lambda}$ for this case is lower triangular. To do this, let α be a repeated eigenvalue of Φ. Further suppose that $\phi_{j+1}(B) = \phi_j(B)(1 - \alpha B)$ and $\phi_{i+1}(B) = \phi_i(B)(1 - \alpha B)$ with $i < j$, so that there is a local reduction corresponding to α in the ith and jth marginal models. Let $\bar{\mathbf{p}}_i$ and $\bar{\mathbf{p}}_j$ be the ith and jth column of $\bar{\mathbf{P}}$, respectively. Using Result 2,

$$\Phi\bar{\mathbf{p}}_i = \alpha\bar{\mathbf{p}}_i + \bar{\mathbf{p}}_j, \tag{5.4}$$

so that $\bar{\mathbf{p}}_i$ is a generalized eigenvector associated with $\bar{\mathbf{p}}_j$. From the equality $\Phi\bar{\mathbf{P}} = \bar{\mathbf{P}}\Lambda$ it follows that

$$\Phi\bar{\mathbf{p}}_i = \bar{\mathbf{P}}\bar{\Lambda}_i \tag{5.5}$$

where $\bar{\Lambda}_i$ is the ith column of $\bar{\Lambda}$. But comparing (5.4) and (5.5), $\bar{\Lambda}_i$ must have α on the diagonal of the ith row, zeros above the diagonal, and 1 below the diagonal in the ith column. Therefore $\bar{\Lambda}$ is a lower triangular matrix and $\Phi = \bar{\mathbf{P}}\bar{\Lambda}\bar{\mathbf{P}}^{-1}$ is lower triangular. □

The following corollary addresses reductions in a bivariate model with a first-order autoregressive component.

COROLLARY. *Suppose* $\mathbf{z}_t$ *follows the model* $(\mathbf{I} - \Phi B)\mathbf{z}_t = \Theta(B)\mathbf{a}_t$ *with* Φ *nonsingular and* $k = 2$. *Then there is an order reduction in the marginal models due to the structure of* Φ *if and only if* Φ *is either diagonal, lower triangular, or upper triangular.*

Proof. The discussion of this section implies that there is an interblock order reduction of at least one marginal if Φ is diagonal, lower triangular, or upper triangular. Conversely, if there is an order reduction in both marginal models, then either both marginals have the same autoregressive components and there was exactly one intrablock reduction, or they have different autoregressive parameters and there was one interblock local reduction in each marginal. In either case, Result 3 implies that Φ is diagonal. If there is an order reduction in only one marginal model, say z_{1t}, then the autoregressive part of z_{1t} is $(1 - \alpha_1 B)$ and that of z_{2t} is $(1 - \alpha_1 B)(1 - \alpha_2 B)$ for some α_1 and α_2, and Result 4 implies that Φ is lower triangular, in fact, with $\phi_{11} = \alpha_1$ and $\phi_{22} = \alpha_2$. □

6. Discussion. The task of modeling multivariate time series is a complicated one. In this paper we have analyzed some properties which are inherent to the model structure. In the modeling phase, these properties, including order reductions of the marginals based upon the known model $\Phi(B)\mathbf{z}_t = \Theta(B)\mathbf{a}_t$ may prove of assistance in selecting the

final model form. One way to use the results is to compare the marginals derived from the multiple model with separately (individually) modeled univariate time series. In practice, this has already been done for bivariate models, where it is relatively easy to identify the marginal models and their order reductions.

More importantly, the results derived in our paper indicate that intrablock and local reductions will cause identifiability problems if the multiple time series model is analyzed using (2.1).

Much of the analysis centers around repeated eigenvalues of Φ^*. It can be argued that in checking implied marginals against fitted univariate models as a goodness of fit, the sampling variability in estimated orders is ignored, and the occurrence of multiple eigenvalues may be an exception rather than the rule. While further research should address this issue, there is one important case where repeated roots are very likely in practice, namely the unit root. Many time series, for instance, virtually all business and economic time series, are nonstationary. Currently there is no universally accepted methodology on how to model nonstationary multiple time series. The theory presented in this paper holds implications for how to approach modeling this class of models. This issue should be investigated in future research.

Acknowledgment. The authors thank two anonymous referees for their valuable comments, which substantially strengthened this paper.

REFERENCES

[1] H. AKAIKE, *Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes*, Ann. Inst. Statist. Math., 26 (1974), pp. 363–387.

[2] ———, *Canonical correlation analysis of time series and the use of an information criterion*, in Advances and Case Studies in System Identification, R. Mehra and D. G. Lainiotis, eds., Academic Press, New York, 1976.

[3] W-Y. T. CHAN AND K. F. WALLIS, *Multiple time series modelling: Another look at the mink-muskrat interaction*, Appl. Statist., 27 (1978), pp. 168–175.

[4] W. A. FULLER, *Introduction to Statistical Time Series*, John Wiley, New York, 1976.

[5] F. R. GANTMACHER, *The Theory of Matrices, Vol.* 1, Chelsea, New York, 1977.

[6] J. GEWEKE, *Measurement of linear dependence and feedback between multiple time series*, J. Amer. Statist. Assoc., 77 (1982), pp. 304–313.

[7] C. W. J. GRANGER, *Investigating causal relations by econometric models and cross-spectral methods*, Econometrica, 37 (1969), pp. 424–438.

[8] E. J. HANNAN, *The identification of vector mixed autoregressive-moving average systems*, Biometrika, 56 (1969), pp. 223–225.

[9] ———, *Multiple Time Series*, John Wiley, New York, 1970.

[10] ———, *The identification and parameterization of* ARMAX *and state space forms*, Econometrica, 44 (1976), pp. 713–723.

[11] D. A. PIERCE AND L. D. HAUGH, *Causality in temporal systems: Characterizations and a survey*, J. Econometrics, 5 (1977), pp. 265–293.

[12] M. B. PRIESTLEY, *Spectral Analysis and Time Series*, Academic Press, New York, 1981.

[13] A. STAM, *Order reductions for the marginals of multivariate* ARIMA *time series models*, Ph.D. Thesis, School of Business, University of Kansas, Lawrence, KS, 1986.

[14] A. STAM AND S. C. HILLMER, *Marginals of multivariate first order autoregressive time series models*, J. Time Ser. Anal., 9 (1988), pp. 89–97.

[15] ———, *Marginals of Multiple ARMA Models*, Working Paper #90-287, College of Business Administration, University of Georgia, Athens, GA, 1990.

[16] G. STRANG, *Linear Algebra and its Applications*, Second Edition, Academic Press, New York, 1980.

[17] K. F. WALLIS, *Multiple time series analysis and the final form of econometric models*, Econometrica, 45 (1977), pp. 1481–1497.

[18] A. ZELLNER AND F. PALM, *Time series analysis and simultaneous equation econometric models*, J. Econometrics, 2 (1974), pp. 17–54.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 4, pp. 690–706, October 1991

007

GLOBALLY AND RAPIDLY CONVERGENT ALGORITHMS FOR SYMMETRIC EIGENPROBLEMS*

GILES AUCHMUTY†

Abstract. Some special variational principles for finding particular eigenpairs of the weighted symmetric eigenproblem $Ax = \lambda Mx$ are described and analyzed. The functions involved are even, fourth-degree polynomials in x. These variational principles are then used to develop two different numerical algorithms for finding eigenvalues and eigenvectors. The first algorithm is almost explicit, requiring only the solution of a single cubic equation at each stage, and is globally convergent. The second algorithm is a modification of Newton's method and is cubically convergent to both the eigenvector and the eigenvalue when the desired eigenvalue is simple.

Key words. symmetric eigenproblems, algorithm, variational principles

AMS(MOS) subject classifications. primary 65F15; secondary 15A18, 49G05

1. Introduction. This paper will describe some new algorithms and methods for finding particular eigenvalues and eigenvectors of the weighted symmetric eigenproblem

$$Ax = \lambda Mx.$$

They are based on some unconstrained variational principles for finding the smallest or largest eigenvalues of (A, M). Similar variational principles were described and analyzed by Auchmuty in [1] and [2]. Here, by restricting attention to a special case where the function becomes a polynomial of degree 4 in the variable x, we can derive a number of special identities which lead to particularly efficient algorithms. Analogously to shifted inverse Rayleigh iteration, we shall also describe some parameterized functions which are minimized precisely at eigenvectors of (A, M) corresponding to the eigenvalue which is closest to and either less, or greater, than a parameter μ. These functions and their analytical properties are described in §§ 3 and 4, and general error estimates are proved in § 5.

We shall describe two classes of numerical algorithms for finding eigenvalues and eigenvectors based on these variational principles. The first class are direct descent methods. We choose directions as in the steepest descent or conjugate gradient methods. The special form of these functions then provides us with explicit formulae for the distance to go in the descent direction. This is much more efficient than the usual inexact line search algorithms. The algorithms are described in § 6, while their global convergence properties and descent estimates are proven in § 7.

The other class is based on Newton's method and is described in § 8. When the desired eigenvalue is simple, this method will be cubically convergent—for both the eigenvalue and the corresponding eigenvector. This method, however, is not global and requires the solution of two linear equations involving the same coefficient matrix at each iteration.

The algorithms described here may well be compared with the Rayleigh quotient method described in Parlett [6]. The dynamical system corresponding to our method is a straightforward gradient system derived from the even, quartic function being minimized. For almost all initial conditions, the algorithm described in § 6 converges to an

* Received by the editors November 20, 1989; accepted for publication (in revised form) July 9, 1990. This research was partially supported by the National Science Foundation and the Welch Foundation.

† School of Mathematics, Institute for Advanced Study, Princeton, New Jersey 08540. Present address, Department of Mathematics, University of Houston, Houston, Texas 77204-3476 (auch@uh.edu).

eigenvector associated with the largest eigenvalue λ_n of (A, M). The corresponding eigenvalue is related to both the value of the function and the 2-norm of the minimizing vector. Other eigenvectors will only be saddle points of this function as described in the analysis in [1]. In comparison, the dynamics of Rayleigh quotient iteration may be extremely complicated. This has recently been analyzed in Batterson and Smillie [6] and some examples have been given in Beattie and Fox [3].

In this paper, in §§ 5–8, we only describe the behavior of the algorithm for the particular function $\mathcal{G}$ introduced for finding the largest eigenvalue λ_n of (A, M). All the results described here, nevertheless, apply to the other functionals introduced for finding other particular eigenvalues of (A, M). All the functions have similar forms; they just involve different matrices.

2. Notation. Here we shall collect some definitions and notation which will be used throughout this paper. Only real arithmetic and the Euclidean inner product and the 2-norm on $\mathbb{R}^n$ defined by

$$\langle x, y\rangle = \sum_{j=1}^{n} x_j y_j \quad \text{and} \quad \|x\|^2 = \langle x, x\rangle$$

will be used. Terms from linear algebra will be defined as in Strang [7].

An $n \times n$ real symmetric matrix A is said to be positive definite if there is a constant $c > 0$ such that

$$\langle Ax, x\rangle \geqq c\|x\|^2 \quad \text{for all } x \text{ in } \mathbb{R}^n.$$

When $f : \mathbb{R}^n \to \mathbb{R}$ is a given function, its derivative, or gradient, at a point x is

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x), \cdots, \frac{\partial f}{\partial x_n}(x)\right)^T$$

and its Hessian, or second derivative, is the $n \times n$ matrix

$$D^2 f(x) = \left(\frac{\partial^2 f}{\partial x_j\, \partial x_k}(x)\right).$$

When f is continuously differentiable (C^1) on $\mathbb{R}^n$, then a point z in $\mathbb{R}^n$ is said to be a critical point of f, provided

$$\nabla f(z) = 0.$$

A critical value of f is the value of f at a critical point.

A critical point z of f is said to be nondegenerate if $D^2 f(z)$ exists and is a nonsingular, symmetric matrix.

When z is a nondegenerate critical point, then its Morse index $i(z)$ is the number of negative eigenvalues of $D^2 f(z)$. In particular, if z is a local minimizer of f and is a nondegenerate critical point, then $i(z) = 0$. If $i(z) \geqq 1$, then z will not be a local minimizer.

The function f is said to be coercive on $\mathbb{R}^n$, provided

$$\lim_{\|x\| \to \infty} \frac{f(x)}{\|x\|} = +\infty.$$

Our interest is in studying the nontrivial solutions of

$$Ax = \lambda Mx \tag{2.1}$$

where

(A1) A, M are real symmetric $n \times n$ matrices; and

(A2) M is a positive-definite matrix.

When x is a nonzero solution of (2.1), it is said to be an eigenvector of (A, M) and the corresponding value of λ is called an eigenvalue. When (A1)–(A2) hold, then it is well known that (2.1) has n real eigenvalues $\lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_n$ and there is a corresponding family of real eigenvectors $\{e^j : 1 \leqq j \leqq n\}$ which may be chosen so that

$$\langle Me^j, e^k \rangle = \delta_{jk} \quad \text{for } 1 \leqq j,k \leqq n. \tag{2.2}$$

Here δ_{jk} is the Kronecker delta. Such a set is said to be M-orthonormal. See Parlett [5, § 15.3] for more information on this. For convenience in some later results, we shall put $\lambda_0 = -\infty$ and $\lambda_{n+1} = \infty$.

When λ_j is an eigenvalue of (2.1) we shall define

$$E_j = \{x \in \mathbb{R}^n : (A - \lambda_j M)x = 0\} \quad \text{and} \quad S_j = \{x \in E_j : \langle Mx, x \rangle = 1\}. \tag{2.3}$$

E_j is the eigenspace of (A, M) and S_j is the set of all normalized eigenvectors of (A, M) corresponding to λ_j. An eigenvalue λ_j is said to be simple if the dimension of E_j is one. In this case S_j consists of exactly two points. If dim $E_j = d \geqq 2$, then S_j will be an infinite, connected, compact set in $\mathbb{R}^n$. When (A2) holds, S_j will be diffeomorphic to a $(d-1)$-dimensional sphere.

3. Unconstrained variational principles for the extreme eigenvalues. We shall first describe and analyze certain functions which are minimized precisely at eigenvectors of (A, M) corresponding to the largest eigenvalue λ_n of (2.1), at least if λ_n is positive. In this case both the M-norm of the minimizer and the minimal value of the function are related to λ_n.

Functions of the type to be described here were studied in § 6 of [1] and in Example 7.3 of [2]. In this last example, certain algorithms for finding the corresponding eigenpairs were described and some general convergence results were proven. Here we shall specialize to the case where $p = 4$ in these general principles. This makes the function a polynomial of degree 4 in the n variables $x_1, x_2, \cdots, x_n$.

Define the function $\mathcal{G} : \mathbb{R}^n \to \mathbb{R}$ by

$$\mathcal{G}(x) = \tfrac{1}{4}\langle Mx, x \rangle^2 - \tfrac{1}{2}\langle Ax, x \rangle \tag{3.1}$$

and consider the variational principle ($\mathcal{P}$) of minimizing $\mathcal{G}$ on $\mathbb{R}^n$ and finding

$$\alpha = \inf_{x \in \mathbb{R}^n} \mathcal{G}(x). \tag{3.2}$$

By straightforward calculations, the gradient of $\mathcal{G}$ is

$$\nabla \mathcal{G}(x) = \langle Mx, x \rangle Mx - Ax \tag{3.3}$$

and its Hessian is

$$D^2\mathcal{G}(x) = \langle Mx, x \rangle M - A + 2Mx \otimes Mx \tag{3.4}$$

where $y \otimes z$ is the rank 1 matrix whose entries are $y_i z_j$.

The basic results about this unconstrained optimization problem are described in the following theorem and its corollaries.

THEOREM 1. *Suppose A, M obey* (A1)–(A2) *and $\mathcal{G}$, α are defined by* (3.1) *and* (3.2). *Then*

(i) *α is finite and $\mathcal{G}$ attains its infimum on $\mathbb{R}^n$.*

(ii) *If* $\lambda_n > 0$, *then* $\alpha = -\frac{1}{4}\lambda_n^2$ *when* λ_n *is the largest eigenvalue of* (A, M). $\mathcal{G}$ *is minimized at the points* $\hat{x} = \sqrt{\lambda_n} e$ *where* e *is any point in* S_n *defined by* (2.3).

(iii) *If* $\lambda_n \leqq 0$, *then* $\alpha = 0$ *and* $\hat{x} = 0$ *is the unique minimizer of* $\mathcal{G}$.

Proof. (i) Since (A2) holds, there exists an $m_0 > 0$ such that

$$\langle Mx, x \rangle \geqq m_0 \|x\|^2. \tag{3.5}$$

Thus $\mathcal{G}(x) \geqq (m_0^2/4)\|x\|^4 - (\|A\|/2)\|x\|^2$. Hence $\mathcal{G}$ is coercive on $\mathbb{R}^n$ and for any positive c, the set $\{x \in \mathbb{R}^n : \mathcal{G}(x) \leqq c\}$ is closed, nonempty, and bounded. Since $\mathcal{G}$ is continuous, it attains a finite infimum on this set.

(ii)–(iii) If $\hat{x}$ minimizes $\mathcal{G}$ on $\mathbb{R}^n$, then it must be a critical point of $\mathcal{G}$, so, from (2.3), it must be a solution of

$$Ax = \langle Mx, x \rangle Mx. \tag{3.6}$$

Now $x = 0$ is a solution of this and $\mathcal{G}(0) = 0$.

If $\tilde{x}$ is a nonzero solution of (3.6), then $\tilde{x}$ will be an eigenvector of (A, M) corresponding to the eigenvalue

$$\tilde{\lambda} = \langle M\tilde{x}, \tilde{x} \rangle. \tag{3.7}$$

$\tilde{\lambda}$ must be positive from (A2). Moreover,

$$\langle A\tilde{x}, \tilde{x} \rangle = \langle M\tilde{x}, \tilde{x} \rangle^2 = \tilde{\lambda}^2 \tag{3.8}$$

and

$$\mathcal{G}(\tilde{x}) = -\tfrac{1}{4}\tilde{\lambda}^2 \tag{3.9}$$

upon taking inner products of (3.6) with $\tilde{x}$ and substituting in (3.1).

When $\lambda_n \leqq 0$, there cannot be any nonzero critical points of (3.6), as such critical points must be eigenvectors corresponding to positive eigenvalues of (A, M). Thus zero is the only critical point of $\mathcal{G}$ and it will be the global minimizer of $\mathcal{G}$ with $\alpha = 0$.

When (A, M) has positive eigenvalues, then $\tilde{x} = \sqrt{\lambda_j} e$, with $\lambda_j > 0$ and e being in S_j defined by (2.3), will be a solution of (3.6). Thus $\mathcal{G}(\tilde{x}) = (-1/4)\lambda_j^2$ from (3.9). $\mathcal{G}$ will be minimized on $\mathbb{R}^n$ precisely at those points of the form $\tilde{x} = \sqrt{\lambda_n} e$ with e in S_n, $\mathcal{G}(\hat{x}) = \alpha = (-1/4)\lambda_n^2$, and λ_n being the largest eigenvalue of (A, M). Hence (ii) holds. □

COROLLARY 1. *Suppose* A, M *obey* (A1)–(A2) *and* $\lambda_n > 0$. *Then*

(i) *If* λ_n *is a simple eigenvalue of* (A, M), *then the set of minimizers of* $\mathcal{G}$ *is* $\{\pm\sqrt{\lambda_n} e^n\}$, *where* e^n *is a normalized eigenvector of* (A, M) *corresponding to* λ_n. *Moreover,* $D^2\mathcal{G}(\hat{x})$ *is positive definite with*

$$\langle D^2\mathcal{G}(\hat{x})h, h \rangle \geqq C_1 \langle Mh, h \rangle \tag{3.10}$$

for all h *in* $\mathbb{R}^n$ *and where*

$$C_1 = \min(2\lambda_n, \lambda_n - \lambda_{n-1}). \tag{3.11}$$

(ii) *If* λ_n *is an eigenvalue of* (A, M) *of multiplicity* $d \geqq 2$, *then the set of minimizers of* $\mathcal{G}$ *is an infinite compact connected set in* $\mathbb{R}^n$ *and* $D^2\mathcal{G}(\hat{x})$ *is singular at each minimizing vector.*

Proof. (i) When λ_n is a simple eigenvalue of (A, M), then E_n is one-dimensional and S_n will consist of exactly two points $\pm e_n$. Thus (ii) of the theorem implies the first part. Substituting for $\hat{x}$ in (3.4) we find that

$$D^2\mathcal{G}(\hat{x}) = \lambda_n M - A + 2\lambda_n Me^n \otimes Me^n. \tag{3.12}$$

Consequently, when $\{e^j : 1 \leqq j \leqq n\}$ is an M-orthonormal set of eigenvectors of (A, M)

it is also an M-orthonormal set of eigenvectors of $(D^2\mathcal{G}(\hat{x}), M)$ and we have

$$D^2\mathcal{G}(\hat{x})e^j = \begin{cases} (\lambda_n - \lambda_j)Me^j & \text{if } 1 \leqq j \leqq n-1, \\ 2\lambda_n Me^n & \text{when } j = n. \end{cases}$$

Thus (3.10) and (3.11) follow upon expanding h in terms of the e_j's.

(ii) When λ_n is an eigenvalue of multiplicity $d \geqq 2$, then S_n will be an infinite, bounded, connected, closed set in $\mathbb{R}^n$ and so part (ii) of the theorem implies the same for the minimizers of $\mathcal{G}$.

Let $\hat{x} = \sqrt{\lambda_n} e^n$ be a minimizer of $\mathcal{G}$ and choose an e^{n-1} which is M-orthogonal to e^n and also is an eigenvector of (A, M) corresponding to the eigenvalue λ_n. Then, from (3.12),

$$D^2\mathcal{G}(\hat{x})e^{n-1} = 0$$

and so the Hessian is singular at any minimizer of $\mathcal{G}$ on $\mathbb{R}^n$. □

The preceding results have described the minimizers of $\mathcal{G}$ on $\mathbb{R}^n$. The function $\mathcal{G}$ may have other nonzero critical points which are not global minimizers when (A, M) has more than one positive eigenvalue. The next corollary describes these; in particular, it shows that $\mathcal{G}$ does not have any local minima which are not global minima.

COROLLARY 2. *Suppose A, M obey* (A1)–(A2) *and $\lambda_n > 0$. If $\tilde{x}$ is a nonzero critical point of $\mathcal{G}$, then*

(i) *There is a positive eigenvalue λ_j of (A, M) and a corresponding e in S_j such that $\tilde{x} = \sqrt{\lambda_j} e$.*

(ii) *$\tilde{x}$ is a nondegenerate critical point of $\mathcal{G}$ if and only if λ_j is a simple eigenvalue of (A, M). In this case, its Morse index $i(\tilde{x}) = n - j$.*

(iii) *If $\lambda_j < \lambda_n$, then $\tilde{x}$ is not a local minimum of $\mathcal{G}$.*

Also, zero is a critical point of $\mathcal{G}$ which is not a local minimum. When A is nonsingular, then zero will be a nondegenerate critical point of $\mathcal{G}$ and its Morse index is J where J is the number of positive eigenvalues of (A, M).

Proof. Part (i) was proven in the last part of the proof of Theorem 1.

(ii) When $\tilde{x} = \sqrt{\lambda_j} e$ then, from (3.4), we have

$$D^2\mathcal{G}(\tilde{x}) = \lambda_j M - A + 2\lambda_j Me \otimes Me.$$

Just as in the proof of Corollary 1, we have

$$D^2\mathcal{G}(\tilde{x})e^k = \begin{cases} (\lambda_j - \lambda_k)Me^k & \text{if } k \neq j, \quad 1 \leqq k \leqq n, \\ 2\lambda_j Me^j & \text{when } j = k, \end{cases}$$

where $\{e^j : 1 \leqq j \leqq n\}$ is an M-orthonormal set of eigenvectors of (A, M). Hence $D^2\mathcal{G}(\tilde{x})$ will be nonsingular if and only if λ_j is a simple eigenvalue of (A, M). In this case the Morse index of $\tilde{x}$ will be the number of eigenvalues λ_k of (A, M) with $\lambda_k > \lambda_j$. Thus (ii) follows.

(iii) For any h in $\mathbb{R}^n$ we have

$$\mathcal{G}(\tilde{x} + h) = \mathcal{G}(\tilde{x}) + \langle D^2\mathcal{G}(\tilde{x} + \tau h)h, h \rangle$$

for some $0 < \tau < 1$. When $\tilde{x} = \sqrt{\lambda_j} e$ with e in S_j, $h = te^n$, and t small enough we have

$$\langle D^2\mathcal{G}(\tilde{x} + \tau h)h, h \rangle = t^2 \langle D^2\mathcal{G}(\tilde{x} + \tau h)e^n, e^n \rangle < 0$$

from the continuity of the Hessian and the formula above. Thus $\tilde{x}$ is not a local minimizer of $\mathcal{G}$ when $\lambda_j < \lambda_n$.

We have $D^2\mathcal{G}(0) = -A$ from (3.4) and thus zero will be a nondegenerate critical point if and only if A is nonsingular. The Morse index computation is similar to the one done above. □

If $\lambda_n \leqq 0$, then from Rayleigh's principle we have that $\langle Ax, x\rangle \leqq 0$ for all x in $\mathbb{R}^n$ and the function $\mathcal{G}$ defined by (3.1) is strictly convex. In this case the problem of minimizing $\mathcal{G}$ is not very interesting. To find λ_n in this case, choose $\mu > |\lambda_n|$ and replace A in (2.1) and (3.1) by $A + \mu M$. The eigenvalues now become $\mu + \lambda_j$ and Theorem 1 implies that

$$\inf_{x\in\mathbb{R}^n} [\tfrac{1}{4}\langle Mx, x\rangle^2 - \tfrac{1}{2}\langle (A+\mu M)x, x\rangle] = -\tfrac{1}{4}(\mu+\lambda_n)^2$$

with this infimum being attained at $\hat{x} = \sqrt{\mu+\lambda_n}e$ with e in S_n. Thus we can always find the largest eigenvalue and the corresponding eigenvectors of (A, M) by using a functional of this form.

The values of this function $\mathcal{G}$ provide lower bounds on λ_n when λ_n is positive. When $\mathcal{G}(x) < 0$, we have, from the theorem, that

$$\lambda_n^2 \geqq -4\mathcal{G}(x)$$

and in fact

$$\lambda_n = 2\cdot \sup_{\mathcal{G}(x)<0} \sqrt{-\mathcal{G}(x)}. \tag{3.13}$$

There is a similar variational principle for finding the least eigenvalue λ_1 of (A, M). Define the function $\mathcal{G}_1 : \mathbb{R}^n \to \mathbb{R}$ by

$$\mathcal{G}_1(x) = \tfrac{1}{4}\langle Mx, x\rangle^2 + \tfrac{1}{2}\langle Ax, x\rangle \tag{3.14}$$

and consider the problem ($\mathcal{P}_1$) of finding

$$\alpha_1 = \inf_{x\in\mathbb{R}^n} \mathcal{G}_1(x). \tag{3.15}$$

The function $\mathcal{G}_1$ differs from $\mathcal{G}$ only by a change of sign. It remains a fourth-degree polynomial in $x_1, \cdots, x_n$, which is again an even function. Using exactly the same proofs as in Theorem 1 and its corollaries, we obtain the following results.

THEOREM 2. *Suppose A, M obey* (A1)–(A2) *and $\mathcal{G}_1$, α_1 are defined by* (3.14)–(3.15). *Then*

(i) *α_1 is finite and $\mathcal{G}_1$ attains its infimum on $\mathbb{R}^n$.*

(ii) *If $\lambda_1 < 0$, then $\alpha_1 = -\frac{1}{4}\lambda_1^2$ where λ_1 is the least eigenvalue of (A, M). $\mathcal{G}_1$ is minimized at $\hat{y} = \sqrt{|\lambda_1|}e$ with e being in S_1.*

(iii) *If $\lambda_1 \geqq 0$, then $\alpha_1 = 0$ and 0 is the unique minimizer of $\mathcal{G}_1$ on $\mathbb{R}^n$.*

Essentially the most negative eigenvalue of $-A$ is the largest positive eigenvalue of A. Thus we can also produce analogues of the Corollaries 1 and 2 and this variational principle provides upper bounds on λ_1 as we have

$$\lambda_1^2 \geqq -4\mathcal{G}_1(x)$$

so that

$$\lambda_1 \leqq -2\sqrt{-\mathcal{G}_1(x)}$$

when $\mathcal{G}_1(x) < 0$.

4. Some parameterized, unconstrained, variational principles. Often we are interested in finding the eigenvalues of (A, M) closest to a preassigned number μ. Such

problems arise in computing condition numbers, in studying problems of resonance, or in finding lower or upper bounds on specific eigenvalues. When μ is not an eigenvalue of (A, M), define $\mathcal{F}(\cdot;\mu):\mathbb{R}^n \to \mathbb{R}$ by

$$\mathcal{F}(x;\mu) = \tfrac{1}{4}\langle M^{-1}x, x\rangle^2 - \tfrac{1}{2}\langle (A-\mu M)^{-1}x, x\rangle. \tag{4.1}$$

Consider the variational principle $(\mathcal{P}_\mu)$ of minimizing $\mathcal{F}(\cdot;\mu)$ on $\mathbb{R}^n$ and finding

$$\gamma(\mu) = \inf_{x\in\mathbb{R}^n} \mathcal{F}(x;\mu). \tag{4.2}$$

This function has the same form as $\mathcal{G}$ but with M^{-1} in place of M and $(A-\mu M)^{-1}$ in place of A. The results about this unconstrained optimization problem may be summarized as follows.

THEOREM 3. *Assume A, M obey* (A1)–(A2) *and μ is not an eigenvalue of (A, M). When $\mathcal{F}$, γ are defined by* (4.1)–(4.2), *then*

(i) *$\gamma(\mu)$ is finite and $\mathcal{F}(\cdot;\mu)$ attains this value;*

(ii) *If $\mu > \lambda_n$, then $\gamma(\mu) = 0$ and zero is the unique minimizer of $\mathcal{F}(\cdot;\mu)$; and*

(iii) *If $\lambda_{k-1} < \mu < \lambda_k$ for some $1 \leqq k \leqq n$, then $\gamma(\mu) = -\frac{1}{4}(\lambda_k - \mu)^{-2}$ and $\mathcal{F}(\cdot;\mu)$ is minimized at $\hat{x} = (\lambda_k - \mu)^{-1/2} Me$ with e in S_k.*

Proof. When M is symmetric and positive definite, so is M^{-1} with

$$\langle M^{-1}x, x\rangle \geqq \|M\|^{-1}\|x\|^2$$

for all x in $\mathbb{R}^n$. Hence (i) here follows just as it did in Theorem 1.

Differentiating (4.1), we find

$$\nabla\mathcal{F}(x;\mu) = \langle M^{-1}x, x\rangle M^{-1}x - (A-\mu M)^{-1}x. \tag{4.3}$$

Thus if $\tilde{x}$ is a critical point of $\mathcal{F}(\cdot;\mu)$, then it is a solution of

$$(A-\mu M)^{-1}x = \langle M^{-1}x, x\rangle M^{-1}x. \tag{4.4}$$

When $\tilde{x} \neq 0$, then $\tilde{y} = M^{-1}\tilde{x}$ is a solution of

$$(A-\mu M)y = \langle My, y\rangle^{-1} My \tag{4.5}$$

or $\tilde{y}$ is an eigenvector of (2.1) corresponding to the eigenvalue

$$\lambda_j = \mu + \langle M\tilde{y}, \tilde{y}\rangle^{-1}$$

for some $1 \leqq j \leqq n$. Since M is positive definite, this implies $\lambda_j > \mu$ and, moreover,

$$\langle M\tilde{y}, \tilde{y}\rangle = \langle M^{-1}\tilde{x}, \tilde{x}\rangle = (\lambda_j - \mu)^{-1}. \tag{4.6}$$

Take inner products of (4.4) with $\tilde{x}$; then

$$\langle (A-\mu M)^{-1}\tilde{x}, \tilde{x}\rangle = \langle M^{-1}\tilde{x}, \tilde{x}\rangle^2 = (\lambda_j - \mu)^{-2}$$

so that

$$\mathcal{F}(\tilde{x}, \mu) = -\tfrac{1}{4}(\lambda_j - \mu)^{-2}.$$

Thus $\mathcal{F}(,;u)$ will be minimized when λ_j is the eigenvalue of (A, M) which is larger than μ and closest to μ. When $\mu > \lambda_n$, there is no such eigenvalue so the only solution of (4.4) is $x = 0$ and it must minimize $\mathcal{F}(\cdot;\mu)$ on $\mathbb{R}^n$. Thus (ii) holds.

When $\lambda_{k-1} < \mu < \lambda_k$ for some $1 \leqq k \leqq n$, then we take $j = k$ above, and $\gamma(\mu) = -\frac{1}{4}(\lambda_k - \mu)^{-2}$. The corresponding critical point is

$$\tilde{y} = M^{-1}\tilde{x} = \frac{e}{\sqrt{\lambda_k - \mu}} \quad \text{with } e \text{ in } S_k,$$

upon using (4.5) and (4.6). Thus the function $\mathscr{F}(\cdot;\mu)$ is minimized at $\hat{x} = (\lambda_k - \mu)^{-1/2} Me$ with e in S_k and (iii) holds. □

When $\lambda_{k-1} < \mu < \lambda_k$, this principle provides upper bounds on λ_k. If $\mathscr{F}(x;\mu) < 0$, then from (iii) of the theorem we have

$$(\lambda_k - \mu)^2 \leqq \tfrac{1}{4}[-\mathscr{F}(x;\mu)]^{-1}$$

so

$$\lambda_k \leqq \mu + \tfrac{1}{2}[-\mathscr{F}(x;u)]^{-1/2}. \tag{4.7}$$

Moreover, we see that $\lim_{\mu \to \lambda_k^-} \gamma(\mu) = -\infty$, and

$$\lim_{\mu \to \lambda_k^+} \gamma(\mu) = -\tfrac{1}{4}(\lambda_{k+1} - \lambda_k)^{-2}.$$

In fact, when (A, M) has n distinct eigenvalues, the graph of $\gamma(\mu)$ may be sketched as in Fig. 1.

To obtain lower bounds on the eigenvalues of (A, M) or to find the eigenvalue closest to but less than μ, consider the function $\mathscr{F}_1(\cdot;\mu) : \mathbb{R}^n \to \mathbb{R}$ defined by

$$\mathscr{F}_1(x;\mu) = \tfrac{1}{4}\langle M^{-1}x, x\rangle^2 + \tfrac{1}{2}\langle (A - \mu M)^{-1}x, x\rangle \tag{4.8}$$

and look at the problem of finding

$$\gamma_1(\mu) = \inf_{x \in \mathbb{R}^n} \mathscr{F}_1(x;\mu). \tag{4.9}$$

$\mathscr{F}_1$ differs from $\mathscr{F}$ solely by a sign change and this problem has similar properties to $(\mathscr{P}_\mu)$.

THEOREM 4. *Assume A, M obey* (A1)–(A2) *and μ is not an eigenvalue of (A, M). If $\mathscr{F}_1$, γ_1 are defined by* (4.8)–(4.9), *then*

(i) *$\gamma_1(\mu)$ is finite and $\mathscr{F}_1(\cdot;\mu)$ attains this value;*

(ii) *If $\mu < \lambda_1$, then $\gamma_1(\mu) = 0$ and zero is the unique minimizer of $\mathscr{F}_1(\cdot;\mu)$ on $\mathbb{R}^n$; and*

(iii) *When $\lambda_k < \mu < \lambda_{k+1}$ for some $1 \leqq k \leqq n$, then $\gamma_1(\mu) = -\frac{1}{4}(\lambda_k - \mu)^{-2}$ and $\mathscr{F}_1(\cdot;\mu)$ is minimized at $\hat{x} = (\mu - \lambda_k)^{-1/2} Me$ for any e in S_k.*

Proof. This follows just as the proof of Theorem 3 with appropriate sign changes. □

This time, instead of (4.7), we obtain

$$4(\mu - \lambda_k)^2 \leqq [-\mathscr{F}(x,\mu)]^{-1}$$

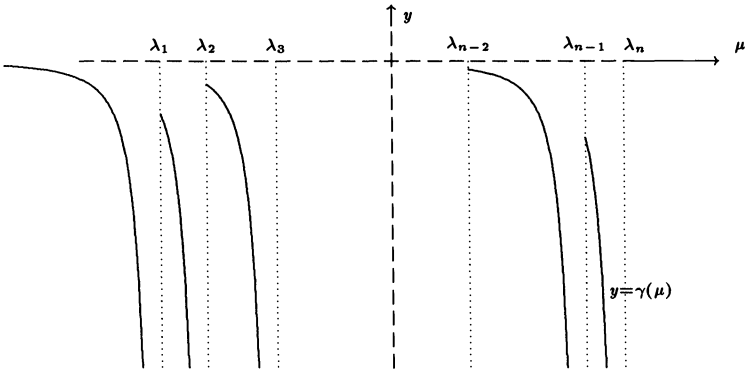


FIG. 1. *Sketch of $y = \gamma(\mu)$, when (A, M) has n distinct eigenvalues.*

whenever $\mathscr{F}(x, \mu) < 0$. Thus

$$\lambda_k \geqq \mu - \tfrac{1}{2}(-\mathscr{F}(x;\mu))^{1/2}$$

for any such x, and this provides lower bounds on λ_k. Also

$$\lim_{\mu \to \lambda_k^-} \gamma_1(\mu) = -\tfrac{1}{4}(\lambda_k - \lambda_{x+1})^{-2}$$

while

$$\lim_{\mu \to \lambda_k^+} \gamma_1(\mu) = -\infty$$

and the graph of $\gamma_1(\mu)$ may be depicted as in Fig. 2.

For both of the functions $\mathscr{F}(\cdot;\mu)$ and $\mathscr{F}_1(\cdot;\mu)$, we have that the global minima are nondegenerate critical points if and only if the corresponding eigenvalues are simple eigenvalues of (A, M), just as was the case for $\mathscr{G}$. Also, if $\tilde{x}$ is a critical point of $\mathscr{F}(\cdot;\mu)$ (or $\mathscr{F}_1(\cdot;\mu)$) which is not a global minimizer of $\mathscr{F}$ (or $\mathscr{F}_1$), then it will not be a local minimizer either, just as was proven in Corollary 2 of Theorem 1 for $\mathscr{G}$.

5. The Hessian and error estimates. From now on we shall restrict our analysis to the function $\mathscr{G}$. Similar results will hold for $\mathscr{G}_1$, $\mathscr{F}(\cdot;\mu)$, and $\mathscr{F}_1(\cdot;\mu)$ when the appropriate substitutions are made. Also assume $\lambda_n > 0$, so the problem remains interesting.

Many of the convergence results in the following sections depend on having information on the Hessian quadratic form

$$\langle D^2\mathscr{G}(x)h, h\rangle = \langle Mx, x\rangle\langle Mh, h\rangle - \langle Ah, h\rangle + 2\langle Mx, h\rangle^2. \tag{5.1}$$

We say that $\mathscr{G}$ is strongly convex on a convex subset K of $\mathbb{R}^n$ if there exists $c_0 > 0$ such that

$$\langle D^2\mathscr{G}(x)h, h\rangle \geqq c_0\|h\|^2 \tag{5.2}$$

for all x in K and h in $\mathbb{R}^n$.

LEMMA 5.1. *Suppose A, M obey* (A1)–(A2) *and h is an M-normalized vector; then*

$$m_0\|x\|^2 - \lambda_n \leqq \langle D^2\mathscr{G}(x)h, h\rangle \leqq 3\langle Mx, x\rangle - \lambda_1 \tag{5.3}$$

where m_0 is defined by (3.5).

Proof. From the generalized Schwarz inequality, when M obeys (A2) and $\langle Mh, h\rangle = 1$, we have

$$0 \leqq \langle Mx, h\rangle^2 \leqq \langle Mx, x\rangle.$$

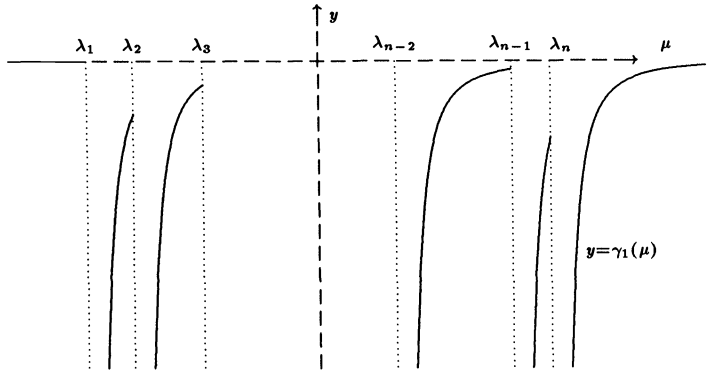


FIG. 2. *Sketch of $y = \gamma_1(\mu)$, when (A, M) has n distinct eigenvalues.*

Also, from Rayleigh's principle for this problem,

$$\lambda_1\langle Mh,h\rangle \leqq \langle Ah,h\rangle \leqq \lambda_n\langle Mh,h\rangle.$$

Substituting these in (5.1), we have

$$\langle D^2\mathcal{G}(x)h,h\rangle \geqq \langle Mx,x\rangle - \lambda_n \geqq m_0\|x\|^2 - \lambda_n,$$

and also $\langle D^2\mathcal{G}(x)h, h\rangle \leqq 3\langle Mx, x\rangle - \lambda_1$, as claimed in (5.3). □

In particular, this shows that $\mathcal{G}$ is strongly convex on sets where $\|x\|$ is large. More importantly, $\mathcal{G}$ will also be strongly convex on a neighborhood of a minimizer $\hat{x}$ of $\mathcal{G}$ provided the corresponding eigenvalue is simple. Specifically, we have Theorem 5.

THEOREM 5. *Suppose A, M obey* (A1)–(A2), $\lambda_n > 0$ *is a simple eigenvalue of* (2.1), *and $\hat{x}$ is a minimizer of $\mathcal{G}$. Then there exists an $\hat{R} > 0$ such that $\mathcal{G}$ is strongly convex on $K_R = \{x \in \mathbb{R}^n : \langle M(x-\hat{x}), x-\hat{x}\rangle \leqq R^2\}$ whenever $R < \hat{R}$.*

Proof. Let $x = \hat{x} + k$; then from (3.4) we have

$$\begin{aligned}\langle D^2\mathcal{G}(x)h,h\rangle &= \langle D^2\mathcal{G}(\hat{x})h,h\rangle + 4\langle M\hat{x},h\rangle\langle Mk,h\rangle + 2\langle M\hat{x},k\rangle\langle Mh,h\rangle \\ &\quad + 2\langle Mk,h\rangle^2 + \langle Mk,k\rangle\langle Mh,h\rangle \\ &\geqq [C_1 - 6\langle M\hat{x},\hat{x}\rangle^{1/2}\langle Mk,k\rangle^{1/2} + \langle Mk,k\rangle]\langle Mh,h\rangle\end{aligned}$$

upon using the generalized Schwarz inequality and (3.10).

Thus $\mathcal{G}$ will be strongly convex on K_R provided

$$c(R) = C_1 - 6\sqrt{\lambda_n}R + R^2 > 0. \tag{5.4}$$

Let $\hat{R} = \sqrt{9\lambda_n} - \sqrt{9\lambda_n - C_1}$; then for $R < \hat{R}$ we have $c(R) > 0$ and $\mathcal{G}$ will obey (5.2) on K_R with $m_0c(R)$ in place of c_0. □

From (3.11), we actually have

$$\hat{R} = \min\left[(\sqrt{9\lambda_n} - \sqrt{8\lambda_n + \lambda_{n-1}}), (3-\sqrt{7})\sqrt{\lambda_n}\right]. \tag{5.5}$$

Using a standard argument from convex optimization theory, this leads to concrete error estimates for $\hat{x}$ and λ_n^2 as follows.

COROLLARY. *Under the assumptions of Theorem* 5, *whenever x is in K_R with $R < \hat{R}$ and $\varepsilon(x) = \|\nabla\mathcal{G}(x)\|/m_0c(R)$, we have*

(i)

$$-\mathcal{G}(x) \leqq \frac{\lambda_n^2}{4} \leqq \frac{1}{2}\varepsilon(x)\|\nabla\mathcal{G}(x)\| - \mathcal{G}(x), \tag{5.6}$$

(ii)

$$\|x - \hat{x}\| \leqq \varepsilon(x), \tag{5.7}$$

where m_0 is defined by (3.5) *and $c(R)$ by* (5.4).

Proof. Apply Taylor's theorem to $\mathcal{G}$ about x. We have that there exists a ξ in (0, 1) such that

$$\begin{aligned}\mathcal{G}(\hat{x}) &= \mathcal{G}(x) + \langle\nabla\mathcal{G}(x), \hat{x}-x\rangle + \tfrac{1}{2}\langle D^2\mathcal{G}(x+\xi(\hat{x}-x))(\hat{x}-x), \hat{x}-x\rangle \\ &\geqq \mathcal{G}(x) - \|\nabla\mathcal{G}(x)\|\|\hat{x}-x\| + \tfrac{1}{2}m_0c(R)\|\hat{x}-x\|^2,\end{aligned}$$

using Cauchy's inequality and Theorem 5.

Now $\mathcal{G}(\hat{x}) = -\frac{1}{4}\lambda_n^2$, so that

$$\mathcal{G}(x) \geqq -\frac{1}{4}\lambda_n^2 \geqq \mathcal{G}(x) + \frac{m_0c(R)}{2}[\|\hat{x}-x\| - \varepsilon(x)]^2 - \frac{\|\nabla\mathcal{G}(x)\|^2}{2m_0c(R)}.$$

Rearranging this, (5.6) follows.

Also, from Taylor's theorem and the strong convexity of $\mathcal{G}$ on K_R, we have

$$\mathcal{G}(x) \geqq \mathcal{G}(\hat{x}) + \tfrac{1}{2} m_0 c(R) \| x - \hat{x} \|^2$$

for all x in K_R. Thus from (5.6)

$$\tfrac{1}{2} m_0 c(R) \| x - \hat{x} \|^2 \leqq \tfrac{1}{2} \varepsilon(x) \| \nabla \mathcal{G}(x) \|,$$

so (5.7) follows. □

6. Descent algorithms for finding eigenpairs. The functions described in §§ 3 and 4 have two particularly nice properties for minimization algorithms. One is the fact that they are quartic polynomials, so that an exact line search only requires that we solve a cubic equation for the steplength. The other is that they do not have any local minima which are not global minima. Thus a local search is sufficient to determine whether we have found a minimizer.

Here we shall describe some descent algorithms for these functions which exploit this special structure of $\mathcal{G}$ in an effective and efficient manner. Recall that a direction h is said to be a descent direction for $\mathcal{G}$ at x, provided

$$\langle \nabla \mathcal{G}(x), h \rangle < 0. \tag{6.1}$$

Whenever x is not a critical point of $\mathcal{G}$ there are such directions. We shall henceforth assume that

$$\alpha = \inf_{x \in \mathbb{R}^n} \mathcal{G}(x) < 0. \tag{6.2}$$

This is equivalent to saying that (A, M) has at least one positive eigenvalue. When (6.2) does not hold and $\mathcal{G}$ is strictly convex, it is minimized at $x = 0$, and there is no need for an algorithm to find it.

When h is normalized so that

$$\langle Mh, h \rangle = 1, \tag{6.3}$$

we have that

$$\mathcal{G}(x + th) = \mathcal{G}(x) + a_1 t + \tfrac{1}{2} a_2 t^2 + a_3 t^3 + \tfrac{1}{4} t^4 \tag{6.4}$$

where

$$a_1 = \langle \nabla \mathcal{G}(x), h \rangle, \qquad a_2 = \langle D^2 \mathcal{G}(x) h, h \rangle, \qquad a_3 = \langle Mx, h \rangle \tag{6.5}$$

with $\nabla \mathcal{G}(x)$ and $D^2 \mathcal{G}(x)$ being defined by (3.3) and (3.4).

A simple descent algorithm for minimizing $\mathcal{G}$ follows.

(1) Given a nonzero x^0 in $\mathbb{R}^n$ use steepest descent or a search method to find x^1 in $\mathbb{R}^n$ with

$$\mathcal{G}(x^1) < 0. \tag{6.6}$$

(2) For $k \geqq 1$, let $y^k = \tau_k x^k$ with

$$\tau_k = \langle Mx^k, x^k \rangle^{-1} \langle Ax^k, x^k \rangle^{1/2}. \tag{6.7}$$

(3) Choose a descent vector h^k for $\mathcal{G}$ at y^k to obey (6.3) and

$$\langle Mx^k, h^k \rangle = 0. \tag{6.8}$$

(4) Find $t_k > 0$ such that

$$\mathcal{G}(y_k + t_k h^k) = \inf_{t \geqq 0} \mathcal{G}(y^k + th^k). \tag{6.9}$$

(5) Put $x^{k+1} = y^k + t_k h^k$ and evaluate

$$r^{k+1} = \nabla \mathcal{G}(x^{k+1}). \tag{6.10}$$

(6) If $\|r^{k+1}\| > \varepsilon$ continue, else stop.

In order for (6.7) to be well defined, we must have that $\langle Ax^k, x^k \rangle > 0$, for all $k \geqq 1$. This is guaranteed if $\mathcal{G}(x^k) < 0$ for all k. Moreover, from (3.9) we have that if $\tilde{x}$ is a critical value of $\mathcal{G}$, then $\mathcal{G}(\tilde{x}) \leqq 0$, so $\mathcal{G}(x) > 0$ implies $\nabla \mathcal{G}(x) \neq 0$. Thus a steepest descent algorithm will lead to the set

$$\mathcal{E}_0 = \{ x \in \mathbb{R}^n : \mathcal{G}(x) < 0 \}.$$

Steps 2–4 divide the minimization into a "radial" minimization in the direction x^k and then into a minimization in a descent direction which is M-orthogonal to x^k. This is done as there are simple explicit formulae for the steplengths τ_k and t_k in each of these directions.

From (3.1) we have

$$\mathcal{G}(\tau x) = \tfrac{1}{4}\tau^4 \langle Mx, x \rangle^2 - \tfrac{1}{2}\tau^2 \langle Ax, x \rangle$$

and this is minimized in τ when $\tau = 0$ if $\langle Ax, x \rangle \leqq 0$, or else when

$$\tau^2 = \langle Mx, x \rangle^{-2} \langle Ax, x \rangle.$$

This leads to the choice of (6.7).

In step (3) we can choose h^k via a steepest descent, or a sequence of conjugate gradient directions, subject to the condition (6.8). The condition (6.8) is imposed to simplify the computation of t_k in step (4). This is done using the following result.

LEMMA 6.1. *Assume A, M obey* (A1)–(A2), *x is not a critical point of $\mathcal{G}$, and h is a normalized descent direction for $\mathcal{G}$ at x obeying $\langle Mx, h \rangle = 0$. Then there is a unique positive $\hat{t}$ which minimizes $\mathcal{G}(x + th)$ on $[0, \infty)$ and it is the unique positive solution of*

$$p(t) = t^3 + a_2 t + a_1 = 0 \tag{6.11}$$

with a_1, a_2 defined by (6.5).

Proof. The condition $\langle Mx, h \rangle = 0$ implies that $a_3 = 0$ in (6.4), so $\mathcal{G}(x + th)$ will be minimized either at $t = 0$ or at the solution of (6.11). Since h is a descent direction, $a_1 < 0$, and $t = 0$ is not a minimizer.

We have $p'(t) = 3t^2 + a_2$, so p' is monotone increasing on $(0, \infty)$. When $a_2 \geqq 0$, then p will be monotone increasing on $(0, \infty)$ and thus there is a unique $\hat{t} > 0$ obeying (6.11).

When $a_2 < 0$, let $T_1 = (-a_2/3)^{1/2}$. Then p decreases on $[0, T_1)$ and p cannot have a zero there. p is strictly monotone increasing on $[T_1, \infty)$, so it has a unique positive zero there. The zero actually minimizes $\mathcal{G}$ on $[0, \infty)$, as $p'(\hat{t})$ is positive. □

From this lemma, we have that t_k in step (4) should be chosen to be the unique positive solution of

$$t^3 + a_2^k t + a_1^k = 0, \tag{6.12}$$

with

$$a_1^k = \langle \nabla \mathcal{G}(y^k), h^k \rangle = -\langle Ay^k, h^k \rangle, \tag{6.13}$$

$$a_2^k = \langle D^2 \mathcal{G}(y^k) h^k, h^k \rangle = \langle My^k, y^k \rangle - \langle Ah^k, h^k \rangle \tag{6.14}$$

from (6.5), (6.8), and (5.1). Note that since $y^k = \tau_k x^k$, these can also be expressed in terms of x^k and τ^k.

To find the solution t_k of (6.12), we could use either an explicit formula or, preferably, a safeguarded Newton method.

The stopping criteria is that the point x^{k+1} be an ε-critical point of $\mathcal{G}$. We know that if $\tilde{x}$ is a nonzero critical point of $\mathcal{G}$, then it will be an eigenvector of (A, M). Hence we expect that, when ε is small enough, x^{k+1} will be an approximate eigenvector of (A, M). Moreover, by evaluating the Hessian of $D^2\mathcal{G}(x^{k+1})$, or by performing a local search, we can see whether x^{k+1} is a local minimizer of $\mathcal{G}$. If so, it will be the desired minimizer of $\mathcal{G}$; otherwise we continue.

7. Global convergence of the descent algorithm. To prove convergence of this algorithm, we need some descent estimates and certain bounds on the iterates. These are collected in the following lemmas.

Let $\{x^k : k \geqq 1\}$ be a sequence of points in $\mathbb{R}^n$ defined by the algorithm in § 6. Define $\alpha_k = \mathcal{G}(x^k)$ and $r^k = \nabla\mathcal{G}(x^k)$. Let

$$\mathcal{E}_k = \{x \in \mathbb{R}^n : \mathcal{G}(x) \leqq \alpha_k\}; \tag{7.1}$$

then since $\alpha_{k+1} < \alpha_k$ we have $\mathcal{E}_{k+1} \subset \mathcal{E}_k$ for all k.

LEMMA 7.1. *Suppose* (A1)–(A2) *hold and* $\lambda_n > 0$. *If* $k \geqq 1$, *and* x *is in* $\mathcal{E}_k$, *then*

$$\frac{2|\alpha_k|}{\lambda_n} \leqq \langle Mx, x\rangle \leqq 2\left(\lambda_n + \frac{\alpha_k}{\lambda_n}\right). \tag{7.2}$$

Proof. To prove this, consider the constrained optimization problem of minimizing and maximizing $m(x) = \frac{1}{2}\langle Mx, x\rangle$ on $\mathcal{E}_k$.

Since $\alpha_k < 0$, $\mathcal{E}_k$ is a closed, bounded set in R^n which does not contain the origin. m attains its infimum and supremum on this set and by using the extremality conditions, these occur at $x_1 = \nu_1 e^n$ and $x_2 = \nu_2 e^n$ where $\nu_1 < \nu_2$ are the solutions of

$$\tfrac{1}{4}\nu^4 - \tfrac{1}{2}\nu^2\lambda_n - \alpha_k = 0$$

and e^n is a normalized eigenvector of (A, M) corresponding to λ_n.

Thus $\langle Mx, x\rangle \geqq \langle Mx_1, x_1\rangle = \nu_1^2$ and similarly $\langle Mx, x\rangle \leqq \nu_2^2$, where

$$\nu_1^2 = \lambda_n\left[1 - \sqrt{1 + \frac{4\alpha_k}{\lambda_n}}\right] \geqq \frac{2|\alpha_k|}{\lambda_n},$$

$$\nu_2^2 = \lambda_n\left[1 + \sqrt{1 + \frac{4\alpha_k}{\lambda_n}}\right] \leqq 2\left(\lambda_n + \frac{\alpha_k}{\lambda_n}\right). \qquad \square$$

LEMMA 7.2. *Under the assumptions of the preceding lemma, we have*

(i)

$$\mathcal{G}(x^k) - \mathcal{G}(\tau_k x^k) = \frac{1}{4}\langle Mx^k, x^k\rangle^{-2}\langle x^k, r^k\rangle^2, \tag{7.3}$$

(ii)

$$\frac{|\alpha_k|\lambda_n^2}{2(\lambda_n^2 + \alpha_k)^2} \leqq \tau_k^2 - \frac{1}{2} \leqq \frac{\lambda_n^2}{2|\alpha_k|} \quad \textit{for all } k. \tag{7.4}$$

Proof. (i) From the definition of G and τ_k^2, we have

$$\mathcal{G}(x^k) - \mathcal{G}(\tau_k x^k) = \tfrac{1}{4}\langle Mx^k, x^k\rangle^2(1 - \tau_k^2)^2.$$

Also, $\langle r^k, x^k\rangle = \langle Mx^k, x^k\rangle^2(1-\tau_k^2)$, so substituting this into the preceding equation, (7.3) follows.

(ii) Let $\tau(x)^2 = \langle Mx, x\rangle^{-2}\langle Ax, x\rangle$ for x in $\mathscr{E}_k$.

Since $\mathscr{G}(x) \leqq \alpha_k$ we have $(\langle Mx, x\rangle^2/4)(1-2\tau(x)^2) = \nu$ for some $\nu \leqq \alpha_k$. This implies that $\tau(x)^2 = \frac{1}{2} - 2\nu/\langle Mx, x\rangle^2$. In particular, $\tau_k^2 = \frac{1}{2} - 2\alpha_k/\langle Mx^k, x^k\rangle^2$ with $\alpha_k < 0$.

Substitute from (7.2); then

$$\frac{|\alpha_k|\lambda_n^2}{2(\lambda_n^2+\alpha_k)^2} \leqq \frac{-2\alpha_k}{\langle Mx^k, x^k\rangle^2} \leqq \frac{\lambda_n^2}{2|\alpha_k|},$$

and thus (7.4) follows. □

LEMMA 7.3. *Suppose $\hat{t}$ minimizes $\varphi(t) = \frac{1}{4}t^4 + \frac{1}{2}at^2 - bt$ on $[0, \infty)$ with $b > 0$ and $a \leqq A$ for some $A > 0$. Then $\varphi(\hat{t}) \leqq \Psi(b)$, where*

$$4\Psi(b) = \begin{cases} -A^{-1}b^2 & \text{for } 0 \leqq b^2 \leqq A^3, \\ -b^{4/3} & \text{for } b^2 \geqq A^3. \end{cases} \tag{7.5}$$

Proof. Consider

$$\varphi(sb^{1/3}) = \frac{1}{2}s^2b^{4/3}\left[ab^{-2/3} + \frac{s^2}{2} - \frac{2}{s}\right]$$
$$\leqq \frac{1}{2}s^2b^{4/3}\left[Ab^{-2/3} + \frac{s^2}{2} - \frac{2}{s}\right]$$

when $a \leqq A$. When $b^2 \geqq A^3$ we have $Ab^{-2/3} \leqq 1$ and then

$$\varphi(\hat{t}) \leqq \varphi(b^{1/3}) \leqq -\tfrac{1}{4}b^{4/3}.$$

When $b^2 \leqq A^3$, take $\hat{s} = A^{-1}b^{2/3}$; then

$$\varphi(\hat{t}) \leqq \varphi(\hat{s}b^{1/3}) \leqq -\frac{A^{-1}b^2}{4},$$

so (7.5) holds as claimed. □

COROLLARY. *Suppose* (A1)–(A2) *hold with $\lambda_n > 0$ and $\{x^k : k \geqq 1\}$ is defined by the preceding algorithm. Then for any $k \geqq 1$ we have*

$$\mathscr{G}(x^{k+1}) - \mathscr{G}(y^k) \leqq \Psi(-\tau_k\langle r^k, h^k\rangle) \tag{7.6}$$

with Ψ defined by (7.5) *and $A = 6\lambda_n - \lambda_1$.*

Proof. We have that $\mathscr{G}(x^{(k+1)}) - \mathscr{G}(y^k) = \varphi(\hat{t})$ where $\hat{t}$ is the solution of (6.12). Thus the b in Lemma 7.3 is $-\langle\nabla\mathscr{G}(y^k), h^k\rangle = -\tau_k\langle r^k, h^k\rangle > 0$ as h^k is a descent direction. a in this lemma is a_2^k, so from (6.14), (5.3), and (7.2), we have $a_2^k \leqq 6\lambda_n - \lambda_1$ for all $k \geqq 1$. Taking $A = 6\lambda_n - \lambda_1$, then the result holds.

To obtain a general convergence theorem for the algorithm described in the last section we need to be slightly more careful in the choice of the direction h^k. At each stage we can write

$$r^k = \nu_k Mx^k + q^k \tag{7.7}$$

with $\langle Mx^k, q^k\rangle = 0$. Then we have $\langle r^k, h^k\rangle = \langle q^k, h^k\rangle$ whenever h^k obeys (6.10).

THEOREM 6. *Assume* (A1)–(A2) *hold and $\lambda_n > 0$. Let $\Gamma = \{x^k : k \geqq 0\}$ be the descent sequence for $\mathscr{G}$ defined in* § 6 *with $\varepsilon = 0$. Assume that for all $k \geqq 1$, the descent*

direction h^k is chosen so that

$$-\langle r^k, h^k\rangle \geqq \delta \|q^k\| \, \|h^k\| \tag{7.8}$$

where $\delta > 0$ and q^k is defined as in (7.7). *When Γ is finite, the last point x^K is a critical point of $\mathscr{G}$. When Γ is infinite, every limit point $\hat{x}$ of Γ is a critical point of $\mathscr{G}$.*

Proof. When $\varepsilon = 0$, the stopping criterion in step (6) is that x^K be a critical point of $\mathscr{G}$, so the first part holds.

When Γ is infinite, then (7.3) and (7.6) imply that

$$\begin{aligned} \mathscr{G}(x^{k+1}) - \mathscr{G}(x^k) &\leqq \frac{-\langle x^k, r^k\rangle^2}{4\langle Mx^k, x^k\rangle} + \Psi(-\tau_k \langle r^k, h^k\rangle) \\ &\leqq \frac{-\langle x^k, r^k\rangle^2}{8\lambda_n} + \Psi(\delta_1 \|q^k\|) \end{aligned} \tag{7.9}$$

as $\langle Mx^k, x^k\rangle \leqq 2\lambda_n$ for all $k \geqq 1$, from (7.2); $\tau_k^2 \geqq \frac{1}{2}$ from (7.4), so

$$\begin{aligned} -\tau_k \langle r^k, h^k\rangle &\geqq \frac{\delta}{\sqrt{2}} \|q^k\| \, \|h^k\| \quad \text{from (7.8)} \\ &\geqq \frac{\delta}{m_0\sqrt{2}} \|q^k\| \end{aligned}$$

as h is normalized and (3.5) holds. Take $\delta_1 = \delta/(m_0\sqrt{2})$ and use the monotonicity of Ψ to obtain (7.9).

Since $\Gamma \subset \mathscr{E}_1$ it is bounded. Let $\hat{x}$ be a limit point of Γ and without loss of generality assume the whole sequence converges to $\hat{x}$. Then $r^k = \nabla\mathscr{G}(x^k)$ converges to $\hat{r} = \nabla\mathscr{G}(\hat{x})$ as $\nabla\mathscr{G}$ is continuous. Also, $\mathscr{G}(x^k)$ converges to $\mathscr{G}(\hat{x})$, so from (7.9)

$$\langle x^k, r^k\rangle \to \langle \hat{x}, \hat{r}\rangle = 0, \tag{7.10}$$

$$\|q^k\| \to 0. \tag{7.11}$$

From (7.11) and (7.7) we must have $\hat{r} = \nu M\hat{x}$ for some ν. Substituting this in (7.10), we have $\nu\langle \hat{x}, M\hat{x}\rangle = 0$, which implies that $\nu = 0$ as $\langle M\hat{x}, \hat{x}\rangle \geqq 2|\alpha_k|/\lambda_n$ for all k from (7.2).

Thus $\hat{r} = 0$, and $\hat{x}$ is a critical point of $\nabla\mathscr{G}$. When Γ has many limit points, this argument carries over for each of them, so the theorem holds. □

It is worth noting that, in this section, we did not have to assume that the largest eigenvalue λ_n of (A, M) was simple. Thus the set of minimizers of G may possibly be a large set and which critical points of $\mathscr{G}$ are found by this algorithm will depend on the choice of descent directions.

8. Cubically convergent algorithms. Since the methods espoused in this paper for finding eigenvalues and eigenvectors of (A, M) are based on the unconstrained minimization of a polynomial, we might well ask about the applicability of Newton's method.

Needless to say, Newton's method will not be a global method. From Corollary 1 to Theorem 1, we also see that if the largest eigenvalue of (A, M) has multiplicity $d \geqq 2$, then the Hessian of $\mathscr{G}$ will be singular at every minimizer of $\mathscr{G}$. In this case convergence will be linear at best.

When λ_n is a simple positive eigenvalue of (A, M), then all the criteria for Newton's method to converge to $\hat{x}$ on a neighborhood of $\hat{x}$ hold. In this case we will obtain quadratic convergence to both the eigenvector and the eigenvalue.

With minimal extra work we can obtain a cubically convergent method. When λ_n is a simple positive eigenvalue, let $\hat{R}$ be defined by (5.5) and K_R with $R < \hat{R}$ be defined as in Theorem 5. This algorithm may be described as follows.

Given x^0 in K_R, $\varepsilon > 0$, for $k \geqq 0$:

(1) Let h^k be the solution of

$$D^2\mathcal{G}(x^k)h = -\nabla\mathcal{G}(x^k). \tag{8.1}$$

(2) Let x^{k+1} be the solution of

$$D^2\mathcal{G}(x^k)(x^{k+1}-x^k) = -\nabla\mathcal{G}(x^k) - \nabla\mathcal{G}(x^k+h^k). \tag{8.2}$$

(3) Put

$$\nu_{k+1} = 2\sqrt{-\mathcal{G}(x^{k+1})}. \tag{8.3}$$

(4) If $\|\nabla\mathcal{G}(x^{k+1})\| > \varepsilon$ continue, else stop.

Note that h^k defined by (8.1) is the usual Newton step and we usually solve it by factorizing the symmetric matrix

$$D^2\mathcal{G}(x^k) = \langle Mx^k, x^k\rangle M - A + 2Mx^k \otimes Mx^k. \tag{8.4}$$

For each k this is a rank 1 perturbation of a matrix of the form

$$D^2\mathcal{G}(x^k) = m_k M - A$$

with $m_k = \langle Mx^k, x^k\rangle$, so it is relatively easy to update the factorization. We then solve (8.2) using the same factorization. That is, (8.1) and (8.2) are two equations involving the same coefficient matrix and for successive k's they do not change very much.

THEOREM 7. *Suppose* (A1)–(A2) *hold and* λ_n *is a simple positive eigenvalue of* (A, M). *If* $\|x^0 - \hat{x}\|$ *is small enough, then the sequence* $\Gamma = \{x^k : k \geqq 0\}$ *defined by this algorithm has a cubic rate of convergence to* $\hat{x} = \sqrt{\lambda_n}e^n$ *and* ν_{k+1} *converges cubically to* λ_n.

Proof. The equation $\nabla\mathcal{G}(x) = 0$ has $\hat{x} = \sqrt{\lambda_n}e^n$ as its unique solution on K_R. Moreover, $\mathcal{G}$ is convex on K_R and $D^2\mathcal{G}$ exists and is Lipschitz continuous there with $D^2\mathcal{G}(\hat{x})$ being nonsingular since λ_n is a simple eigenvalue of (A, M). Then from Theorem 10.2.4 of Ortega and Rheinholdt [4], there is a neighborhood of $\hat{x}$ on which this algorithm has an R-convergence order of at least three.

We have $\mathcal{G}(\hat{x}) = -\frac{1}{4}\lambda_n^2$ and $\mathcal{G}(x^k) = -\frac{1}{4}\nu_k^2$. Upon using the Taylor expansion of $\mathcal{G}$ about $\hat{x}$, we find that ν_k also converges to λ_n cubically as there exist positive c_1, c_2 such that for $\|x - \hat{x}\|$ small enough,

$$c_1\|x-\hat{x}\|^2 \leqq \mathcal{G}(x) - \mathcal{G}(\hat{x}) \leqq c_2\|x-\hat{x}\|^2$$

from Theorem 5 and Lemma 5.1. Substituting x^k for x we have

$$4c_1\|x^k-\hat{x}\|^2 \leqq \lambda_n^2 - \nu_k^2 \leqq 4c_2\|x^k-\hat{x}\|^2.$$

The same inequality holds with x^{k+1}, ν_{k+1} replacing x^k, ν_k and then the cubic convergence of ν_k follows from that of the x^k. □

This shows that this algorithm converges cubically in a similar manner to the Rayleigh quotient iteration. There are a number of ways of modifying this algorithm to ensure that (i) the sequence generated is a descent sequence for $\mathcal{G}$ and (ii) the sequence remains in K_R. With such modifications, we can increase its domain of convergence. It is not clear to us at this time which method is preferable, but the estimates for the domain of convergence of the method based on this analysis are usually quite small.

As described before, we can equally well apply this algorithm to the functions $\mathscr{F}(x; \mu)$ (or $\mathscr{F}_1(x; \mu)$), defined in § 4, provided that the eigenvalue λ_k closest to μ and greater (less) than μ is a simple eigenvalue of (A, M). In the case of $\mathscr{F}$ we obtain a cubically convergent algorithm converging to $(\mu - \lambda_k)^{-1/2} Me$, where e^k is a corresponding normalized eigenvector of (A, M).

9. Computational observations. The author has implemented some versions of the algorithm described in § 6. We have not, however, attempted to make a systematic comparison of these algorithms with any of the standard methods for finding specific eigenvalues of symmetric matrices—although this is obviously of interest.

It was found that if we simply use a standard steepest descent or conjugate gradient routine to minimize the function, then with random matrices and initial conditions the convergence could be painfully slow. Introducing step (2) in the algorithm made a big difference. We implemented steepest descent and conjugate gradient methods with both the Fletcher–Reeves and Pollak–Ribière updating formulae in the choice of the descent vector h^k in step (3). When $M = I$, A a random symmetric matrix, and with n up to 90, the algorithm described in § 6 converged quite rapidly, and stably, to the desired eigenvector. Usually it required $O(n)$ iterations where the constant depends on the size of the matrix entries and ε.

Close to the minimizer, we may improve the rate of convergence by switching to the algorithm described in § 8. Alternatively, we could use a Newton or quasi-Newton algorithm. These are easy to implement because of the special form of the Hessian in (3.4). Each of these steps, however, involves significantly more computation at each iteration and it is worthwhile enforcing a descent condition as we only have local convergence. Thus a hybrid method using the globally convergent algorithm from § 6 initially and then a higher-order method such as that described in § 8 when we are close to the answer, appears to provide a good, stable method for finding particular eigenvectors and eigenvalues of symmetric eigenproblems.

Acknowledgments. I would like to thank Albert W.-K. Chan and Guo Lei for their work on various computations and implementations and C. Lenard for helpful discussions and references.

REFERENCES

[1] G. AUCHMUTY, *Unconstrained variational principles for eigenvalues of real symmetric matrices*, SIAM J. Math. Anal., 20 (1989), pp. 1186–1207.

[2] ———, *Duality algorithms for nonconvex variational principles*, Numer. Funct. Anal. Optim., 10 (1989), pp. 211–264.

[3] C. BEATTIE AND D. W. FOX, *Localization criteria and containment for Rayleigh quotient iteration*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 80–93.

[4] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, San Diego, 1970.

[5] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[6] S. BATTERSON AND J. SMILLIE, *The dynamics of Rayleigh quotient iteration*, SIAM J. Numer. Anal., 26 (1989), pp. 624–636.

[7] G. STRANG, *Linear Algebra and Its Applications*, Third Edition, Harcourt, Brace, Jovanovich, San Diego, CA, 1988.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 4, pp. 707–712, October 1991

008

THE EFFECTIVE COMPUTATION OF EQUILIBRIUM POINT FOR N-PERSON GAMES CYCLIC TO THE NEXT PERSON*

EZIO MARCHI†

Abstract. This paper introduces, in an effective way, the neat computation of equilibrium points for certain classes of n-person games that were introduced as cyclic to the next person.

Key words. equilibrium, computation, N-person games

AMS(MOS) subject classification. 90D05

1. Introduction. This paper develops the effective computation of equilibrium points of a special kind of n-person noncooperative games. These games generalize those completely mixed games studied recently by Cohen, Marchi, and Oviedo [1], which are a natural extension of those introduced by Kaplansky [2] for zero-sum two-person games.

The payoff function of each player is formed by adding a function that depends on his own actions and the actions of the following player (with respect to some ordering defined on the set of players) and another function that depends on the actions of the other players but not on his own strategies.

In this note, we prove a general characterization of equilibrium points. For the subclass of completely mixed n-person games, that characterization provides a simple formula for computing all equilibrium points. That is studied in § 2.

In § 3 we present some particular classes of games for which the set of equilibria turns out to be unique.

2. Equilibrium points and their characterization. Consider a finite n-person noncooperative game

$$\Gamma = (S_1, \cdots, S_n; A_1, \cdots, A_n)$$

where S_i are the strategies of player $i \in N = \{1, \cdots, n\}$. A_i is his payoff function. The mixed extension is given by

$$\bar{\Gamma} = (\Delta(S_1), \cdots, \Delta(S_n); E_1, \cdots, E_n)$$

where

$$\Delta(S_i) = \left\{ \sigma_i \in R^{|S_i|} : \sigma_i(S_i) \geqq 0 \ \forall s_i \in S_i, \sum_{s_i} \sigma_i(s_i) = 1 \right\}$$

is the set of mixed strategies for player $i \in N$. $|S_i|$ denotes the cardinality of S_i. E_i is the expected value of A_i, that is,

$$E_i(\sigma_1, \cdots, \sigma_n) = E_i(\sigma_i, \sigma_{-i}) = \sum_{s_i} \cdots \sum_{s_n} A_i(s_1, \cdots, s_n) \prod_{j=1}^{n} \sigma_j(s_j)$$

where $\sigma_{-i} = (\sigma_1, \cdots, \sigma_{i-1}, \sigma_{i+1}, \cdots, \sigma_n)$.

* Received by the editors October 26, 1988; accepted for publication (in revised form) July 19, 1990.

† Instituto de Matemática Aplicada, Universidad Nacional de San Luis, San Luis, Argentina (rizzotto@imasl.edu.ar). This paper was written at the Department of Applied Mathematics and Analysis, University of Barcelona, Barcelona, Spain, during a visit sponsored by the Ministerio de Educación y Ciencias of Spain.

An equilibrium point of $\bar{\Gamma}$ is a point $\bar{\sigma} = (\bar{\sigma}_i, \bar{\sigma}_{-i}) = (\bar{\sigma}_1, \cdots, \bar{\sigma}_n)$ such that for each $i \in N$

$$E_i(\bar{\sigma}_i, \bar{\sigma}_{-i}) \geqq E_i(\sigma_i, \bar{\sigma}_{-i}) \quad \forall \sigma_i \in \Delta(S_i).$$

By a classical theorem due to Nash, we always know that any $\bar{\Gamma}$ has an equilibrium point. The great problem is its computation. The following result generalizes for n-person games that were presented in Marchi and Tarazaga [6].

PROPOSITION 1. *A point $\bar{\sigma}$ is an equilibrium point of $\bar{\Gamma}$ if and only if $\bar{\sigma}$ is a solution of the system*

$$\lambda_i - E_i(s_i, \sigma_{-i}) = 0 \quad \forall s_i \in \operatorname{supp} \sigma_i$$

$$\lambda_i - E_i(s_i, \sigma_{-i}) \geqq 0 \quad \forall s_i \notin \operatorname{supp} \sigma_i$$

$$\sum_{s_i} \sigma_i(s_i) = 1 \quad \forall i$$

$$\sigma_i(s_i) \geqq 0 \quad \forall_i \quad \forall s_i \in S_i$$

with $\lambda_i = E_i(\sigma_i, \sigma_{-i})$.

The proof of this result is simple and we omit it here. We refer the reader to Marchi [5] for a similar one.

3. Cyclic games and computation of the equilibrium. In this section we deal with the class of games described in the Introduction. We say that the n-person game Γ is cyclic to the next player or briefly cyclic if

$$A_i(s_i, s_{-i}) = a_i^1(s_i, s_{i+1}) + a_i^2(s_{-i}) \bmod n.$$

By mod n we mean, once and for all, that

$$A_n(s_n, s_{-n}) = a_n^1(s_n, s_i) + a_n^2(s_{-n}),$$

and for $i \neq n$, $A_i(s_i, s_{-i}) = a_i^1(s_i, s_{i+1}) + a_i^2(s_{-i})$.

Then it is obvious that

$$E_i(\sigma_i, \sigma_{-i}) = e_i^1(\sigma_i, \sigma_{i+1}) + e_i^2(\sigma_{-i}) \bmod n,$$

where e_i^1 and e_i^2 are the respective expectations of a_i^1 and a_i^2. As an immediate result, we have the following proposition.

PROPOSITION 2. *A point $\bar{\sigma}$ is an equilibrium point of $\bar{\Gamma}$ if and only if it is an equilibrium point of the n-person game*

$$\Gamma' = (\Delta(S_1), \cdots, \Delta(S_n); e_1^1, \cdots, e_n^1).$$

The proof is trivial and therefore we omit it.

For each $i \in N$, we can consider the zero-sum two-person game

$$\Gamma_i' = (\Delta(S_i), \Delta(S_{i+1}); e_i^1) \bmod n.$$

Therefore we can apply the Kaplansky theory. We recall that a completely mixed zero-sum two-person game is a game where all the optimal strategies of both players are completely mixed. That is,

$$\bar{\sigma}_i(s_i) > 0, \qquad \bar{\sigma}_{i+1}(s_{i+1}) > 0$$

for all $s_i \in S_i$ and $s_{i+1} \in S_{i+1}$. Kaplansky [2] proved that the matrix associated with a completely mixed game is square and nonsingular.

Let the matrix ${}_iA$ be the matrix associated with a completely mixed two-person game Γ'_i. Then the optimal strategies might be computed by using the equality

$$ {}_iA\bar{\sigma}_{i+1} = \mu_i \square, $$

where $\bar{\sigma}_{i+1}$ is now considered a column vector of dimension r. $\square$ is the column vector having all the components equal to one. Its dimension is r. All matrices ${}_iA$ are $r \times r$-matrices and are nonsingular. Then

$$ \bar{\sigma}_{i+1} = \mu_{ii} A^{-1} \square $$

and

$$ \square^T \bar{\sigma}_{i+1} = 1 = \mu_i \square^T {}_iA \square $$

or

$$ \mu_i = \frac{1}{\square^T {}_iA^{-1} \square}, $$

substituting

$$ \bar{\sigma}_{i+1} = \frac{{}_iA^{-1}\square}{\square^T {}_iA^{-1}\square} > 0. $$

Moreover, Kaplansky [2] has shown that this optimal strategy is unique. This fact describes the zero-sum two-person game with matrix ${}_iA$.

We now present the following theorem.

THEOREM 3. *The point $\bar{\sigma}$, obtained above, is an equilibrium point of $\bar{\Gamma}'$.*

Proof. Since we have

$$ {}_iA\bar{\sigma}_{i+1} = \mu_i \square $$

for each $i \in N$, then for any σ_i, we always have

$$ e_i^1(\bar{\sigma}_i, \bar{\sigma}_{i+1}) = \bar{\sigma}_{ii}^T A \bar{\sigma}_{i+1} = \mu_i = \sigma_{ii}^T A \bar{\sigma}_{i+1} = \bar{e}_i(\sigma_i, \bar{\sigma}_{i+1}) \quad \forall \sigma_i, $$

which says that the point $\bar{\sigma}$ is an equilibrium point of the game Γ'. $\square$

Now, using Proposition 2, we have that the point $\bar{\sigma}$ constructed above is also an equilibrium point of the game $\bar{\Gamma}$. The corresponding value at this equilibrium point for player $i \in N$ is

$$ \lambda_i = \frac{1}{\square^T {}_iA^{-1}\square} + e_i^2\left(\frac{{}_nA^{-1}}{\square^T_n A^{-1}\square}, \frac{{}_1A^{-1}}{\square^T_1 A^{-1}\square}, \cdots, \frac{{}_{i-2}A^{-1}}{\square^T_{i-2} A^{-1}\square}, \frac{{}_iA^{-1}}{\square^T_i A^{-1}\square}, \cdots, \frac{{}_{n-1}A^{-1}}{\square^T_{n-1} A^{-1}\square}\right). $$

At this point we are tempted to ask about the uniqueness of this point in the game Γ'. Unfortunately, the structure of the set of equilibrium points is so complex that we do not have uniqueness in Γ'. Indeed, consider the game with n players where the matrix ${}_iA$ is the identity matrix I_r of dimension r. Now the point constructed above is an equilibrium point. Such a point is

$$ \bar{\sigma}_{i+1} = \left(\frac{1}{r}, \cdots, \frac{1}{r}\right) $$

for each player $i \in N$ and $\mu_i = 1/r$. However, the point

$$ \tilde{\sigma}_i = (1, 0, \cdots, 0) \quad \forall i \in N $$

is also an equilibrium point. Indeed, we have

$$e_i^1(\tilde{\sigma}_i, \tilde{\sigma}_{i+1}) = \tilde{\sigma}_{ii}^T A \tilde{\sigma}_{i+1} = \tilde{\sigma}_i^T I_r \tilde{\sigma}_{i+1} = 1 \geqq \sigma_i^T I_r \tilde{\sigma}_{i+1} = e_i^1(\sigma_i, \tilde{\sigma}_{i+1}) \quad \forall \sigma_i \in \Delta(S_i).$$

Moreover, it is possible to prove, as has been done in Marchi [4], that the number of equilibrium points for such a game is

$$\sum_{i=1}^{r} \binom{r}{i} = 2^r - 1.$$

Now we would like to give some insight into uniqueness.

We define a cyclic n-person game as completely mixed if and only if all its equilibrium points are completely mixed.

We generalize Theorem 1 of Cohen, Marchi, and Oviedo [1] as follows.

THEOREM 4. *Suppose that the n-person game* Γ' *is nonsingular (all* ${}_iA$ *are nonsingular) and completely mixed. Then the game has a unique solution and nonzero values. The solution and values of the game for each player are given by*

$$\bar{\sigma}_{i+1} = \frac{{}_iA^{-1}\square}{\square^T {}_iA^{-1}\square}, \qquad \mu_i = \frac{1}{\square^T {}_iA^{-1}\square}.$$

Proof. Since the game is completely mixed, any equilibrium point $\bar{\sigma} = (\bar{\sigma}_1, \cdots, \bar{\sigma}_n)$ has the form

$$\tilde{\sigma}_i(s_i) > 0 \quad \forall i \quad \forall s_i \in S_i = \{1, \cdots, r\}.$$

This implies that for each $i \in N$,

$${}_iA\tilde{\sigma}_{i+1} = \mu_i \square.$$

Since there was a row s_i with

$$({}_iA\bar{\sigma}_{i+1})s_i < \mu_i,$$

then $\bar{\sigma}_i(s_i) = 0$. This is impossible since the game is completely mixed.

Then, repeating the argument given before Theorem 3, the result holds true. □

As in Cohen, Marchi, and Oviedo [1], we have that, although it is not surprising that ${}_iA$ uniquely determines μ_i, it is somewhat surprising that ${}_iA$ uniquely determines $\bar{\sigma}_{i+1}$.

For real matrices $\{{}_iA\}$ satisfying the hypothesis of Theorem 4, it follows that

$$\frac{{}_iA^{-1}\square}{\square^T {}_iA^{-1}\square} > O_r^T$$

where O is the column of r elements O and the inequalities apply element-by-element.

These inequalities do not guarantee that the game Γ' is completely mixed, as the example of the game given above shows. However, we point out that the set of completely mixed n-person games similar to Γ' is nonempty.

To see this, we present the following diagonal-cyclic game.

Fix $a_{ij} > 0$ for all $i = 1, \cdots, n, j = 1, \cdots, r, 2 \leqq n < r$.

$$ {}_1A = \begin{bmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{1r} \end{bmatrix}, \qquad {}_2A = \begin{bmatrix} 0 & a_{21} & & 0 \\ & \ddots & \ddots & \\ & & \ddots & a_{2r-1} \\ a_{2r} & & & 0 \end{bmatrix} $$

$$ {}_iA = \begin{bmatrix} 0 & \cdots & 0 & a_{i1} & \cdots & 0 \\ & & & & \ddots & \\ & & & & & a_{ir-i+1} \\ a_{ir-i+2} & & & 0 & & \\ & \ddots & & & & \\ 0 & & a_{ir} & & & 0 \end{bmatrix} $$

$$ {}_nA = \begin{bmatrix} 0 & \cdots & 0 & a_{n1} & \cdots & 0 \\ & & & & \ddots & \\ & & & & & a_{inr-n+1} \\ a_{nr-n+2} & & & 0 & & \\ & \ddots & & & & \\ 0 & & a_{nr} & & & 0 \end{bmatrix}. $$

THEOREM 5. *If Γ' is the diagonal-cyclic n-person game given above, then Γ' is nonsingular and completely mixed. Moreover, the game Γ' has a unique solution*

$$ \bar{\sigma}_{i+1}(j) = (\mu_{ii}A^{-1}\square)_j = 1 \Big/ \sum_{j=1}^{r} a_{ij} = 1/a_{ir-i+j+1} \bmod r $$

$$ \mu_i = 1 \Big/ \sum_{j=1}^{r} a_{ij}. $$

Proof. Clearly ${}_iA$ is nonsingular. It is immediate that

$$ {}_nA = \begin{bmatrix} 0 & \cdots & 0 & 1/a_{ir-i+2} & & 0 \\ & & & & \ddots & \\ & & & & & 1/a_{ir} \\ 1/a_{i1} & & & 0 & & \\ & \ddots & & & & \\ 0 & & a_{ir-i+1} & & & 0 \end{bmatrix}, $$

and by the formula

$$ \mu_i = 1/\square^T{}_iA^{-1}\square = 1 \Big/ \sum_{j=1}^{r} a_{ij} $$

$$ \bar{\sigma}_{i+1}(j) = 1 \Big/ \sum_{j=1}^{r} a_{ij} = 1/a_{ir-i+j+1} \bmod r, $$

we have the real computation of an equilibrium point, and no other solution with all positive elements is possible.

To prove that every solution has all elements positive, an argument similar to that developed in the proof of Theorem 3 of [1] applies here with n steps instead of two. Thus the game Γ' is completely mixed. □

As an application of the previous theorem, we have that a cyclic game whose first component e_i^1 is diagonal-cyclic is also completely mixed.

It seems of some importance to develop further the theory of completely mixed n-person games, as well as to apply the theory of perturbation developed in [1] for two-person games to the cyclic games. We leave this subject for a further study.

Acknowledgment. I acknowledge the warm invitation and the hospitality of all the people of the Department of Applied Mathematics and Analysis, University of Barcelona, Barcelona, Spain. In particular I thank Professor J. E. Martinez Legaz for his interest in my visit, as well as different comments.

REFERENCES

[1] J. COHEN, E. MARCHI, AND J. A. OVIEDO, *Perturbation theory of completely mixed bimatrix games*, J. Linear Algebra Appl., 114/115 (1989), pp. 169–180.

[2] I. KAPLANSKY, *A contribution to von Neumann's theory of games*, Ann. of Math., 46 (1945), pp. 474–479.

[3] S. KARLIN, *Mathematical Methods and Theory in Programming and Economics*, Vol. 1, Addison–Wesley, Reading, MA, 1959.

[4] E. MARCHI, *On equilibrium points of diagonal n-person games*, J. Optim. Theory Appl., 64 (1990).

[5] ———, *Una nota acerca de los puntos E-estables perfectos y propios*, Collect. Math., 39 (1988), pp. 9–19.

[6] E. MARCHI AND P. TARAZAGA, *Relevant aspects in two person games*, J. Optim. Theory Appl., 53 (1987), pp. 125–131.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 4, pp. 713–725, October 1991

009

REDUCING THE COMPUTATIONS OF THE SINGULAR VALUE DECOMPOSITION ARRAY GIVEN BY BRENT AND LUK*

B. YANG† AND J. F. BÖHME†

Abstract. A new, efficient, two-plane rotation (TPR) method for computing two-sided rotations involved in singular value decomposition (SVD) is presented. It is shown that a two-sided rotation can be evaluated by only two plane rotations and a few additions. This leads to significantly reduced computations. Moreover, if coordinate rotation digital computer (CORDIC) processors are used for realizing the processing elements (PEs) of the SVD array given by Brent and Luk, the computational overhead of the diagonal PEs due to angle calculations can be avoided. The resulting SVD array has a homogeneous structure with identical diagonal and off-diagonal PEs. Similar results can also be obtained if the TPR method is applied to Luk's triangular SVD array and to Stewart's Schur decomposition array.

Key words. singular value decomposition, systolic arrays, CORDIC, two-sided rotations, VLSI

AMS(MOS) subject classification. 15A18

1. Introduction.

One important problem in linear algebra and digital signal processing is the singular value decomposition (SVD). Typical applications arise in beamforming and direction finding, spectrum analysis, digital image processing, etc. [1]. Recently, there has been a massive interest in parallel architectures for computing SVD because of the high computational complexity of SVD, the growing importance of real-time signal processing, and the rapid advances in very large scale integration (VLSI) that make low-cost, high-density and fast processing memory devices available.

There are different numerically stable methods for computing complete singular value and singular vector systems of dense matrices, for example, the Jacobi SVD method, the QR method, and the one-sided Hestenes method. For parallel implementations, the Jacobi SVD method is far superior in terms of simplicity, regularity, and local communications. Brent, Luk, and Van Loan have shown how the Jacobi SVD method with parallel ordering can be implemented by a two-dimensional systolic array [2], [3]. Various coordinate rotation digital computer (CORDIC) realizations of the SVD array have been reported by Cavallaro and Luk [4] and Delosme [5], [6].

The Jacobi SVD method is based on, as common for all two-sided approaches, applying a sequence of two-sided rotations to 2×2 submatrices of the original matrix. The computational complexity is thus determined by how to compute the two-sided rotations. In most previous works, a two-sided rotation is evaluated in a straightforward manner by four plane rotations, where two of them are applied from left to the two column vectors of the 2×2 submatrix and the other ones are applied from right to the row vectors, respectively. In the diagonal processing elements (PEs), additional operations for calculating rotation angles are required. This leads to an inhomogeneous array architecture containing two different types of PEs.

In this paper, we develop a two-plane rotation (TPR) method for computing two-sided rotations. We show that the above computational complexity can be reduced significantly because each two-sided rotation can be evaluated by only two plane rotations and a few additions. Moreover, the SVD array given by Brent and Luk becomes homogeneous with identical diagonal and off-diagonal PEs when CORDIC processors are

* Received by the editors September 28, 1989; accepted for publication (in revised form) August 2, 1990.
† Department of Electrical Engineering, Ruhr-Universität Bochum, 4630 Bochum, Germany.

used. In a recent work [6], Delosme has also indicated this possibility in connection with "rough rotations" independently. He has taken, however, a different approach that is based on encoding the rotation angles. He has still required four plane rotations on the off-diagonal PEs while diagonal and off-diagonal operations can be overlapped.

Our paper is organized as follows. In § 2, we briefly reexamine Jacobi's SVD method and Brent and Luk's SVD array. Then, we develop the TPR method in § 3. The CORDIC algorithm is described in § 4, where in particular CORDIC scaling correction techniques are discussed and examples of scaling-corrected CORDIC sequences are given. In § 5, a unified CORDIC SVD module for all PEs of the SVD array is presented. This module is compared to those proposed by Cavallaro, Luk, and Delosme in § 6. Finally, we stress the applicability of the TPR method to several other problems.

2. Jacobi SVD method. In this paper, we consider real, square, and nonsymmetric matrices. Let $M \in \mathbb{R}^{N\times N}$ be a matrix of dimension N. The SVD is given by

$$M = U\Sigma V^T, \tag{1}$$

where $U \in \mathbb{R}^{N\times N}$ and $V \in \mathbb{R}^{N\times N}$ are orthogonal matrices containing the left and right singular vectors, and $\Sigma \in \mathbb{R}^{N\times N}$ is a diagonal matrix of singular values, respectively. The superscript T denotes matrix transpose. Based on an extension of the Jacobi eigenvalue algorithm [7], Kogbetliantz [8] and Forsythe and Henrici [9] proposed to diagonalize M by a sequence of two-sided rotations,

$$M_0 = M, \qquad M_{k+1} = U_k^T M_k V_k \qquad (k = 0, 1, 2, \cdots). \tag{2}$$

U_k and V_k describe two rotations in the (i, j)-plane ($1 \leqq i < j \leqq N$), where the rotation angles are chosen to annihilate the elements of M_k at the positions (i, j) and (j, i). Usually, several sweeps are necessary to complete the SVD, where a sweep is a sequence of $N(N-1)/2$ two-sided rotations according to a special ordering of the $N(N-1)/2$ different index pairs (i, j).

For sequential computing on a uniprocessor system, possibly the most frequently used orderings are the cyclic orderings, namely, the cyclic row ordering

$$(i,j) = (1,2),(1,3),\cdots,(1,N),(2,3),\cdots,(2,N),\cdots,(N-1,N) \tag{3}$$

or the equivalent cyclic column ordering. Sameh [10] and Schwiegelshohn and Thiele [11] have shown how to implement the cyclic row ordering on a ring-connected or a mesh-connected processor array. Recently, a variety of parallel orderings have been developed. Luk and Park [12] have shown that these parallel orderings are essentially equivalent to the cyclic orderings and thus share the same convergence properties.

Brent and Luk have suggested a particular parallel ordering and developed a square systolic array consisting of $\lceil N/2 \rceil \times \lceil N/2 \rceil$ PEs for implementing the Jacobi SVD method (Fig. 1). To do this, the matrix M is partitioned into 2×2 submatrices. Each PE contains one submatrix and performs a two-sided rotation

$$B = R(\theta_1)^T A R(\theta_2), \tag{4}$$

where

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \tag{5}$$

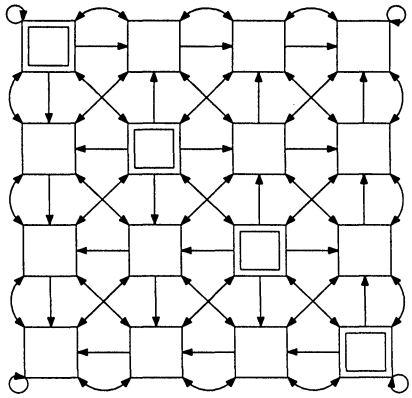


FIG. 1. *The* SVD *array given by Brent and Luk.*

denote the submatrix before and after the two-sided rotation, respectively, and

$$R(\theta)=\begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix} \tag{6}$$

describes a plane rotation through the angle θ. At first, the diagonal PEs (symbolized by a double square in Fig. 1) generate the rotation angles to diagonalize the 2×2 submatrices ($b_{12}=b_{21}=0$) stored in them. This means that θ_1 and θ_2 are first calculated from the elements of A and then relation (4) is used to compute b_{11} and b_{22}. We call this the generation mode. Then, the rotation angles are sent to all off-diagonal PEs in the following way: the angles associated to the left-side rotations propagate along the rows while the angles associated to the right-side rotations propagate along the columns. Once these angles are received, the off-diagonal PEs perform the two-sided rotations (4) on their stored data. We call this the rotation mode. Clearly, if we compute the rotation mode straightforwardly, we require four plane rotations. For the generation mode, additional operations for calculating θ_1 and θ_2 are required.

3. TPR method for computing two-sided rotations. In order to develop the TPR method for computing two-sided rotations more efficiently, we first discuss the commutative properties of two special types, the rotation-type and the reflection-type, of 2×2 matrices. We define

$$\mathcal{M}^{\mathrm{rot}}=\left\{\begin{pmatrix}x & y\\ -y & x\end{pmatrix}\middle|\, x,y\in\mathbb{R}\right\}\quad\text{and}\quad \mathcal{M}^{\mathrm{ref}}=\left\{\begin{pmatrix}x & y\\ y & -x\end{pmatrix}\middle|\, x,y\in\mathbb{R}\right\}. \tag{7}$$

The former is called rotation-type because it has the same matrix structure as a 2×2 plane rotation matrix. Similarly, the latter is called reflection-type because it has the same matrix structure as a 2×2 Givens reflection matrix [13]. Note that x and y must not be normalized to $x^2+y^2=1$. Using the above definitions, the following results can be shown by some elementary manipulations.

LEMMA 1. *If* $A_1\in\mathcal{M}^{\mathrm{rot}}$ *and* $A_2\in\mathcal{M}^{\mathrm{rot}}$, *then* $A_1A_2=A_2A_1\in\mathcal{M}^{\mathrm{rot}}$.

LEMMA 2. *If* $A_1\in\mathcal{M}^{\mathrm{ref}}$ *and* $A_2\in\mathcal{M}^{\mathrm{rot}}$, *then* $A_1A_2=A_2^TA_1\in\mathcal{M}^{\mathrm{ref}}$.

In particular, if we consider two plane rotations, we know the following.

LEMMA 3. *If* $R(\theta_1)$ *and* $R(\theta_2)$ *are plane rotations described by* (6), *then* $R(\theta_1)R(\theta_2)=R(\theta_1+\theta_2)$ *and* $R(\theta_1)^TR(\theta_2)=R(\theta_2-\theta_1)$.

Now, we give a theorem describing the rotation mode of the TPR method.

THEOREM. *If the* 2×2 *matrix* A *and the two rotation angles* θ_1 *and* θ_2 *are given, then the two-sided rotation* (4) *can be computed by two plane rotations, ten additions,*

and four scalings by $\frac{1}{2}$:

$$p_1=(a_{22}+a_{11})/2, \qquad p_2=(a_{22}-a_{11})/2, \tag{8}$$
$$q_1=(a_{21}-a_{12})/2, \qquad q_2=(a_{21}+a_{12})/2,$$

$$\theta_-=\theta_2-\theta_1, \qquad \theta_+=\theta_2+\theta_1, \tag{9}$$

$$\begin{pmatrix} r_1 \\ t_1 \end{pmatrix}=R(\theta_-)\begin{pmatrix} p_1 \\ q_1 \end{pmatrix}, \qquad \begin{pmatrix} r_2 \\ t_2 \end{pmatrix}=R(\theta_+)\begin{pmatrix} p_2 \\ q_2 \end{pmatrix}, \tag{10}$$

$$b_{11}=r_1-r_2, \qquad b_{12}=-t_1+t_2, \tag{11}$$
$$b_{21}=t_1+t_2, \qquad b_{22}=r_1+r_2.$$

Proof. Using (8), the matrix A can be reformulated as

$$A=A_1+A_2=\begin{pmatrix} p_1 & -q_1 \\ q_1 & p_1 \end{pmatrix}+\begin{pmatrix} -p_2 & q_2 \\ q_2 & p_2 \end{pmatrix}.$$

Clearly, $R(\theta_1)$, $R(\theta_2)$ in (4) and A_1 are elements of $\mathcal{M}^{\text{rot}}$ while A_2 belongs to $\mathcal{M}^{\text{ref}}$. This leads to the following reformulation of the matrix B by using Lemmas 1–3:

$$\begin{aligned} B&=R(\theta_1)^TAR(\theta_2)\\ &=R(\theta_1)^TA_1R(\theta_2)+R(\theta_1)^TA_2R(\theta_2)\\ &=R(\theta_1)^TR(\theta_2)A_1+R(\theta_1)^TR(\theta_2)^TA_2\\ &=R(\theta_2-\theta_1)A_1+R(\theta_2+\theta_1)^TA_2\\ &=R(\theta_-)\begin{pmatrix} p_1 & -q_1 \\ q_1 & p_1 \end{pmatrix}+R(\theta_+)^T\begin{pmatrix} -p_2 & q_2 \\ q_2 & p_2 \end{pmatrix}\\ &=\begin{pmatrix} r_1 & -t_1 \\ t_1 & r_1 \end{pmatrix}+\begin{pmatrix} -r_2 & t_2 \\ t_2 & r_2 \end{pmatrix}. \end{aligned}$$

This completes the proof.

The generation mode of the TPR method follows directly from the above theorem.

COROLLARY. *If the* 2×2 *matrix* A *is given, we can diagonalize* A *and calculate the corresponding rotation angles* θ_1 *and* θ_2 *by two Cartesian-to-polar coordinates conversions, eight additions, and four scalings by* $\frac{1}{2}$:

$$p_1=(a_{22}+a_{11})/2, \qquad p_2=(a_{22}-a_{11})/2, \tag{12}$$
$$q_1=(a_{21}-a_{12})/2, \qquad q_2=(a_{21}+a_{12})/2,$$

$$r_1=\operatorname{sign}(p_1)\sqrt{p_1^2+q_1^2}, \qquad r_2=\operatorname{sign}(p_2)\sqrt{p_2^2+q_2^2}, \tag{13}$$
$$\theta_-=\arctan(q_1/p_1), \qquad \theta_+=\arctan(q_2/p_2),$$

$$\theta_1=(\theta_+-\theta_-)/2, \qquad \theta_2=(\theta_++\theta_-)/2, \tag{14}$$

$$b_{11}=r_1-r_2, \qquad b_{22}=r_1+r_2. \tag{15}$$

Proof. Regarding (11), $b_{12}=b_{21}=0$ is equivalent to $t_1=t_2=0$. Equation (13) follows then from (10). This completes the proof.

In equation (13), we choose the rotation through the smaller angle. All vectors lying in the first or the fourth quadrant are rotated onto the positive x-axis, and all vectors lying in the second and the third quadrant are rotated onto the negative x-axis. For vectors on the y-axis, the rotation direction is arbitrary. Thus, the generated rotation

angles θ_- and θ_+ satisfy $|\theta_-|, |\theta_+| \leqq 90°$. This results in

$$|\theta_1| \leqq 90° \quad \text{and} \quad |\theta_2| \leqq 90°, \tag{16}$$

due to (14).

Equation (16) is important with respect to the convergence of the Jacobi SVD method. Forsythe and Henrici [9] have proven the convergence for cyclic orderings if the rotation angles θ_1 and θ_2 are restricted to a closed interval inside the open interval $(-90°, 90°)$. They have also demonstrated that this condition may fail to hold, i.e., θ_1 and θ_2 may be $\pm 90°$, if the off-diagonal elements b_{12} and b_{21} in (5) have to be exactly annihilated. As a remedy, they suggested an under- or overrotation by computing the two-sided rotation (4) with angles $(1 - \gamma)\theta_1$ and $(1 - \gamma)\theta_2$ $(-1 < \gamma < 1)$ and proved its convergence. In practice, however, the finite machine accuracy in the real arithmetic allows only an approximative computation of the rotation angles and implies under- or overrotations. So the Jacobi SVD method converges without using under- or overrotations as shown by the experimental results of Brent, Luk, and Van Loan [3]. In case of CORDIC implementations, the effect of implicit under- or overrotations is more apparent. The angles $\pm 90°$ can never be exactly calculated because of the limited angle resolution arctan (2^{-p}) of the CORDIC algorithm, where p denotes the mantissa length.

4. The CORDIC algorithm. In the previous section, we have seen that the main operations of the TPR-method are plane rotations and Cartesian-to-polar coordinates conversions. These operations can be carried out by multiplier–adder-based processors supported by software or special hardware units. An alternative approach is the use of dedicated processors that usually map algorithms more effectively to hardware. The CORDIC processor is such a powerful one for calculating trigonometric functions.

The CORDIC algorithm was originally designed by Volder [14] as an iterative procedure for computing plane rotations and Cartesian-to-polar coordinates conversions. It was later generalized and unified by Walther [15], enabling a CORDIC processor to calculate more functions, including hyperbolic functions, as well as multiplications and divisions. In the following, we consider Volder's CORDIC algorithm because only trigonometric functions are involved in SVD applications.

The CORDIC algorithm consists of iterative shift-add operations on a three-component vector,

$$\begin{pmatrix} x_{i+1} \\ y_{i+1} \end{pmatrix} = \begin{pmatrix} x_i - \sigma_i \delta_i y_i \\ y_i + \sigma_i \delta_i x_i \end{pmatrix} = \frac{1}{\cos(\alpha_i)} \begin{pmatrix} \cos(\alpha_i) & -\sigma_i \sin(\alpha_i) \\ \sigma_i \sin(\alpha_i) & \cos(\alpha_i) \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix}, \tag{17}$$

$$z_{i+1} = z_i - \varepsilon \sigma_i \alpha_i \qquad (0 < \delta_i < 1;\ \sigma_i = \pm 1;\ \varepsilon = \pm 1;\ i = 0, 1, \cdots, n-1), \tag{18}$$

in which the iteration stepsize δ_i is defined by

$$\delta_i = \tan(\alpha_i) = 2^{-S(i)}. \tag{19}$$

The set of integers $\{S(i)\}$ parametrizing the iterations is called CORDIC sequence. Equation (17) can be interpreted, except for a scaling factor of

$$k_i = \frac{1}{\cos(\alpha_i)} = \sqrt{1 + \delta_i^2}, \tag{20}$$

as a rotation of $(x_i, y_i)^T$ through the angle α_i, where the sign $\sigma_i = \pm 1$ gives the rotation direction. After n iterations, the results are given by

$$\begin{pmatrix} x_n \\ y_n \end{pmatrix} = K \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}, \tag{21}$$

$$z_n = z_0 - \varepsilon\alpha, \tag{22}$$

with the overall scaling factor $K = \prod_i k_i$ and the total rotation angle $\alpha = \sum_i \sigma_i \alpha_i$. Now, if the CORDIC sequence satisfies the following convergence condition

$$\alpha_i - \sum_{j=i+1}^{n-1} \alpha_j \leqq \alpha_{n-1} \qquad (i = 0, 1, \cdots, n-2), \tag{23}$$

we can choose the sign parameter

$$\sigma_i = \begin{cases} -\text{sign}\,(x_i y_i) & \text{for } y_n \to 0, \\ \text{sign}\,(\varepsilon z_i) & \text{for } z_n \to 0 \end{cases} \tag{24}$$

to force y_n or z_n to zero, provided that the input data x_0, y_0, and z_0 lie in the convergence region

$$C = \sum_{i=0}^{n-1} \alpha_i \geqq \begin{cases} |\arctan\,(y_0/x_0)| & \text{for } y_n \to 0, \\ |z_0| & \text{for } z_n \to 0. \end{cases} \tag{25}$$

In this way, two different types of CORDIC trigonometric functions can be computed (Table 1). In the mode $y_n \to 0$, the Cartesian coordinate (x_0, y_0) of a plane vector is converted to its polar representation, where the parameter $\varepsilon = \pm 1$ determines the sign of the phase angle calculated. When $z_n \to 0$, a given plane vector is rotated through the angle z_0, where $\varepsilon = \pm 1$ controls the rotation direction.

In Table 1, the principal value $|\arctan\,(y_0/x_0)| \leqq 90°$ of the inverse tangent function is calculated when computing Cartesian-to-polar coordinates conversions. Correspondingly, x_n may be positive or negative according to the sign of x_0. So, it is guaranteed that a vector is always rotated through the smaller angle onto the x-axis in accordance with (13). In this case, a convergence region of $C \geqq 90°$ is sufficient for the generation mode of the two-sided rotation.

One main drawback of the CORDIC algorithm is the need of correcting the scaling factor K that arises during the iterations (17). For example, if we use Volder's CORDIC sequence

$$\{S(i)\} = \{0, 1, 2, 3, \cdots, p-1, p\}, \tag{26}$$

with $n = p + 1$ CORDIC iterations for a mantissa accuracy of 2^{-p}, the scaling factor is $K \approx 1.64676$. Compensating this undesired scaling effect with a minimum number of computations is of particular importance.

Clearly, multiplying x_n and y_n in Table 1 by K^{-1} will degrade the algorithm performance substantially. Most of the scaling correction issues are based on shift-add operations. For a two-sided rotation that is implemented by four plane rotations, each matrix element undergoes two plane rotations so that the total scaling factor to be corrected is K^2. In this case, Cavallaro and Luk [16] have pointed out that there is a simple systematic approach for scaling correction when using the CORDIC sequence (26). They proposed to use $\lceil p/4 \rceil$ scaling iterations of the type $x \leftarrow x - 2^{-2j}x$ with $j \in J = \{1, 3, 5, \cdots, 2\lceil p/4 \rceil - 1\}$ and one shift operation 2^{-1}. The remaining scaling error is

TABLE 1
CORDIC *trigonometric functions* ($\varepsilon = \pm 1$).

$y_n \to 0$	$z_n \to 0$
$x_n = K\,\text{sign}\,(x_0)\sqrt{x_0^2 + y_0^2}$ $z_n = z_0 + \varepsilon \arctan\,(y_0/x_0)$	$x_n = K(x_0 \cos z_0 - \varepsilon y_0 \sin z_0)$ $y_n = K(\varepsilon x_0 \sin z_0 + y_0 \cos z_0)$

bounded by 2^{-p-1},[1]

$$(27)\quad \left|1-2^{-1}\prod_{j\in J}(1-2^{-2j})\cdot K^2\right| = \left|1-2^{-1}\prod_{j\in J}(1-2^{-2j})\cdot\prod_{i=0}^{p}(1+2^{-2i})\right| < 2^{-p-1}.$$

This approach, however, fails in the TPR method. Here, each matrix element undergoes only one plane rotation. The scaling factor to be corrected is thus K rather than K^2. In order to solve this more difficult problem, different techniques have been developed in the literature. Haviland and Tuszynski [17] used similar scaling iterations as Cavallaro and Luk. Ahmed [18] repeated some CORDIC iterations to force K to a power of the machine radix. Delosme [19] combined both methods of Haviland, Tuszynski, and Ahmed for minimizing the number of computations. Deprettere, Dewilde, and Udo [20] suggested the double-shift concept.

We designed a computer program [21] for a systematic search of CORDIC sequences. We allow shifts parameters $S(i)$ $(i = 0, 1, \cdots, n-1)$ with differences $S(i+1) - S(i) \in \{0, 1, 2\}$ to provide more optimization freedom. For an efficient scaling correction, we require that the scaling factor K be corrected by a sequence of n_k shift-add operations,

$$(28)\quad 2^{-T(0)}\prod_{j=1}^{n_k}(1+\eta(j)2^{-T(j)})\cdot K = 1+\Delta K \qquad (T(j)\,\text{integers}, \eta(j)=\pm 1).$$

These additional scaling iterations are parametrized by the set of signed integers $\{T(0), \eta(1)T(1), \cdots, \eta(n_k)T(n_k)\}$. The total number of iterations is $L = n + n_k$.

In (28), ΔK denotes the remaining relative scaling error after the scaling correction. We emphasize that this is a systematic error with a constant sign. By contrast, the other two types of CORDIC errors, the angular error due to the limited angle resolution and the rounding error, are of statistical nature because they may be positive or negative. The scaling error is thus more critical with respect to error accumulation when repeated CORDIC operations on the same data have to be computed as in SVD applications. Roughly speaking, the total scaling error after k CORDIC function calls increases linearly with k, a fact that has been verified by our numerical experiments. For this reason, we require $|\Delta K|$ to be much smaller than 2^{-p}.

We found catalogues of CORDIC sequences with complexity comparable to those of Cavallaro and Luk. In the following, five examples for different mantissa lengths $p = 16, 20, 24, 28$, and 32, including the total number of iterations $L = n + n_k$, the convergence region C, and the remaining scaling error ΔK are given:

$p=16$: $\{S(i)\} = \{0\,1\,2\,3\,\cdots\,15\,16\}$,

$\{\eta(j)T(j)\} = \{1+2-5+9+10\}$, $L = 17+4$, $C \approx 100°$, $\Delta K \approx -2^{-16.01}$,[2]

$p=20$: $\{S(i)\} = \{0\,1\,2\,3\,\cdots\,19\,20\}$,

$\{\eta(j)T(j)\} = \{1+2-5+9+10+16\}$,

$L = 21+5$, $C \approx 100°$, $\Delta K \approx 2^{-23.05}$,

[1] When replacing $\lceil p/4 \rceil$ by $\lceil (p-1)/4 \rceil$ or $\lceil (p+1)/4 \rceil$, the upper bound in (27) becomes 2^{-p} or 2^{-p-2}, respectively.

[2] When appending an additional scaling iteration with $\eta(5)T(5) = +16$, the scaling accuracy can be enhanced to $\Delta K \approx 2^{-23}$.

$p=24$: $\{S(i)\}=\{1\,1\,2\,3\,3\,4\,5\,5\,6\,6\,7\,8\,8\,9\,10\cdots 23\,24\}$,

$\{\eta(j)T(j)\}=\{0-2+6\}$, $L=29+2$, $C\approx 91°$, $\Delta K\approx -2^{-29.13}$,

$p=28$: $\{S(i)\}=\{1\,1\,2\,3\,3\,4\,5\,5\,6\,6\,7\,8\,8\,9\,10\,11\,12\,13\,14\,14\,15\cdots 27\,28\}$,

$\{\eta(j)T(j)\}=\{0-2+6\}$, $L=34+2$, $C\approx 91°$, $\Delta K\approx 2^{-32.53}$,

$p=32$: $\{S(i)\}=\{0\,0\,1\,3\,3\,3\,4\,5\,6\,7\,8\,9\,9\,10\cdots 31\,32\}$,

$\{\eta(j)T(j)\}=\{1-3-8+16-25-27\}$,

$L=36+5$, $C\approx 145°$, $\Delta K\approx -2^{-39.93}$.

Remember that in order to meet the convergence condition (23) and to provide a convergence region $C \geqq 90°$, the minimum number of CORDIC iterations is $p+1$. So, for all CORDIC sequences given above, the number $L-(p+1)$ of additional iterations for scaling corrrection is $p/4$. Moreover, except for the first CORDIC sequence, the remaining scaling error $|\Delta K|$ is significantly smaller than 2^{-p}. This leads to improved numerical properties compared with other CORDIC sequences reported in the literature. We also remark that if the symmetric eigenvalue problem is considered for which a convergence region of $C \geqq 45°$ is sufficient [2], the total number of CORDIC iterations L can be further reduced. An example that is nearly identical to the last CORDIC sequence given above is

$p=32$: $\{S(i)\}=\{1\,3\,3\,3\,4\,5\,6\,7\,8\,9\,9\,10\cdots 31\,32\}$,

$\{\eta(j)T(j)\}=\{0-3-8+16-25-27\}$,

$L=34+5$, $C\approx 55°$, $\Delta K\approx -2^{-39.93}$.

For comparison, Delosme [5] has also given an optimized CORDIC sequence for the same situation. His sequence requires one iteration more ($L=40$) and achieves a scaling accuracy of $\Delta K\approx 2^{-33.16}$.

We suspect that similar results can also be obtained by using Deprettere's double-shift concept. However, this method requires a slightly increased hardware complexity and will not be discussed in this paper.

5. CORDIC implementation of the SVD PEs. For easy illustration, we first introduce a CORDIC processor symbol as shown in Fig. 2. The descriptions inside the box determine uniquely the function mode of the CORDIC processor according to Table 1. The output data x and y are assumed to be scaling corrected.

It is now simple to map the operations (8)–(11) and (12)–(15) of the TPR method onto a two CORDIC processor architecture. In Fig. 3, the diagonal PEs of the SVD array are implemented by two CORDIC processors and eight adders. The dotted inputs of the adders represent negated inputs. Because the diagonal PEs work in the generation mode,

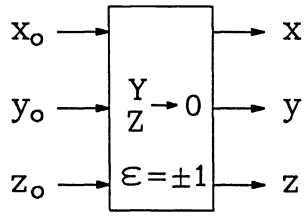


FIG. 2. CORDIC *symbol*.

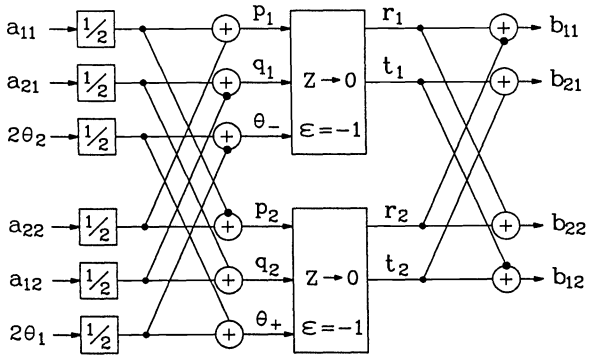


FIG. 3. CORDIC *implementation of the diagonal* PE *of the* SVD *array*.

both CORDIC processors are driven in the "$y \rightarrow 0$" mode for computing Cartesian-to-polar coordinates conversions. In Fig. 4, the off-diagonal PEs working in the rotation mode are implemented by two CORDIC processors and ten adders. Here, the CORDIC processors are driven in the "$z \rightarrow 0$" mode for performing plane rotations.

Obviously, both CORDIC implementations have nearly the same architecture. All PEs of the SVD array can thus be implemented by one unified CORDIC SVD module (Fig. 5) without considerably increased hardware cost. The different computation modes of the diagonal and off-diagonal PEs are easily "programmed" by one control bit. The resulting SVD array is similar to that in Fig. 1, but homogeneous with identical PEs.

We remark that Fig. 5 is more a "graphic program" describing the sequence of operations to be computed rather than a hardware block diagram. We show in the following that the 12 adders that are paired into three pre-butterflys and three post-butterflys can be integrated into the two CORDIC processors without separate hardware realizations. The Jacobi SVD method is a recursive method. Each PE of the SVD array has to exchange data with its diagonal neighbors. Because of this data dependency, only recursive CORDIC processors can be used here. This is an arithmetic unit consisting of mainly three adders and two barrel-shifters. It carries out the L iterations of the CORDIC algorithm in L cycles by using data feedback. The two CORDIC processors contained in one CORDIC SVD module require six adders altogether. So, it is natural to modify the CORDIC processor architecture slightly and to use the existing six adders for computing both the pre-butterfly and the post-butterfly operations. The resulting CORDIC SVD module has

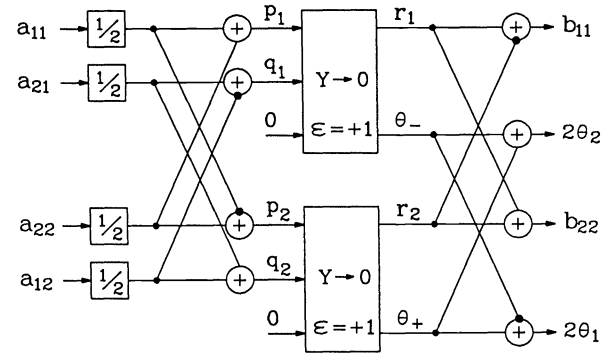


FIG. 4. CORDIC *implementation of the off-diagonal* PE *of the* SVD *array*.

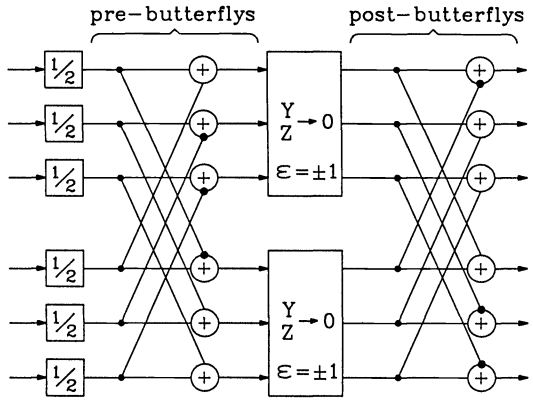


FIG. 5. *A unified* CORDIC SVD *module for implementing all* PE*s of the* SVD *array.*

the hardware complexity of two recursive CORDIC processors and requires a total computation time of $L + 2$ iterations.

In Fig. 6, the principal architecture of such a two CORDIC processor SVD module is shown. The dashed lines and boxes represent the additional hardware components

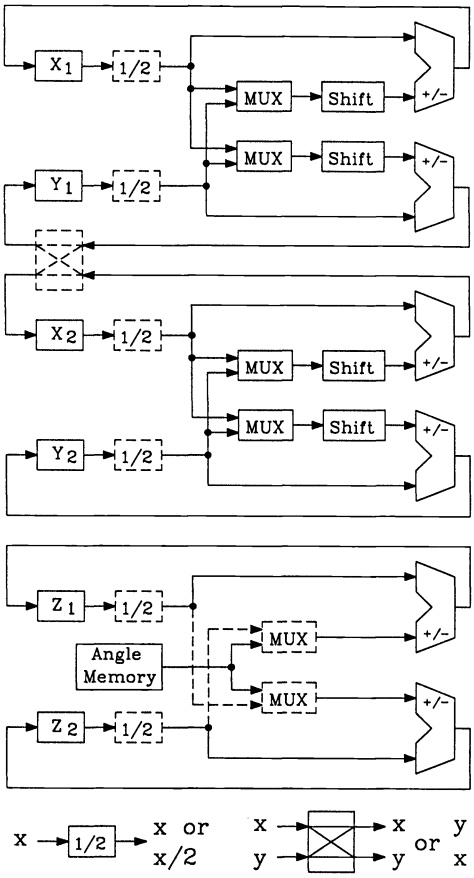


FIG. 6. *The principal architecture of the unified* CORDIC SVD *module.*

enabling the CORDIC processors to compute the butterfly operations. It is easily verified that the upper four adders devoted to x and y can perform the following types of operations $2^{-1}x \pm 2^{-1}y$ (pre-butterfly), $x \pm y$ (post-butterfly), $x \pm 2^{-s}y$ (CORDIC iteration) and $x \pm 2^{-s}x$ (scaling iteration) while the lower two adders devoted to z can compute $2^{-1}z_1 \pm 2^{-1}z_2$ (pre-butterfly), $z_1 \pm z_2$ (post-butterfly), and $z \pm \alpha$ (CORDIC iteration), respectively. The cross switch between the registers "Y_1" and "X_2" is needed to exchange data when the CORDIC SVD module switches from the pre-butterfly operations into the CORDIC iterations or from the CORDIC iterations into the post-butterfly operations, respectively. Then, we see from Fig. 3 that the output data pairs of the pre-butterflys are (p_1, p_2) and (q_1, q_2), while the desired input data pairs for the CORDIC iterations are (p_1, q_1) and (p_2, q_2), respectively. So, p_2 and q_1 have to be exchanged.

6. Comparisons. We now compare the new CORDIC SVD module with those proposed by Cavallaro and Luk [4] and Delosme [5]. Let A_{csvd} and T_{csvd} denote the area and time complexity of a CORDIC SVD module and A_{cordic} and T_{cordic} those of a CORDIC processor, respectively. Cavallaro and Luk have shown that their most efficient parallel diagonalization method requires $A_{\text{csvd}} \approx 2A_{\text{cordic}}$ and $T_{\text{csvd}} \approx 3T_{\text{cordic}}$ for the diagonal PEs and $A_{\text{csvd}} \approx 2A_{\text{cordic}}$ and $T_{\text{csvd}} \approx 2T_{\text{cordic}}$ for the off-diagonal PEs. By using the TPR method, we require $T_{\text{csvd}} \approx 2A_{\text{cordic}}$ and $T_{\text{csvd}} \approx T_{\text{cordic}}$ for all PEs. In other words, having approximately the same hardware complexity, the computation time is reduced by more than 50 percent.

A comparison to Delosme's method is more difficult because he follows a quite different approach. Therefore, only rough performance estimates are given here. In our method, we compute the rotation angles explicitly. After these computations have been completed in the diagonal PEs, the angles propagate to the off-diagonal ones. We assume that the propagation from one PE to its neighbors takes one cycle T_{cycle}, the time required for computing one CORDIC iteration. This implies local communications without broadcasting data. At the beginning of the second propagation cycle, the angles reach the diagonal neighbors of the diagonal PEs which complete their computations after T_{csvd}. This means that the diagonal PEs have to wait for a delay time $T_{\text{delay}} = T_{\text{cycle}} + T_{\text{csvd}}$ before they can exchange data with their diagonal neighbors.[3] The total time elapsed between two adjacent activities at each PE is thus $T_{\text{csvd}} + T_{\text{delay}} = 2T_{\text{csvd}} + T_{\text{cycle}} \approx 2T_{\text{cordic}}$ because T_{cycle} is negligible with respect to $T_{\text{cordic}} = L \cdot T_{\text{cycle}}$.

Delosme does not compute the rotation angles explicitly. He rather calculates encodings of the angles, i.e., sequences of signs ± 1, and sends them to the off-diagonal PEs. This enables overlap of diagonal and off-diagonal rotations because the encoding signs are recursively obtained and become available before the completion of diagonal operations. Accordingly, no delay time is required ($T_{\text{delay}} = 0$), provided that the SVD array size (the half of the matrix size) is smaller than the number of CORDIC scaling iterations n_k (for details, see [5]). The drawback is, however, that the TPR method cannot be applied to the off-diagonal PEs. Four plane rotations are hence required, resulting in $T_{\text{cvsd}} = 2T_{\text{cordic}}$ for two CORDIC processors in one module. In other words, the time complexities $T_{\text{csvd}} + T_{\text{delay}}$ of both methods are nearly identical and equal $2T_{\text{cordic}}$. If, however, multiple problems are interleaved, the fraction of idle time that is 50 percent in our case can be reduced to almost zero. In such a situation, our method provides the double speed compared with Delosme's one.

[3] If the propagation time is assumed to be T_{csvd}, we get the well-known result $T_{\text{delay}} = 2T_{\text{csvd}}$ given by Brent and Luk [2].

In terms of area complexity, both CORDIC SVD modules contain two CORDIC processors. Our module consists of essentially six adders, four barrel shifters, and one ROM table containing n angle values. Delosme's architecture requires four carry-save adders, four adders, and eight barrel shifters. So, as a rough estimate, both SVD modules have the same order of area complexities.

Perhaps the most important advantage of Delosme's approach is the 2-bit wide horizontal and vertical data connections for sending angle encodings serially rather than sending the full angle values in parallel. The prices are the upper bound of the SVD array size depending on the number of CORDIC scaling iterations, the relatively complicated timing, and a nonregular CORDIC architecture design. We also mention that while Delosme's method presumes a CORDIC implementation, the TPR method is applicable to other computing architectures.

7. Other applications of the TPR method. Another advantage of the TPR method seems to be the relatively wide range of applications. We indicate some of them in the following.

For the SVD of a rectangular matrix, a well-known method is first to triangularize the matrix by QR decomposition and then to apply the Jacobi SVD procedure to the triangular factor. Luk [22] has shown that both steps can be implemented by one triangular systolic array. Each PE contains a 2×2 submatrix. It applies two plane rotations (through the same angle) to the two column vectors at the QR step and a two-sided rotation at the SVD step. For computing the SVD step, the PE can be realized by the CORDIC SVD module, as before. On the other side, the two CORDIC processors contained in the module are also appropriate to perform the two-plane rotations of the QR step. The CORDIC SVD module presented in this paper thus provides a suitable PE for Luk's triangular SVD array.

Stewart [23] has proposed a square systolic array for computing the Schur decomposition (SD) of a non-Hermitian matrix which, for example, is useful for evaluating the eigenvalues of the matrix. His approach is similar to the Jacobi SVD method. It is based on applying a sequence of two-sided rotations to 2×2 submatrices, where the left and right rotation angles are identical to make the diagonal submatrices upper triangular. While the diagonal PEs perform operations different from those in SVD, the off-diagonal PEs have exactly the same computational task as in SVD computing. Therefore, the CORDIC SVD module can also be used in Stewart's SD array.

Even in sequential computations on a uniprocessor system, one can still apply the TPR method to reduce the computational complexity of two-sided rotations.

8. Conclusion. We have investigated a novel algorithm for computing two-sided rotations requiring only two plane rotations and a few additions. This results in significantly reduced computations of various SVD and SD methods. For parallel implementations, we have presented a unified CORDIC SVD module for implementing all PEs of the SVD array given by Brent and Luk. This leads to a homogeneous array architecture that is simpler in hardware and offers twice the computational speed of that of Cavallaro and Luk. Moreover, we have pointed out that the same CORDIC SVD module can be efficiently used in other array architectures, such as Luk's triangular SVD array and Stewart's SD array.

Acknowledgments. The authors thank the referees and the editor for their valuable comments on this paper.

REFERENCES

[1] J. M. SPEISER, *Signal processing computational needs*, in Proc. SPIE Advanced Algorithms and Architectures for Signal Processing I, Vol. 696, Society of Photo-optical Instrumentation Engineers, 1986, pp. 2–6.

[2] R. P. BRENT AND F. T. LUK, *The solution of singular value and symmetric eigenvalue problems on multiprocessor arrays*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 69–84.

[3] R. P. BRENT, F. T. LUK, AND C. F. VAN LOAN, *Computation of the singular value decomposition using mesh-connected processors*, J. VLSI Comput. Syst., 1 (1985), pp. 242–270.

[4] J. R. CAVALLARO AND F. T. LUK, *Architectures for a* CORDIC SVD *processor*, Proc. SPIE Real-Time Signal Processing IX, Vol. 698, Society of Photo-optical Instrumentation Engineers, 1986, pp. 45–53.

[5] J. M. DELOSME, *A processor for two-dimensional symmetric eigenvalue and singular value arrays*, in Proc. 21st Asilomar Conference on Circuits, Systems and Computers, November 1987.

[6] ——, CORDIC *algorithms: Theory and extensions*, Proc. SPIE Advanced Algorithms and Architectures for Signal Processing IV, Vol. 1152, Society of Photo-optical Instrumentation Engineers, August 1989, pp. 131–145.

[7] C. G. J. JACOBI, *Über ein leichtes Verfahren die in der Theorie der Säkularstörungen vorkommenden Gleichungen numerisch aufzulösen*, J. Reine Angew. Math., 30 (1846), pp. 51–95.

[8] E. G. KOGBETLIANTZ, *Solution of linear equations by diagonalization of coefficient matrices*, Quart. Appl. Math., 13 (1955), pp. 123–132.

[9] G. E. FORSYTHE AND P. HENRICI, *The cyclic Jacobi method for computing the principal values of a complex matrix*, Trans. Amer. Math. Soc., 94 (1960), pp. 1–23.

[10] A. H. SAMEH, *Solving the linear least squares problem on a linear array of processors*, in Algorithmically Specialized Parallel Computers, L. Synder et al., eds., Academic Press, New York, 1985, pp. 191–200.

[11] U. SCHWIEGELSHOHN AND L. THIELE, *A systolic algorithm for cyclic-by-rows* SVD, in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 1987, pp. 768–770.

[12] F. T. LUK AND H. PARK, *On the equivalence and convergence of parallel Jacobi* SVD *algorithms*, in Proc. SPIE Advanced Algorithms and Architectures for Signal Processing II, Vol. 826, Society of Photo-optical Instrumentation Engineers, 1987, pp. 152–159.

[13] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983, p. 44.

[14] J. E. VOLDER, *The* CORDIC *trigonometric computing technique*, IRE Trans. Electronic Comput., 8 (1959), pp. 330–334.

[15] J. S. WALTHER, *A unified algorithm for elementary functions*, in Proc. Spring Joint Computer Conference, AFIPS Press, New Jersey, 1971, pp. 379–385.

[16] J. R. CAVALLARO AND F. T. LUK, CORDIC *arithmetic for an* SVD *processor*, J. Parallel Distributed Comput., 5 (1988), pp. 271–290.

[17] G. L. HAVILAND AND A. A. TUSZYNSKI, *A* CORDIC *arithmetic processor chip*, IEEE Trans. Comput., 29 (1980), pp. 68–79.

[18] H. M. AHMED, *Signal processing algorithms and architectures*, Ph.D. thesis, Department of Electrical Engineering, Stanford University, Stanford, CA, December 1981.

[19] J. M. DELOSME, VLSI *implementation of rotations in pseudo-Euclidean spaces*, in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Boston, 1983, pp. 927–930.

[20] E. F. DEPRETTERE, P. DEWILDE, AND R. UDO, *Pipelined* CORDIC *architectures for fast* VLSI *filtering and array processing*, in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, San Diego, CA, March 1984, pp. 41A.6.1–41A.6.4.

[21] D. KÖNIG AND J. F. BÖHME, *Optimizing the* CORDIC *algorithm for processors with pipeline architecture*, in Proc. EUSIPCO, North-Holland, Elsevier Science Publishers, Amsterdam, 1990, pp. 1391–1394.

[22] F. T. LUK, *A triangular processor array for computing singular values*, Linear Algebra Appl., 77 (1986), pp. 259–273.

[23] G. W. STEWART, *A Jacobi-like algorithm for computing the Schur decomposition of a non-Hermitian matrix*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 853–864.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 4, pp. 726–729, October 1991

010

ON MONOTONE LINEAR OPERATORS AND THE SPECTRAL RADIUS OF THEIR REPRESENTING MATRICES*

HARRY H. TIGELAAR†

Abstract. In this paper, linear operators on the space of $p \times p$ matrices are considered. Such linear operators can be represented by $p^2 \times p^2$ matrices. In particular, sums of Kronecker products occur as representing matrices. Let the linear operators $\mathscr{L}_S$ and $\mathscr{L}_U$ be represented by the matrices S and U, where U is of the form $U = \sum A_k \otimes \bar{A}_k$. It is shown that, in order that $\mathscr{L}_U(X) \leqq \mathscr{L}_S(X)$ for all positive-semidefinite X, it is necessary that the spectral radii of U and S satisfy the inequality $\rho(U) \leqq \rho(S)$.

Key words. linear operators, positive-semidefinite matrices, Kronecker products, spectral radius of a square matrix

AMS(MOS) subject classifications. 15A69, 15A45, 15A18

1. Introduction. Let $\mathbb{C}^{p^*p}$ denote the p^2-dimensional space of all complex $p \times p$ matrices and $\mathbb{C}^N$ the N-dimensional vector space of complex N-vectors. With every $A \in \mathbb{C}^{p^*p}$ we can associate a vector in $\mathbb{C}^N$ where $N = p^2$, by considering vec (A). The vec-operation is a linear one-to-one transformation, but when special matrix properties are involved, it is not always adequate for interpreting a $p \times p$ matrix as a p^2-vector. In particular, it is difficult to translate a property like positive-definiteness into terms of the space $\mathbb{C}^N$. In this paper, however, we consider linear operators from $\mathbb{C}^{p^*p}$ onto itself. In particular, we are interested in linear operators that leave specific matrix properties as positive-(semi)definiteness invariant. It turns out that in that case it makes sense to analyze the corresponding linear operators from $\mathbb{C}^N$ onto itself, which are represented by $p^2 \times p^2$ matrices. Let $\mathscr{L}_S$ denote the linear operator corresponding to the $p^2 \times p^2$ matrix S. Then we have the basic relation

$$\text{vec}\,(\mathscr{L}_S(X)) = S\,\text{vec}\,(X), \qquad X \in \mathbb{C}^{p^*p}. \tag{1.1}$$

In calculating the matrix S for a given linear operator, Kronecker products of matrices are frequently encountered. Therefore, we briefly outline their properties and the relation to the vec-operator. When A and B are arbitrary matrices and $(A)_{ij}$ denotes the (i, j)th element of A, then the Kronecker product $A \otimes B$ is defined as the (partitioned) matrix, which is obtained by replacing $(A)_{ij}$ by the matrix $(A)_{ij}B$. When a partition of a matrix S in equal-sized blocks is given, it is convenient to denote the (r, s)th element of the (i, j)th block by $(S)_{rs;ij}$. Thus when $S = A \otimes B$, we have $(S)_{rs;ij} = (A)_{ij}(B)_{rs}$. In this paper we only deal with $p^2 \times p^2$ matrices, partitioned into $p \times p$ blocks. For details on Kronecker products we refer to [1]. We shall list here some properties used in the sequel:

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD), \tag{1.2}$$

$$\text{vec}\,(ABC) = (C^T \otimes A)\,\text{vec}\,(B). \tag{1.3}$$

Here C^T denotes the transpose of the matrix C (when transposition is combined with complex conjugation, we write C^*).

Let $\rho(S)$ denote the *spectral radius* of the square matrix S, i.e., the maximum of the absolute values of the eigenvalues of S. Then we have, for square matrices A and B,

$$\rho(A \otimes B) = \rho(A)\rho(B), \tag{1.4}$$

* Received by the editors March 7, 1990; accepted for publication (in revised form) August 17, 1990.
† Tilburg University B831, Postbox 90153, 5000 LE Tilburg, the Netherlands.

which follows easily from the fact that the eigenvalues of $A \otimes B$ are precisely the products of the eigenvalues of A and B.

The following lemma, which is not easily found in the literature, is a combination of known results. It relates the spectral radius of A to $\|A^n\|$, where $\|\cdot\|$ may be any of the following three natural matrix norms:

$$(1.5) \qquad \|A\|_2 = [\rho(A^*A)]^{1/2}, \qquad \|A\|_1 = \max_j \sum_i |(A)_{ij}|, \qquad \|A\|_\infty = \|A^T\|_1.$$

LEMMA 1. *For every $A \in \mathbb{C}^{p^* p}$, there exists a constant c such that*

$$(\rho(A))^n \leqq \|A^n\| \leqq cn^{p-1}(\rho(A))^n, \qquad n \in \mathbb{N}.$$

Proof. The first inequality is well known (and holds for all natural matrix norms), so we only prove the second. From the Jordan decomposition theorem it follows that A can be written in the form

$$A = H(\Lambda + N)H^{-1},$$

where $N^p = 0$ and Λ is a diagonal matrix of eigenvalues of A that commutes with N. Hence

$$(1.6) \qquad A^n = H\left(\sum_{k=0}^{p-1} \binom{n}{k} \Lambda^{n-k} N^k\right) H^{-1}.$$

Since for any of the three matrix norms we have $\|\Lambda\| = \rho(A)$, it follows from (1.6) that

$$\|A^n\| \leqq \|H\| \, \|H^{-1}\| \sum_{k=0}^{p-1} \binom{n}{k} (\rho(A))^{n-k} \|N\|^k$$

$$\leqq \|H\| \, \|H^{-1}\| (\|N\|^{p-1} + 1)(\rho(A)^{-p+1} + 1) n^{p-1} (\rho(A))^n,$$

where the last inequality follows from $x^a \leqq x^b + 1$ for $0 \leqq a \leqq b$ and $x \geqq 0$. Hence it follows that

$$\|A^n\| \leqq cn^{p-1}(\rho(A))^n,$$

where c does not depend on n. □

From the lemma follows the result that the matrix power series $\sum A^k$ converges absolutely (elementwise) if and only if $\rho(A) < 1$. This is a special case of a known result on a more general power series. See, e.g., Theorem 49 of [2].

2. Monotone linear operators. For $X \in \mathbb{C}^{p^* p}$ we denote $X \geqq 0$ when X is positive semidefinite, and for X, $Y \in \mathbb{C}^{p^* p}$ we denote $X \geqq Y$ when $X - Y \geqq 0$.

DEFINITION 2.1. The linear operator $\mathscr{L} : \mathbb{C}^{p^* p} \to \mathbb{C}^{p^* p}$ is *monotone* when, for all $X \geqq 0$, we have $\mathscr{L}(X) \geqq 0$.

From this definition, it follows immediately that for a monotone linear operator $\mathscr{L}$, the inequality $X \geqq Y$ implies $\mathscr{L}(X) \geqq \mathscr{L}(Y)$; hence the partial ordering in $\mathbb{C}^{p^* p}$ is invariant.

Examples of monotone linear operators are easily obtained. Let $A \in \mathbb{C}^{p^* p}$ be arbitrary and $S = \bar{A} \otimes A$, where $\bar{A}$ denotes the complex conjugate of the matrix A. Then the linear operator $\mathscr{L}_S$ defined by (1.1) is monotone since for $X \geqq 0$ we have $\mathscr{L}_S(X) = AXA^* \geqq 0$. More generally, S may be of the form

$$(2.1) \qquad S = \sum_{k=1}^{m} (\bar{A}_k \otimes A_k), \qquad A_k \in \mathbb{C}^{p^* p}.$$

DEFINITION 2.2. The monotone linear operator $\mathscr{L}_S$ is said to be of the *Kronecker-type* when S is of the form (2.1).

The following identity is a useful tool when powers of sums of matrices must be calculated. Let $Q_i \in \mathbb{C}^{p^* p}$ for $i = 1, 2, \cdots, m$. Then we have

$$\left(\sum_{i=1}^{m} Q_i\right)^n = \sum_{(m_j)} \prod_{j=1}^{n} Q_{m_j}, \qquad n \in \mathbb{N}, \tag{2.2}$$

where the summation on the right-hand side is over all m^n sequences $(m_j)_{j=1}^n$, $m_j \in \{1, \cdots, m\}$. It is used twice in the proof of the following lemma.

LEMMA 2. *Let S be given by* (2.1). *Then the elements of S^n satisfy the following inequality*:

$$|(S^n)_{ij,rs}| \leqq \tfrac{1}{2}[(S^n)_{ij,ij} + (S^n)_{rs,rs}]. \tag{2.3}$$

Proof. By (2.2) we have

$$|(S^n)_{ij,rs}| = \left|\left(\left(\sum_{k=1}^{m} \bar{A}_k \otimes A_k\right)^n\right)_{ij,rs}\right| = \left|\left(\sum_{(m_k)} \left(\prod_{k=1}^{n} \bar{A}_{m_k} \otimes A_{m_k}\right)_{ij,rs}\right)\right|.$$

Using (1.2) the last expression equals

$$\begin{aligned}
\left|\left(\sum_{(m_k)} \left(\prod_{k=1}^{n} \bar{A}_{m_k}\right) \otimes \left(\prod_{k=1}^{n} A_{m_k}\right)\right)_{ij,rs}\right| &= \left|\sum_{(m_k)} \left(\overline{\prod_{k=1}^{n} A_{m_k}}\right)_{ij} \left(\prod_{k=1}^{n} A_{m_k}\right)_{rs}\right| \\
&\leqq \sum_{(m_k)} \left|\left(\overline{\prod_{k=1}^{n} A_{m_k}}\right)_{ij} \left(\prod_{k=1}^{n} A_{m_k}\right)_{rs}\right| \\
&\leqq \frac{1}{2} \sum_{(m_k)} \left(\left|\left(\prod_{k=1}^{n} A_{m_k}\right)_{ij}\right|^2 + \left|\left(\prod_{k=1}^{n} A_{m_k}\right)_{rs}\right|^2\right) \\
&= \tfrac{1}{2}[(S^n)_{ij,ij} + (S^n)_{rs,rs}],
\end{aligned}$$

where the last equality follows from (2.2) in reverse direction. This proves the lemma. □

DEFINITION 2.3. The monotone linear operator $\mathscr{L}_S$ *dominates* the monotone linear operator $\mathscr{L}_U$, denoted by $\mathscr{L}_S \geqq \mathscr{L}_U$, when for all $X \geqq 0$, $X \in \mathbb{C}^{p^* p}$, we have $\mathscr{L}_S(X) \geqq \mathscr{L}_U(X)$.

Let $\mathscr{L}_S^n$ denote the linear operator defined by

$$\mathscr{L}_S^1 = \mathscr{L}_S \quad \text{and} \quad \mathscr{L}_S^n = \mathscr{L}_S \mathscr{L}_S^{n-1} \quad \text{for } n = 2, 3, \cdots.$$

From (1.1) it follows easily that $\mathscr{L}_S^n = \mathscr{L}_T$ for $T = S^n$.

LEMMA 3. *When $\mathscr{L}_S$ and $\mathscr{L}_U$ are monotone linear operators, then $\mathscr{L}_U \leqq \mathscr{L}_S$ implies $\mathscr{L}_U^n \leqq \mathscr{L}_S^n$, $n \in \mathbb{N}$.*

Proof. By induction it follows that $\mathscr{L}_S^n$ is monotone for all n. Hence for $X \geqq 0$ we have $\mathscr{L}_S^n(X) \geqq 0$. Supposing that $\mathscr{L}_S^n$ dominates $\mathscr{L}_U^n$, we obtain

$$\mathscr{L}_S^{n+1}(X) \geqq \mathscr{L}_S \mathscr{L}_U^n(X) \geqq \mathscr{L}_U \mathscr{L}_U^n(X) = \mathscr{L}_U^{n+1}(X),$$

and so, by induction, the lemma is proved. □

LEMMA 4. $\mathscr{L}_U \leqq \mathscr{L}_S \Rightarrow (U)_{ij,ij} \leqq (S)_{ij,ij}$.

Proof. Let e_i denote the (i)th unit vector in $\mathbb{C}^p$. Put $X = e_j e_j^T$. Then $X \geqq 0$. Furthermore, we have

$$(U)_{ij,ij} = (e_i \otimes e_i)^T U (e_j \otimes e_j) = e_i^T \mathscr{L}_U(X) e_i$$

and a similar expression for $(S)_{ij,ij}$. The result then follows easily. □

We are now ready to state and prove the main theorem of this paper. It relates dominance of certain monotone linear operators on $\mathbb{C}^{p^*p}$ to inequalities between the spectral radii of their representing matrices on $\mathbb{C}^N$.

THEOREM. *If $\mathscr{L}_K$ is of the Kronecker type, and is dominated by $\mathscr{L}_S$, then $\rho(K) \leqq \rho(S)$.*

Proof. Let

$$K = \sum_{k=1}^{m} (A_k \otimes \bar{A}_k) \quad \text{and} \quad r = \rho(S).$$

Let $\varepsilon > 0$ be arbitrary and put

$$U = (r+\varepsilon)^{-1}K, \qquad V = (r+\varepsilon)^{-1}S.$$

From $\mathscr{L}_K \leqq \mathscr{L}_S$ it follows that $\mathscr{L}_U \leqq \mathscr{L}_V$ and so by Lemma 3 we have $\mathscr{L}_U^n \leqq \mathscr{L}_V^n$ for all $n \in \mathbb{N}$. Hence, by Lemma 4 we obtain

$$(U^n)_{ij,ij} \leqq (V^n)_{ij,ij}.$$

Using Lemma 2 it follows that

$$|(U^n)_{ij,rs}| \leqq \tfrac{1}{2}[(V^n)_{ij,ij} + (V^n)_{rs,rs}]. \tag{2.4}$$

Since $\rho(V) = r(r+\varepsilon)^{-1} < 1$, it follows from (2.4) that the series $\sum_n U^n$ converges absolutely (elementwise). But then we must have $\rho(U) < 1$, or equivalently $\rho(K) < r + \varepsilon$.

Since ε was arbitrary it follows that $\rho(K) \leqq r$. □

3. Application. As an application of the theory developed in the previous sections, we shall prove the following interesting result. Let A and B be real $p \times p$-matrices satisfying the condition

$$\rho(A \otimes A + B \otimes B) < 1. \tag{3.1}$$

Then both $I - A$ and $I - B$ are nonsingular. (It is not essential that A and B are real.)

The proof of the statement is simple. Put $K = A \otimes A$ and $S = A \otimes A + B \otimes B$. Then, clearly, we have $\mathscr{L}_K \leqq \mathscr{L}_S$ and so Theorem 1 implies $\rho(K) = \rho(A \otimes A) < 1$. But this is equivalent to $\rho(A) < 1$, which implies that $I - A$ is nonsingular. Because of symmetry, $I - B$ must also be nonsingular.

The statement can of course be generalized for more than two matrices. It is not easy to see how such results can be proved without Theorem 1. Conditions like (3.1) arise in a probabilistic context in the theory of stationary *bilinear processes* (see [4]).

REFERENCES

[1] A. GRAHAM, *Kronecker Products and Matrix Calculus with Applications*, John Wiley, New York, 1981.

[2] C. C. MACDUFFEE, *The Theory of Matrices*, Chelsea, New York, 1956.

[3] M. MARCUS AND H. MINC, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, 1964.

[4] M. B. RAO, T. S. RAO, AND A. M. WALKER, *On the existence of some bilinear time series models*, J. Time Ser. Anal., 4 (1983), pp. 95–110.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 4, pp. 730–746, October 1991

011

ON THE IDENTIFICATION OF LOCAL MINIMIZERS IN INERTIA-CONTROLLING METHODS FOR QUADRATIC PROGRAMMING*

A. L. FORSGREN†, P. E. GILL‡, AND W. MURRAY§

Abstract. The verification of a local minimizer of a general (i.e., nonconvex) quadratic program is in general an NP-hard problem. The difficulty concerns the optimality of certain points (which we call *dead points*) at which the first-order necessary conditions for optimality are satisfied, but strict complementarity does not hold. Inertia-controlling quadratic programming (ICQP) methods form an important class of methods for solving general quadratic programs. We derive a computational scheme for proceeding at a dead point that is appropriate for a general ICQP method.

Key words. quadratic programming, local minimizer, NP-hardness, optimality conditions

AMS(MOS) subject classifications. 49D40, 65K05, 90C20

1. Introduction.

The general quadratic programming (QP) problem is to find a local minimizer of a quadratic function subject to linear inequality constraints. The form of QP problem to be considered in this paper is

$$\begin{aligned} &\underset{x\in\Re^n}{\text{minimize}} && \varphi(x) = c^T x + \tfrac{1}{2} x^T H x \\ &\text{subject to} && \mathcal{A}x \geq \beta, \end{aligned} \tag{1.1}$$

where H, the *Hessian matrix*, is symmetric, and $\mathcal{A}$ is an $m_L \times n$ matrix. Of particular interest is the nonconvex case, which is characterized by an arbitrary distribution of positive, negative, and zero eigenvalues in H.

This paper will focus on a specific class of methods for general quadratic programming. *Inertia-controlling quadratic programming* (ICQP) *methods* use a linearly independent subset of the constraints, known as the *working set*, to define a search direction and multiplier estimates. A distinctive feature of ICQP methods is that constraint deletions are restricted in order to control the inertia of the *reduced Hessian matrix*, which is never permitted to have more than one nonpositive eigenvalue. The first ICQP method was proposed by Fletcher [FLE71]; other methods include those of Gill et al. [GMSW84] and Gould [GOU86].

For any nonconvex quadratic program there may exist certain *dead points* at which all quadratic programming methods will find it difficult to proceed (see §2.5 for a precise definition of a dead point). The difficulty arises because the verification of such a point as a local minimizer of (1.1) is an NP-hard problem—see Murty and Kabadi [MK87] and Pardalos and Schnitger [PS88]. However, even if lower values of φ exist in the neighborhood of a dead point, it may be necessary to delete many constraints simultaneously to find a direction of improvement. Since existing ICQP methods are only able to delete one constraint at a time, they may be unable to proceed from a dead point. This behavior may be contrasted with that of the simplex

* Received by the editors September 13, 1989; accepted for publication (in revised form) August 24, 1990. This research was partially supported by National Science Foundation grant ECS-8715153, Office of Naval Research grant N00014-90-J-1242, and the Göran Gustafsson Foundation.

† Optimization and Systems Theory, Department of Mathematics, The Royal Institute of Technology, S-100 44 Stockholm, Sweden (andersf@math.kth.se).

‡ Department of Mathematics, University of California, San Diego, La Jolla, California 92093 (peg@optimal.ucsd.edu).

§ Department of Operations Research, Stanford University, Stanford, California 94305 (walter@sol-walter.stanford.edu).

method at a degenerate vertex. The simplex method is able to keep iterating at a degenerate vertex, but a large number of iterations may be performed, during which the working set changes but the vertex remains the same. By contrast, existing ICQP methods usually terminate at a dead point, irrespective of whether or not the dead point is a local minimizer.

If progress is to be made at a dead point, a scheme must be devised that allows the possibility of more than one constraint being deleted from the working set at one time. Moreover, it must be possible to implement the scheme within the general framework of an ICQP method. Such a scheme is presented in this paper, together with its computational and theoretical properties. We show that the behavior of the method at a dead point is similar to that of the simplex method at a degenerate vertex: the algorithm is able to proceed, but there exists the danger of cycling. Our method is not guaranteed to prevent cycling at a dead point. However, since the verification of optimality is NP-hard, *no* known scheme can be guaranteed to make progress in a reasonable amount of computational effort.

In an ICQP method, it is sometimes necessary to impose *artificial constraints* to ensure that the reduced Hessian is positive definite at the starting point. (The role of artificial constraints is discussed further in §2.4.) Unfortunately, artificial constraints may introduce dead points that are not present in the original problem. In §5 we give a computational scheme that is able to treat artificial constraints without difficulty.

In order to describe the new scheme, we first review results on necessary and sufficient conditions for optimality in general quadratic programming. For a discussion of these conditions, see, for example, Majthay [MAJ71], Mangasarian [MAN80], Contesse [CON80], or Borwein [BOR82]. The results presented here allow the presence of artificial constraints in the working set.

2. Background.

2.1. Notation and terminology. Throughout the paper, x will denote the feasible point of (1.1) to be examined. It will be assumed that m constraints are in the working set at x, and that the $m \times n$ working-set matrix and associated m-vector of right-hand sides are A and b. The vector $g(x)$ is the gradient of φ at x, i.e., $g(x) = c + Hx$ (we shall omit the argument x when the meaning is clear). The columns of the matrix Z form a basis for the null space of A; the reduced gradient and reduced Hessian of φ with respect to A are then $Z^T g(x)$ and $Z^T H Z$. The vector e_i will be used to denote the ith unit vector of the appropriate dimension.

A vector p is said to be a *descent direction* if $g^T p < 0$; a *direction of positive curvature* if $p^T H p > 0$; a *direction of negative curvature* if $p^T H p < 0$; a *direction of zero curvature* if $p^T H p = 0$; and a *feasible direction* if $Ap \geq 0$. A matrix D is said to be *copositive* if $v^T D v \geq 0$ for all $v \geq 0$. A constraint $a_i^T x \geq \beta_i$ is said to be *active* at x if $a_i^T x = \beta_i$; *inactive* at x if $a_i^T x > \beta_i$; and *violated* at x if $a_i^T x < \beta_i$.

2.2. Assumptions. The following assumptions are used:

A1. The objective function, φ, is bounded from below in the feasible region.

A2. All constraints active at x are in the working set.

A3. The working-set matrix A has full row rank.

A4. The point x satisfies the first-order necessary conditions for optimality, i.e., there exists a nonnegative Lagrange multiplier vector μ such that x and μ satisfy the Karush–Kuhn–Tucker (KKT) system

$$\begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ -\mu \end{pmatrix} = \begin{pmatrix} -c \\ b \end{pmatrix}. \tag{2.1}$$

A5. The reduced Hessian matrix is positive definite.

2.3. The inertia of a matrix. Let M be any symmetric matrix. We denote by $i_p(M)$, $i_n(M)$, and $i_z(M)$, respectively, the (nonnegative) numbers of positive, negative, and zero eigenvalues of M. The *inertia* of M—denoted by $\mathrm{In}\,(M)$—is the associated integer triple (i_p, i_n, i_z). The following lemma states an important relationship between the inertia of the KKT-*matrix*

$$K = \begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix}$$

and the reduced Hessian.

LEMMA 2.1. *Given assumptions* A3 *and* A5, *the inertia of the* KKT *matrix* K *is* $(n, m, 0)$.

Proof. See Gould [GOU85, Lem. 3.4]. □

Lemma 2.1 implies that K is nonsingular, so that the Lagrange multipliers in (2.1) are unique.

2.4. Inertia-controlling methods for quadratic programming. Associated with each iteration of an ICQP method is a linearly independent subset of the constraints known as the working set. The working set at the initial point x_0 must be chosen so that the reduced Hessian is positive definite. Thereafter, the working set changes by only one constraint at each iteration and the reduced Hessian is never permitted to have more than one nonpositive eigenvalue.

ICQP methods depend critically on a procedure for finding an initial working set with an associated positive-definite reduced Hessian. In order to ensure that the reduced Hessian is positive definite, the initial working set may need to include "artificial" constraints that are not specified in the original problem. (The original constraints of the problem will be referred to as *regular constraints*.) The only requirement for an artificial constraint is that it be linearly independent of the other constraints in the working set. An artificial constraint does not restrict the feasible region, since the direction of the inequality is not defined. As soon as an artificial constraint is removed from the working set, it is eliminated from the problem.

The type of artificial constraint depends on the form of the particular QP method. For example, the initial working set will often vary with the method used to solve the KKT system. A simple example of a problem requiring artificial constraints is given in §5. We emphasize that artificial constraints are not part of the original problem, but are an artifact of the solution method.

Once a constrained minimizer has been found, an ICQP algorithm deletes *one* constraint from the working set and finds either a feasible descent direction or a feasible direction of negative curvature. Constraint deletion is permitted only when the reduced Hessian matrix is positive definite.

All ICQP methods generate search vectors and multipliers that satisfy the KKT equations. However, since the equations may be solved either implicitly or explicitly, one ICQP algorithm may appear to be very different from another. In this paper, only the form of equations to be solved is stated. For a discussion on the relationship between different ICQP methods, see Gill et al. [GMSW88].

Dead points associated with only regular constraints are discussed first. In §5, we consider the case when artificial constraints are present.

2.5. Dead points. If A contains only regular constraints, a *dead point* is defined to be a point that satisfies assumptions A4 and A5 with one or more zero components

in the Lagrange multiplier vector μ. We emphasize that such a point may not be a local minimizer.

Since a dead point satisfies the first-order necessary conditions for optimality, there exists no feasible descent direction. Therefore, it is necessary to find a feasible direction of negative curvature if an ICQP method is to proceed to find a local minimizer. Unfortunately, it may be impossible to compute a feasible direction of negative curvature by deleting only one constraint at a time, as can be seen from the following problem.

$$\begin{aligned} &\underset{x\in\Re^2}{\text{minimize}} && -x_1x_2 \\ &\text{subject to} && 0 \le x_1 \le 1, \qquad 0 \le x_2 \le 1. \end{aligned} \tag{2.2}$$

If the starting point is the origin, and both active constraints are in the working set, assumptions A1–A5 are satisfied. However, if either of the constraints is deleted from the working set, the resulting reduced Hessian is positive semidefinite and singular. Therefore, no feasible direction of negative curvature may be computed by deleting only one constraint. Since constraint deletion is permitted only when the reduced Hessian is positive definite, no more than one constraint may be deleted from the working set and an ICQP method cannot continue at this point, even though the origin is not a local minimizer.

In this situation—where neither a feasible descent direction nor a feasible direction of negative curvature may be found by deleting only one constraint—it is necessary to develop a scheme for deleting more than one constraint simultaneously if an ICQP method is to proceed.

2.6. Optimality conditions. In this section we review the necessary and sufficient conditions for x to be a local minimizer under assumptions A1–A5. It will be necessary to distinguish between constraints with positive and zero multipliers. Without loss of generality we may assume that the rows of A are ordered such that

$$A = \begin{pmatrix} A_+ \\ A_0 \end{pmatrix},$$

where A_+ corresponds to rows with positive Lagrange multipliers and A_0 corresponds to rows with zero Lagrange multipliers. Let m_+ denote the number of rows in A_+, and let m_0 denote the number of rows in A_0. Also, let μ_+ denote the vector containing the m_+ positive components of μ.

The following two necessary and sufficient conditions for x being a local minimizer for (1.1) when assumption A2 holds are given by Majthay [MAJ71] and Contesse [CON80].

C1. *The point x satisfies the first-order necessary conditions for optimality, i.e., there exists a nonnegative Lagrange multiplier vector μ, such that x and μ satisfy the KKT equations*

$$\begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x \\ -\mu \end{pmatrix} = \begin{pmatrix} -c \\ b \end{pmatrix}.$$

C2. *It holds that $d^THd \ge 0$ for all d such that $A_+d = 0$ and $A_0d \ge 0$.*

In his proof, Contesse derives an alternative formulation of condition C2 involving the set of generators for the finite cone

$$\{p \mid A_+p = 0,\ A_0p \ge 0\}.$$

This formulation is described in Theorem 3.6 below. For completeness, Contesse's proof is included with a notation relevant to our assumptions.

3. A proof of the optimality conditions. Let Y_+ denote the $n \times m_+$ matrix whose jth column y_{+j} is defined to be the unique vector satisfying the equation

$$\begin{pmatrix} H & A_+^T & A_0^T \\ A_+ & 0 & 0 \\ A_0 & 0 & 0 \end{pmatrix} \begin{pmatrix} y_{+j} \\ -\rho_j \\ -\eta_j \end{pmatrix} = \begin{pmatrix} 0 \\ e_j \\ 0 \end{pmatrix}, \tag{3.1}$$

and let Y_0 denote the $n \times m_0$ matrix whose jth column y_{0j} is defined to be the unique vector such that

$$\begin{pmatrix} H & A_+^T & A_0^T \\ A_+ & 0 & 0 \\ A_0 & 0 & 0 \end{pmatrix} \begin{pmatrix} y_{0j} \\ -\lambda_j \\ -\theta_j \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ e_j \end{pmatrix}. \tag{3.2}$$

Equations (3.1) and (3.2) imply that the computation of y_{+j} and y_{0j} involves solving the KKT equations with a unit right-hand side. For a detailed discussion of the properties of the KKT equations in this context, see Gould [GOU86, Thm. 2.3].

Given Y_+ and Y_0, let M denote the $n \times n$ matrix $M = (Z \quad Y_+ \quad Y_0)$.

LEMMA 3.1. *The matrix M is nonsingular.*

Proof. It is enough to show that the columns of M are linearly independent. Assume that

$$Mv = Zv_Z + Y_+v_+ + Y_0v_0 = 0.$$

Successive premultiplication of Mv by A_+ and A_0 gives $v_+ = 0$ and $v_0 = 0$. Since the columns of Z are independent, it follows that $v_Z = 0$. □

LEMMA 3.2. *The sets*

$$\{p \mid Ap \geq 0\} \quad \textit{and} \quad \{p \mid p = Zv_Z + Y_+v_+ + Y_0v_0,\ v_+ \geq 0,\ v_0 \geq 0\}$$

are identical.

Proof. From Lemma 3.1 it follows that the columns of Z, Y_+, and Y_0 span $\Re^n$. Consequently, any p in $\Re^n$ may be written in the form $p = Zv_Z + Y_+v_+ + Y_0v_0$, for some suitably dimensioned vectors v_Z, v_+, and v_0. Premultiplication of p by A yields

$$Ap = \begin{pmatrix} A_+ \\ A_0 \end{pmatrix} p = \begin{pmatrix} v_+ \\ v_0 \end{pmatrix}.$$

Hence, the vector Ap is nonnegative if and only if v_+ and v_0 are nonnegative. □

We now show that verification of the optimality of x is equivalent to finding a local solution of a special quadratic program.

LEMMA 3.3. *The vector x is a local minimizer of* (1.1) *if and only if $p = 0$ is a local minimizer of the quadratic program*

$$\begin{aligned} &\underset{p \in \Re^n}{\text{minimize}} && g^Tp + \tfrac{1}{2}p^THp \\ &\text{subject to} && Ap \geq 0. \end{aligned} \tag{3.3}$$

Proof. The Taylor-series expansion of φ gives $g^Tp + \frac{1}{2}p^THp = \varphi(x+p) - \varphi(x)$. The vector Ap is nonnegative if and only if $A(x+p) \geq Ax$. Since every active constraint

is included in A, the point x will not be a local minimizer of (1.1) if and only if there exists an infinite sequence $\{x^k\}_{k=1}^{\infty}$ converging to x such that $Ax^k \geq b$ and $\varphi(x^k) < \varphi(x)$. We need consider only those constraints in the working set because assumption A2 guarantees that if $\{x^k\}_{k=1}^{\infty}$ converges to x, all other constraints will be satisfied for k sufficiently large. Similarly, the zero vector will not be a local minimizer of (3.3) if and only if there exists an infinite sequence $\{p^k\}_{k=1}^{\infty}$ converging to zero such that $Ap^k \geq 0$ and $g^T p^k + \frac{1}{2} p^{kT} H p^k < 0$. The proof is complete if we let $x^k = x + p^k$. □

LEMMA 3.4. *All elements of the matrices $Z^T H Y_+$ and $Z^T H Y_0$ are zero.*

Proof. Direct substitution into (3.1) yields $Z^T H y_{+j} = 0$ for $j = 1, \cdots, m_+$ and direct substitution into (3.2) yields $Z^T H y_{0j} = 0$ for $j = 1, \cdots, m_0$. □

Lemmas 3.3 and 3.4 are now combined to show that the verification of x as a local minimizer is achieved by solving the QP problem

$$\begin{aligned} \underset{v \in \Re^n}{\text{minimize}} \quad & \mu_+^T v_+ + \tfrac{1}{2} v_z^T Z^T H Z v_z + \tfrac{1}{2} v_+^T Y_+^T H Y_+ v_+ \\ & + v_+^T Y_+^T H Y_0 v_0 + \tfrac{1}{2} v_0^T Y_0^T H Y_0 v_0 \\ \text{subject to} \quad & v_+ \geq 0, \qquad v_0 \geq 0. \end{aligned} \tag{3.4}$$

LEMMA 3.5. *The vector x is a local minimizer of* (1.1) *if and only if zero is a local minimizer of* (3.4).

Proof. Problem (3.4) is derived from problem (3.3) by using the transformation

$$p = Mv = Zv_z + Y_+ v_+ + Y_0 v_0.$$

Assumption A4, equations (3.1) and (3.2), and Lemma 3.4 are used to simplify the objective function. The feasible region is obtained by using Lemma 3.2. Finally, Lemma 3.3 implies that zero is a local minimizer of (1.1) if and only if it is a local minimizer of (3.4). □

Using these results it is possible to pose the problem of determining local optimality as a copositivity problem, as the following theorem shows.

THEOREM 3.6 (Contesse [CON80]). *The point x is a local minimizer of* (1.1) *if and only if $Y_0^T H Y_0$ is copositive.*

Proof. Assume that $Y_0^T H Y_0$ is not copositive. Then there exists a nonnegative vector v_0 such that $v_0^T Y_0^T H Y_0 v_0$ is negative, and zero is not a local minimizer of (3.4). Lemma 3.5 implies that x is not a local minimizer of (1.1).

Assume that $Y_0^T H Y_0$ is copositive. If zero is not a local minimizer of (3.4), there must exist an infinite sequence $\{v^k\}_{k=1}^{\infty}$ converging to zero such that

$$\mu_+^T v_+^k + \tfrac{1}{2} v_z^{kT} Z^T H Z v_z^k + \tfrac{1}{2} v_+^{kT} Y_+^T H Y_+ v_+^k + v_+^{kT} Y_+^T H Y_0 v_0^k + \tfrac{1}{2} v_0^{kT} Y_0^T H Y_0 v_0^k < 0,$$

where v_0^k and v_+^k are nonnegative. Since $Z^T H Z$ is positive definite and $Y_0^T H Y_0$ is copositive, it must hold that

$$\mu_+^T v_+^k + \tfrac{1}{2} v_+^{kT} Y_+^T H Y_+ v_+^k + v_+^{kT} Y_+^T H Y_0 v_0^k < 0.$$

At least one component of v_+^k must be positive, since the left-hand side is zero when v_+^k is zero. Since μ_+ is a positive vector, it must have a positive least component $\mu_{\min}$, and we may write

$$\mu_{\min} e^T v_+^k + \tfrac{1}{2} v_+^{kT} Y_+^T H Y_+ v_+^k + v_+^{kT} Y_+^T H Y_0 v_0^k < 0,$$

where e is a suitably dimensioned vector with unit components. If both sides of this last equation are divided by the positive quantity $e^T v_+^k$, we obtain the inequality

$$\mu_{\min} + \frac{1}{2e^T v_+^k} v_+^{kT} Y_+^T H Y_+ v_+^k + \frac{1}{e^T v_+^k} v_+^{kT} Y_+^T H Y_0 v_0^k < 0. \tag{3.5}$$

If we now consider this inequality as k goes to infinity, we note that $\mu_{\min}$ must be nonpositive, which contradicts the assumption that μ_+ is a positive vector. Hence, the zero vector is a local minimizer of (3.4) and Lemma 3.5 implies that x is a local minimizer of (1.1). □

From this theorem, it follows that if we are able to check the $m_0 \times m_0$ matrix $Y_0^T H Y_0$ for copositivity, we are able to determine if x is a local minimizer.

4. On the copositivity of a matrix. It was shown in the previous section that the verification of optimality of a dead point x is equivalent to checking if the $m_0 \times m_0$ matrix $Y_0^T H Y_0$ is copositive. Once Y_0 has been computed, the matrix $Y_0^T H Y_0$ may be calculated by direct matrix multiplication. However, the following lemma shows that the solutions of the equation (3.2) for $j = 1, \cdots, m_0$ provide the columns of the matrix $Y_0^T H Y_0$.

LEMMA 4.1. *If θ_j satisfies* (3.2), *then $Y_0^T H Y_0 e_j = \theta_j$ for $j = 1, \cdots, m_0$.*

Proof. Direct substitution into (3.2) yields $y_{0i}^T H y_{0j} = e_i^T \theta_j$. □

Copositive matrices have been studied extensively (see, e.g., Cottle, Habetler, and Lemke [CHL70] and Pereira [PER72]). The problem of deciding if a given matrix is copositive has been shown to be NP-hard (see Murty and Kabadi [MK87] and Pardalos and Schnitger [PS88]). Therefore, no computationally tractable method for solving the general problem is known.

However, there are special situations in which a matrix may be simply checked for copositivity. Two such situations are discussed in the following lemmas.

LEMMA 4.2. *If the elements of the matrix $Y_0^T H Y_0$ are nonnegative, it is copositive.*

Proof. If $Y_0^T H Y_0$ is not copositive, there must exist a nonnegative vector v_0 such that $v_0^T Y_0^T H Y_0 v_0 < 0$. This is clearly impossible if all elements of $Y_0^T H Y_0$ are nonnegative. □

LEMMA 4.3. *If a diagonal element of $Y_0^T H Y_0$, say $y_{0i}^T H y_{0i}$, is negative, the matrix is not copositive. Moreover, the vector y_{0i} is a feasible direction of negative curvature.*

Proof. If $y_{0i}^T H y_{0i} < 0$, then clearly y_{0i} is a direction of negative curvature. Lemma 3.2 implies that y_{0i} is a feasible direction, as required. □

It is also straightforward to check for copositivity when $Y_0^T H Y_0$ is a 2×2 matrix with nonnegative diagonal elements.

LEMMA 4.4. *A 2×2 real symmetric matrix P with nonnegative diagonal elements is not copositive if and only if its determinant is negative and its off-diagonal elements are negative. Moreover, if P is not copositive, the eigenvector corresponding to the negative eigenvalue is a positive vector.*

Proof. See Cottle, Habetler, and Lemke [CHL70, Thm. 3.1]. □

As a consequence of Lemma 4.4, the following lemma is immediate.

LEMMA 4.5. *Let $Y_0^T H Y_0$ have nonnegative diagonal elements and a 2×2 principal submatrix P with negative determinant and negative off-diagonal elements. A feasible direction of negative curvature for $Y_0^T H Y_0$ is the n_0-vector whose nonzero elements are the components of the eigenvector corresponding to the negative eigenvalue of P.*

Proof. It follows from Lemma 4.4 that the eigenvector corresponding to the negative eigenvalue of P has nonnegative components. This eigenvector, when extended

by adding zeros in the remaining $(n_0 - 2)$ positions, is a feasible direction of negative curvature. □

Now we propose a scheme for the verification of local optimality based on the lemmas above. First, it is shown that artificial constraints cause no additional difficulties.

5. Artificial constraints in the working set. From the earlier discussion, it is clear that there may exist certain dead points at which the verification of local optimality is very difficult. In this section we demonstrate that this inherent difficulty need not be exacerbated by the presence of artificial constraints.

To simplify the discussion, it will be necessary to distinguish between artificial and regular constraints. Accordingly, we partition A_0 and Y_0 such that

$$A_0 = \begin{pmatrix} A_R \\ A_A \end{pmatrix} \quad \text{and} \quad Y_0 = \begin{pmatrix} Y_R & Y_A \end{pmatrix},$$

where the subscript "R" denotes regular constraints and the subscript "A" denotes artificial constraints. Let m_R denote the number of rows of A_R and let m_A denote the number of rows of A_A. Also let y_{Rj} denote the jth column of Y_R and let y_{Aj} denote the jth column of Y_A. When artificial constraints are present, we must redefine a feasible direction to be a vector p such that $A_+ p \geq 0$ and $A_R p \geq 0$ (note that the sign of $A_A p$ is not restricted).

It is also necessary to use the following slightly modified version of assumption A4:

A4′. The point x satisfies the first-order necessary conditions for optimality, i.e., there exists a Lagrange multiplier vector $\mu = (\mu_+^T \quad \mu_A^T \quad \mu_R^T)^T$, with $\mu_+ \geq 0$, $\mu_R \geq 0$, and $\mu_A = 0$, such that x and μ satisfy the KKT equations

$$\begin{pmatrix} H & A_+^T & A_R^T & A_A^T \\ A_+ & 0 & 0 & 0 \\ A_R & 0 & 0 & 0 \\ A_A & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ -\mu_+ \\ -\mu_R \\ -\mu_A \end{pmatrix} = \begin{pmatrix} -c \\ b_+ \\ b_R \\ b_A \end{pmatrix}.$$

The difference between assumptions A4 and A4′ is that the Lagrange multipliers of the artificial constraints are required to be zero. If an artificial constraint has a nonzero multiplier, it could be deleted from the working set to yield a feasible descent direction. Therefore, assumption A4′ is appropriate for x being a constrained stationary point. Consequently, a point x is said to be a dead point if it satisfies assumptions A4′ and A5.

Unfortunately, dead points may be added to the problem by imposing artificial constraints. Consider the problem

$$\begin{aligned} &\underset{x \in \Re^2}{\text{minimize}} && -x_1 x_2 \\ &\text{subject to} && -1 \leq x_1 \leq 1, \qquad -1 \leq x_2 \leq 1. \end{aligned} \tag{5.1}$$

If the starting point is $(0, 0)$, no regular constraints are active, and artificial constraints are needed to define a positive-definite reduced Hessian. If artificial bound constraints $x_1 = 0$ and $x_2 = 0$ are imposed, assumptions A1–A3, A4′, and A5 are satisfied. However, as in problem (2.2), the origin is not a local minimizer and no feasible direction of negative curvature may be obtained by deleting only one artificial bound.

It might appear that an arbitrary (unknown) number of artificial constraints must be deleted to give a feasible direction of negative curvature (if one exists). However,

we shall show below that such a direction may be computed by making only *one or two* artificial constraints leave the working set. Our approach is to use the solution of (3.1) and (3.2) to identify those constraints in the working set that may be deleted to yield a positive-definite reduced Hessian. A similar approach is given by Gould [GOU86, Thm. 2.3] and reviewed in the following lemma.

LEMMA 5.1. *If a constraint corresponding to a positive diagonal element of $Y_0^T H Y_0$ is deleted from A, the resulting reduced Hessian is positive definite.*

Proof. Let y_{0i} correspond to the deleted constraint $a_i^T x \geq \beta_i$. Lemma 3.1 implies that y_{0i} is independent of the columns of Z and it follows from (3.2) that $A_0 y_{0i} = e_i$ and $A_+ y_{0i} = 0$. Therefore, a basis for the new null space is obtained by adding the column y_{0i} to Z. Lemma 3.4 implies that $Z^T H y_{0i}$ is zero. Hence, the fact that $y_{0i}^T H y_{0i}$ is positive implies that the new reduced Hessian remains positive definite. □

In order to distinguish between artificial and regular constraints we partition $Y_0^T H Y_0$ such that

$$Y_0^T H Y_0 = \begin{pmatrix} Y_R^T H Y_R & Y_R^T H Y_A \\ Y_A^T H Y_R & Y_A^T H Y_A \end{pmatrix}.$$

It follows from Lemma 4.3 that if a diagonal element of $Y_A^T H Y_A$ is negative, a feasible direction of negative curvature can be computed. By Lemma 5.1 it follows that if a diagonal element of $Y_A^T H Y_A$ is positive, the corresponding artificial constraint can be deleted and the new reduced Hessian will be positive definite. Clearly, unless all diagonal elements of $Y_A^T H Y_A$ are zero, either a feasible direction of negative curvature can be computed or an artificial constraint can be deleted.

LEMMA 5.2. *If two diagonal elements of $Y_A^T H Y_A$, say $y_{Ai}^T H y_{Ai}$ and $y_{Aj}^T H y_{Aj}$, are zero, and $y_{Ai}^T H y_{Aj}$ is nonzero, the point x is not a local minimizer. Moreover, either $y_{Ai} - y_{Aj}$ or $y_{Ai} + y_{Aj}$ is a feasible direction of negative curvature.*

Proof. Direct calculation yields

$$(y_{Ai} + y_{Aj})^T H (y_{Ai} + y_{Aj}) = -(y_{Ai} - y_{Aj})^T H (y_{Ai} - y_{Aj}) = 2 y_{Ai}^T H y_{Aj} \neq 0.$$

Hence, either $y_{Ai} + y_{Aj}$ or $y_{Ai} - y_{Aj}$ is a direction of negative curvature. Feasibility follows from the relations $A_+(y_{Ai} \pm y_{Aj}) = 0$ and $A_R(y_{Ai} \pm y_{Aj}) = 0$. □

This lemma demonstrates that unless the matrix $Y_A^T H Y_A$ is zero, either a feasible direction of negative curvature can be computed or an artificial constraint can be deleted.

LEMMA 5.3. *If the diagonals of $Y_A^T H Y_A$ are zero, and an element of $Y_R^T H Y_A$ (say $y_{Ri}^T H y_{Aj}$) is nonzero, the point x is not a local minimizer and a feasible direction of negative curvature may be computed.*

Proof. Let p be a vector of the form $\alpha_i y_{Ri} + \alpha_j y_{Aj}$. Direct calculation yields that p is feasible if α_i is nonnegative. The quantity $p^T H p$ may be expressed as

$$p^T H p = \begin{pmatrix} \alpha_i & \alpha_j \end{pmatrix} \begin{pmatrix} y_{Ri}^T H y_{Ri} & y_{Ri}^T H y_{Aj} \\ y_{Ri}^T H y_{Aj} & 0 \end{pmatrix} \begin{pmatrix} \alpha_i \\ \alpha_j \end{pmatrix}.$$

Consider the 2×2 matrix T given by

$$T = \begin{pmatrix} y_{Ri}^T H y_{Ri} & y_{Ri}^T H y_{Aj} \\ y_{Ri}^T H y_{Aj} & 0 \end{pmatrix}.$$

Since $y_{Ri}^T H y_{Aj}$ is nonzero, T has one negative and one positive eigenvalue. It has orthogonal eigenvectors, since it is a real symmetric matrix. Hence, α_i and α_j may

be chosen so that p is the eigenvector corresponding to the negative eigenvalue, with α_i nonnegative. For those values of α_i and α_j, the vector p will be a feasible direction of negative curvature. □

Clearly, whenever a component of $Y_R^T H Y_A$ is nonzero, either an artificial constraint can be deleted or a feasible direction of negative curvature can be computed. To summarize, the following result holds when artificial constraints are present in the working set.

THEOREM 5.4. *If $Y_A^T H Y_A$ has nonpositive diagonal elements, then x is a local minimizer of* (1.1) *if and only if $Y_R^T H Y_R$ is copositive and $Y_R^T H Y_A$ and $Y_A^T H Y_A$ are zero.*

Proof. If $Y_R^T H Y_A$ or $Y_A^T H Y_A$ are nonzero, there exists a feasible direction of negative curvature and x cannot be a local minimizer.

Assume that $Y_R^T H Y_R$ is not copositive. In this case, a feasible direction of negative curvature may be computed and the local optimality of x is contradicted.

Assume that $Y_R^T H Y_R$ is copositive and $Y_R^T H Y_A$ and $Y_A^T H Y_A$ are zero. Using a similar analysis to that for the regular-constraint case, we can make the following assertions. As in Lemma 3.2, partition the vector v_0 such that

$$v_0 = \begin{pmatrix} v_R \\ v_A \end{pmatrix}$$

and replace the constraint $v_0 \geq 0$ in (3.4) by $v_R \geq 0$. If x is not a local minimizer of (1.1), there must exist an infinite sequence $\{v^k\}_{k=1}^{\infty}$ converging to zero such that

$$\mu_{\min} + \frac{1}{2e^T v_+^k} v_+^{kT} Y_+^T H Y_+ v_+^k + \frac{1}{e^T v_+^k} v_+^{kT} Y_+^T H Y_R v_R^k + \frac{1}{e^T v_+^k} v_+^{kT} Y_+^T H Y_A v_A^k < 0.$$

Again, if we consider this inequality as k goes to infinity, we obtain the required contradiction. □

Consequently, if assumptions A1–A3, A4′, and A5 hold, the artificial constraints will cause no extra problem in determining if x is a local minimizer. There remains the hard question of verifying that the matrix $Y_R^T H Y_R$ is copositive.

6. Computation of directions of negative curvature. In this section, we propose an extension to ICQP methods that will allow progress to be made at a dead point. Algorithm 6.1 provides a means of computing a direction of negative curvature by making one or two active constraints inactive. Lemma 6.1 below indicates that the algorithm will terminate with either a direction of negative curvature or the conclusion that x is a local minimizer.

LEMMA 6.1. *Algorithm* 6.1 *will terminate in at most m_0 steps. Moreover, if termination occurs without the computation of a direction of negative curvature, x is a local minimizer of* (1.1).

Proof. At each step, either the algorithm terminates or a constraint is deleted from the working set. Since there are only m_0 constraints to delete, the algorithm must stop in at most m_0 steps.

If $Y_A^T H Y_A$ has a positive diagonal element, the corresponding artificial constraint is deleted. Since this deletion will be repeated until every diagonal element of $Y_A^T H Y_A$ is nonpositive, we may assume that $Y_A^T H Y_A$ has nonpositive diagonal elements. At this point, if no direction of negative curvature is computed, the matrices $Y_R^T H Y_A$ and $Y_A^T H Y_A$ will be zero at each subsequent step of the algorithm. Either the algorithm detects that the matrix $Y_R^T H Y_R$ is copositive, or a constraint corresponding to a

positive diagonal element of $Y_R^T H Y_R$ is deleted. If the algorithm terminates without having computed a direction of negative curvature, the algorithm has determined that a local minimizer has been found with respect to the constraints that are still present in A_R. However, this conclusion still holds if the deleted constraints are added again, since deletion of constraints may only increase the size of the feasible region. □

ALGORITHM 6.1. *An algorithm for finding a direction of negative curvature*

repeat
  Compute $Y_0^T H Y_0$; Initialize m_A and m_R;
  if $(m_A > 0)$ **then**
    $k \leftarrow$ argument satisfying $\max_i y_{Ai}^T H y_{Ai}$;
    if $(y_{Ak}^T H y_{Ak} > 0)$ **then**
      Delete artificial constraint k; **go to** again; (see Lemma 5.1)
    end if;
    $k \leftarrow$ argument satisfying $\min_i y_{Ai}^T H y_{Ai}$;
    if $(y_{Ak}^T H y_{Ak} < 0)$ **then**
      $p \leftarrow y_{Ak}$; **go to** exit; (see Lemma 4.3)
    end if;
    $k, l \leftarrow$ arguments satisfying $\max_{i,j} |y_{Ai}^T H y_{Aj}|$;
    if $(y_{Ak}^T H y_{Al} \neq 0)$ **then**
      $p \leftarrow y_{Ak} \pm y_{Al}$; **go to** exit; (see Lemma 5.2)
    else if $(m_R > 0)$ **then**
      $k, l \leftarrow$ arguments satisfying $\max_{i,j} |y_{Ai}^T H y_{Rj}|$;
      if $(y_{Ak}^T H y_{Rl} \neq 0)$ **then**
        Compute p; **go to** exit; (see Lemma 5.3)
      end if;
    end if;
  end if;
  if $(m_R = 0)$ **or** $(\min_{i,j} y_{Ri}^T H y_{Rj} \geq 0)$ **then**
    x is a local minimizer; **go to** exit; (see Lemma 4.2)
  end if;
  $k \leftarrow$ argument satisfying $\min_i y_{Ri}^T H y_{Ri}$;
  if $(y_{Rk}^T H y_{Rk} < 0)$ **then**
    $p \leftarrow y_{Rk}$; **go to** exit; (see Lemma 4.3)
  end if;
  for $i \leftarrow 1$ **to** m_R **do**
    for $j \leftarrow i+1$ **to** m_R **do**
      $negdet \leftarrow (y_{Ri}^T H y_{Ri}\ y_{Rj}^T H y_{Rj} - y_{Ri}^T H y_{Rj}{}^2 < 0)$;
      if $(negdet)$ **and** $(y_{Ri}^T H y_{Rj} < 0)$ **then**
        Compute p; **go to** exit; (see Lemma 4.5)
      else if $(m_R = 2)$ **then**
        x is a local minimizer; **go to** exit; (see Lemma 4.4)
      end if;
    end do;
  end do;
  $k \leftarrow$ argument satisfying $\max_i y_{Ri}^T H y_{Ri}$;
  Delete regular constraint k; (see Lemma 5.1)
  label again:
until exit occurs;
label exit:

Hence, if Algorithm 6.1 does not terminate at a given step, a constraint with a positive diagonal element of $Y_0^T H Y_0$ is deleted. Recall that Lemma 5.1 implies that the new reduced Hessian is positive definite whenever a constraint corresponding to

a positive element of $Y_0^T H Y_0$ is deleted.

The amount of work needed at each step may be reduced by updating Y_0 and $Y_0^T H Y_0$. To show this, we assume that the normal of the constraint $a_i^T x \geq \beta_i$ is deleted from A_0 corresponding to a positive diagonal element of $Y_0^T H Y_0$. Partition A_0 such that

$$A_0 = \begin{pmatrix} A_{01} \\ a_i^T \end{pmatrix}.$$

In order to state the results in compact form, let Λ denote the matrix whose jth column is λ_j in (3.2) and let Θ denote the matrix whose jth column is θ_j. With this partition of A_0, let the induced partition of Y_0, Λ, and Θ be given by

$$Y_0 = \begin{pmatrix} Y_{01} & y_{0i} \end{pmatrix}, \qquad \Lambda = \begin{pmatrix} \Lambda_1 & \lambda_i \end{pmatrix} \quad \text{and} \quad \Theta = \begin{pmatrix} \Theta_{11} & \theta_{1i} \\ \Theta_{i1} & \theta_{ii} \end{pmatrix}.$$

With this partition equation, (3.2) may be written in compact form as

$$\begin{pmatrix} H & A_+^T & A_{01}^T & a_i \\ A_+ & 0 & 0 & 0 \\ A_{01} & 0 & 0 & 0 \\ a_i^T & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} Y_{01} & y_{0i} \\ -\Lambda_1 & -\lambda_i \\ -\Theta_{11} & -\theta_{1i} \\ -\Theta_{i1} & -\theta_{ii} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ I & 0 \\ 0 & 1 \end{pmatrix}. \tag{6.1}$$

Let $\bar{Y}_0$, $\bar{\Lambda}$, and $\bar{\Theta}$ denote the solution of (3.2) in the next step of Algorithm 6.1. Then, $\bar{Y}_0$, $\bar{\Lambda}$, and $\bar{\Theta}$ will satisfy the equation

$$\begin{pmatrix} H & A_+^T & A_{01}^T \\ A_+ & 0 & 0 \\ A_{01} & 0 & 0 \end{pmatrix} \begin{pmatrix} \bar{Y}_0 \\ -\bar{\Lambda} \\ -\bar{\Theta} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ I \end{pmatrix}.$$

LEMMA 6.2. *The quantities $\bar{\Theta}$ and $\bar{Y}_0$ may be obtained from the solution of* (6.1) *as*

$$\bar{\Theta} = \Theta_{11} - \frac{\theta_{1i}\theta_{1i}^T}{\theta_{ii}} \quad \text{and} \quad \bar{Y}_0 = Y_{01} - \frac{y_{0i}\theta_{1i}^T}{\theta_{ii}}.$$

Proof. The matrices $\bar{\Theta}$ and $\bar{Y}_0$ satisfy the equation

$$\begin{pmatrix} H & A_+^T & A_{01}^T & a_i \\ A_+ & 0 & 0 & 0 \\ A_{01} & 0 & 0 & 0 \\ a_i^T & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \bar{Y}_0 \\ -\bar{\Lambda} \\ -\bar{\Theta} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ I \\ a_i^T \bar{Y}_0 \end{pmatrix}. \tag{6.2}$$

Equations (6.1) and (6.2) imply that the barred quantities may be obtained from the equations

$$\bar{Y}_0 = Y_{01} + y_{0i} a_i^T \bar{Y}_0, \tag{6.3a}$$

$$\bar{\Lambda} = \Lambda_1 + \lambda_i a_i^T \bar{Y}_0, \tag{6.3b}$$

$$\bar{\Theta} = \Theta_{11} + \theta_{1i} a_i^T \bar{Y}_0, \tag{6.3c}$$

$$0 = \Theta_{i1} + \theta_{ii} a_i^T \bar{Y}_0. \tag{6.3d}$$

It follows from Lemma 4.1 that $\Theta = Y_0^T H Y_0$. Hence, Θ is a symmetric matrix with $\Theta_{i1} = \theta_{1i}^T$. Equation (6.3d) implies that $\theta_{1i}^T + \theta_{ii} a_i^T \bar{Y}_0 = 0$. The fact that $a_i^T x \geq \beta_i$

is associated with a positive diagonal element of $Y_0^T H Y_0$ implies that θ_{ii} is positive. Substitution in (6.3a) and (6.3c) yields the desired result. □

Hence, only a rank-one modification of Y_0 and $Y_0^T H Y_0$ is needed at each step of Algorithm 6.1.

LEMMA 6.3. *Assume that $y_{0i}^T H y_{0j}$ is zero and $y_{0i}^T H y_{0i}$ is positive at one step of Algorithm* 6.1. *Also assume that the constraint with normal $A_0^T e_i$ is deleted at this step. At the next step, the column of $Y_0^T H Y_0$ corresponding to the constraint with normal $A_0^T e_j$ is modified only by deletion of the zero element $y_{0i}^T H y_{0j}$.*

Proof. Lemma 6.2 implies that the rank-one modification of column j is zero when $y_{0i}^T H y_{0j}$ is zero. □

LEMMA 6.4. *If, in one step of Algorithm* 6.1, *it holds that the matrices $Y_R^T H Y_A$ and $Y_A^T H Y_A$ are zero, then they will remain zero.*

Proof. Since both the matrices $Y_R^T H Y_A$ and $Y_A^T H Y_A$ are zero, the only way the algorithm does not terminate is when a regular constraint corresponding to a positive diagonal element of $Y_R^T H Y_R$ is deleted. Lemma 6.3 implies that the matrices $Y_R^T H Y_A$ and $Y_A^T H Y_A$ will remain zero. Only one column of zeros from $Y_R^T H Y_A$ is deleted at each step. □

Hence, once the matrices $Y_R^T H Y_A$ and $Y_A^T H Y_A$ are zero, they remain zero.

LEMMA 6.5. *If $Y_R^T H Y_R$ is positive semidefinite and $Y_R^T H Y_A$ and $Y_A^T H Y_A$ are zero, then Algorithm* 6.1 *will determine that x is a local minimizer in at most m_R steps.*

Proof. Lemma 6.4 implies that the matrices $Y_R^T H Y_A$ and $Y_A^T H Y_A$ will remain zero until the algorithm terminates. Hence, the only iteration when the algorithm does not halt is when regular constraints corresponding to positive diagonal elements of $Y_R^T H Y_R$ are deleted. Therefore, at most m_R steps may be taken in the algorithm.

Assume that the algorithm terminates without determining that x is a local minimizer. It follows that a direction of negative curvature must have been computed. But Lemma 6.2 implies that the matrix $\bar{\Theta}$ of the next step is obtained as

$$\bar{\Theta} = \Theta_{11} - \frac{\theta_{1i}\theta_{1i}^T}{\theta_{ii}}.$$

Sylvester's law of inertia, (see, e.g., [GV89, p. 416]) implies that $\text{In}(\bar{\Theta}) = \text{In}(\Theta) - \text{In}(\theta_{ii})$. At the initial iteration, Θ is positive semidefinite. The value of the scalar θ_{ii} is positive. Hence, $\bar{\Theta}$ will have no negative eigenvalues. It follows by induction that no direction of negative curvature can be computed. □

Hence, if $Y_R^T H Y_R$ is positive semidefinite, Algorithm 6.1 determines that x is a local minimizer.

7. Changes in the working set. In this section the changes in the working set are described. In the proposed algorithm, either one or two constraints in A will become inactive. In an ICQP method, only one constraint is added or deleted at a time. However, we shall give a scheme that allows deletion of two constraints at a dead point, maintaining the properties of an ICQP method, i.e., the reduced Hessian having at most one nonpositive eigenvalue and the working-set matrix having full row rank.

When a direction of negative curvature, p, is computed, the objective function is strictly decreasing along that direction. The boundedness of φ in the feasible region guarantees that a sufficiently large step along p will violate a constraint. Let a_k denote the normal of the first constraint that is violated. In order to determine how to update

A, it is necessary to know if a_k is dependent on the rows of A. The following lemma, given by Gill et al. [GMSW88], shows how linear independence may be checked.

LEMMA 7.1. *Consider the equations*

$$\begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} \omega \\ v \end{pmatrix} = \begin{pmatrix} a_k \\ 0 \end{pmatrix}. \tag{7.1}$$

The vector a_k is dependent on the rows of A if and only if the vector ω is zero in the solution of (7.1).

Proof. Suppose that a_k is dependent on the rows of A. In this case, there must exist a vector v such that $a_k = A^T v$, and ω is zero in the solution of (7.1).

Assume that ω is zero in the solution of (7.1). It follows that $a_k = A^T v$, and a_k is dependent on the rows of A. □

When the algorithm is applied, either one or two constraints leave the working set. The following sections show how to update the working-set matrix.

7.1. One constraint becomes inactive. Assume that p is given by $p = y_{0i}$ and let $a_i^T x \geq \beta_i$ denote the constraint that leaves the working set.

LEMMA 7.2. *Assume that p is computed by deleting one constraint from the working set. If a_k is independent of the rows of A, it is added to A, while a_i is maintained in A as an artificial constraint. If a_k is dependent on the rows of A, a_k and a_i are exchanged. In either case, the resulting reduced Hessian is positive definite and working-set matrix has full row rank.*

Proof. If a_k is independent of the rows of A, the new reduced Hessian remains positive definite since only one more constraint is added to the working set. Also, the new working-set matrix has full row rank.

Now assume that a_k is dependent on the rows of A. If a_k and a_i are exchanged, the rows of the new working set will span the same space as the rows of A. Hence, the new reduced Hessian is positive definite. Also, the new working-set matrix has the same number of rows as the old one, and therefore it has full row rank. □

Hence, after having either added a_k or exchanged a_k and a_i, the new reduced Hessian is positive definite and the new working-set matrix has full row rank.

7.2. Two constraints become inactive. Assume that p is given by $p = \alpha_i y_{0i} + \alpha_j y_{0j}$, where α_i and α_j are both nonzero. Let a_i and a_j denote the normals of the constraints which leave the working set, and let A_2 denote the submatrix of A that remains when a_i and a_j are removed.

LEMMA 7.3. *Assume that $a_k = A_2^T v_2 + a_i v_i + a_j v_j$. Then it cannot hold that $v_i = v_j = 0$.*

Proof. Assume that $a_k = A_2^T v_2$. Premultiplication by p^T yields $p^T a_k = 0$. But this could not hold since a_k becomes violated when a sufficiently large step along p is taken. □

LEMMA 7.4. *Assume that p is computed by making two active constraints inactive. If a_k is independent of the rows of A, it is added to A, while a_i and a_j are maintained as artificial constraints. If a_k is dependent on the rows of A and $|v_i| > |v_j|$, a_k and a_i are exchanged. If a_k is dependent on the rows of A and $|v_i| \leq |v_j|$, a_k and a_j are exchanged. In each case, the new reduced Hessian is positive definite and the new working-set matrix has full row rank.*

Proof. Assume that a_k is independent of the rows of A. The new reduced Hessian remains positive definite since only one more constraint is added to the working set. Also, the new working-set matrix has full row rank.

Assume that a_k is dependent on the rows of A. Lemma 7.3 implies that at least one of the scalars v_i and v_j is nonzero. Hence, by performing the specified exchange, the rows of the new working set will span the same space as the rows of A. Hence, the new reduced Hessian will be positive definite. Also, the new working-set matrix has the same number of rows as the old one, and therefore it has full row rank. □

Hence, after having either added a_k or exchanged either a_k and a_i or a_k and a_j, the new reduced Hessian is positive definite and the working-set matrix has full row rank.

8. Verification of local optimality. In this section we describe a complete algorithm for checking if a given dead point x is a local minimizer. In Algorithm 8.1, a direction of negative curvature is found by making one or two constraints leave the working set at a time. If no such direction exists, Algorithm 6.1 yields the result that x is a local minimizer.

ALGORITHM 8.1. *An algorithm checking for local optimality*

repeat
  Apply Algorithm 6.1;
  local_minimizer ← Algorithm 6.1 implies that x is a local minimizer;
  if (*local_minimizer*) **then**
    go to exit;
  else
    a_k ← constraint that is first violated along p;
    α_F ← maximum feasible step along p;
    Solve $\begin{pmatrix} H & A^T \\ A & 0 \end{pmatrix}\begin{pmatrix} \omega \\ v \end{pmatrix} = \begin{pmatrix} a_k \\ 0 \end{pmatrix}$; (see Lemma 7.1)
    indep ← $(\|\omega\| > 0)$;
    nr_inactiv ← number of constraints that become inactive;
    if (*nr_inactiv* = 1) **then**
      a_i ← constraint that becomes inactive;
      if (*indep*) **then**
        Add a_k;
      else
        Exchange a_k and a_i; (see Lemma 7.2)
      end if;
    else
      a_i, a_j ← constraints that become inactive;
      if (*indep*) **then**
        Add a_k;
      else
        if $(|v_i| > |v_j|)$ **then**
          Exchange a_k and a_i; (see Lemma 7.4)
        else
          Exchange a_k and a_j;
        end if;
      end if;
    end if;
    if $(\alpha_F > 0)$ **then**
      $x \leftarrow x + \alpha_F p$;
      local_minimizer ← **false**;
      go to exit;
    end if;
  end if;
until too many iterations;
label exit:

If constraints corresponding to positive diagonal elements of $Y_R^T H Y_R$ are deleted

in Algorithm 6.1, assumption A2 will no longer hold. In this case, if a direction of negative curvature is computed, the resulting maximum feasible step could be zero and there is a danger of cycling.

However, if Algorithm 8.1 terminates, it will provide either a feasible direction of negative curvature along which a nonzero step may be taken or the information that x is a local minimizer. As shown in §6, the algorithm will terminate with the information that x is a local minimizer in the special case when the matrix $Y_R^T H Y_R$ is positive semidefinite.

9. Conclusions. When solving a general quadratic programming problem there may exist certain dead points at which it is very difficult to verify optimality. We emphasize that this difficulty is inherent to the problem, and is independent of the solution method.

In this paper, the verification of optimality has been discussed within the context of an inertia-controlling method. We have derived a computational method appropriate for general ICQP methods that will attempt to determine if a dead point is a local minimizer. The use of artificial constraints may introduce additional dead points. It has been shown that the new procedure does not terminate at such points, unless they are local minimizers.

However, the verification of optimality in the general case is an NP-hard problem, so we would not expect to find a procedure capable of solving a general problem in a reasonable amount of computational effort. In our scheme, there is a potential danger of cycling, and a more elaborate scheme is needed to guarantee the solution of the problem in a finite number of iterations.

Acknowledgments. We would like to thank Richard Cottle for bibliographical assistance and helpful discussions on the properties of copositive matrices. We also thank the referees for their careful reading of the paper.

REFERENCES

[BOR82] J. M. BORWEIN, *Necessary and sufficient conditions for quadratic minimality*, Numer. Funct. Anal. Optim., 5 (1982), pp. 127–140.

[CHL70] R. W. COTTLE, G. J. HABETLER, AND C. E. LEMKE, *On classes of copositive matrices*, Linear Algebra Appl., 3 (1970), pp. 295–310.

[CON80] L. B. CONTESSE, *Une caractérisation complète des minima locaux en programmation quadratique*, Numer. Math., 34 (1980), pp. 315–332.

[FLE71] R. FLETCHER. *A general quadratic programming algorithm*, J. Inst. Math. Appl., 7 (1971), pp. 76–91.

[GMSW84] P. E. GILL, W. MURRAY, M. A. SAUNDERS, AND M. H. WRIGHT, *User's guide for* SOL/QPSOL (*Version* 3.2), Report SOL 84-6, Department of Operations Research, Stanford University, Stanford, CA, 1984.

[GMSW88] ———, *Inertia-controlling methods for quadratic programming*, Report SOL 88-3, Department of Operations Research, Stanford University, Stanford CA, 1988; SIAM Rev., 33 (1991), pp. 1–36.

[GOU85] N. I. M. GOULD, *On practical conditions for the existence and uniqueness of solutions to the general equality quadratic programming problem*, Math. Programming, 32 (1985), pp. 90–99.

[GOU86] ———, *An algorithm for large scale quadratic programming*, Tech. Report CSS 219, Computer Science and Systems Division, AERE Harwell, Oxford, England, 1986.

[GV89] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Second Edition, The Johns Hopkins University Press, Baltimore, MD, 1989.

[MAJ71] A. MAJTHAY, *Optimality conditions for quadratic programming*, Math. Programming, 1 (1971), pp. 359–365.

[MAN80] O. L. MANGASARIAN, *Locally unique solutions of quadratic programs, linear and nonlinear complementarity problems*, Math. Programming, 19 (1980), pp. 200–212.

[MK87] K. G. MURTY AND S. N. KABADI, *Some* NP*-complete problems in quadratic and nonlinear programming*, Math. Programming, 39 (1987), pp. 117–129.

[PER72] F. J. PEREIRA, *On characterizations of copositive matrices*, Ph.D. thesis, Department of Operations Research, Stanford University, Stanford, CA, 1972.

[PS88] P. M. PARDALOS AND G. SCHNITGER, *Checking local optimality in constrained quadratic programming is* NP*-hard*, Oper. Res. Lett., 7 (1988), pp. 33–35.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 4, pp. 747–765, October 1991

012

ON THE SMITH NORMAL FORM OF STRUCTURED POLYNOMIAL MATRICES*

KAZUO MUROTA†

Abstract. The Smith normal form of a polynomial matrix $D(s) = Q(s) + T(s)$ is investigated, where $D(s)$ is "structured" in the sense that (i) the coefficients of the entries of $Q(s)$ belong to a field $\mathbf{K}$, (ii) the nonzero coefficients of the entries of $T(s)$ are algebraically independent over $\mathbf{K}$, and (iii) every minor of $Q(s)$ is a monomial in s. Such matrices have been useful in the structural approach in control theory. It is shown that all the invariant polynomials except for the last are monomials in s and the last invariant polynomial is expressed in terms of the combinatorial canonical form (CCF) of a layered mixed matrix associated with $D(s)$. On the basis of this, the Smith form of $D(s)$ can be computed by means of an efficient (polynomial-time) matroid-theoretic algorithm that involves arithmetic operations in the base field $\mathbf{K}$ only.

Key words. Smith normal form, structural controllability, matroid-theoretic algorithm, mixed matrix, combinatorial canonical form (CCF)

AMS(MOS) subject classifications. 15A21, 15A54, 05C50, 93B

1. Introduction.

The Smith normal form (e.g., [4], [18]) of a polynomial matrix $D(s)$ is of fundamental importance in many fields of mathematical sciences. In control theory (e.g., [20]), for example, the controllability of a linear time-invariant descriptor system

$$F\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + B\mathbf{u}(t) \tag{1}$$

with state-vector $\mathbf{x}$ and input-vector $\mathbf{u}$ is known to be equivalent to the condition that the Smith form of $D(s) = (A - sF \mid B)$ is equal to $(I \mid O)$.

In the structural approach to control systems, as initiated by Lin [7], graph-theoretic methods have been developed under the assumption that all the nonzero numbers characterizing a dynamical system (e.g., the nonzero entries of the matrices A, B, and F in (1)) are algebraically independent parameters. As a refinement of such a graph-theoretic approach, matroid-theoretic methods for structural analyses of dynamical systems have been developed under a more realistic assumption that the coefficients are classified into independent physical parameters and dimensionless fixed constants (see [9]–[11], [12], [14], [17]).

In the matroid-theoretic methods, we encounter a class of polynomial matrices $D(s) = Q_D(s) + T_D(s)$ which are "structured" in the sense that (i) the coefficients of the entries of $Q_D(s)$ belong to a field $\mathbf{K}$, (ii) the nonzero coefficients of the entries of $T_D(s)$ are algebraically independent over $\mathbf{K}$, and (iii) every minor of $Q_D(s)$ is a monomial in s. This means in particular that $D(s)$ is a mixed matrix with respect to $\mathbf{K}(s)$. (See Example 3.1 for a concrete instance of such a matrix and §2 for the definition of a mixed matrix.) In applications to control systems, we usually have $\mathbf{K} = \mathbf{Q}$ (the field of rational numbers) and the third condition reflects the consistency of equations with respect to physical dimensions (see Remark 3.2 for the physical background). Note that the so-called structured matrix that has independent nonzero coefficients is the special case where $Q_D(s) = O$.

* Received by the editors January 8, 1990; accepted for publication (in revised form) August 27, 1990.

† Department of Mathematical Engineering and Information Physics, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan (murota@tansei.cc.u-tokyo.ac.jp).

In this paper we investigate the Smith normal form of such a polynomial matrix $D(s)$. First, in §2, we show some preliminary results on mixed matrices. The main theorems are given in §3, which state that all the invariant polynomials of $D(s)$ except for the last are monomials in s and the last invariant polynomial can be expressed in terms of the combinatorial canonical form (CCF) of a layered mixed matrix associated with $D(s)$. On the basis of this, we present in §4 an efficient matroid-theoretic algorithm using weighted matroid-intersection algorithms (e.g., [6], [21]) for computing the Smith form of $D(s)$; in particular, the degrees of invariant polynomials can be computed in polynomial time. The algorithm is practical, involving arithmetic operations in the base field $\mathbf{K}$ only. As an application of the present results to control theory, we obtain an alternative efficient algorithm for testing for the structural controllability in the sense of [11].

2. Preliminaries on mixed matrices. This section gives some properties of mixed matrices to be used in this paper (see [10], [13], and [14] for other properties). The notion of mixed matrix was introduced by [15].

For a matrix A, the row set and the column set of A are denoted by $\mathrm{Row}(A)$ and $\mathrm{Col}(A)$. For $I \subseteq \mathrm{Row}(A)$ and $J \subseteq \mathrm{Col}(A)$, $A[I, J]$ means the submatrix of A with row set I and column set J. The (multi)set of nonzero entries of A is denoted by $\mathcal{N}(A)$. The zero/nonzero structure of a matrix A is represented by a bipartite graph $G(A) = (\mathrm{Row}(A), \mathrm{Col}(A), \mathcal{N}(A))$ with vertex set $\mathrm{Row}(A) \cup \mathrm{Col}(A)$ and arc set $\mathcal{N}(A)$; $G(A)$ has an arc from $i \in \mathrm{Row}(A)$ to $j \in \mathrm{Col}(A)$ if A has a nonzero entry at the position (i, j). The term-rank [19] of A is equal to the maximum size of a matching in $G(A)$.

Let $\mathbf{K}$ be a subfield of a field $\mathbf{F}$. A matrix A is called a *mixed matrix* with respect to $\mathbf{K}$ if

$$A = Q_A + T_A, \tag{2}$$

where

(i) Q_A is a matrix over $\mathbf{K}$, and
(ii) T_A is a matrix over $\mathbf{F}$ such that the set $\mathcal{N}(T_A)$ of its nonzero entries is algebraically independent over $\mathbf{K}$.

The following identity is fundamental. It can be translated nicely into the matroid-theoretic language and enables us to compute the rank of A by an efficient algorithm using arithmetic operations in the subfield $\mathbf{K}$ only.

THEOREM 2.1 (rank identity [15]). *For a mixed matrix* $A = Q_A + T_A$,

$$\operatorname{rank} A = \max\{\operatorname{rank} Q_A[R - I, C - J] + \text{term-rank}\, T_A[I, J] \mid I \subseteq R,\ J \subseteq C\},$$

where $R = \mathrm{Row}(A)$ *and* $C = \mathrm{Col}(A)$. □

A matrix A is called a *layered mixed matrix* (or an LM-*matrix*) with respect to $\mathbf{K}$ if it takes the following form (possibly after a permutation of rows):

$$A = \begin{pmatrix} Q \\ T \end{pmatrix} \tag{3}$$

such that

(i) $Q = (Q_{ij})$ is a matrix over $\mathbf{K}$, and
(ii) $T = (T_{ij})$ is a matrix over $\mathbf{F}$ such that the set $\mathcal{N}(T)$ of its nonzero entries is algebraically independent over $\mathbf{K}$.

In other words, an LM-matrix is a mixed matrix (2) such that the nonzero rows of Q_A and T_A are disjoint.

An LM-matrix A of (3) is associated with a set function p as follows. Set $\mathrm{Row}(Q) = R_Q$, $\mathrm{Row}(T) = R_T$, and $\mathrm{Row}(A) = R$; then $R = R_Q \cup R_T$. The column sets of A, Q and T, being identified with one another, are denoted by C; $\mathrm{Col}(A) = \mathrm{Col}(Q) = \mathrm{Col}(T) = C$. Put

$$
\begin{aligned}
&\rho(I,J) = \operatorname{rank} Q[I,J], \qquad I \subseteq R_Q,\ J \subseteq C,\\
&\Gamma(I,J) = \bigcup_{j\in J}\{i \in I \mid T_{ij} \neq 0\}, \qquad I \subseteq R_T,\ J \subseteq C,\\
&\gamma(I,J) = |\Gamma(I,J)|, \qquad I \subseteq R_T,\ J \subseteq C,\\
&p(I,J) = \rho(I \cap R_Q, J) + \gamma(I \cap R_T, J) - |J|, \qquad I \subseteq R,\ J \subseteq C.
\end{aligned} \tag{4}
$$

The function $p : 2^R \times 2^C \to \mathbf{Z}$ is known to be bisubmodular:

$$
p(I_1 \cup I_2, J_1 \cap J_2) + p(I_1 \cap I_2, J_1 \cup J_2) \leq p(I_1, J_1) + p(I_2, J_2),
\qquad I_i \subseteq R,\ J_i \subseteq C \qquad (i = 1,2).
$$

Put

$$
L(I) = \{J \subseteq C \mid p(I,J) \leq p(I,J'), \forall\, J' \subseteq C\}, \quad I \subseteq R. \tag{5}
$$

Based on the rank identity in Theorem 2.1 and the min-max formula for the matroid-union of Edmonds [2], we can prove the following identity, an extension of the well-known min-max characterization (e.g., [6], [8]) of the term rank of a matrix or the maximum matching in a bipartite graph, which is ascribed to Egerváry, König, Hall, Rado, Ore, and others.

THEOREM 2.2 ([10], [16]). *For an* LM-*matrix* A,

$$
\operatorname{rank} A[I,J] = \min\{p(I,J') \mid J' \subseteq J\} + |J|, \qquad I \subseteq R,\ J \subseteq C. \qquad \square
$$

By the admissible transformation for an LM-matrix A of (3), we mean the transformation of the form:

$$
P_{\mathrm{r}} \begin{pmatrix} S & O \\ O & I \end{pmatrix} \begin{pmatrix} Q \\ T \end{pmatrix} P_{\mathrm{c}}, \tag{6}
$$

where S is a nonsingular matrix over the subfield $\mathbf{K}$, and P_{r} and P_{c} are permutation matrices. The admissible transformation brings an LM-matrix into another LM-matrix and two LM-matrices are said to be LM-*equivalent* if and only if they are connected by an admissible transformation.

Remark 2.1. An electrical network is typically described by means of a layered mixed matrix when currents in and voltages across branches are chosen as the elementary variables (see, e.g., [5]). In that case, the Q-part represents Kirchhoff's current and voltage laws, which, as is well known, can be written down in a number of different ways. The LM-equivalence accounts exactly for the degree of freedom in expressing these conservation laws.

It is known that there exists a finest block-triangular matrix, say $\bar{A}$, among the matrices which are LM-equivalent to A (cf. Theorem 2.3 below). This is called the *combinatorial canonical form* (or CCF for short) of A. Since the transformation

(6) is more general than mere permutations of rows and columns, the CCF is a generalization of the canonical decomposition due to Dulmage and Mendelsohn [1] of a bipartite graph.

The column set $\mathrm{Col}(\bar{A})$ of the CCF is partitioned into blocks as

$$\{C_0; C_1, \cdots, C_b; C_\infty\} \tag{7}$$

with reference to $L(R)$ (cf. (5) for notation), which is a sublattice of 2^C. We denote by

$$\{R_0; R_1, \cdots, R_b; R_\infty\} \tag{8}$$

the partition of the row set $\mathrm{Row}(\bar{A})$. Note that

$$R_k \cap R_l = \emptyset \quad \text{if } k \neq l, \{k,l\} \subseteq \{0, 1, \cdots, b, \infty\},$$

$$C_k \cap C_l = \emptyset \quad \text{if } k \neq l, \{k,l\} \subseteq \{0, 1, \cdots, b, \infty\},$$

and

$$R_k \neq \emptyset, \qquad C_k \neq \emptyset \quad \text{for } k = 1, \cdots, b,$$

whereas R_0, R_∞, C_0, and C_∞ can be empty.

A partial order, denoted as $\preceq$, is induced among the blocks of (7) from the lattice $L(R)$. We assume here that the blocks are indexed so that $C_k \preceq C_l$ implies $k \leq l$ $(1 \leq k, l \leq b)$. $C_k \prec C_l$ will mean that $C_k \preceq C_l$ and $C_k \neq C_l$; and $C_k \prec \cdot\, C_l$ will mean that $C_k \prec C_l$ and there does not exist C_m such that $C_k \prec C_m \prec C_l$.

THEOREM 2.3 ([10], [16]). *The* CCF $\bar{A}$ *of an* LM*-matrix* A *has the following properties:*

(a) $\bar{A}$ *is block-triangularized with respect to the partitions* (7) *and* (8)*, i.e.,*

$$\bar{A}[R_k, C_l] = O \quad \text{if} \quad 0 \leq l < k \leq \infty.$$

Moreover, the partial order on $\{C_k \mid k = 1, \cdots, b\}$ *induced by the zero/nonzero structure of* $\bar{A}$ *agrees with the partial order* $\preceq$ *defined by the lattice* $L(R)$*; i.e.,*

$$\begin{aligned} &\bar{A}[R_k, C_l] = O \quad \text{unless } C_k \preceq C_l \qquad (1 \leq k, l \leq b), \\ &\bar{A}[R_k, C_l] \neq O \quad \text{if } C_k \prec \cdot\, C_l \qquad (1 \leq k, l \leq b). \end{aligned}$$

(b)

$$\begin{aligned} &|R_0| < |C_0| \quad \text{if} \quad R_0 \neq \emptyset, \\ &|R_k| = |C_k| \quad (>0) \quad \text{for} \quad k = 1, \cdots, b, \\ &|R_\infty| > |C_\infty| \quad \text{if} \quad C_\infty \neq \emptyset. \end{aligned}$$

(c)

$$\begin{aligned} &\operatorname{rank} \bar{A}[R_0, C_0] = |R_0|, \\ &\operatorname{rank} \bar{A}[R_k, C_k] = |R_k| = |C_k| \quad \text{for} \quad k = 1, \cdots, b, \\ &\operatorname{rank} \bar{A}[R_\infty, C_\infty] = |C_\infty|. \end{aligned}$$

(d) *$\bar{A}$ has the finest block-triangular form among the matrices that have the properties* (b) *and* (c) *and are* LM-*equivalent to* A. □

The submatrices $\bar{A}[R_0, C_0]$ and $\bar{A}[R_\infty, C_\infty]$ are called the *horizontal tail* and the *vertical tail*, respectively.

Remark 2.2. The CCF is uniquely determined so far as the partitions of the row and column sets, as well as the partial order among the blocks, are concerned, whereas there remains some indeterminacy in the numerical values of the entries in the Q-part. We sometimes refer to *a* CCF, instead of *the* CCF, in order to emphasize this numerical indeterminacy.

An LM-matrix A is called LM-*irreducible* or simply *irreducible* if its CCF does not split into more than one nonempty block, that is, if (a) $b = 1$ and $C_0 = R_\infty = \emptyset$, (b) $b = 0$ and $R_\infty = \emptyset$, or (c) $b = 0$ and $C_0 = \emptyset$. Each block $\bar{A}[R_k, C_k]$ of the CCF above is irreducible for $k = 0, 1, \cdots, b, \infty$. The irreducibility of an LM-matrix is characterized by the function p of (4) as follows.

THEOREM 2.4 ([10], [16]). *Let A be an* LM-*matrix with $R = \mathrm{Row}(A)$ and $C = \mathrm{Col}(A)$.*

(a) *In the case where $|R| = |C|$ (> 0):*

$$A \text{ is irreducible} \iff p(R, J) > p(R, \emptyset) = p(R, C)\ (= 0), \quad \forall J \neq \emptyset, C.$$

(b) *In the case where $|R| < |C|$:*

$$A \text{ is irreducible} \iff p(R, J) > p(R, C), \quad \forall J \neq C.$$

(c) *In the case where $|R| > |C|$:*

$$A \text{ is irreducible} \iff p(R, J) > p(R, \emptyset)\ (= 0), \quad \forall J \neq \emptyset. \qquad \square$$

A minor (subdeterminant) of A is a polynomial in $\mathcal{T} = \mathcal{N}(T)$ over $\mathbf{K}$. Let $d_k(\mathcal{T}) \in \mathbf{K}[\mathcal{T}]$ denote the kth determinantal divisor of A, i.e., the greatest common divisor of all minors of order k in A as polynomials in $\mathcal{T}$ over $\mathbf{K}$.

THEOREM 2.5. *Let A be an irreducible* LM-*matrix with respect to* $\mathbf{K}$*; put $R = \mathrm{Row}(A)$, $C = \mathrm{Col}(A)$.*

(a) *In the case where $|R| = |C|$ (> 0):*

$$d_k(\mathcal{T}) \in \mathbf{K} - \{0\} \quad \textit{for } k = 1, \cdots, |R| - 1$$

and

$$d_{|R|}(\mathcal{T}) = \det A \neq 0.$$

(b) *In the case where $|R| < |C|$:*

$$d_k(\mathcal{T}) \in \mathbf{K} - \{0\} \quad \textit{for } k = 1, \cdots, |R|.$$

(c) *In the case where $|R| > |C|$:*

$$d_k(\mathcal{T}) \in \mathbf{K} - \{0\} \quad \textit{for } k = 1, \cdots, |C|.$$

Proof. First note that $\mathrm{rank}\, A = \min(|R|, |C|)$ if A is LM-irreducible.

Case (a). It suffices to show that $d_{|R|-1} \in \mathbf{K}$, since d_{k-1} divides d_k for $k = 1, \cdots, |R|$. We will show that no $t \in \mathcal{T}$ can appear in $d_{|R|-1}$. Suppose that t is contained as the (i,j) entry of A, where $i \in R_T$ and $j \in C$. By Theorem 2.4(a) we see that

$$p(R-i, J) \geq 0 \quad \forall\, J \subseteq C - j,$$

since

$$p(R-i, J) \geq p(R, J) - 1 \geq 0 \quad \text{if } \emptyset \neq J \subseteq C - j$$

and

$$p(R-i, J) = 0 \quad \text{if } J = \emptyset.$$

It follows from Theorem 2.2 that $\det A[R-i, C-j] \neq 0$. Obviously, $\det A[R-i, C-j]$ does not contain t and is a multiple of $d_{|R|-1}$. Therefore, $d_{|R|-1}$ does not contain t.

Case (b). Though the claim in this case follows easily from the argument in [11], we give a simpler direct proof here. Similarly to Case (a), it suffices to show that no $t \in \mathcal{T}$ can appear in $d_{|R|}$. Suppose that t is contained in column $j \in C$. By Theorem 2.4(b) and Theorem 2.2 we see that $\operatorname{rank} A[R, C-j] = |R|$. In other words, there exists $J \subseteq C - j$ with $|J| = |R|$ such that $\det A[R, J] \neq 0$. The rest of the proof is similar to the one for Case (a).

Case (c). Suppose that $t \in \mathcal{T}$ is contained in row $i \in R_T$. By Theorem 2.4(c) we see, just as in Case (a), that

$$p(R-i, J) \geq 0, \quad \forall\, J \subseteq C.$$

Then $\operatorname{rank} A[R-i, C] = |C|$ by Theorem 2.2 and the rest of the proof is similar to the one for Case (a). □

It is known [13] that $\det A$ is an irreducible polynomial in the ring $\mathbf{K}[\mathcal{T}]$ if A is a nonsingular and irreducible LM-matrix. Combining Theorem 2.5 with this result we obtain the following statement. The special case of this statement when $\operatorname{rank} A = |R|$ has been noted in [14].

THEOREM 2.6. *Let A be an* LM*-matrix of rank r with respect to $\mathbf{K}$. The decomposition of the rth determinantal divisor $d_r(\mathcal{T})$ of A into irreducible factors in the ring $\mathbf{K}[\mathcal{T}]$ is given by*

$$d_r(\mathcal{T}) = \alpha \cdot \prod_{k=1}^{b} \det \bar{A}[R_k, C_k],$$

where $\bar{A}[R_k, C_k]$ $(k = 1, \cdots, b)$ are the irreducible square blocks in the CCF *of A, and $\alpha \in \mathbf{K} - \{0\}$.* □

With an $m \times n$ mixed matrix $A = Q_A + T_A$ with respect to $\mathbf{K}$, we associate a $(2m) \times (m+n)$ LM-matrix

$$\tilde{A} = \begin{pmatrix} I_m & Q_A \\ -\operatorname{diag}[t_1, \cdots, t_m] & T_A \end{pmatrix} = \begin{pmatrix} \tilde{Q} \\ \tilde{T} \end{pmatrix}, \tag{9}$$

where $t_1, \cdots, t_m$ are new indeterminates (in $\mathbf{F}$). Note that $\operatorname{rank} \tilde{A} = \operatorname{rank} A + m$.

3. The Smith normal form. In the first section, the main results of this paper are presented as Theorems 3.1 and 3.2, and are illustrated in Example 3.1. Their proofs are postponed to the second section.

3.1. Theorems. Let $D(s)$ be an $m \times n$ polynomial matrix in indeterminate s over a field $\mathbf{F}$ represented as

$$D(s) = Q_D(s) + T_D(s), \tag{10}$$

where

(i) the coefficients of the entries of $Q_D(s)$ belong to a subfield $\mathbf{K}$ of $\mathbf{F}$,
(ii) the nonzero coefficients $\mathcal{T}$ ($\subseteq \mathbf{F}$) of the entries of $T_D(s)$ are algebraically independent over $\mathbf{K}$, and
(iii) every minor of $Q_D(s)$ is a monomial in s over $\mathbf{K}$.

The first two conditions imply that $D(s)$ is a mixed matrix with respect to the rational function field $\mathbf{K}(s)$. The last condition (iii) is known (cf. [9], [10]) to imply that

$$Q_D(s) = \operatorname{diag}[s^{r_1}, \cdots, s^{r_m}] \cdot Q_D(1) \cdot \operatorname{diag}[s^{-c_1}, \cdots, s^{-c_n}] \tag{11}$$

for some integers r_i ($i = 1, \cdots, m$) and c_j ($j = 1, \cdots, n$). This means in particular that each entry of $Q_D(s)$ is a monomial in s over $\mathbf{K}$. See Remark 3.2 at the end of this section for the physical motivations for such classes of matrices.

Let $d_k(s)$ denote the kth determinantal divisor of $D(s)$ over $\mathbf{F}$ for $k = 1, \cdots, r$, where $r = \operatorname{rank} D(s)$. Then the kth invariant polynomial $e_k(s)$ of $D(s)$ is expressed as

$$e_k(s) = \frac{d_k(s)}{d_{k-1}(s)} \quad \text{for } k = 1, \cdots, r, \tag{12}$$

where $d_0(s) = 1$ by convention. The Smith normal form of $D(s)$ is given by

$$\Sigma(s) = \operatorname{diag}[e_1(s), \cdots, e_r(s), 0, \cdots, 0].$$

We choose $d_k(s)$ and $e_k(s)$ to be monic.

The main objective of this paper is to show that the Smith form $\Sigma(s)$ of $D(s)$ has special properties, as stated in Theorems 3.1 and 3.2 below. The former refers to $e_k(s)$ for $k = 1, \cdots, r-1$, whereas the latter refers to $e_r(s)$. Based on these theorems we shall give in §4 an efficient algorithm for computing the Smith form of $D(s)$. For a polynomial we denote by "ord" its order, i.e., the lowest degree of a nonvanishing term in it.

THEOREM 3.1. *The determinantal divisors of $D(s)$, except for the last, are monomials:*

$$d_k(s) = s^{p_k} \quad \textit{for } k = 1, \cdots, r-1, \tag{13}$$

where

$$p_k = \min\{\operatorname{ord} \det D(s)[I, J] \mid |I| = |J| = k\} \quad \textit{for } k = 1, \cdots, r-1. \tag{14}$$

Hence the invariant polynomials, except for the last, are also monomials:

$$e_k(s) = s^{p_k - p_{k-1}} \quad \textit{for } k = 1, \cdots, r-1, \tag{15}$$

where $p_0 = 0$ by convention. □

We now turn to the last invariant polynomial $e_r(s)$. Just as in (9), we associate with $D(s)$ an augmented $(2m) \times (m+n)$ matrix

$$\tilde{D}(s) = \tilde{D}(s;t) = \begin{pmatrix} I_m & Q_D(s) \\ -\mathrm{diag}\,[t_1, \cdots, t_m] & T_D(s) \end{pmatrix} = \begin{pmatrix} \tilde{Q}(s) \\ \tilde{T}(s;t) \end{pmatrix} \tag{16}$$

with indeterminates $t = (t_1, \cdots, t_m)$ in $\mathbf{F}$, where

$$\tilde{Q}(s) = (I_m \mid Q_D(s)), \qquad \tilde{T}(s;t) = (-\mathrm{diag}\,[t_1, \cdots, t_m] \mid T_D(s)).$$

It is not difficult to see that

$$\tilde{D}(s;1) = \tilde{D}(s;t)\Big|_{t_1=\cdots=t_m=1}$$

is equivalent, as a polynomial matrix in s over $\mathbf{F}$, to

$$\begin{pmatrix} I_m & O \\ O & D(s) \end{pmatrix},$$

and hence the Smith form of $D(s)$ is embedded in the Smith form of $\tilde{D}(s;1)$ as

$$\begin{pmatrix} I_m & O \\ O & \Sigma(s) \end{pmatrix}. \tag{17}$$

In particular, $e_r(s)$ is equal to the last invariant polynomial of $\tilde{D}(s;1)$.

Since $\tilde{D}(s;t)$ is an LM-matrix with respect to $\mathbf{K}(s)$, we can talk of its CCF, say $\bar{D}(s;t)$; let $\{\bar{D}_l(s;t) \mid l = 0, 1, \cdots, b, \infty\}$ denote the family of its irreducible diagonal blocks. As mentioned in Remark 2.2, there is some indeterminacy in the entries of $\bar{D}_l(s;t)$. The following theorem states that $e_r(s)$ is characterized by those diagonal blocks when they are appropriately chosen; in particular, the statement of the theorem presupposes that we can choose $\bar{D}_l(s;t)$ in such a way that the diagonal blocks are polynomial matrices in s. See Lemma 3.1 in §3.2 for the concrete choice of $\bar{D}(s;t)$.

THEOREM 3.2. *The rth determinantal divisor of $D(s)$, where $r = \mathrm{rank}\, D(s)$, is given by*

$$d_r(s) = \alpha_r \cdot s^{p_r} \prod_{l=1}^{b} \det \bar{D}_l(s;1) \tag{18}$$

for some $\alpha_r \in \mathbf{F} - \{0\}$, where $p_r = p_r^0 + p_r^\infty$ with

$$\begin{aligned} p_r^0 &= \min\{\mathrm{ord}\, \det \bar{D}_0(s;t)[R_0, J] \mid |J| = |R_0|\}, \\ p_r^\infty &= \min\{\mathrm{ord}\, \det \bar{D}_\infty(s;t)[I, C_\infty] \mid |I| = |C_\infty|\}, \end{aligned} \tag{19}$$

and $\{\bar{D}_l(s;t) \mid l = 0, 1, \cdots, b, \infty\}$ denotes the family of the diagonal blocks of an appropriately chosen CCF *$\bar{D}(s;t)$ of $\tilde{D}(s;t)$. Hence the last invariant polynomial is*

$$e_r(s) = \alpha_r \cdot s^{p_r - p_{r-1}} \prod_{l=1}^{b} \det \bar{D}_l(s;1) \tag{20}$$

with p_{r-1} given by (14). □

Remark 3.1. In the case where $D(s)$ itself is an LM-matrix, there is no need to introduce the augmented LM-matrix $\tilde{D}(s;t)$. The claims in Theorem 3.2 remain valid when we redefine $\bar{D}$ to be an appropriately chosen CCF of D. □

Remark 3.2. The class of polynomial matrices considered in this paper has been proposed by Murota in the context of structural approach to dynamical systems (see [9], [10]). Here we will explain, only briefly, the physical observations which motivate such classes of matrices.

The decomposition (10) into two parts with the properties (i) and (ii) reflects the way we recognize the structure of a physical/engineering system. It is based on the distinction between "generic system parameters" and "fixed constants," which are dubbed as "accurate" and "inaccurate" numbers in [15] as follows:

(1) Inaccurate numbers. Numbers representing independent physical parameters such as resistances in electrical networks which, being contaminated with noise and other errors, take values independent of one another, and therefore can be modeled as algebraically independent numbers; and

(2) Accurate numbers. Numbers accounting for various sorts of conservation laws such as Kirchhoff's laws which, stemming from topological incidence relations, are precise in value (often ± 1), and therefore cause no serious numerical difficulty in arithmetic operations on them.

The last condition (iii), being concerned with the "accurate numbers," represents the consistency with respect to physical dimensions. The "accurate numbers" usually represent topological and/or geometrical incidence coefficients, which have no physical dimensions, so that it is natural to expect that the coefficients of the entries of $Q_D(s)$ are dimensionless constants. On the other hand, the indeterminate s corresponds, in the context of dynamical system theory, to the differentiation with respect to time and therefore should have the physical dimension of the inverse of time.

If the matrix $D(s)$ is to represent a physical system at all, relevant physical dimensions are associated with the columns and rows of $D(s)$. Choosing time as one of the fundamental dimensions, we denote by $-c_j$ and $-r_i$ the exponent to the dimension of time associated, respectively, with the jth column and the ith row. The principle of dimensional homogeneity then requires that the (i,j) entry of $D(s)$ should have the dimension of time with exponent $c_j - r_i$. Combining this fact with the observations on the nondimensionality of the coefficients of $Q_D(s)$ and on the dimension of s, we are led to the condition (iii). □

Example 3.1. The theorems above are illustrated here for a 5×5 matrix

$$D(s) = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{pmatrix} 0 & 0 & 1+p_1 s & 3s & p_2 \\ s & 1 & 1 & 0 & p_3 + p_4 s \\ 2s^2 & 2s & 2s & 0 & p_5 s \\ 0 & 0 & 0 & s^2 & p_6 \\ 2s^3 & 2s^2 & 2s^2 & 0 & s + p_7 s^2 \end{pmatrix} \end{array} \tag{21}$$

with

$$\mathrm{Row}(D) = \{w_1, w_2, w_3, w_4, w_5\}, \qquad \mathrm{Col}(D) = \{x_1, x_2, x_3, x_4, x_5\}.$$

This matrix is expressed as $D(s) = Q_D(s) + T_D(s)$ in the form of (10) for $\mathbf{K} = \mathbf{Q}$ with

$$(22)\qquad Q_D(s) = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 3s & 0 \\ s & 1 & 1 & 0 & 0 \\ 2s^2 & 2s & 2s & 0 & 0 \\ 0 & 0 & 0 & s^2 & 0 \\ 2s^3 & 2s^2 & 2s^2 & 0 & s \end{pmatrix} \end{array}$$

and

$$(23)\qquad T_D(s) = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{pmatrix} 0 & 0 & p_1 s & 0 & p_2 \\ 0 & 0 & 0 & 0 & p_3 + p_4 s \\ 0 & 0 & 0 & 0 & p_5 s \\ 0 & 0 & 0 & 0 & p_6 \\ 0 & 0 & 0 & 0 & p_7 s^2 \end{pmatrix} \end{array}.$$

Here

$$\mathcal{T} = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7\}$$

is the set of algebraically independent parameters, and $Q_D(s)$ satisfies (11) with

$$r_1 = r_2 = 1, \quad r_3 = r_4 = 2, \quad r_5 = 3; \qquad c_1 = c_4 = 0, \quad c_2 = c_3 = 1, \quad c_5 = 2.$$

The augmented LM-matrix $\tilde{D}(s;t)$ of (16) is given by

$$(24)\qquad \tilde{Q}(s) = \begin{array}{c} \begin{array}{ccccccccccc} w_1 & & w_2 & & w_3 & & w_4 & & w_5 & x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{pmatrix} 1 & & & & & & & & & 0 & 0 & 1 & 3s & 0 \\ & & 1 & & & & & & & s & 1 & 1 & 0 & 0 \\ & & & & 1 & & & & & 2s^2 & 2s & 2s & 0 & 0 \\ & & & & & & 1 & & & 0 & 0 & 0 & s^2 & 0 \\ & & & & & & & & 1 & 2s^3 & 2s^2 & 2s^2 & 0 & s \end{pmatrix} \end{array}$$

and

$$(25)\quad \tilde{T}(s;t) = \begin{array}{c} \begin{array}{ccccccccccc} w_1 & & w_2 & & w_3 & & w_4 & & w_5 & x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{pmatrix} -t_1 & & & & & & & & & 0 & 0 & p_1 s & 0 & p_2 \\ & & -t_2 & & & & & & & 0 & 0 & 0 & 0 & p_3 + p_4 s \\ & & & & -t_3 & & & & & 0 & 0 & 0 & 0 & p_5 s \\ & & & & & & -t_4 & & & 0 & 0 & 0 & 0 & p_6 \\ & & & & & & & & -t_5 & 0 & 0 & 0 & 0 & p_7 s^2 \end{pmatrix} \end{array}.$$

We find (by the algorithm of §4) the following CCF $\bar{D}(s;t)$ of $\tilde{D}(s;t)$; note that the diagonal blocks are polynomial matrices in s, whereas a fraction "$-3/s$" is contained

in an off-diagonal block:

$$
(26)\quad \bar{D}(s;t) = \left(\begin{array}{cc|cc|cc|ccc|c}
x_1 & x_2 & w_1 & x_3 & x_4 & w_4 & w_3 & w_5 & w_2 & x_5 \\
\hline
s & 1 & & 1 & & & & & 1 & \\
\hline
 & & 1 & 1 & & -3/s & & & & \\
 & & -t_1 & p_1 s & & & & & & p_2 \\
\hline
 & & & & s^2 & 1 & & & & \\
 & & & & & -t_4 & & & & p_6 \\
\hline
 & & & & & & 1 & 0 & -2s & 0 \\
 & & 0 & & & & 0 & 1 & -2s^2 & s \\
 & & & & & & -t_3 & 0 & 0 & p_5 s \\
 & & & & & & 0 & -t_5 & 0 & p_7 s^2 \\
 & & & & & & 0 & 0 & -t_2 & p_3 + p_4 s
\end{array}\right).
$$

We have nonempty tails:

$$
\bar{D}_0(s;t) = \begin{array}{c} \begin{array}{cc} x_1 & x_2 \end{array} \\ \begin{pmatrix} s & 1 \end{pmatrix} \end{array},
$$

$$
\bar{D}_\infty(s;t) = \begin{array}{c} \begin{array}{cccc} w_3 & w_5 & w_2 & x_5 \end{array} \\ \begin{pmatrix} 1 & 0 & -2s & 0 \\ 0 & 1 & -2s^2 & s \\ -t_3 & 0 & 0 & p_5 s \\ 0 & -t_5 & 0 & p_7 s^2 \\ 0 & 0 & -t_2 & p_3 + p_4 s \end{pmatrix} \end{array}
$$

and $b = 3$ square diagonal blocks:

$$
\bar{D}_1(s;t) = \begin{array}{c} \begin{array}{cc} w_1 & x_3 \end{array} \\ \begin{pmatrix} 1 & 1 \\ -t_1 & p_1 s \end{pmatrix} \end{array}, \quad \bar{D}_2(s;t) = \begin{array}{c} x_4 \\ \begin{pmatrix} s^2 \end{pmatrix} \end{array}, \quad \bar{D}_3(s;t) = \begin{array}{c} w_4 \\ \begin{pmatrix} -t_4 \end{pmatrix} \end{array}.
$$

The CCF reveals that

$$
r = \operatorname{rank} D(s) = 4 \ (< 5).
$$

Then, according to Theorem 3.2, we see that

$$
d_4(s) = \alpha_4 \cdot s^{p_4} \cdot (p_1 s + 1) \cdot s^2 \cdot (-1),
$$

where $\alpha_4 = -1/p_1$ to make $d_4(s)$ a monic polynomial. We can compute $p_4^0 = 0$ and $p_4^\infty = 1$ (by the algorithm given in §4) to obtain $p_4 = p_4^0 + p_4^\infty = 1$. Therefore,

$$
d_4(s) = s^3 \cdot (s + 1/p_1).
$$

As for the other determinantal divisors, we obtain

$$
d_1(s) = d_2(s) = d_3(s) = 1
$$

(again by the algorithm given in §4). Hence the Smith form $\Sigma(s)$ of $D(s)$ is given by

$$
\Sigma(s) = \operatorname{diag}[1, 1, 1, s^3(s + 1/p_1), 0].
$$

This example will be considered again in Example 3.2. □

3.2. Proofs. This section gives the proofs for Theorems 3.1 and 3.2. Example 3.2 at the end of this section will serve to give concrete ideas for the following derivation.

The matrix $\tilde{D}(s;t)$ of (16) plays the primary role here, since the Smith form of $D(s)$ is obtained from that of $\tilde{D}(s;1)$, as noted in (17). Recall that $\tilde{D}(s;t)$ is an LM-matrix with respect to $\mathbf{K}(s)$ and note that (11) implies

$$\tilde{Q}(s) = \mathrm{diag}\,[s^{r_1}, \cdots, s^{r_m}] \cdot \tilde{Q}(1) \cdot \mathrm{diag}\,[s^{-r_1}, \cdots, s^{-r_m}; s^{-c_1}, \cdots, s^{-c_n}]. \tag{27}$$

The CCFs of $\tilde{D}(s;t)$ and $\tilde{D}(1;t)$ are closely related as follows, where it should be noted that $\tilde{D}(1;t)$ is an LM-matrix with respect to $\mathbf{K}$. By virtue of (27), $\tilde{D}(s;t)$ and $\tilde{D}(1;t)$ share the same function p of (4), so that they have the same partition of the column set $C = \mathrm{Col}(\tilde{D})$ in their CCFs. We denote by $\{C_0; C_1, \cdots, C_b; C_\infty\}$ the common partition of C; it is assumed as before that $C_k \preceq C_l$ implies $k \leq l$ $(1 \leq k, l \leq b)$.

Suppose that S is a nonsingular matrix over $\mathbf{K}$ such that

$$P_{\mathrm{r}} \begin{pmatrix} S & O \\ O & I \end{pmatrix} \begin{pmatrix} \tilde{Q}(1) \\ \tilde{T}(1;t) \end{pmatrix} P_{\mathrm{c}} \tag{28}$$

is the CCF of $\tilde{D}(1;t)$. If we define

$$S(s) = \mathrm{diag}\,[s^{r_1}, \cdots, s^{r_m}] \cdot S \cdot \mathrm{diag}\,[s^{-r_1}, \cdots, s^{-r_m}] \tag{29}$$

with reference to (27), we see

$$\bar{D}(s) = \bar{D}(s;t) = P_{\mathrm{r}} \begin{pmatrix} S(s) & O \\ O & I \end{pmatrix} \begin{pmatrix} \tilde{Q}(s) \\ \tilde{T}(s;t) \end{pmatrix} P_{\mathrm{c}} \tag{30}$$

is the CCF of $\tilde{D}(s;t)$.

In general, the transformation (30) is not qualified as an equivalence transformation of $\tilde{D}(s;t)$ as a polynomial matrix in s, since $S(s)$ can involve negative powers of s. The following lemma claims, however, that we may restrict ourselves to a unimodular transformation of the form (30) if we do not care about the upper off-diagonal blocks.

LEMMA 3.1. *There exists a unit lower-triangular polynomial matrix $L(s)$ over $\mathbf{K}[s]$ such that*

$$\hat{D}(s;t) = P_{\mathrm{r}} \begin{pmatrix} L(s) & O \\ O & I \end{pmatrix} \tilde{D}(s;t) P_{\mathrm{c}}$$

is in the same block-triangular form as a CCF $\bar{D}(s;t)$ *of $\tilde{D}(s;t)$, and that the diagonal blocks of $\hat{D}(s;t)$ coincide with those of $\bar{D}(s;t)$. We can choose $L(s)$ in the form*

$$L(s) = \mathrm{diag}\,[s^{r_1}, \cdots, s^{r_m}] \cdot L \cdot \mathrm{diag}\,[s^{-r_1}, \cdots, s^{-r_m}]$$

with a unit lower-triangular matrix L over $\mathbf{K}$. □

To prove this lemma we will briefly review the construction of CCF. The matrix S in (28), which is most important, is obtained through Gauss–Jordan-type row-wise elimination operations on $\tilde{Q}(1)$ as follows. We fix an arbitrary ordering of $R_Q = \mathrm{Row}(\tilde{Q})$ and set $\tilde{Q}^{(0)} = \tilde{Q}(1)$.

First we find a basis of the row vectors of the submatrix $\tilde{Q}^{(0)}[R_Q, C_0]$ by collecting independent row vectors according to the fixed ordering of R_Q. That is, let R_{Q0} be

the subset of R_Q that minimizes $\max(R_{Q0})$ (the maximum row index in R_{Q0} with respect to the ordering of R_Q) subject to

$$|R_{Q0}| = \operatorname{rank} \tilde{Q}^{(0)}[R_{Q0}, C_0] = \operatorname{rank} \tilde{Q}^{(0)}[R_Q, C_0].$$

Then the row vectors of $\tilde{Q}^{(0)}[R_Q - R_{Q0}, C_0]$ can be expressed as linear combinations of those of $\tilde{Q}^{(0)}[R_{Q0}, C_0]$. Hence, for a unit lower-triangular matrix $L^{(0)}$, the modified matrix $\tilde{Q}^{(1)} = L^{(0)}\tilde{Q}^{(0)}$ satisfies

$$\tilde{Q}^{(1)}[R_Q - R_{Q0}, C_0] = O.$$

Next, let R_{Q1} be the subset of $R_Q - R_{Q0}$ that minimizes $\max(R_{Q1})$ subject to

$$|R_{Q1}| = \operatorname{rank} \tilde{Q}^{(1)}[R_{Q1}, C_1] = \operatorname{rank} \tilde{Q}^{(1)}[R_Q - R_{Q0}, C_1].$$

Then, for another unit lower-triangular matrix $L^{(1)}$, the matrix $\tilde{Q}^{(2)} = L^{(1)}\tilde{Q}^{(1)}$ satisfies

$$\tilde{Q}^{(2)}[R_Q - (R_{Q0} \cup R_{Q1}), C_1] = O.$$

Continuing in this way, we find a sequence of unit lower-triangular matrices

$$L^{(0)}, L^{(1)}, \cdots, L^{(b)}$$

such that

$$\hat{Q} = L^{(b)} \cdots L^{(1)} L^{(0)} \tilde{Q}^{(0)} = L\tilde{Q}(1)$$

satisfies

$$\hat{Q}[R_{Qk}, C_l] = O \quad \text{if } 0 \le l < k \le \infty,$$

where

$$L = L^{(b)} \cdots L^{(1)} L^{(0)}$$

is a unit lower-triangular matrix over $\mathbf{K}$.

To obtain the CCF $\bar{D}(1;t)$ of $\tilde{D}(1;t)$, we further eliminate the blocks of $\hat{Q}[R_{Qk}, C_l]$ for $k < l$ as far as possible by premultiplying $\hat{Q}$ with a unit upper block-triangular matrix U over $\mathbf{K}$. Then the matrix S of (28) is given by

$$S = UL.$$

As for the T-part, we define the partition $\{R_{T0}; R_{T1}, \cdots, R_{Tb}; R_{T\infty}\}$ of $\operatorname{Row}(\tilde{T})$ by

$$R_{Tk} = Y_{Tk} - Y_{T,k-1} \quad \text{for } k = 0, 1, \cdots, b, \infty,$$

where

$$Y_{Tk} = \bigcup_{l=0}^{k} \{i \in \operatorname{Row}(\tilde{T}) \mid \tilde{T}_{ij} \neq 0,\ j \in C_l\} \quad \text{for } k = 0, 1, \cdots, b, \infty,$$

and we set $Y_{T,-1} = \emptyset$ by convention. Then the partition $\{R_0; R_1, \cdots, R_b; R_\infty\}$ of $R = \mathrm{Row}(\tilde{D})$ in its CCF is given by $R_k = R_{Qk} \cup R_{Tk}$. The permutation matrices P_r and P_c are introduced to bring the CCF into an explicit block-triangular form.

Based on the construction of $\bar{D}(s;t)$ described above, we can prove Lemma 3.1. Define

$$L(s) = \mathrm{diag}\,[s^{r_1}, \cdots, s^{r_m}] \cdot L \cdot \mathrm{diag}\,[s^{-r_1}, \cdots, s^{-r_m}] \tag{31}$$

and

$$\hat{D}(s;t) = P_\mathrm{r} \begin{pmatrix} L(s) & O \\ O & I \end{pmatrix} \begin{pmatrix} \tilde{Q}(s) \\ \tilde{T}(s;t) \end{pmatrix} P_\mathrm{c}. \tag{32}$$

By the construction above, $\hat{D}(s;t)$ thus defined has the same block-triangular form and the same diagonal blocks as the CCF $\bar{D}(s;t)$; i.e.,

$$\hat{D}(s;t)[R_k, C_l] = O \quad \text{if } 0 \le l < k \le \infty, \tag{33}$$

$$\hat{D}(s;t)[R_k, C_k] = \bar{D}(s;t)[R_k, C_k] \quad \text{for } k = 0, l, \cdots, b, \infty. \tag{34}$$

Note that $\hat{D}(s;t)$ and $\bar{D}(s;t)$ do not coincide in the upper off-diagonal blocks.

In the construction of CCF we have assumed that an ordering of R_Q is given arbitrarily. Now we choose this ordering with reference to r_i of (27) in such a manner that

$$r_1 \le r_2 \le \cdots \le r_m. \tag{35}$$

Since L is lower-triangular, this ordering guarantees that $L(s)$ of (31) be a polynomial matrix in s. Thus Lemma 3.1 is established.

Lemma 3.1 shows that $\tilde{D}(s;t)$ and $\hat{D}(s;t)$ share the same Smith form, since they are connected by a unimodular transformation. On the other hand, (17) shows that the Smith form $\Sigma(s)$ of $D(s)$ is embedded in the Smith form of $\tilde{D}(s;1)$. We will show here that $\tilde{D}(s;t)$ and $\tilde{D}(s;1)$ are related by "scaling" and their Smith forms are essentially identical. This will establish a link among the Smith forms, which may be schematically displayed as

$$D(s) \overset{(17)}{\longleftrightarrow} \tilde{D}(s;1) \overset{(37)}{\longleftrightarrow} \tilde{D}(s;t) \overset{\text{Lemma 3.1}}{\longleftrightarrow} \hat{D}(s;t). \tag{36}$$

We write $\tilde{D}(s;t,\mathcal{T})$ for $\tilde{D}(s;t)$ to explicitly indicate its dependence on the coefficients $\mathcal{T}$ in $T_D(s)$. By the definition (16) we see

$$\tilde{D}(s;t,\mathcal{T}) = \begin{pmatrix} I_m & O \\ O & \mathrm{diag}\,[t_1, \cdots, t_m] \end{pmatrix} \tilde{D}(s;1,\mathcal{T}/t), \tag{37}$$

where the expression $\mathcal{T}/t$ in the last factor means the substitution of $\tilde{t}/t_i$ for $\tilde{t}$ if $\tilde{t} \in \mathcal{T}$ is contained in row i of $\tilde{D}(s;1,\mathcal{T})$. This means that the determinantal divisors of $\tilde{D}(s;t,\mathcal{T})$ and $\tilde{D}(s;1,\mathcal{T}/t)$, as polynomials in s over $\mathbf{F}$, are identical. Therefore, the Smith form of $\tilde{D}(s;1,\mathcal{T})$ is obtained from that of $\tilde{D}(s;t,\mathcal{T})$ by setting $t_1 = \cdots = t_m = 1$, and conversely, the Smith form of $\tilde{D}(s;t,\mathcal{T})$ is obtained from that of $\tilde{D}(s;1,\mathcal{T})$ by replacing $\tilde{t}$ with $\tilde{t}/t_i$ if $\tilde{t} \in \mathcal{T}$ is contained in row i.

Based on (36) we may concentrate on the Smith form of $\hat{D}(s;t)$. Regarding $\hat{D}(s;t) = \hat{D}(s;t,\mathcal{T})$ as a matrix over the ring $\mathbf{K}[s,t,\mathcal{T}]$, we denote by $\hat{d}_k(s;t)$ ($\in \mathbf{K}[s,t,\mathcal{T}]$) the kth determinantal divisor of $\hat{D}(s;t)$ for $k = 1,\cdots,r+m$. Then

$$d_k(s) = \alpha_k \cdot \hat{d}_{k+m}(s;1) \quad \text{for } k = 1,\cdots,r, \tag{38}$$

where α_k ($\in \mathbf{K}(\mathcal{T}) \subseteq \mathbf{F}$) is introduced since the determinantal divisor $d_k(s)$ was defined to be a monic polynomial in $\mathbf{F}[s]$. Since $\hat{D}$ is in the block-triangular form (33) with full-rank diagonal blocks (cf. (34) and Theorem 2.3(c)), we have

$$r+m = \operatorname{rank} \hat{D} = \sum_{l=0}^{b} |R_l| + |C_\infty|,$$

and therefore a nonvanishing minor of $\hat{D}$ of order $r+m$ is expressed as

$$\begin{aligned} &\det \hat{D}[R_0,J] \cdot \det \hat{D}[I,C_\infty] \cdot \prod_{l=1}^{b} \det \hat{D}[R_l,C_l] \\ &= \det \bar{D}[R_0,J] \cdot \det \bar{D}[I,C_\infty] \cdot \prod_{l=1}^{b} \det \bar{D}[R_l,C_l] \end{aligned} \tag{39}$$

for some $J \subseteq C_0$ and $I \subseteq R_\infty$. Then Theorem 3.2 follows from (38) and (39) and the lemma below.

LEMMA 3.2.

$$\begin{aligned} \gcd\{\det \bar{D}(s)[R_0,J] \mid |J| = |R_0|,\ J \subseteq C_0\} &= \alpha_0 \cdot s^{p_r^0}, \\ \gcd\{\det \bar{D}(s)[I,C_\infty] \mid |I| = |C_\infty|,\ I \subseteq R_\infty\} &= \alpha_\infty \cdot s^{p_r^\infty}, \end{aligned}$$

where $\alpha_0, \alpha_\infty \in \mathbf{F}$, *and*

$$\begin{aligned} p_r^0 &= \min\{\operatorname{ord} \det \bar{D}(s)[R_0,J] \mid |J| = |R_0|,\ J \subseteq C_0\}, \\ p_r^\infty &= \min\{\operatorname{ord} \det \bar{D}(s)[I,C_\infty] \mid |I| = |C_\infty|,\ I \subseteq R_\infty\}. \end{aligned}$$

Proof. Regarding $\bar{D}(s)[R_0,C_0]$ as an irreducible LM-matrix with respect to $\mathbf{K}(s)$, we obtain the first identity from Theorem 2.5(b). The second follows similarly from Theorem 2.5(c). □

Finally, Theorem 3.1 follows from (38), (39), Theorem 3.2, and the lemma below. Note that the essential claim of Theorem 3.1 is that $d_{r-1}(s)$ is a monomial in s over $\mathbf{F}$.

LEMMA 3.3. *For* $l = 1,\cdots,b$,

$$\gcd\{\det \bar{D}(s)[I,J] \mid |I| = |J| = |R_l| - 1,\ I \subseteq R_l,\ J \subseteq C_l\}$$

is a monomial in s *over* $\mathbf{F}$.

Proof. Regarding $\bar{D}(s)[R_l,C_l]$ as an irreducible LM-matrix with respect to $\mathbf{K}(s)$, we can apply Theorem 2.5(a) to establish the claim. □

Example 3.2. The derivations above are illustrated here for the 5×5 matrix $D(s)$ considered in Example 3.1. First note that (35) is satisfied. The CCF $\bar{D}(s;t)$, given

in (26), of $\tilde{D}(s;t)$ is obtained by means of $S(s) = U(s)L(s)$ with

$$L(s) = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \begin{array}{c} \begin{array}{ccccc} w_1 & w_2 & w_3 & w_4 & w_5 \end{array} \\ \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & -2s & 1 & & \\ & & & 1 & \\ & -2s^2 & & & 1 \end{pmatrix} \end{array}, \tag{40}$$

$$U(s) = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{array} \begin{array}{c} \begin{array}{ccccc} w_1 & w_2 & w_3 & w_4 & w_5 \end{array} \\ \begin{pmatrix} 1 & & & -3/s & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{pmatrix} \end{array}. \tag{41}$$

Note that $L(s)$ is a unimodular polynomial matrix in s over $\mathbf{Q}$. Using $L(s)$ we obtain the block-triangular matrix $\hat{D}(s;t)$ of (32) as follows:

$$\hat{D}(s;t) = \begin{array}{c} \begin{array}{cccccccccc} x_1 & x_2 & w_1 & x_3 & x_4 & w_4 & w_3 & w_5 & w_2 & x_5 \end{array} \\ \left(\begin{array}{cc|cc|c|c|ccc|c} s & 1 & & 1 & & & & & 1 & \\ \hline & & 1 & 1 & 3s & & & & & \\ & & -t_1 & p_1 s & & & & & & p_2 \\ \hline & & & & s^2 & 1 & & & & \\ & & & & & -t_4 & & & & p_6 \\ \hline & & & & & & 1 & 0 & -2s & 0 \\ & & 0 & & & & 0 & 1 & -2s^2 & s \\ & & & & & & -t_3 & 0 & 0 & p_5 s \\ & & & & & & 0 & -t_5 & 0 & p_7 s^2 \\ & & & & & & 0 & 0 & -t_2 & p_3 + p_4 s \end{array}\right) \end{array}. \tag{42}$$

As claimed, the matrix $\hat{D}$ is a polynomial matrix in s and it agrees with $\bar{D}$ in the diagonal blocks. Also, notice the difference between the zero/nonzero structures of $\bar{D}$ and $\hat{D}$. In particular, we can exchange the positions of the two blocks $\{w_1, x_3\}$ and $\{x_4\}$ in $\bar{D}$ without destroying the block-triangular structure if we accordingly exchange the corresponding rows, whereas these two blocks must be arranged in this order in $\hat{D}$ to make it into a block-triangular form. In other words, the square diagonal blocks are partially ordered as

$$\{w_1, x_3\} \prec \{w_4\}, \qquad \{x_4\} \prec \{w_4\}$$

with respect to the zero/nonzero structure in $\bar{D}$, whereas they are totally ordered as

$$\{w_1, x_3\} \prec \{x_4\} \prec \{w_4\}$$

in $\hat{D}$.

4. Algorithms. This section describes efficient matroid-theoretic algorithms for computing the Smith form of $D(s)$ on the basis of Theorems 3.1 and 3.2.

First we point out that the degree p_k of $d_k(s)$ given in Theorem 3.1 can be computed by solving an independent assignment (or equivalently, weighted matroid-intersection) problem [6], [21]. This is an obvious adaptation of the result of [17],

in which the maximum degree, instead of the minimum order in (14), of kth order minors is investigated.

We associate with $D(s) = Q_D(s) + T_D(s)$ a bipartite graph $G = G(D) = (V, E)$ having vertex bipartition $V = V^+ \cup V^-$ with

$$V^+ = R_T \cup R_Q \cup C_Q, \qquad V^- = R \cup C,$$

where R_T and R_Q are disjoint copies of $R = \mathrm{Row}(D)$, and C_Q is a disjoint copy of $C = \mathrm{Col}(D)$. By $\varphi_Q : R \cup C \to R_Q \cup C_Q$ and $\varphi_T : R \to R_T$ we will denote the natural correspondences between the copies. The edge set E is defined by

$$\begin{aligned} E = &\{(\varphi_Q(i), i) \mid i \in R\} \cup \{(\varphi_Q(j), j) \mid j \in C\} \\ &\cup \{(\varphi_T(i), i) \mid i \in R\} \cup \{(\varphi_T(i), j) \mid T_{ij}(s) \not\equiv 0, i \in R, j \in C\}, \end{aligned}$$

where $T_{ij}(s)$ means the (i, j) entry of $T_D(s)$.

We introduce matroid structures on V^+ and V^-. First define

$$\begin{aligned} \mathcal{L} &= \{(I, J) \mid Q_D[I, J] \text{ is nonsingular}, I \subseteq R, J \subseteq C\}, \\ \mathcal{L}_Q &= \{(\varphi_Q(I), \varphi_Q(J)) \mid (I, J) \in \mathcal{L}\}, \end{aligned}$$

and then consider two families of subsets of V^+ and V^-, respectively, defined as

$$\begin{aligned} \mathcal{B}_k^+ &= \{U^+ \subseteq V^+ \mid (R_Q - U^+, C_Q \cap U^+) \in \mathcal{L}_Q, |U^+| = |R| + k\}, \\ \mathcal{B}_k^- &= \{U^- \subseteq V^- \mid U^- \supseteq R, |U^-| = |R| + k\}. \end{aligned}$$

These two families define matroids on V^+ and V^-, respectively. To be more precise, $\mathcal{B}_k^+$ forms the base family of a matroid of rank $|R| + k$ which is the direct sum of a linear matroid of rank $|R|$ representable over $\mathbf{K}$ and a uniform matroid of rank k; and $\mathcal{B}_k^-$ forms the base family of another matroid of rank $|R| + k$ which is the direct sum of a uniform matroid of rank k and a free matroid of rank $|R|$.

A matching M in G is called an independent assignment if $\partial^+ M \in \mathcal{B}_k^+$ and $\partial^- M \in \mathcal{B}_k^-$, where $\partial^+ M$ (respectively, $\partial^- M$) denotes the set of end vertices of M in V^+ (respectively, V^-).

In addition, we will introduce a weight function $\zeta : E \to \mathbf{Z}$ with reference to the degrees of the entries of $Q_D(s)$ and $T_D(s)$. Using the numbers r_i and c_j in (11), we define, for $e \in E$,

$$\zeta(e) = \begin{cases} -r_i & \text{if } e = (\varphi_Q(i), i), \quad i \in R, \\ -c_j & \text{if } e = (\varphi_Q(j), j), \quad j \in C, \\ \mathrm{ord}\, T_{ij}(s) & \text{if } e = (\varphi_T(i), j), \quad i \in R,\ j \in C, \\ 0 & \text{if } e = (\varphi_T(i), i), \quad i \in R, \end{cases} \tag{43}$$

where $\mathrm{ord}\, T_{ij}(s)$ denotes the order of $T_{ij}(s)$, i.e., the minimum degree of a nonvanishing term in $T_{ij}(s)$.

We now give a combinatorial characterization of p_k in terms of an independent assignment problem.

THEOREM 4.1. *The degree of the kth determinantal divisor $d_k(s)$ (for $k \leq r - 1$) is given by*

$$p_k = \min_M \zeta(M) + \sum_{i \in R} r_i \quad \text{for } k = 1, \cdots, r - 1,$$

where the minimum on the right-hand side is taken over all independent assignments M in $G(D)$ with matroids defined by $\mathcal{B}_k^+$ and $\mathcal{B}_k^-$.

Proof. It follows from the result of [17] that the right-hand side above is equal to the right-hand side of (14) in Theorem 3.1. □

Now we describe an efficient algorithm for computing the rth determinantal divisor $d_r(s)$ of $D(s)$ given in Theorem 3.2. From the argument in §3.2 the following procedure suggests itself.

1. Find the partition $\{C_0; C_1, \cdots, C_b; C_\infty\}$ of $\mathrm{Col}(\tilde{D})$ in the CCF of the augmented matrix $\tilde{D}$ of (16) associated with $D(s)$ (by the algorithm of [10], [16]).
2. Find the lower-triangular matrix L over $\mathbf{K}$ with reference to the ordering (35) (by means of a variant of Gaussian elimination described in §3.2).
3. Compute the diagonal blocks $\bar{D}_k = \bar{D}[R_k, C_k] = \hat{D}[R_k, C_k]$ $(k = 0, 1, \cdots, b, \infty)$ of the CCF of $\tilde{D}$ (according to Lemma 3.1).
4. Compute p_r^0 and p_r^∞ in Lemma 3.2 (by applying Theorem 4.1) and set

$$p_r = p_r^0 + p_r^\infty.$$

Then

$$d_r(s) = \alpha_r \cdot s^{p_r} \prod_{l=1}^{b} \det \bar{D}_l(s; 1)$$

(cf. (18)).

The combinatorial characterizations established in Theorem 4.1 and the procedure above provide us with an efficient and practical way of computing the Smith form of $D(s)$ by means of well-established algorithms for the independent assignment problem. See also [3] for a recently developed faster algorithm for this problem.

Note also that $\deg d_r(s)$ can be determined from Theorem 3.2 by further computing $\deg \det \bar{D}_k$ for $k = 1, \cdots, b$ by the algorithm of [9], [10], and [17].

5. Conclusion. In this paper we have revealed the simple structure of the Smith normal form of a "structured" polynomial matrix. The proposed algorithm for computing the Smith form gives an alternative way of testing for the structural controllability of a control system (1) if it is applied to $D(s) = (A - sF \mid B)$, as already mentioned in the Introduction. The result obtained in this paper will find applications to many other control-theoretic problems. For example, it reveals some nice properties of the Smith–McMillan form of the transfer matrix of a structured descriptor system.

Acknowledgments. The author thanks Professor S. Shin of the University of Tsukuba for indicating relevant literature and A. Sugimoto for pointing out flaws in an earlier version of this paper.

REFERENCES

[1] A. L. Dulmage and N. S. Mendelsohn, *A structure theory of bipartite graphs of finite exterior dimension*, Trans. Roy. Soc. Canada (3), 53 (1959), pp. 1–13.

[2] J. Edmonds, *Submodular functions, matroids and certain polyhedra*, in Combinatorial Structures and Their Applications, R. Guy, H. Hanai, N. Sauer, and J. Schönheim, eds., Gordon and Breach, 1970, pp. 69–87.

[3] H. N. GABOW AND Y. XU, *Efficient theoretic and practical algorithms for linear matroid intersection problems*, CU-CS-424-89, Department of Computer Science, University of Colorado, Boulder, CO, 1989.
[4] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1959.
[5] M. IRI, *Applications of matroid theory*, in Mathematical Programming—State of the Art, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, New York, 1983, pp. 158–201.
[6] E. L. LAWLER, *Combinatorial Optimization: Networks and Matroids*, Holt, Rinehart and Winston, New York, 1976.
[7] C. -T. LIN, *Structural controllability*, IEEE Trans. Automat. Control, 19 (1974), pp. 201–208.
[8] L. LOVÁSZ AND M. PLUMMER, *Matching Theory*, North-Holland, Amsterdam, 1986.
[9] K. MUROTA, *Use of the concept of physical dimensions in the structural approach to systems analysis*, Japan J. Appl. Math., 2 (1985), pp. 471–494.
[10] ———, *Systems Analysis by Graphs and Matroids—Structural Solvability and Controllability*, Algorithms and Combinatorics, Vol. 3, Springer-Verlag, Berlin, New York, 1987.
[11] ———, *Refined study on structural controllability of descriptor systems by means of matroids*, SIAM J. Control Optim., 25 (1987), pp. 967–989.
[12] ———, *A matroid-theoretic approach to structurally fixed modes of control systems*, SIAM J. Control Optim., 27 (1989), pp. 1381–1402.
[13] ———, *On the irreducibility of layered mixed matrices*, Linear and Multilinear Algebra, 24 (1989), pp. 273–288.
[14] ———, *Some recent results in combinatorial approaches to dynamical systems*, Linear Algebra Appl., 122/123/124 (1989), pp. 725–759.
[15] K. MUROTA AND M. IRI, *Structural solvability of a system of equations—a mathematical formulation for distinguishing accurate and inaccurate numbers in structural analysis of systems*, Japan J. Appl. Math., 2 (1985), pp. 247–271.
[16] K. MUROTA, M. IRI, AND M. NAKAMURA, *Combinatorial canonical form of layered mixed matrices and its application to block-triangularization of systems of equations*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 123–149.
[17] K. MUROTA AND J. VAN DER WOUDE, *Structure at infinity of structured descriptor systems and its applications*, SIAM J. Control Optim., 29 (1991), pp. 878–894.
[18] M. NEWMAN, *Integral Matrices*, Academic Press, New York, 1972.
[19] O. ORE, *Graphs and matching theorems*, Duke Math. J., 22 (1955), pp. 625–639.
[20] H. H. ROSENBROCK, *State-Space and Multivariable Theory*, Nelson, London, 1970.
[21] D. J. A. WELSH, *Matroid Theory*, Academic Press, New York, 1976.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 4, pp. 766–789, October 1991

013

MINIMAX POLYNOMIAL PRECONDITIONING FOR HERMITIAN LINEAR SYSTEMS*

STEVEN F. ASHBY†

Abstract. This paper explores the use of polynomial preconditioning for Hermitian positive definite and indefinite linear systems $Ax = b$. Unlike preconditioners based on incomplete factorizations, polynomial preconditioners are easy to employ and well suited to vector and/or parallel architectures. It is shown that any polynomial iterative method may be used to define a preconditioning polynomial, and that the optimum polynomial preconditioner is obtained from a minimax approximation problem. A variety of preconditioning polynomials are then surveyed, including the Chebyshev, de Boor and Rice, Grcar, and bilevel polynomials. Adaptive procedures for each of these polynomials are also discussed, and it is shown that the new bilevel polynomial is particularly well suited for use in adaptive CG algorithms.

Key words. conjugate gradient methods, polynomial preconditioning, minimax approximation, adaptive procedure

AMS(MOS) subject classifications. 65F10, 41A10

1. Introduction.

This paper surveys recent and ongoing research in polynomial preconditioning for Hermitian linear systems $Ax = b$. Such systems arise in many scientific applications. For example, the matrices resulting from finite difference and finite element methods are often Hermitian positive definite (hpd), whereas the matrices arising in the numerical solution of Stokes flow and constrained minimization problems are typically Hermitian indefinite (hid). Since these matrices are usually large and sparse, an iterative method is required. If A is hpd, the classical conjugate gradient method of Hestenes and Stiefel [26] is applicable. This method, which we call CGHS, minimizes the A-norm of the error over a Krylov subspace. If A is hid, CGHS is not applicable, but the conjugate residual method is. This method minimizes the Euclidean norm of the residual over a Krylov subspace.

Unfortunately, these CG methods may converge slowly if A is ill-conditioned, in which case a preconditioner is needed. Preconditioners based on incomplete factorizations [10], [20], [30] are particularly popular, and especially effective on scalar machines. On more advanced vector and vector/parallel machines, however, these preconditioners usually perform poorly because of their sequential nature. To obtain efficient preconditioned CG methods for these new architectures, we will employ *polynomial preconditioning.* That is, we will apply a CG method to

$$C(A)Ax = C(A)b \tag{1.1}$$

where $C(\lambda)$ is a *preconditioning polynomial* and $C(A)$ is the associated *polynomial preconditioner.* We will assume that $C(\lambda)$ has real coefficients, in which case both $C(A)$ and $C(A)A$ are Hermitian. The matrix A is assumed to be nonsingular.

Polynomial preconditioning has several advantages. First, it is simple: there are only two intrinsic operations, matrix-vector multiplication (*matvec*) and vector

* Received by the editors November 1, 1989; accepted for publication (in revised form) November 12, 1990. This paper was presented at the Symposium on Sparse Matrices at Salishan Resort, Gleneden Beach, Oregon, May 22-24, 1989, which was sponsored by the SIAM Activity Group on Linear Algebra. This work was supported in part by the Applied Mathematical Sciences subprogram of the Office of Energy Research, Department of Energy, under contract W-7405-ENG-48.

† Computing and Mathematics Research Division, Lawrence Livermore National Laboratory, P.O. Box 808, Livermore, California 94551 (ashby@llnl.gov).

addition (*saxpy*). The user need only specify the polynomial degree and initialize a few parameters; the preconditioning may be implemented automatically. There is no complicated programming as with an incomplete factorization, nor is there any expensive preprocessing. Since polynomial preconditioning requires only matrix-vector multiplication, it is ideally suited to "matrix-free" computations [9].

Polynomial preconditioning is also versatile, the key to which is commutativity: a polynomial in A commutes with A. Moreover, it is possible to choose the polynomial so that the preconditioned matrix, $C(A)A$, is hpd. In particular, it is possible to transform an indefinite matrix A into a positive definite $C(A)A$. This makes practicable several CG methods; see [4], [6], and [7]. Polynomial preconditioning also may be combined with other preconditionings. For example, if an incomplete factorization is effective, it can be further accelerated with a polynomial preconditioner. Specifically, one applies CG to

$$C(M^{-1}A)M^{-1}Ax = C(M^{-1}A)M^{-1}b \tag{1.2}$$

where M represents the incomplete factorization.

The main advantage of polynomial preconditioning is its suitability for vector and/or parallel architectures. If the matvec is vectorizable, as when A has a regular sparsity pattern, polynomial preconditioning is effective on vector machines [5], [6], [16], [27], [28], [31]. In contrast, incomplete factorizations are difficult to vectorize, especially for the nonexpert. It is also possible to chain the matvecs implicit in the preconditioning, thereby enhancing data locality and reducing memory traffic [12], [13], [38], [41]. Polynomial preconditioning is also effective on parallel machines [11], [31], especially those on which inner products are a bottleneck. This is so because polynomial preconditioned CG methods converge in fewer steps than unpreconditioned CG, and thus compute fewer inner products, albeit at the cost of several matvecs per step instead of one. However, in many applications the matvec is parallelizable, and so we can expect an overall reduction in CPU time by substituting matvecs for inner products on some architectures.

1.1. Purpose of paper. Since the effectiveness of polynomial preconditioning has been demonstrated in a variety of recent papers [1], [2], [4], [5], [6], [11], [16], [27], [28], [31], [37], [38], we will not present numerical results. Instead, this paper endeavors to explain the design of optimum preconditioning polynomials for Hermitian linear systems. Our main purpose is to survey various minimax preconditioning polynomials, discuss their relative merits, and describe an adaptive procedure for each. (See [4] and [7] for a discussion of the many ways in which polynomial preconditioning can be used in CG methods.) We will first show (§ 2) that any residual polynomial, and hence any polynomial iterative method, may be used to define a preconditioning polynomial. To obtain an *optimum* $C(\lambda)$ we will consider a minimax approximation problem. The resulting preconditioner is optimum in that it minimizes a bound on the condition number of the preconditioned matrix $C(A)A$. In the hpd case (§ 3), $C(\lambda)$ is obtained from a scaled and translated Chebyshev polynomial; in the hid case (§ 4), it is obtained from the de Boor and Rice (DR) residual polynomial [15]. We will also discuss the related Grcar preconditioning polynomial. We then introduce (§ 5) the *bilevel* polynomial for hid A and contrast it with the DR polynomial. Since each of these polynomials requires estimates for the extreme eigenvalues of A, *adaptive procedures* for dynamically determining them are needed. Adaptive procedures are described at length in [4]–[6] for the Chebyshev, DR, and Grcar polynomials, and so we will discuss only their salient features. We then propose an adaptive procedure for

the bilevel polynomial and discuss its advantages over those for the DR and Grcar polynomials.

Note that much of this material first appeared in [4].

2. Preconditioning polynomials. In this section we show that any residual polynomial, and therefore any polynomial iterative method, may be used to define a preconditioning polynomial $C(\lambda)$. Moreover, we show that the preconditioning may be effected by executing m steps of the polynomial iterative method. This leads to an inner/outer formulation for polynomial preconditioned CG methods. To obtain an *optimum* preconditioning polynomial we consider a minimax approximation problem; the polynomial is optimum in that it minimizes a bound on the condition number of the preconditioned matrix $C(A)A$. This result is obtained for both Hermitian positive definite and indefinite matrices A.

2.1. Richardson's method. To illustrate the connection between residual and preconditioning polynomials we first consider Richardson's method [3], [15], [23], [29], [35]. Let x_0 be an initial guess vector and let $\tau_0, \tau_1, \ldots$ be nonzero iteration parameters. *Richardson's iteration* is defined by

$$r_j = b - Ax_j \tag{2.1}$$

$$x_{j+1} = x_j + \tau_j r_j. \tag{2.2}$$

In practice, only m iteration parameters are used and the iteration is restarted every m steps. This yields a *cyclic* iteration called *Richardson's method*, the *period* of which is m. Of course, to execute the method one must have the τ_j. These parameters may be obtained from the roots of a residual polynomial, as now shown.

The residual at step m, $r_m = b - Ax_m$, is first expressed in terms of the desired iteration parameters. Equation (2.2) yields $r_m = (I - \tau_{m-1}A)r_{m-1}$, and thus

$$r_m = \prod_{j=0}^{m-1} (I - \tau_j A) r_0. \tag{2.3}$$

If we introduce the polynomial

$$R_m(\lambda) = \prod_{j=0}^{m-1} (1 - \tau_j \lambda), \tag{2.4}$$

equation (2.3) may be rewritten as

$$r_m = R_m(A) r_0. \tag{2.5}$$

Note that $R_m(0) = 1$. Such a polynomial is called a *residual polynomial*. Any residual polynomial defines a Richardson method: the τ_j are the reciprocals of the roots of $R_m(\lambda)$. The goal is to find that $R_m(\lambda)$ for which Richardson's method converges most rapidly in some sense.

One way to obtain optimum convergence is to make $\|r_j\|_2$ as small as possible. Although the method of conjugate residuals minimizes $\|r_j\|_2$ over a Krylov subspace, it does so by dynamically computing a different residual polynomial at each step. In the present context, the polynomial is sought a priori. If A is Hermitian, it is possible to minimize a *bound* on $\|r_m\|_2/\|r_0\|_2$ *after* one period. To see this, consider the Euclidean norm of (2.5); it gives

$$\frac{\|r_m\|_2}{\|r_0\|_2} \leq \|R_m(A)\|_2 = \rho(R_m(A)) \tag{2.6}$$

where $\rho(G)$ is the spectral radius of the matrix G. Next let S be a compact set, and for f a continuous function on S, define

$$\|f\|_S = \max_{\lambda \in S} |f(\lambda)|. \tag{2.7}$$

This norm is called the *uniform norm*; note its dependence on the set S. If the spectrum of A, denoted $\sigma(A)$, lies in S, then $\rho(R_m(A)) \leq \|R_m\|_S$, and (2.6) gives

$$\frac{\|r_m\|_2}{\|r_0\|_2} \leq \|R_m\|_S. \tag{2.8}$$

If $\|R_m\|_S < 1$, Richardson's method converges. Moreover, the "smaller" S is, the more rapid the convergence.

The *minimax residual polynomial* is defined to be that residual polynomial for which $\|R_m\|_S$ is minimized. It is the solution of the following minimax approximation problem:

$$\min_{\substack{R \in \pi_m \\ R(0)=1}} \|R\|_S = \min_{\substack{R \in \pi_m \\ R(0)=1}} \max_{\lambda \in S} |R(\lambda)| \tag{2.9}$$

where $\pi_m = \{p : p \text{ is a real polynomial of degree } m \text{ or less}\}$. In other words, the minimax residual polynomial is that residual polynomial deviating least from zero on S. The roots of this polynomial define an *optimum* Richardson's method in the sense that the bound (2.8) is minimized.

2.2. The minimax preconditioning polynomial. We will next show how one may use the minimax residual polynomial to define an optimum preconditioning polynomial. Specifically, we will show that the optimum $C(\lambda)$ is derived from the solution of a constrained minimax approximation problem, the error in which is a minimax residual polynomial. The derivation will illustrate the connection between residual and preconditioning polynomials. We begin with first principles.

We wish to chose C to accelerate convergence of the conjugate gradient method. One way of doing this is to choose $C(A) \approx A^{-1}$, for example, by choosing $C(\lambda) \approx \lambda^{-1}$ on some set $S \supset \sigma(A)$. Since A is Hermitian and nonsingular, we may assume that S is a subset of the real line that excludes the origin. Adopting the minimax definition of "$\approx$," we shall seek the polynomial that minimizes $\|1 - C(\lambda)\lambda\|_S$. That is, we seek the best polynomial approximation to 1 from among all polynomials of degree m or less having a root at zero. If $p_m(\lambda) \equiv C(\lambda)\lambda$ is this polynomial, the problem may be recast as a constrained minimax approximation problem:

$$\min_{\substack{p \in \pi_m \\ p(0)=0}} \|1 - p\|_S. \tag{2.10}$$

In other words, we wish to choose p_m so that the eigenvalues of the preconditioned matrix $p_m(A)$ are as tightly clustered around 1 as possible. In this way we expect to accelerate convergence of the CG method. Since $p_m(A)$ is the preconditioned matrix, we call $p_m(\lambda)$ the *preconditioned polynomial.*

Note that the error in (2.10), $e_m = 1 - p_m$, is a residual polynomial. Problems (2.9) and (2.10) are therefore equivalent, and so the minimax preconditioned polynomial may be obtained from the minimax residual polynomial. The associated preconditioning polynomial is given by

$$C(\lambda) = \frac{1 - e_m(\lambda)}{\lambda}, \tag{2.11}$$

which is indeed a polynomial in λ because $e_m(0) = 1$. It may be shown that this $C(\lambda)$ is optimum in that it minimizes a bound on the condition number of $p_m(A)$; see §§ 3 and 4. It also minimizes [40] the *relative* error in the related approximation problem $\min_{C \in \pi_{m-1}} \|\lambda^{-1} - C\|_S$.

The minimax preconditioning polynomial is attractive for several reasons. First, its behavior is well understood. For example, the minimax preconditioned polynomial $p_m(\lambda)$ equioscillates about 1 over the set S. This means that the preconditioning polynomial $C(\lambda)$ is unbiased in its suppression of the error: no portion of the set S is preferred over another. Second, if $\sigma(A) \subset S$, then $\sigma(p_m(A)) \subset [1 - \epsilon_m, 1 + \epsilon_m]$, where $\epsilon_m = \|1 - p_m\|_S$. Since $\epsilon_m < 1$, the minimax preconditioned matrix, $p_m(A)$, is hpd—even if the original matrix A is indefinite. This makes possible a variety of CG methods [7]. Also note that the spectral condition number of $p_m(A)$, $\kappa(p_m(A))$, satisfies

$$\kappa(p_m(A)) \leq \frac{1 + \epsilon_m}{1 - \epsilon_m} \tag{2.12}$$

when $\sigma(A) \subset S$. This bound yields an estimate of the number of CG steps required for convergence. One needs approximately

$$\frac{\ln(\delta/2)}{\ln(CF)} \tag{2.13}$$

steps to reduce the error by an amount δ [25], where

$$CF = CF(p_m(A)) = \frac{\sqrt{\kappa(p_m(A))} - 1}{\sqrt{\kappa(p_m(A))} + 1} \tag{2.14}$$

is the *CG convergence factor* for the hpd matrix $p_m(A)$. If the eigenvalues of $p_m(A)$ are uniformly distributed throughout $[1 - \epsilon_m, 1 + \epsilon_m]$, then (2.13) is fairly accurate. For other properties of $p_m(\lambda)$, see §§ 3 and 4.

2.3. The need for adaptive procedures. Recall that the uniform norm is defined with respect to a compact set S, which we have assumed contains the spectrum of A, $\sigma(A)$. Unfortunately, such a set S is seldom known a priori, and so we need to determine it dynamically. We will do this by way of an *adaptive procedure*, which dynamically determines an $S \supset \sigma(A)$ by computing eigenvalue estimates for A from the CG iteration parameters. To make the task a bit easier, we will make some assumptions about the set S. Since A is Hermitian nonsingular, it is reasonable to take S to be the union of a finite number of disjoint closed intervals, each excluding the origin. In particular, if A is hpd, we will assume $S = [c, d]$, $0 < c \leq d$. If A is hid, we will assume $S = [a, b] \cup [c, d]$, $a \leq b < 0 < c \leq d$. These choices for S are attractive because only a few extreme eigenvalues of A are needed. Since the interior eigenvalues of an indefinite matrix are the most difficult to ascertain, we will pay special attention to the behavior of the DR, Grcar, and bilevel polynomials when b and/or c is inaccurate.

2.4. Inner/outer iterations. Equation (2.11) may be used to derive a preconditioning polynomial from *any* residual polynomial. This implies that any polynomial iterative method may be used to define a preconditioning polynomial. It also suggests an inner/outer formulation for the implementation of polynomial preconditioned CG methods: One uses an inner iteration to implement the preconditioning required in the outer CG iteration.

To elucidate, consider a *polynomial iterative method* for the linear system $A\tilde{x} = v$. Let $\tilde{x}_m$ be the mth iterate and let $\tilde{r}_m = v - A\tilde{x}_m$ be the corresponding residual. By definition, these residuals satisfy $\tilde{r}_m = R_m(A)\tilde{r}_0$, where R_m is a residual polynomial. (In general, a polynomial iterative method is defined by a family of residual polynomials, which may depend on the right-hand side vector v.) If we begin with an initial guess of $\tilde{x}_0 = 0$, we obtain $\tilde{r}_m = R_m(A)v$. Next let p_m and C be defined by $R_m(A) = I - p_m(A) = I - C(A)A$. This gives

$$\tilde{r}_m = R_m(A)v = (I - C(A)A)v \tag{2.15}$$

and

$$\tilde{e}_m = \tilde{x} - \tilde{x}_m = (A^{-1} - C(A))v, \tag{2.16}$$

which implies $\tilde{x}_m = C(A)v$. Suppose we now wish to use $C(A)$ as a preconditioner in a conjugate gradient method. To effect the preconditioning step in this *outer* iteration, we must compute $w = C(A)v$ for some vector v, usually the residual. If the polynomial iterative method is independent of the right-hand side vector, as are Chebyshev and SSOR, this may be done by carrying out m steps of the polynomial iterative method. The mth inner iterate is the desired preconditioned vector $C(A)v$. Note that we need only $m-1$ matrix-vector multiplications because the corresponding residual need not be computed. If the polynomial iterative method depends on the right-hand side vector, as do CG methods, each inner iteration will yield a different polynomial preconditioner, one that is dependent on the vector being preconditioned. This may or may not be allowable.

2.5. Other preconditioning polynomials. A simple preconditioning polynomial is based on the Neumann series. Let $A = M - N$ and consider

$$A^{-1} = (M - N)^{-1} = (I + G + G^2 + G^3 + \cdots)M^{-1} \tag{2.17}$$

where $G = M^{-1}N$. If $\rho(G) < 1$, the series converges. We obtain our approximation to A^{-1} by truncating the Neumann series [2], [11], [16], [28]. The advantage of this approach is its simplicity: there are no parameters to estimate. Unfortunately, it may yield a poor preconditioner. For example, one can usually do much better with the minimax preconditioning polynomial [28].

Although we have chosen to work in the uniform norm, other norms are of interest. For example, one might consider a weighted least squares norm, which is induced by the inner product

$$\langle f, g\rangle_w = \int_S f(\lambda)\overline{g(\lambda)}w(\lambda)d\lambda \tag{2.18}$$

where w is a nonnegative weight function (not identically equal to zero). The weighted least squares polynomial is that one minimizing $\|1 - p\|_w$. Since least squares polynomials are orthogonal, they satisfy a three-term recursion, which permits efficient and stable computations. Saad [37], [38], [39] has advocated these polynomials, as have Smolarski and Saylor [42].

Finally, we remark that polynomial preconditioning may be combined with other preconditionings. For example, if an incomplete factorization is effective, it can be further accelerated with a polynomial preconditioner. Specifically, one applies the CG method to

$$C(M^{-1}A)M^{-1}Ax = C(M^{-1}A)M^{-1}b \tag{2.19}$$

where M^{-1} represents the inner preconditioning. Note that if M and A are Hermitian, then so is the total preconditioner $C(M^{-1}A)M^{-1}$. Several CG methods are applicable under certain conditions; see [1], [2], and [7]. Observe that the m-step method of Adams [1], [2] may be viewed in this light: here M is the SSOR splitting and the preconditioning step is effected by carrying out m steps of SSOR.

3. The Chebyshev preconditioning polynomial. In this section we discuss the Chebyshev preconditioning polynomial for hpd matrices A. Unlike the other polynomials we will survey, this polynomial is explicitly known: it is obtained from a shifted and scaled Chebyshev polynomial. We will assume $\sigma(A) \subset S = [c, d]$, where $0 < c \leq d$ are given. Ideally, $c = \lambda_c$ and $d = \lambda_d$, the extreme eigenvalues of A, in which case the Chebyshev preconditioning polynomial is optimum. Since λ_c and λ_d are seldom known a priori, we need an adaptive procedure for dynamically determining them. Manteuffel has devised such a procedure and we will describe it. We remark that Rutishauser [36] was the first to propose Chebyshev polynomial preconditioning for the conjugate gradient method; his motive was to mitigate the rounding errors in CG. We advocate polynomial preconditioning because it is well suited to vector and/or parallel machines.

Recall that the minimax preconditioned polynomial $p_m(\lambda)$ is derived from the approximation problem (2.10). The solution to this problem is well known for $S = [c, d]$:

$$p_m(\lambda) = 1 - \frac{T_m(\frac{d+c-2\lambda}{d-c})}{T_m(\frac{d+c}{d-c})} \tag{3.1}$$

where $T_m(x)$ is the mth Chebyshev polynomial of the first kind. The associated Chebyshev preconditioning polynomial is given by $C(\lambda) = p_m(\lambda)/\lambda$. This polynomial may be evaluated via a three-term recursion, which is computationally efficient and stable. Thus, to implement the preconditioning, we need only take m steps of the Chebyshev iteration. Neither the powers of A nor the polynomial coefficients are formed explicitly.

3.1. Some properties. A Chebyshev preconditioned polynomial is illustrated in Fig. 3.1. Observe that it satisfies the interpolatory constraint, $p_m(0) = 0$, and equioscillates about 1 over the interval $[c, d]$. If $\sigma(A) \subset [c, d]$, then $\sigma(p_m(A)) \subset [1 - \epsilon_m, 1 + \epsilon_m]$, and so (2.12) holds. Moreover, ϵ_m is minimized (with respect to S) when $c = \lambda_c$ and $d = \lambda_d$.

From (3.1) we find $\epsilon_m = \|1 - p_m\|_S = |T_m^{-1}((d+c)/(d-c))|$, which is a monotonically increasing function of $\kappa = d/c$. Thus, the optimum Chebyshev preconditioning polynomial depends only on the condition number of A: if $\kappa(A_1) = \kappa(A_2)$, then $\kappa(p_m(A_1)) = \kappa(p_m(A_2))$. This is not true in the indefinite case, where the condition number is a poor indicator of the CG rate of convergence. The relationship between the condition numbers (i.e., convergence factors) of hpd A and $p_m(A)$ is illustrated in Figs. 3.2 and 3.3. In the first figure we plot the convergence factor (2.14) of $p_m(A)$ as a function of m for several $\kappa(A)$; in the second figure we plot $CF(p_m(A))$ against $CF(A)$ for various m. Note that $\kappa(p_m(A))$ may be made as small as desired by taking m large enough: For any $\delta > 1$, if

$$m \geq \frac{\cosh^{-1}\left(\frac{\delta+1}{\delta-1}\right)}{\cosh^{-1}\left(\frac{\kappa(A)+1}{\kappa(A)-1}\right)}, \tag{3.2}$$

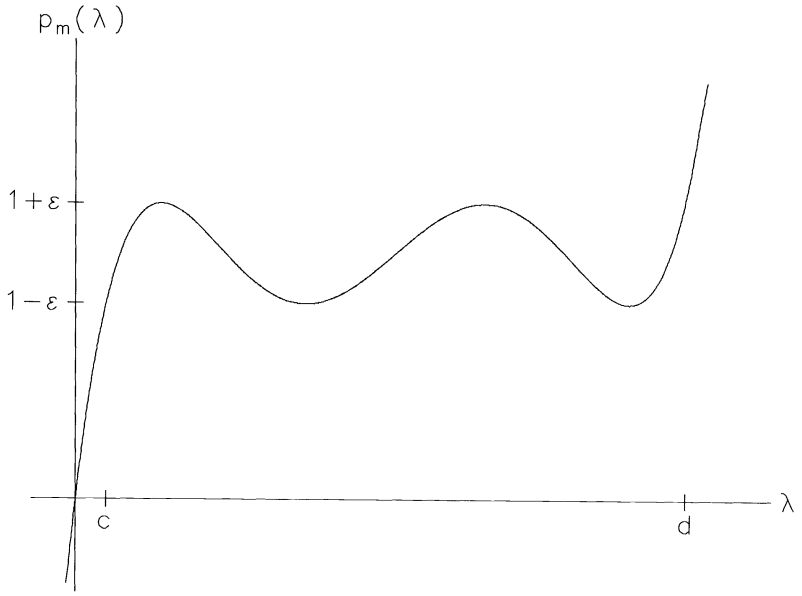


FIG. 3.1. *Chebyshev preconditioned polynomial* ($m = 5$) *for* $S = [1, 20]$.

then $\kappa(p_m(A)) < \delta$. Of course, as m increases, each CG iteration becomes more expensive, requiring m matvecs instead of one. The optimum m is that one for which the total CPU time required to solve the linear system is minimized. Numerical experiments [4], [5], [31] suggest that low degree (2–16) preconditioning polynomials are usually best for hpd A, and that the optimum m tends to increase with $\kappa(A)$.

An attractive feature of the Chebyshev polynomial preconditioner is its optimality: it minimizes a bound on $\kappa(p_m(A))$. This is a consequence of the following.

THEOREM 3.1. *A solution to*

$$\min_{C \in \pi_{m-1}} \frac{\max_{\lambda \in S} |C(\lambda)\lambda|}{\min_{\lambda \in S} |C(\lambda)\lambda|} \tag{3.3}$$

is given by the Chebyshev preconditioning polynomial [5], [27].

If $\sigma(A) \subset S$, the ratio in (3.3) gives a bound on the condition number of $p_m(A)$. Moreover, this bound is minimized (with respect to S) when $S = \Sigma(A) \equiv [\lambda_c, \lambda_d]$. A similar results holds for the de Boor and Rice polynomial of § 4.

The discussion and theorem above assume that the smallest and largest eigenvalues of A, λ_c and λ_d, are known. If one bases the Chebyshev polynomial on any other endpoints, the resulting preconditioner is not optimum. It is even possible to choose c and d so that $\kappa(p_m(A)) > \kappa(A)$, thereby slowing convergence. It is therefore important to have accurate estimates for λ_c and λ_d. Such estimates may be obtained from the CG iteration parameters. This is equivalent to dynamically determining the optimum polynomial preconditioner. The resulting *adaptive* CG algorithm works remarkably well in practice in that it quickly and accurately determines λ_c and λ_d. We describe this idea next.

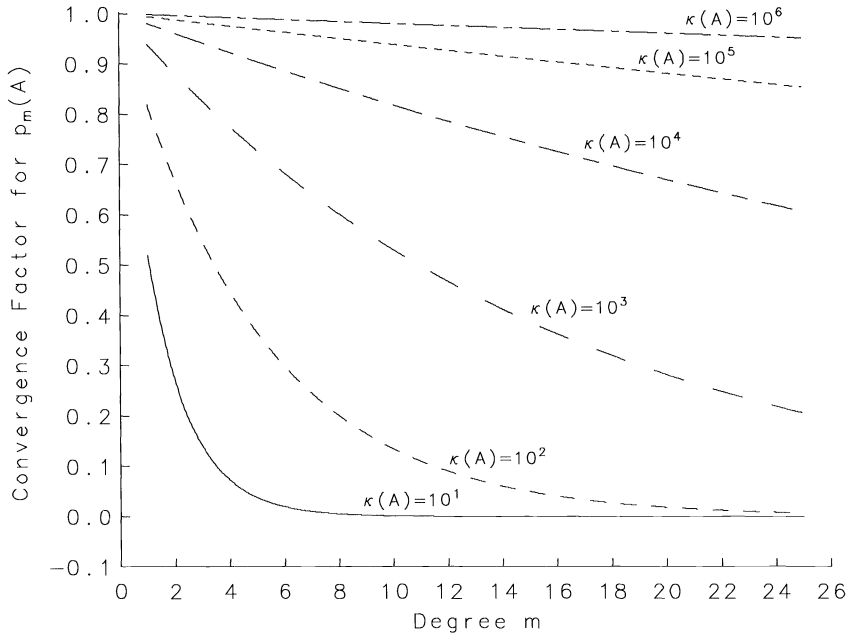


FIG. 3.2. *Degree m versus $CF(p_m(A))$ for hpd A.*

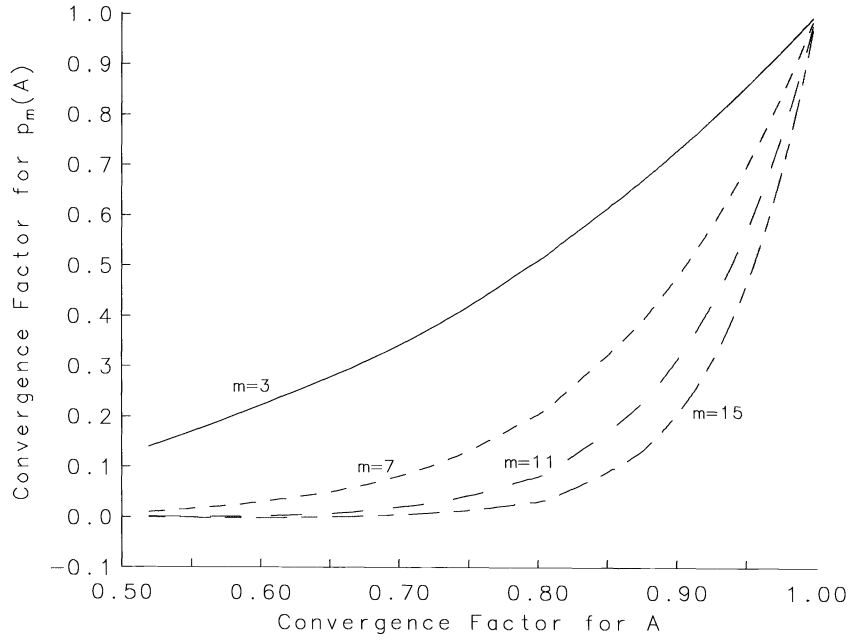


FIG. 3.3. *$CF(A)$ versus $CF(p_m(A))$ for hpd A.*

3.2. An adaptive procedure. In this section we discuss *adaptive* CG algorithms. Such an algorithm performs a sequence of iterations, each of which uses the same CG method, but a different preconditioner. Within a given iteration, the CG method is applied to $C(A)Ax = C(A)b$, where $C(\lambda)$ is the current preconditioning polynomial. Information from the current iteration is used to obtain a better preconditioner $\tilde{C}(A)$, which is then used in the next iteration. In this way the adaptive algorithm dynamically determines the optimum polynomial preconditioner for A.

Determining this optimum preconditioner is equivalent to determining a set $S = [c, d]$ that contains the spectrum of A, $\sigma(A)$. Ideally, $S = \Sigma(A) \equiv [\lambda_c, \lambda_d]$. Since the extreme eigenvalues of A are seldom known a priori, S is only an approximation to $\Sigma(A)$. The purpose of the adaptive procedure is to improve this approximation. Since a detailed discussion and numerical results are given in [4] and [5], we will sketch only the essential elements.

Given a set $S \subset \Sigma(A)$ and a minimax preconditioning polynomial $C(\lambda)$, the CG method is applied to $C(A)Ax = C(A)b$. After a prescribed number of steps, say ℓ, the adaptive procedure is called:

(1) Compute eigenvalue estimates for $p_m(A) = C(A)A$.
(2) Extract eigenvalue estimates for A and update S.
(3) Determine the new preconditioning polynomial.
(4) Resume or restart the iteration, whichever is appropriate.

After another ℓ steps, the adaptive procedure is called again, and so on until convergence.

Eigenvalue estimates for $p_m(A)$ are easily obtained from the CG iteration parameters by exploiting the equivalence of the CG and Lanczos algorithms [7], [14], [21], [22]. In particular, after k steps, one may obtain k eigenvalue estimates for $p_m(A)$ from a $k \times k$ Hermitian tridiagonal matrix of iteration parameters, $\tilde{T}_k$. If (μ, y) is an eigenpair for $\tilde{T}_k$, then (μ, x) is an approximate eigenpair for $p_m(A)$, where $x = P_k D_k^{-1/2} y$, P_k is a matrix whose columns are the B-orthogonal CG direction vectors, and D_k is a diagonal matrix. See [7] for details. One may show that $\mu \in \mathcal{H}(p_m(A))$, the convex hull of $\sigma(p_m(A))$. Moreover, as k increases, μ converges to an eigenvalue of $p_m(A)$.

After an eigenvalue estimate μ for $p_m(A)$ is computed, we obtain an eigenvalue estimate for A by determining the inverse image(s) of μ. If μ has several inverse images, it is important to choose one that lies in $\Sigma(A)$. Since $\mu \in \mathcal{H}(p_m(A))$, this may always be done. Otherwise, the set S might be irrevocably and improperly expanded, which would slow convergence of subsequent CG iterations.

As shown in Fig. 3.4, this is easy to do when m is odd. Since we wish to expand S, we may discard any eigenvalue estimate μ for $p_m(A)$ lying in $[1 - \epsilon_m, 1 + \epsilon_m]$ because each of its inverse images lie in S. Thus, suppose there is an eigenvalue estimate $\mu \notin [1 - \epsilon_m, 1 + \epsilon_m]$. Since $p_m(\lambda)$ is monotonically increasing for m odd and $\lambda \notin S$, μ has a unique inverse image, which is our eigenvalue estimate for A. Specifically, if we compute $\mu_1 < 1 - \epsilon_m$, the left endpoint c is decreased to $\lambda_1 \in (0, c)$. If we compute $\mu_2 > 1 + \epsilon_m$, the right endpoint d is increased to $\lambda_2 \in (d, \infty)$. In this way the set S is dynamically enlarged until it captures the spectrum of A. Once the new S is determined, a decision is made whether to resume the outer iteration using the current polynomial, or to restart using the new polynomial. This decision is based on (2.13) using the convergence factors for the current and new preconditioning polynomials.

So far we have assumed that m is odd, which is important for two reasons. First, if m were even, an eigenvalue $\mu_1 < 1 - \epsilon_m$ would have *two* inverse images, only one of which must be an eigenvalue estimate for A. If we choose the wrong one (or take

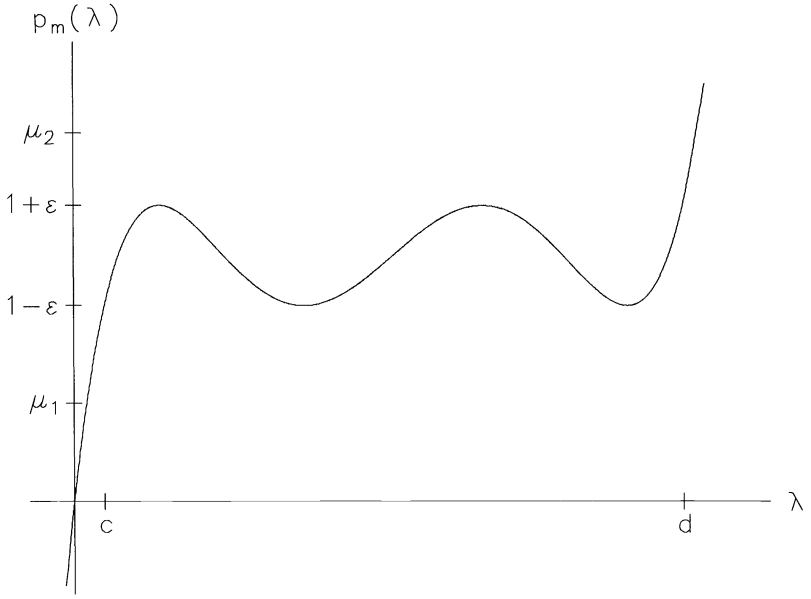


FIG. 3.4. *The Chebyshev adaptive procedure for hpd A.*

both), we might enlarge S beyond $\Sigma(A)$. Second, if d is too small, as is usually the case in the early stages of the adaptive algorithm, then the preconditioned matrix might be indefinite—which can cause difficulties for many CG methods. By choosing m odd we avoid both these difficulties. In particular, note that $p_{odd}(A)$ is hpd for *any* set S. This makes possible robust adaptive CG algorithms.

3.3. Other adaptive algorithms. The adaptive procedure described above uses information from only the extreme eigenvalues of $p_m(A)$ to update the set S. Freund [19] has recently proposed approximating the distribution of the eigenvalues of A by computing all the Lanczos eigenvalue estimates. (Recall that the CG rate of convergence is determined by the eigenvalue distribution, not the condition number of A.) He then uses this approximate distribution to obtain a *weighted* minimax preconditioning polynomial. Preliminary results indicate that the resulting preconditioner is often superior to the one based on the Chebyshev polynomial. Unfortunately, there is no guarantee that the preconditioned matrix will be hpd for all S. It is also unclear whether this idea can be developed into an adaptive algorithm.

The hybrid algorithm of O'Leary [14], [32] is an interesting alternative to the adaptive algorithms described above. The idea is to iterate with CG, compute estimates for λ_c and λ_d, and then switch to a cheaper Chebyshev iteration based on these estimates. The initial guess for the Chebyshev iteration is the most recent CG iterate. The adaptive Chebyshev algorithm of Hageman and Young [25] is yet another alternative.

4. The DR and Grcar preconditioning polynomials. In this section we discuss the de Boor and Rice (DR) and Grcar preconditioning polynomials for Hermitian indefinite (hid) matrices A. We will now assume $\sigma(A) \subset [a,b] \cup [c,d]$, where

$a \leq b < 0 < c \leq d$. Ideally, $a = \lambda_a$, $b = \lambda_b$, $c = \lambda_c$, and $d = \lambda_d$, the four extreme eigenvalues of A. As we will see, the dynamic determination of these eigenvalues is much more difficult than in the hpd case. In particular, we must contend with the problem of *ambiguity*. An adaptive procedure for this case has been proposed in [6]; we will outline it below.

Once again consider the minimax approximation problem (2.10), which seeks to cluster the eigenvalues of $p_m(A)$ around 1. De Boor and Rice [15] were the first to study this problem for $S = [a, b] \cup [c, d]$, but they were interested in using the roots of the related residual polynomial to define an optimum Richardson's method (recall § 2). We use the polynomial p_m to define an optimum polynomial preconditioner $C(A)$ for conjugate gradient methods.

4.1. Behavior of the DR polynomial. When $d - c = b - a$, the DR minimax polynomial is obtained from a Chebyshev polynomial [29]:

$$p_m(\lambda) = 1 - \frac{T_k(q_2(\lambda))}{T_k(q_2(0))} \tag{4.1}$$

where $k = \lfloor m/2 \rfloor$ and

$$q_2(\lambda) = 1 + \frac{2(\lambda - b)(\lambda - c)}{ad - bc} \tag{4.2}$$

maps both intervals of S to $[-1, 1]$, each one monotonically. Note that p_m has even degree. Thus, for equal length intervals, the DR polynomial of odd degree has leading coefficient zero, which cannot happen in the hpd case. If we fix $S_1 = [a, b]$ and let $S_2 = [c, d]$ move away from the origin, we find that ϵ_m decreases. In particular, $\epsilon_m \to T_k^{-1}(-(b + a)/(b - a))$ as $c \to \infty$.

In practice, the intervals of S seldom have the same length, in which case the DR polynomial is not explicitly known. However, by simply extending an endpoint until the two new intervals have the same length, one may use (4.1). Unfortunately, this idea has a serious drawback. If the two intervals of S differ greatly in length, the polynomial based on the extended interval pair will yield an inferior preconditioner. For this reason we seek the DR preconditioned polynomial for S consisting of two intervals of unequal length.

De Boor and Rice found this polynomial as the solution of an extremal problem [15]. One may also use the *Grcar characterization theorem* (GCT) [23]. This result, which generalizes the well-known theorem of Chebyshev, characterizes the equioscillation property of the error in constrained (interpolatory) minimax approximation. (Although the GCT may be used to characterize preconditioned polynomials for any number of disjoint intervals, we consider just two intervals to facilitate the development of adaptive procedures.) In general, there is no explicit formula for the DR polynomial, but it may be found, for example, by a Remez algorithm. Rather than compute the polynomial coefficients with respect to the usual power representation, it is better to express the DR polynomial in terms of Chebyshev polynomials. One may then use Clenshaw's rule to stably and efficiently evaluate the polynomial. The powers of A are not computed. See [4] for a discussion of the Remez algorithm in this context.

The GCT may be used to show that the behavior of the DR polynomial illustrated in Fig. 4.1 is typical. Note that the polynomial equioscillates about 1 over $[a, b]$, satisfies the interpolatory constraint by passing through the origin, and then equioscillates

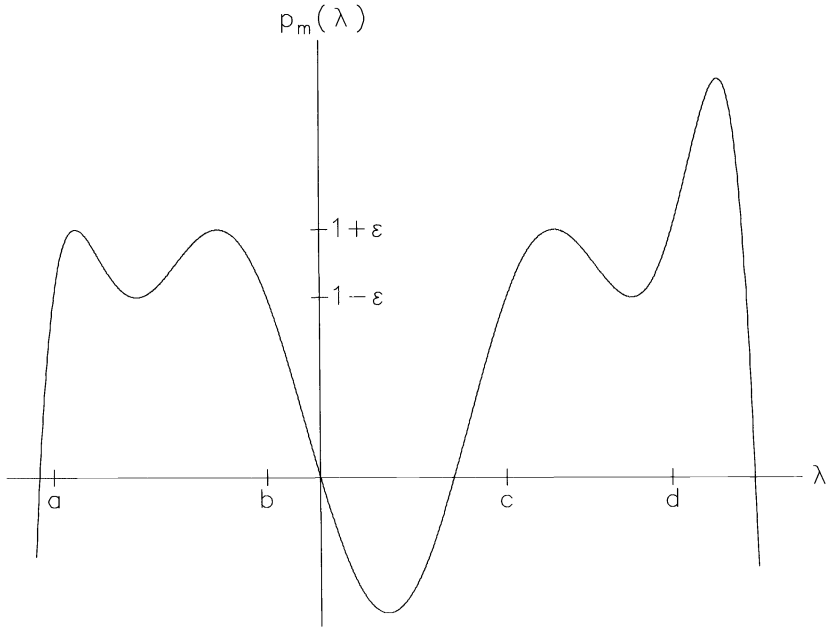


FIG. 4.1. *DR preconditioned polynomial* ($m = 8$) *for* $S = [-10, -2] \cup [7, 13.2]$.

about 1 over $[c, d]$. The relative minimum between b and c is a consequence of having disjoint intervals. There also may be an *outside extremum*, depending on the interval pair. For example, in Fig. 4.1 there is a relative maximum to the right of d. If $d - c = b - a$, this "outside" extremum would be inside S. Finally, we note that the same polynomial can be the minimax polynomial for more than one set S, something that cannot happen in the hpd case for $m > 1$. See [4] for a detailed characterization.

4.2. Some properties. If $\sigma(A) \subset S$, then $\sigma(p_m(A)) \subset [1 - \epsilon_m, 1 + \epsilon_m]$, where $\epsilon_m < 1$ for $m > 1$ [15]. In other words, we have *transformed* the Hermitian indefinite matrix A into an hpd $p_m(A)$. This makes possible several CG methods [6]. As with the Chebyshev polynomial, ϵ_m is minimized (with respect to S) when $S = \Sigma(A) \equiv [\lambda_a, \lambda_b] \cup [\lambda_c, \lambda_d]$. Since $p_m(A)$ is hpd, (2.12) and (2.13) hold. Unfortunately, there is no explicit formula for ϵ_m, and so it is impossible to accurately predict how many CG steps are needed for convergence. (To determine ϵ_m, one must first determine the polynomial.) It is possible, however, to obtain a crude estimate of the number of CG steps required for convergence by using (4.1) to bound ϵ_m. We remark that one can also use this bound to estimate the rate of convergence for the unpreconditioned conjugate residual (CR) method.

Since ϵ_m is a nonincreasing function of m, we can make the condition number of $p_m(A)$ as small as desired by making m large enough. Unlike the hpd case, however, ϵ_m does not depend solely on the condition number of A. We might have $\kappa(A_1) = \kappa(A_2)$, and yet $\kappa(p_m(A_1)) \neq \kappa(p_m(A_2))$. What matters is the relative lengths of the intervals, as well as their location relative to the origin. This is not completely unexpected because $\kappa(A)$ is a poor indicator of the CR rate of convergence for hid matrices. One may show that $\kappa(p_m(A)) \leq \kappa(A^2)$ for $m > 1$, and so a polynomial preconditioned CG

method will converge faster than CGHS applied to the normal equations. In general, high degree (20–50) polynomials are best for hid A [4], [6]. This is in contrast to the hpd case, where low degree polynomials are usually best.

The DR preconditioning polynomial enjoys an optimality property similar to that of the Chebyshev polynomial, but with a subtle difference.

THEOREM 4.1. *A solution to*

$$\min_{\substack{C\in\pi_{m-1}\\ C(\lambda)\lambda>0,\ \lambda\in S}} \frac{\max_{\lambda\in S} C(\lambda)\lambda}{\min_{\lambda\in S} C(\lambda)\lambda} \tag{4.3}$$

is given by the DR preconditioning polynomial [6].

In other words, the DR preconditioning polynomial minimizes a bound on $\kappa(p_m(A))$ among those polynomials for which $p_m(A)$ is hpd. We must impose this condition because the intervals are disjoint. If one simply seeks to minimize $\kappa(p_m(A))$, one would not necessarily obtain the DR polynomial. We note that $\kappa(p_m(A))$ is minimized with respect to S when $S = \Sigma(A)$.

4.3. The Grcar preconditioning polynomial. In this section we consider the Grcar preconditioned polynomial. It is obtained from the following minimax approximation problem:

$$\min_{\substack{p\in\pi_m\\ p(0)=0\\ p'(0)=0}} \|1-p\|_S. \tag{4.4}$$

By adding the second constraint, $p'(0) = 0$, we may devise CG methods that minimize the Euclidean norm of the true error without resorting to some form of the normal equations [7]. The implication of this added constraint on the behavior of the Grcar polynomial is even more important, as we shall see.

One may again use the GCT to characterize the behavior of the Grcar preconditioned polynomial, which is illustrated in Fig. 4.2. The polynomial, $p_m(\lambda) = C(\lambda)\lambda = \Gamma(\lambda)\lambda^2$, equioscillates about 1 over $[a, b]$, has a double root at the origin, and then equioscillates about 1 over $[c, d]$. In general, there is a relative maximum between the intervals, which is a consequence of the double root. As with the DR polynomial, there may be an outside extremum, depending on the interval pair. We note that the DR and Grcar polynomials behave similarly when $b \approx -c$. To understand why, consider Figs. 4.1 and 4.2. As $c \to -b$, the DR relative minimum moves toward the origin, and so the DR polynomial begins to look like the Grcar polynomial. Meanwhile, the Grcar relative maximum moves into S, and so the Grcar polynomial begins to look like the DR polynomial. When $a = -d$ and $b = -c$, the two polynomials coincide. See Table 4.1.

Although motivated by the desire for a specific CG method, the real advantage of the Grcar polynomial is its utility for robust adaptive CG algorithms. During the early stages of an adaptive algorithm, estimates for the extreme eigenvalues are likely to be poor; the inner endpoints are especially difficult to ascertain. Because of this, the DR preconditioned matrix might be indefinite, which is a problem for many CG methods. The Grcar preconditioned matrix, on the other hand, is hpd for *any* b and c, assuming $a = \lambda_a$ and $d = \lambda_d$ (see below). This makes possible robust adaptive CG algorithms.

Of course, there is a drawback to the Grcar polynomial: In general, it is a poorer preconditioner than the DR polynomial because it has larger oscillations, a result of

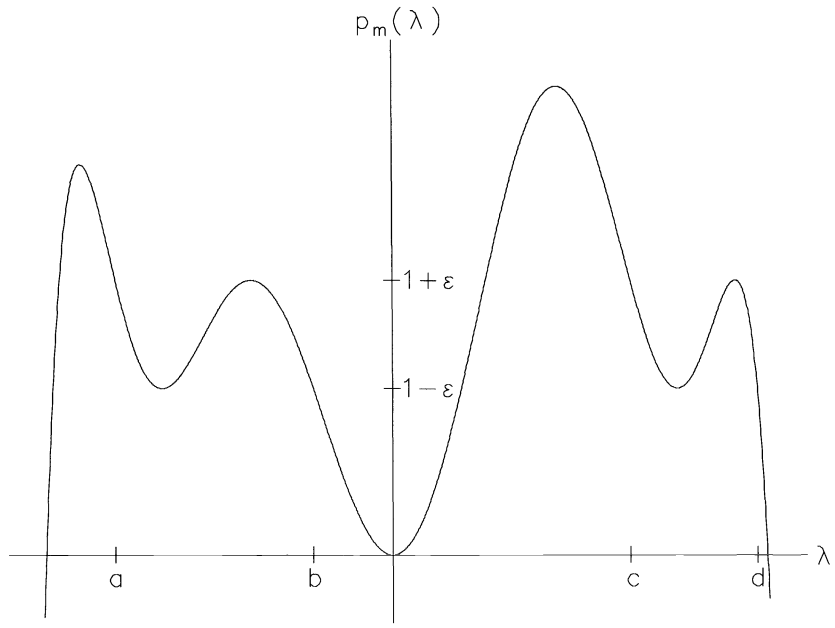


FIG. 4.2. *Grcar preconditioned polynomial* ($m = 8$) *for* $S = [-7, -2] \cup [6, 9.2]$.

TABLE 4.1
DR versus Grcar: $c = -b$, d *increasing.*

$m = 14$		$S = [a,b] \cup [c,d]$		Norm of residual poly		Convergence factor	
a	b	c	d	DR	Grcar	DR	Grcar
−10.0	−1.0	1.0	2.0	0.1037e+00	0.1301e+00	0.5200e−01	0.6534e−01
−10.0	−1.0	1.0	5.0	0.3082e+00	0.3214e+00	0.1579e+00	0.1651e+00
−10.0	−1.0	1.0	10.0	0.4630e+00	0.4630e+00	0.2454e+00	0.2454e+00
−10.0	−1.0	1.0	100.0	0.9245e+00	0.9270e+00	0.6693e+00	0.6741e+00
−10.0	−1.0	1.0	1000.0	0.9923e+00	0.9926e+00	0.8832e+00	0.8852e+00
−10.0	−1.0	1.0	10000.0	0.9999e+00	0.9999e+00	0.9877e+00	0.9877e+00

adding the second constraint. See Table 4.2. Note that the superiority of the DR polynomial increases as c moves away from $-b$. (We remark that the one interval Grcar polynomial [4] has little merit compared to the Chebyshev polynomial.)

4.4. An adaptive procedure. The task of dynamically determining S for hid A is more difficult because there are four extreme eigenvalues to estimate: λ_a, λ_b, λ_c, and λ_d. Nonetheless, the adaptive procedure is basically unchanged: we iterate, compute eigenvalue estimates for $p_m(A)$, recover eigenvalue estimates for A, update the set S, and then resume or restart, whichever is appropriate. The difficulty lies in recovering estimates for the extreme eigenvalues of A from those for $p_m(A)$. Since the adaptive procedures for the DR and Grcar polynomials are essentially the same, we shall focus on the former.

As before, let μ be an eigenvalue estimate for $p_m(A)$. We wish to determine an inverse image $\lambda \in \Sigma(A)$, which is possible because $\mu \in \mathcal{H}(p_m(A))$. Since $\mu \in [1-\epsilon_m, 1+\epsilon_m]$ has an inverse image in the currently known set S, we cannot confidently

TABLE 4.2
DR versus Grcar: c moving away from −b.

$m = 14$		$S = [a,b] \cup [c,d]$		Norm of residual poly		Convergence factor	
a	b	c	d	DR	Grcar	DR	Grcar
−10.0	−1.0	1.0	105.0	0.9303e+00	0.9325e+00	0.6806e+00	0.6852e+00
−10.0	−1.0	1.5	105.5	0.8986e+00	0.9283e+00	0.6245e+00	0.6767e+00
−10.0	−1.0	2.0	106.0	0.8686e+00	0.9288e+00	0.5808e+00	0.6777e+00
−10.0	−1.0	5.0	109.0	0.7157e+00	0.9317e+00	0.4214e+00	0.6834e+00
−10.0	−1.0	10.0	114.0	0.5445e+00	0.9357e+00	0.2961e+00	0.6917e+00
−10.0	−1.0	100.0	204.0	0.1768e+00	0.7595e+00	0.8912e−01	0.4602e+00

expand S, and so we will assume $\mu \notin [1-\epsilon_m, 1+\epsilon_m]$. Depending on the parity of m and the nature of the outside extremum, there may be as many as five inverse images, only one of which is necessarily an eigenvalue estimate for A. We will now briefly describe an adaptive procedure for resolving this ambiguity. This description, which is culled from [6], is given here so that the reader may better understand the adaptive procedure proposed in the next section for the bilevel polynomial.

The essence of the adaptive procedure is the extraction of eigenvalue estimates for A from those for $p_m(A)$. To simplify this task, we will assume $a = \lambda_a$ and $d = \lambda_d$. That is, we will assume that the algebraically smallest and largest eigenvalues of A are known. This is not unreasonable since these are the easiest to estimate, for example, via the power method, Gershgorin's theorem, or conjugate residuals. An eigenvalue estimate μ for $p_m(A)$ now has at most two inverse images of interest.

To see how we might choose between these two inverse images, consider Fig. 4.3. First notice that $\mu < 1 - \epsilon_m$ because $\sigma(A) \subset [a,d]$. If $\mu = \mu_1 < 0$, there are two inverse images of interest, $\lambda_1 < \lambda_2$, both of which lie in $(0,c)$. To guarantee that the new set $\tilde{S}$ lies in $\Sigma(A)$, we will decrease c to λ_2. If the eigenvalue of A is nearer λ_1, subsequent calls to the adaptive procedure will converge to it. Next suppose $\mu = 0$. There is a single inverse image in $(0,c)$, and this is our eigenvalue estimate for A. Finally, suppose $\mu = \mu_2 \in (0, 1-\epsilon_m)$. There are again two inverse images of interest, λ_n and λ_p, but now one is negative and the other is positive.

A heuristic scheme for choosing the proper inverse image in this case is given in [6]. The basic idea is this: We find an approximate eigenvector for A, calculate the corresponding Rayleigh quotient, compute an error bound for this approximate eigenpair, and then use this bound to determine which inner endpoint to move, and by how much. An approximate eigenvector x for $p_m(A)$ is easily obtained from the CG iteration parameters; recall § 3.2. This vector x is either an approximate eigenvector of A, or the linear combination of two approximate eigenvectors, the eigenvalues of which have opposite sign. Once x and its corresponding Rayleigh quotient, λ_r, are determined, we may calculate an interval known to contain a true eigenvalue of A, λ_t [34]:

$$|\lambda_t - \lambda_r| \leq \frac{\|Ax - \lambda x\|_2}{\|x\|_2} \equiv \delta. \tag{4.5}$$

To be specific in the sequel, we will assume $\lambda_r > 0$. If $[\lambda_r - \delta, \lambda_r + \delta] \subset (0,c)$, c may be decreased to $\lambda_r + \delta$, and the new set is $\tilde{S} = [a,b] \cup [\lambda_r + \delta, d] \subset \Sigma(A)$. But suppose $\lambda_r - \delta \leq 0$ or $\lambda_r + \delta \geq c$. In neither case may we confidently expand S, so the current CG iteration is resumed. If subsequent calls to the adaptive procedure continue to find μ, but fail to expand S, we will take $\tilde{S} = [a, \lambda_n] \cup [\lambda_p, d]$.

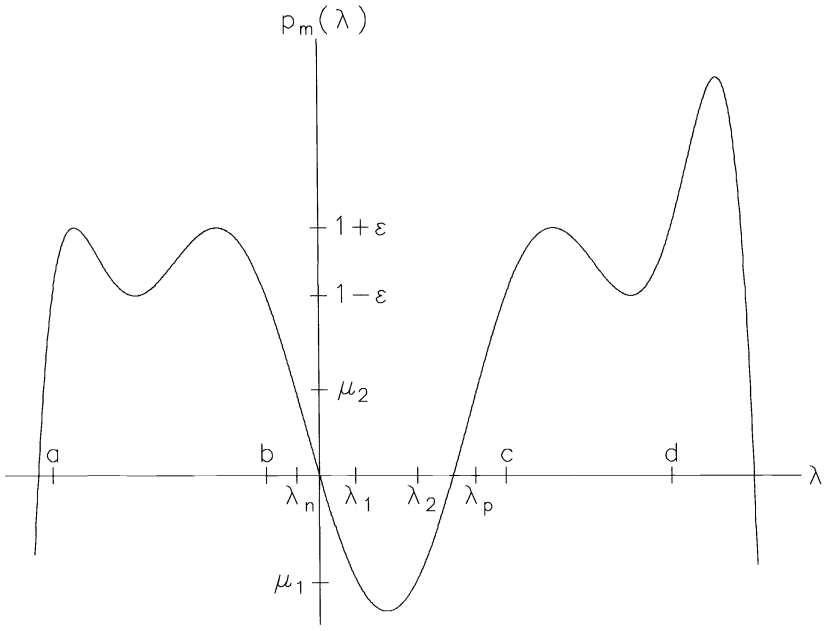


FIG. 4.3. *The DR adaptive procedure for hid A.*

This heuristic is deficient in several respects. First, to compute $x = P_k D_k^{-1/2} y$, we must store several past CG direction vectors. If storage is limited, only a few may be kept, which may impair the accuracy of μ and x. Second, the calculation of δ requires one A-matvec, one saxpy, and one norm computation. This expense is nontrivial. It is also possible to update S incorrectly. In the next section we will see how the bilevel polynomial leads to an adaptive procedure that avoids the first two problems.

Finally, a few words on an adaptive procedure for the Grcar polynomial. The situation is similar to that for the DR polynomial. Once again, the main difficulty is with an eigenvalue estimate μ for $p_m(A)$ that lies in $(0, 1-\epsilon)$, in which case there are two inverse images of opposite sign. The above strategy may be used to resolve this ambiguity. The case $\mu > 1+\epsilon$ is analogous to the case $\mu < 0$ for the DR polynomial.

5. The bilevel preconditioning polynomial. In this section we introduce the *bilevel* preconditioning polynomial for hid A. This polynomial, which leads to a new class of preconditioners for Hermitian indefinite matrices, is fundamentally different from those considered thus far. We have heretofore chosen the preconditioning polynomial $C(\lambda)$ so that the eigenvalues of the preconditioned matrix are clustered around 1. In the case of the DR polynomial, this means transforming an indefinite matrix A into an hpd $p_m(A)$. Let us now take a different approach: Instead of forcing the preconditioned matrix to be hpd, we will allow it to be indefinite, but choose $C(\lambda)$ so that the eigenvalues of $p_m(A)$ are clustered. For example, we might choose $C(\lambda)$ so that the negative eigenvalues of A are clustered around -1 and the positive eigenvalues of A are clustered around $+1$.

This idea, which is due to Freund [17] and Grcar [24], is motivated by the following well-known property of conjugate gradient methods: they converge in at most k steps,

where k is the number of distinct eigenvalues of the preconditioned matrix. By clustering the eigenvalues of $p_m(A)$ around -1 and $+1$, we hope to speed the convergence of several CG methods. A numerical study of bilevel polynomial preconditioned CG methods is planned. In this paper we wish only to introduce the polynomial, discuss its essential features, and mention some important variants. We will also propose an adaptive procedure for the bilevel polynomial and show that it has several advantages over those for the DR and Grcar polynomials.

5.1. The approximation problem. We will again assume $\sigma(A) \subset S = [a, b] \cup [c, d]$, $a \leq b < 0 < c \leq d$. The bilevel polynomial is obtained from the following minimax approximation problem:

$$\min_{\substack{p \in \pi_m \\ p(0)=0}} \|f - p\|_S \tag{5.1}$$

where

$$f(\lambda) = \begin{cases} -1 & \text{if } \lambda \in [a, b], \\ +1 & \text{if } \lambda \in [c, d]. \end{cases} \tag{5.2}$$

Unlike our previous approximation problems, the error in (5.1) is not a residual polynomial.

We may use the GCT to characterize the equioscillation property of the error in (5.1), which allows us to characterize the behavior of the bilevel preconditioned polynomial. The polynomial in Fig. 5.1 is typical. It equioscillates about -1 over $S_1 = [a, b]$, passes through the origin, and then equioscillates about 1 over $S_2 = [c, d]$. As with the DR and Grcar polynomials, there may be an outside extremum, depending on the interval pair. There is also at most one extremum in $[b, c]$. If this extremum is positive, it is a relative maximum; if it is negative, it is a relative minimum. Thus, any $\mu \in [-1 + \epsilon_m, 1 - \epsilon_m]$ has a unique inverse image in (b, c), which is important in the adaptive procedure described below.

If $\sigma(A) \subset S$, then $\sigma(p_m(A)) \subset [-1-\epsilon_m, -1+\epsilon_m] \cup [1-\epsilon_m, 1+\epsilon_m]$. If ϵ_m is small, the eigenvalues of $p_m(A)$ are tightly clustered, and the CG method will converge rapidly. Although we can make ϵ_m as small as desired by taking m large enough, the cost per CG step increases with m. The optimum value of m is likely to depend not only on the particular problem, but also on the computer architecture (cf. [6], [8]).

If $\epsilon_m \geq 1$, there is no clustering of eigenvalues, and $p_m(A)$ might even be singular. Fortunately, it is easy to show that

$$\epsilon_m \leq \epsilon_1 = \frac{\rho - \upsilon}{\rho + \upsilon} < 1 \tag{5.3}$$

where $\upsilon = \min\{|b|, c\}$ and $\rho = \max\{|a|, d\}$. This follows from the next lemma.

LEMMA 5.1. *The linear bilevel polynomial for* $[a, b] \cup [c, d]$ *is* $p_1(\lambda) = 2\lambda/(\rho + \upsilon)$.

Combining this with our earlier characterization of the bilevel polynomial gives the following result.

THEOREM 5.2. *Let* $p_m(\lambda)$ *be the solution to* (5.1). *Then* $p_m(\lambda) < 0$ *for* $\lambda \in [a, 0)$ *and* $p_m(\lambda) > 0$ *for* $\lambda \in (0, d]$.

Although the bilevel polynomial preconditioned matrix $p_m(A)$ is hid, Theorem 5.2 implies that $C(A)$ is hpd when $\sigma(A) \subset [a, d]$. This observation, which is due to Otto [33], means that the preconditioned conjugate residual method is applicable [7].

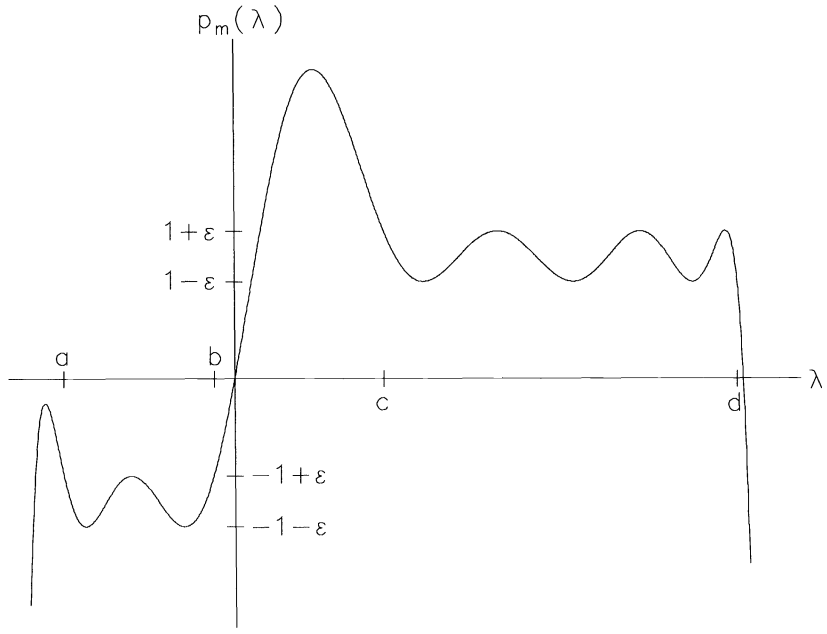


FIG. 5.1. *Bilevel preconditioned polynomial* ($m = 12$) *for* $S = [a, b] \cup [c, d]$.

The DR preconditioning polynomial is optimum in the sense of minimizing a bound on the condition number of the preconditioned matrix, which is required to be hpd. It is unclear whether a similar result holds for the bilevel polynomial. To obtain an optimum preconditioning polynomial for indefinite matrices, Freund has suggested minimizing the asymptotic rate of convergence [18].

5.2. Comparison with the DR polynomial. It is natural to ask which is better, the DR or bilevel preconditioning polynomial. A precise answer depends on several factors, including the eigenvalue distribution before and after preconditioning, the relative location and lengths of S_1 and S_2, and the degree m. However, we may perform an a priori analysis based on convergence factors. Recall that the DR convergence factor is

$$CF_{dr} = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}, \qquad \kappa = \frac{1 + \delta_m}{1 - \delta_m} \tag{5.4}$$

where $2\delta_m$ is the magnitude of the oscillations in the DR preconditioned polynomial. Since the bilevel polynomial preconditioned matrix is indefinite, $\kappa(p_m(A))$ is a poor indicator of the CG rate of convergence. Notice, however, that the bilevel preconditioning polynomial $C(\lambda)$ maps S_1 and S_2 onto two new intervals of equal length, $I_1 = [-1 - \epsilon_m, -1 + \epsilon_m]$ and $I_2 = [1 - \epsilon_m, 1 + \epsilon_m]$. We may therefore obtain an estimate of the bilevel convergence factor from (4.1) [4, p. 14]. In particular, we may show

$$CF_{bi} \leq \sqrt{\frac{\nu - 1}{\nu + 1}} = \sqrt{\epsilon_m}, \qquad \nu = \frac{1 + \epsilon_m}{1 - \epsilon_m}. \tag{5.5}$$

TABLE 5.1
δ_m and ϵ_m for which $CF_{bi} = CF_{dr}$.

δ_m	ϵ_m	CF
0.9	0.3929	0.6268
0.99	0.7527	0.8676
0.999	0.9144	0.9562
0.9999	0.9721	0.9860

Thus, the bilevel preconditioning polynomial is better when

$$\sqrt{\epsilon_m} \leq \frac{\sqrt{1+\delta_m} - \sqrt{1-\delta_m}}{\sqrt{1+\delta_m} + \sqrt{1-\delta_m}}. \tag{5.6}$$

In Table 5.1 we list a few values of δ_m and give the corresponding value of ϵ_m for which $CF_{bi} = CF_{dr}$. When the DR preconditioned matrix is well-conditioned, ϵ_m must be much smaller than δ_m before the bilevel polynomial bests the DR polynomial. As δ_m increases, the bilevel polynomial is more likely to be superior.

The key question is this: for which interval pairs is the bilevel polynomial superior to the DR polynomial? Numerical experiments in [4] and [6] suggest that the DR preconditioning polynomial performs best when $d - c \approx b - a$, in which case the DR polynomial is obtained from a Chebyshev polynomial; recall (4.1). Preliminary numerical results suggest that the bilevel polynomial performs best when $b \approx -c$ and one interval, say S_1, is much shorter than the other. In this case the bilevel polynomial *is* a Chebyshev polynomial; see below.

5.3. A family of bilevel polynomials. In (5.1) we chose to cluster the negative eigenvalues of A around -1 and the positive eigenvalues of A around $+1$. An obvious generalization is to cluster the negative eigenvalues around some constant $\gamma < 0$. This flexibility might lead to a smaller ϵ_m, and consequently faster convergence of the CG method. It also allows one to take into consideration the location of S_1 and S_2 relative to the origin. For example, Freund [18] has suggested $\gamma = (d+c)/(b+a)$, the ratio of the interval midpoints. Alternatively, one might take $\gamma = c/b$. Unfortunately, Theorem 5.2 does not necessarily hold for $\gamma \neq -1$, which is a serious deficiency. An important open question is this: for which S and γ does Theorem 5.2 hold?

When one interval, say S_1, is much shorter than the other, one may choose γ so that the bilevel and Chebyshev preconditioned polynomials coincide. To elucidate, let $p_m(\lambda)$ be the Chebyshev preconditioned polynomial for S_2. If S_1 is sufficiently short, one may choose γ so that $|\gamma - p_m(\lambda)| < \epsilon_m$ for $\lambda \in S_1$. Thus, by the GCT, $p_m(\lambda)$ is the bilevel preconditioned polynomial for $S_1 \cup S_2$.

The bilevel preconditioning polynomial $C(\lambda)$ maps two intervals of arbitrary lengths, S_1 and S_2, into two new intervals of the *same* length, I_1 and I_2. One might obtain faster convergence of the CG method by preserving the relative interval lengths. For example, Freund has suggested [18] employing the weight function $w = (b-a)/(d-c)$ on S_1. This idea leads to a weighted minimax approximation problem:

$$\min_{\substack{p \in \pi_m \\ p(0)=0}} \|w(f-p)\|_S \tag{5.7}$$

where

$$f(\lambda) = \begin{cases} \gamma & \text{if } \lambda \in [a,b], \\ 1 & \text{if } \lambda \in [c,d], \end{cases} \qquad w(\lambda) = \begin{cases} \omega & \text{if } \lambda \in [a,b], \\ 1 & \text{if } \lambda \in [c,d]. \end{cases} \tag{5.8}$$

Note that (2.10) and (5.1) are both instances of (5.7).

By introducing a weight function we change the equioscillation property of the bilevel preconditioned polynomial. It still equioscillates over S_1 and S_2, but with different magnitudes. Specifically, if the polynomial equioscillates about 1 with magnitude ϵ_m, it equioscillates about γ with magnitude $\omega\epsilon_m$. The Roloff polynomial [35] is characterized by a similar quasi-equioscillation property. (The Roloff residual polynomial is obtained from (2.9) under the additional constraint that its roots lie in S.) Note that the analysis of § 5.2 no longer applies because I_1 and I_2 now have different lengths.

We remark that (5.7) and (5.8) define a *family* of bilevel polynomials. A particular member is obtained by specifying γ and ω. If we minimize (5.7) over all γ and ω, we obtain an optimum preconditioner. Alternatively, one might choose γ and ω to minimize the asymptotic convergence factor associated with S. An in-depth theoretical examination of this problem is given in [18].

5.4. An adaptive procedure. In this section we propose an adaptive procedure for bilevel polynomials. Unlike the DR and Grcar adaptive procedures, this procedure requires little work or storage. The reason: the bilevel polynomial avoids the ambiguity inherent in the DR and Grcar polynomials. As we will see, this allows us to fully exploit the orthogonality of the CG direction vectors. Although we will concentrate on the polynomial obtained from (5.1), our procedure is easily modified for any bilevel polynomial for which Theorem 5.2 holds.

As with the DR adaptive procedure, the main difficulty is to extract an eigenvalue estimate λ for A from an eigenvalue estimate μ for $p_m(A)$. To simplify this task, we will again assume $a = \lambda_a$ and $d = \lambda_d$. As before, we will discard any $\mu \in I_1 \cup I_2$ because it has an inverse image in S. See Fig. 5.1. If $\mu > 1 + \epsilon_m$, there is a relative maximum in $(0, c)$, and we will decrease c to the nearest inverse image of μ. (If there were a relative minimum in $(b, 0)$, we might have $\mu < -1 - \epsilon_m$, in which case we would increase b to the nearest inverse image of μ.) This situation is similar to the case $\mu < 0$ in the DR adaptive procedure.

Let us now suppose $\mu \in [-1 + \epsilon_m, 1 - \epsilon_m]$. At first glance the adaptive procedure appears trivial: simply determine the unique inverse image of μ and update the set S. Unfortunately, since we know only that $\mu \in \mathcal{H}(p_m(A))$, this idea is flawed: a positive (negative) μ might be a poor estimate for a negative (positive) eigenvalue of $p_m(A)$. Instead, we shall compute an interval J_μ that contains a true eigenvalue of $p_m(A)$. The inverse image of this interval must contain a true eigenvalue of A.

To elaborate, let (μ, y) be an eigenpair for the Hermitian tridiagonal matrix $\tilde{T}_k$ (§ 3.2), and let $x = P_k D_k^{-1/2} y$ be an approximate eigenvector for $p_m(A)$. Then there is a true eigenvalue μ_t of $C(A)A$ satisfying

$$|\mu - \mu_t| \le \frac{\|C(A)Ax - \mu x\|_B}{\|x\|_B} \equiv \eta \tag{5.9}$$

where B is the hpd inner product matrix defining the polynomial preconditioned CG method [7]. By exploiting the B-orthogonality of the CG direction vectors, we may compute the right-hand side of (5.9) without explicitly computing x. This obviates

the need to store the CG direction vectors, which was a drawback of the DR adaptive procedure. In particular, we may show that

$$\eta = |y^{(k)}|/\|y\|_2 \tag{5.10}$$

where $y^{(k)}$ is the last component of the eigenvector y. We may assume $\|y\|_2 = 1$, and so (5.9) becomes

$$|\mu - \mu_t| \leq |y^{(k)}|. \tag{5.11}$$

Thus, to determine the interval $J_\mu = [\mu - \eta, \mu + \eta]$, we need only compute the last component of the eigenvector y, which is fairly inexpensive. This is a generalization of a similar result that holds for the Euclidean ($B = I$) norm; see, e.g., [34, p. 260].

If $J_\mu \subset (0, 1-\epsilon_m)$, c is decreased to the inverse image of $\mu+\eta$; if $J_\mu \subset (-1+\epsilon_m, 0)$, b is increased to the inverse image of $\mu - \eta$. Otherwise no new spectral information is available, and we resume the CG iteration using the current polynomial. Since μ is an estimate for an interior eigenvalue of the indefinite matrix $p_m(A)$, this is likely to happen in the early stages of the algorithm.

Although this strategy is similar to the one described in § 4, there are important differences. In the DR adaptive procedure, we cannot exploit the B-orthogonality of the CG direction vectors via (5.10) because of the ambiguity problem. Instead, we compute an interval on the horizontal λ-axis that is known to contain a true eigenvalue of A; recall (4.5). This requires the storage of several past CG direction vectors and nontrivial expense. In the bilevel adaptive procedure, on the other hand, we compute the interval J_μ on the vertical μ-axis, which allows us to exploit the CG orthogonality properties. Finally, we remark that as a consequence of Theorem 5.2, the bilevel polynomial is well suited to robust adaptive CG algorithms.

6. Summary. In this paper we have examined the use of polynomial preconditioning for Hermitian matrices. Polynomial preconditioning is simple, versatile, and effective. If the matrix has a regular sparsity pattern, it is also well suited to vector and vector/parallel architectures. We have shown that any residual polynomial, and hence any polynomial iterative method, may be used to define a preconditioning polynomial. If $r(\lambda)$ is the residual polynomial, $C(\lambda) = (1 - r(\lambda))/\lambda$ is the preconditioning polynomial. This suggests an inner/outer formulation for polynomial preconditioned CG methods: one uses an inner iteration to implement the preconditioning required in the outer CG iteration. We then surveyed the Chebyshev, de Boor and Rice, and Grcar preconditioning polynomials. In each case the polynomial is obtained from a minimax approximation problem, the goal of which is to cluster the eigenvalues of the preconditioned matrix around 1. In general, these polynomials are optimum in that they minimize a bound on the condition number of the hpd preconditioned matrix. We also described an adaptive procedure for each of these polynomials. Such a procedure enables one to dynamically compute the optimum preconditioning polynomial from the CG iteration parameters.

In the last section we introduced bilevel preconditioning polynomials for Hermitian indefinite matrices. These polynomials result from a radically different approach to the design of preconditioning polynomials: Instead of forcing the preconditioned matrix to be hpd, we allow it to be indefinite, but cluster its eigenvalues. The simplest bilevel polynomial is obtained from a minimax problem in which we cluster the negative eigenvalues of A around -1 and the positive eigenvalues around $+1$. We next considered clustering the negative eigenvalues around a constant γ and using a

weight function to preserve relative interval lengths. This two-parameter family of polynomials has been extensively studied by Freund [18]. Finally, we proposed an adaptive procedure for these bilevel polynomials and discussed its advantages over those for the DR and Grcar polynomials.

Acknowledgments. I wish to thank Tom Manteuffel and Paul Saylor for their interest in my research, for their guidance, and most of all, for their friendship. I also wish to thank the organizers of the SIAM Symposium on Sparse Matrices (May 1989) for inviting me to present the lecture on which this paper is based. Finally, I thank the referees for their many helpful comments.

REFERENCES

[1] L. M. ADAMS, *Iterative Algorithms for Large, Sparse Linear Systems on Parallel Computers*, Ph.D. thesis, Dept. of Applied Mathematics, University of Virginia, Charlottesville, VA, 1982.

[2] ———, *m-step preconditioned conjugate gradient methods*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 452–463.

[3] R. S. ANDERSSEN AND G. H. GOLUB, *Richardson's non-stationary matrix iterative procedure*, Tech. Report 304, Computer Science Dept., Stanford University, Stanford, CA, 1972.

[4] S. F. ASHBY, *Polynomial Preconditioning for Conjugate Gradient Methods*, Ph.D. thesis, Dept. of Computer Science, University of Illinois, Urbana, IL, December 1987. Available as Tech. Report 1355.

[5] S. F. ASHBY, T. A. MANTEUFFEL, AND J. S. OTTO, *Adaptive polynomial preconditioning for HPD linear systems*, in Proc. Ninth International Conference on Computing Methods in Applied Sciences and Engineering, R. Glowinski and A. Lichnewsky, eds., Paris, 1990, pp. 3–23.

[6] S. F. ASHBY, T. A. MANTEUFFEL, AND P. E. SAYLOR, *Adaptive polynomial preconditioning for Hermitian indefinite linear systems*, BIT, 29 (1989), pp. 583–609.

[7] ———, *A taxonomy for conjugate gradient methods*, SIAM J. Numer. Anal., 27 (1990), pp. 1542–1568.

[8] O. AXELSSON, *A survey of preconditioned iterative methods for linear systems of algebraic equations*, BIT, 25 (1985), pp. 166–187.

[9] P. N. BROWN AND A. C. HINDMARSH, *Matrix-free methods for stiff systems of ODE's*, SIAM J. Numer. Anal., 23 (1986), pp. 610–638.

[10] J. R. BUNCH AND B. N. PARLETT, *Direct methods for solving symmetric indefinite systems of linear equations*, SIAM J. Numer. Anal., 8 (1971), pp. 639–655.

[11] T. F. CHAN, C. J. KUO, AND C. TONG, *Parallel elliptic preconditioners: Fourier analysis and performance on the Connection Machine*, Tech. Report CAM 88-22, Dept. of Mathematics, University of California, Los Angeles, 1988.

[12] A. CHRONOPOULOS, *A Class of Parallel Iterative Methods Implemented on Multiprocessors*, Ph.D. thesis, Dept. of Computer Science, University of Illinois, Urbana, IL, November 1986. Available as Tech. Report 1267.

[13] A. T. CHRONOPOULOS AND C. W. GEAR, *Implementation of s-step methods on parallel vector architectures*, Tech. Report 1346, Dept. of Computer Science, University of Illinois, Urbana, IL, June 1987.

[14] P. CONCUS, G. H. GOLUB, AND D. P. O'LEARY, *A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations*, in Sparse Matrix Computations, J. R. Bunch and D. J. Rose, eds., Academic Press, New York, 1976, pp. 309–332.

[15] C. DE BOOR AND J. R. RICE, *Extremal polynomials with application to Richardson iteration for indefinite linear systems*, SIAM J. Sci. Statist. Comput., 3 (1982), pp. 47–57.

[16] P. F. DUBOIS, A. GREENBAUM, AND G. H. RODRIGUE, *Approximating the inverse of a matrix for use on iterative algorithms on vector processors*, Computing, 22 (1979), pp. 257–268.

[17] R. FREUND, *private communication*, 1986.

[18] ———, *On polynomial preconditioning for indefinite Hermitian matrices*, Tech. Report 89.32, Research Institute for Advanced Computer Science, Moffet Field, CA, August 1989.

[19] ———, *Polynomial Preconditioners for Hermitian and Certain Nonhermitian Matrices*, paper presented at SIAM Annual Meeting, San Diego, CA, July 1989.

[20] P. GILL, W. MURRAY, AND M. SAUNDERS, *Preconditioning Indefinite Systems with the Bunch–Parlett Factorization*, paper presented at SIAM Symposium on Sparse Matrices, Gleneden Beach, OR, May 1989.

[21] G. H. GOLUB AND M. D. KENT, *Estimates of eigenvalues for iterative methods*, Math. Comp., 53 (1989), pp. 619–626.

[22] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Second Edition, The Johns Hopkins University Press, Baltimore, MD, 1989.

[23] J. F. GRCAR, *Analyses of the Lanczos Algorithm and of the Approximation Problem in Richardson's Method*, Ph.D. thesis, Dept. of Computer Science, University of Illinois, Urbana, IL, December 1981. Available as Tech. Report 1074.

[24] ———, *private communication*, 1987.

[25] L. A. HAGEMAN AND D. M. YOUNG, *Applied Iterative Methods*, Academic Press, New York, 1981.

[26] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, 49 (1952), pp. 409–435.

[27] O. G. JOHNSON, C. A. MICCHELLI, AND G. PAUL, *Polynomial preconditioning for conjugate gradient calculations*, SIAM J. Numer. Anal., 20 (1983), pp. 362–376.

[28] T. L. JORDAN, *Conjugate gradient preconditioners for vector and parallel processors*, in Proc. Conference on Elliptic Problem Solvers, G. Birkhoff and A. Schoenstadt, eds., New York, 1984, Academic Press.

[29] V. I. LEBEDEV, *Iterative methods for the solution of operator equations with their spectrum lying on several intervals*, Zh. Vychisl. Mat. i Mat. Fiz., 9 (1969), pp. 1247–1252. An English translation appears in [3].

[30] J. A. MEIJERINK AND H. A. VAN DER VORST, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. Comp., 31 (1977), pp. 148–162.

[31] P. D. MEYER, A. J. VALOCCHI, S. F. ASHBY, AND P. E. SAYLOR, *A numerical investigation of the conjugate gradient method as applied to three-dimensional groundwater flow problems in randomly heterogeneous porous media*, Water Resources Res., 25 (1989), pp. 1440–1446.

[32] D. P. O'LEARY, *Hybrid Conjugate Gradient Algorithms*, Ph.D. thesis, Computer Science Dept., Stanford University, Stanford, CA, 1976. Available as Tech. Report 548.

[33] J. S. OTTO, *private communication*, 1989.

[34] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[35] R. R. ROLOFF, *Iterative Solution of Matrix Equations for Symmetric Matrices Possessing Positive and Negative Eigenvalues*, Ph.D. thesis, Dept. of Computer Science, University of Illinois, Urbana, IL, 1979. Available as Tech. Report 1018.

[36] H. RUTISHAUSER, *Theory of gradient methods*, in Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-Adjoint Boundary Value Problems, Mitt. Inst. angew. Math. ETH Zürich, Nr. 8, Birkhäuser, Basel, 1959, pp. 24–49.

[37] Y. SAAD, *Iterative solution of indefinite symmetric linear systems by methods using orthogonal polynomials over two disjoint intervals*, SIAM J. Numer. Anal., 20 (1983), pp. 784–810.

[38] ———, *Practical use of polynomial preconditionings for the conjugate gradient method*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 865–881.

[39] ———, *Least squares polynomials in the complex plane and their use for solving nonsymmetric linear systems*, SIAM J. Numer. Anal., 24 (1987), pp. 155–169.

[40] P. E. SAYLOR, *Preconditioning symmetric indefinite matrices*, in Preconditioning Methods: Analysis and Applications, D. J. Evans, ed., Gordon and Breach, New York, 1983, pp. 295–319.

[41] ———, *Leapfrog variants of iterative methods for linear algebraic equations*, J. Comput. Appl. Math., 24 (1988), pp. 169–193.

[42] D. C. SMOLARSKI, S. J. AND P. E. SAYLOR, *An optimum semi-iterative method for solving any linear set with a square matrix*, Tech. Report 1218, Dept. of Computer Science, University of Illinois, Urbana, IL, July 1985.

SIAM J. MATRIX ANAL. APPL.
Vol. 12, No. 4, pp. 790–801, October 1991

014

SOME INEQUALITIES ON THE DECOMPOSABLE NUMERICAL RADII OF MATRICES *

CHI-KWONG LI†

Abstract. Let m and n be positive integers such that $1 \leq m \leq n$. Denote by $\mathbb{C}_{n\times m}$ the set of all $n \times m$ complex matrices. For a matrix $A \in \mathbb{C}_{n\times n}$, its mth decomposable numerical radius is defined and denoted by

$$r_m^\wedge(A) = max\{|\, det(X^*AX)| : X \in \mathbb{C}_{n\times m},\ X^*X = I_m\}.$$

If $m = 1$, it reduces to the classical numerical radius of A, which is denoted by $r(A)$; and $r_n^\wedge(A) = |\, det(A)|$. In this note we prove the inequalities

$$r(A) \equiv r_1^\wedge(A) \geq r_2^\wedge(A)^{1/2} \geq \cdots \geq r_{n-1}^\wedge(A)^{1/(n-1)} \geq r_n^\wedge(A)^{1/n} \equiv |\, det(A)|^{1/n},$$

and

$$\binom{n}{m} r_m^\wedge(A) \geq E_m(\sigma_1(A), \cdots, \sigma_n(A)),$$

where $E_m(\cdot)$ denotes the mth elementary symmetric function, and $\sigma_1(A) \geq \cdots \geq \sigma_n(A)$ are the singular values of A. Complete characterizations of the matrices for which any one of the equalities holds are given.

Key words. decomposable numerical radius, unitary similarity, compound matrix

AMS(MOS) subject classification. 15A60

1. Introduction.

Let m and n be positive integers such that $1 \leq m \leq n$. Denote by $\mathbb{C}_{n\times m}$ the set of all $n \times m$ complex matrices. For a matrix $A \in \mathbb{C}_{n\times n}$, its mth *decomposable numerical radius* $(1 \leq m \leq n)$ is defined and denoted by

$$r_m^\wedge(A) = \max\{|\det(X^*AX)| : X \in \mathbb{C}_{n\times m},\ X^*X = I_m\}.$$

Let $\mathcal{U}_n$ be the set of all $n \times n$ unitary matrices, and let $Q_{m,n}$ be the set of all strictly increasing sequences of m integers chosen from the set $\{1, \cdots, n\}$.

Then one easily verifies that

$$r_m^\wedge(A) = \max\{|\det\ U^*AU[\omega]| : U \in \mathcal{U}_n,\ \omega \in Q_{m,n}\},$$

where for any $\omega \in Q_{m,n}$, $X[\omega]$ denotes the principal submatrix of $X \in \mathbb{C}_{n\times n}$ lying in rows and columns $\omega(1), \cdots, \omega(m)$. If $m = 1$, $r_m^\wedge(A)$ reduces to the *classical numerical radius* of A, which is denoted by $r(A)$, and $r_n^\wedge(A) = |\det(A)|$. It is known (e.g., see [2], [3], [4], [9] and their references) that the classical numerical radius is a norm on $\mathbb{C}_{n\times n}$ which is not submultiplicative, and is useful in study of various subjects. There has been a great deal of interest in studying inequalities involving $r(\cdot)$. For example, it is known (e.g., see [3], [4], [13]) that

$$\rho(A) \leq r(A) \leq \sigma_1(A) \leq 2r(A) \tag{1}$$

*Received by the editors March 20, 1990; accepted for publication (in revised form) November 13, 1990. This research was supported by the National Science Foundation grant DMS 89 00922.

†Department of Mathematics, The College of William and Mary, Williamsburg, Virginia 23185 (ckli@cma.math.wm.edu).

and

$$nr(A) \geq \sigma_1(A) + \cdots + \sigma_n(A), \tag{2}$$

where $\rho(A)$ is the spectral radius and $\sigma_1(A) \geq \cdots \geq \sigma_n(A)$ are the singular values of A. Moreover, characterizations of the matrices for which any one of the equalities holds are known. It is worth noting that these inequalities are useful in the study of other subjects, such as unitarily invariant norms (e.g., see [5]). The decomposable numerical radius is one of the many interesting generalizations of the classical concept. For example, another generalization of $r(\cdot)$ is the mth *higher numerical radius* of A defined and denoted by

$$r_m(A) = \max\{|\mathrm{tr}(X^*AX)| : X \in \mathbb{C}_{n\times m},\ X^*X = I_m\},$$

and it is known (see [2], [3], [4], [7]) that

$$r(A) \equiv r_1(A) \geq \frac{1}{2}r_2(A) \geq \cdots \geq \frac{1}{n}r_n(A) \equiv \frac{1}{n}|\mathrm{tr}\ A|, \tag{3}$$

and for $n > m+1$,

$$\left(1 + \frac{2}{\min\{m, n-m-1\}}\right) r_{m+1}(A) \geq \left(1 + \frac{1}{m}\right) r_m(A). \tag{4}$$

As pointed out in [12] (see also [1], [11]), the mth decomposable numerical radius can be considered in the context of the mth exterior space $\wedge^m\mathbb{C}^n$ over $\mathbb{C}^n$, and defined as

$$\begin{aligned} r_m^{\wedge}(A) &= \max\{|\langle C_m(A)v, v\rangle| : v \text{ is a decomposable unit vector in } \wedge^m \mathbb{C}^n\} \\ &= \max\{|\det(X^*AX)| : X \in \mathbb{C}_{n\times m},\ \det(X^*X) = 1\}, \end{aligned}$$

where $C_m(A)$ denote the mth compound matrix of A (e.g., see [10] for the definition and properties). This makes the subject more interesting, and in fact, it attracted the attention of many authors in recent years (e.g., see [1], [6], [8], [11], [14]). Other interesting inequalities related to the subject include (e.g., see [1], [6], [11])

$$\rho(C_m(A)) \leq r_m^{\wedge}(A) \leq r(C_m(A)) \leq \sigma_1(C_m(A)) \equiv \prod_{j=1}^{m} \sigma_j(A). \tag{5}$$

The conditions on A for which the equalities hold have been obtained. These results can be viewed as an extension of those concerning the inequalities in (1). In this paper we generalize inequality (2) to

$$\binom{n}{m} r_m^{\wedge}(A) \geq E_m(\sigma_1(A), \ldots, \sigma_n(A)), \tag{6}$$

where $E_m(\cdot)$ denotes the mth elementary symmetric function, and obtain the following analog of (3) for decomposable numerical radii

$$r(A) \equiv r_1^{\wedge}(A) \geq r_2^{\wedge}(A)^{1/2} \geq \cdots \geq r_{n-1}^{\wedge}(A)^{1/(n-1)} \geq r_n^{\wedge}(A)^{1/n} \equiv |\det(A)|^{1/n}. \tag{7}$$

Examples are given to show that there is no hope to obtain inequalities like those in (4). Moreover, complete characterizations of the matrices for which any one of the equalities in (6) or (7) holds are obtained.

Note that if $A \in \mathbb{C}_{n\times n}$ has rank smaller than m, then $r_m^\wedge(A) = \cdots = r_n^\wedge(A) = 0$. So we always assume that $\operatorname{rank}(A) \geq m$ in our results to avoid trivial consideration. For $A \in \mathbb{C}_{n\times n}$, denote by A_{ij} the (i,j) entry and denote by $\operatorname{adj}(A)$ the adjoint of A. In our discussion, we shall frequently use the following proposition whose proof can be verified readily.

PROPOSITION 1.1. *Let $Q \in \mathbb{C}_{n\times n}$ be such that*

$$Q_{ij} = \begin{cases} (-1)^i & \text{if } i+j = n+1, \\ 0 & \text{otherwise.} \end{cases}$$

Then $C_{n-1}(A) = (-1)^{n+1}Q(\operatorname{adj}(A)^t)Q$. As a result, $C_{n-1}(A)$ is unitarily similar to $\operatorname{adj}(A)^t$, and

$$r_{n-1}^\wedge(A) = r(C_{n-1}(A)) = r(\operatorname{adj}(A)).$$

2. Inequalities relating decomposable numerical radii.

THEOREM 2.1. *Let $1 \leq m < n$. Suppose $A \in \mathbb{C}_{n\times n}$ has rank at least m. Then*

$$r_m^\wedge(A)^{1/m} \geq r_{m+1}^\wedge(A)^{1/(m+1)}.$$

The equality holds if and only if there exists $X \in \mathbb{C}_{n\times(m+1)}$ with $X^*X = I_{m+1}$ such that $\mu^{-1}(X^*AX)$ is unitary where $\mu = r_m^\wedge(A)^{1/m}$.

Proof. We may assume $\operatorname{rank}(A) \geq m+1$. Let $X \in \mathbb{C}_{n\times(m+1)}$ satisfy $X^*X = I_{m+1}$ and $|\det(X^*AX)| = r_{m+1}^\wedge(A)$. Let $U \in \mathcal{U}_{m+1}$ be such that the matrix U^*X^*AXU is in lower triangular form with diagonal entries $\lambda_1, \cdots, \lambda_{m+1}$. Suppose $X_k \in \mathbb{C}_{n\times m}$ is obtained from XU by deleting its kth column for $k = 1, \cdots, m+1$. Then

$$\nu_k = \left| \prod_{j=1}^{m+1} \lambda_j/\lambda_k \right| = |\det(X_k^*AX_k)| \leq r_m^\wedge(A).$$

Thus

$$r_{m+1}^\wedge(A)^m = \prod_{k=1}^{m+1} \nu_k \leq r_m^\wedge(A)^{m+1}.$$

If the equality holds, then $\nu_k = r_m^\wedge(A)$ for $k = 1, \cdots, m+1$. It follows that $|\lambda_k| = \mu \equiv r_m^\wedge(A)^{1/m}$ for $k = 1, \cdots, m+1$. Let $A' = \mu^{-1}(X^*AX)$. Then $1 = |\det A'| \leq r_m^\wedge(A')$. Since $r_m^\wedge(A') \leq r_m^\wedge(\mu^{-1}A) = 1$, it follows that $r_m^\wedge(A') = |\det A'| = 1$. By the Corollary in [8], A' is unitary. Conversely, if there exists $X \in \mathbb{C}_{n\times(m+1)}$ with $X^*X = I_{m+1}$ such that $\mu^{-1}(X^*AX)$ is unitary where $\mu = r_m^\wedge(A)^{1/m}$, then $r_{m+1}^\wedge(A) \geq |\det(X^*AX)| = r_m^\wedge(A)^{(m+1)/m}$ and hence $r_{m+1}^\wedge(A)^{1/(m+1)} = r_m^\wedge(A)^{1/m}$. □

By Theorem 2.1, we have the following corollary (cf. [13, Cor. 1]).

COROLLARY 2.2. *Let $A \in \mathbb{C}_{n\times n}$.*

(a) *If νA is unitary for some $\nu \in \mathbb{C}$, then*

$$r(A) \equiv r_1^\wedge(A) = r_2^\wedge(A)^{1/2} = \cdots = r_{n-1}^\wedge(A)^{1/(n-1)} = r_n^\wedge(A)^{1/n} \equiv |\det(A)|^{1/n}.$$

(b) *If there exists m with $1 \leq m < n$ such that $\operatorname{rank}(A) \geq m$ and $r_m^\wedge(A)^{1/m} = r_n^\wedge(A)^{1/n}$, then νA is unitary for some nonzero $\nu \in \mathbb{C}$.*

It is always the case (e.g., see [1], [6], [11]) that if equality holds for certain inequalities involving $r_m^\wedge(A)$, then A is unitarily similar to the direct sum of matrices of smaller sizes. In Theorem 2.1, the matrix X^*AX can be regarded as $U^*AU[1,\cdots,m+1]$ for some $U \in \mathcal{U}_n$. One may ask whether we can further prove that if the equality in Theorem 2.1 holds, then A is unitarily similar to the direct sum of X^*AX and a matrix $B \in \mathbb{C}_{(n-m-1)\times(n-m-1)}$. The following example shows that there is no hope to obtain such a result.

Example 2.3. Let $1 \le m < n-1$ and let $A = I_m \oplus \left[\begin{smallmatrix} 1 & 1 \\ -1 & -1 \end{smallmatrix}\right] \oplus 0_{n-m-2}$. Then (see [6, Example 2]) $1 = r_m^\wedge(A)^{1/m} = r_{m+1}^\wedge(A)^{1/(m+1)}$, but A is not unitarily similar to $A_1 \oplus A_2$ such that $A_1 \in \mathbb{C}_{(m+1)\times(m+1)}$ with $|\det(A_1)| = r_{m+1}^\wedge(A)$.

Note that if $A \in \mathbb{C}_{n\times n}$ has rank m, then $r_m^\wedge(A) > 0 = r_{m+1}^\wedge(A)$. Even for nonsingular A, we show that there is no hope to find positive real numbers η and ν such that $\nu r_{m+1}^\wedge(A)^\eta \ge r_m^\wedge(A)$ in the following proposition.

PROPOSITION 2.4. *Let $1 \le m < n$. Then for any positive real numbers η and ν, there exists a matrix $A = I_m \oplus \varepsilon I_{n-m}$ with $\varepsilon > 0$ such that $r_m^\wedge(A) > \nu r_{m+1}^\wedge(A)^\eta$.*

Proof. For any positive real numbers η and ν, we can find $\varepsilon > 0$ small enough such that $1 > \nu\varepsilon^\eta$. It is well known that for a positive-definite hermitian matrix A, $r_k^\wedge(A) = \rho(C_k(A))$, which is the product of the k largest eigenvalues of A. Hence, if $A = I_m \oplus \varepsilon I_{n-m}$, then $r_m^\wedge(A) = 1 > \nu\varepsilon^\eta = \nu r_{m+1}^\wedge(A)^\eta$. □

3. Relation with singular values. In this section we study the inequality

$$\binom{n}{m} r_m^\wedge(A) \ge E_m(\sigma_1(A),\cdots,\sigma_n(A)),$$

and the conditions on those matrices A for which the equality holds. When $m = n$, the equality holds for any matrix. When $m = 1$, Marcus and Sandy [13] have obtained the following result.

THEOREM 3.1. *Let $A \in \mathbb{C}_{n\times n}$ be a nonzero matrix. Then*

$$nr(A) \ge \sigma_1(A) + \cdots + \sigma_n(A).$$

The equality holds if and only if $r(A)^{-1}A$ is unitarily similar to a direct sum of unit multiples of 2×2 matrices of the form

$$\begin{bmatrix} 1 & d \\ -\overline{d} & -1 \end{bmatrix},$$

with $0 < |d| \le 1$, together with a diagonal unitary matrix.

When $m = n-1$, we have the following result.

THEOREM 3.2. *Let $n > 2$ and $A \in \mathbb{C}_{n\times n}$ have rank at least $n-1$. Then*

$$nr_{n-1}^\wedge(A) \ge E_{n-1}(\sigma_1(A),\cdots,\sigma_n(A)).$$

The equality holds if and only if any one of the following conditions holds.

(a) *$r(C_{n-1}(A))^{-1}C_{n-1}(A)$ is unitarily similar to a direct sum of unit multiples of 2×2 matrices of the form*

$$\begin{bmatrix} 1 & d \\ -\overline{d} & -1 \end{bmatrix},$$

with $0 < |d| < 1$, together with a diagonal unitary matrix.

(b) *The matrix A is nonsingular and $r(A^{-1})^{-1}A^{-1}$ is unitarily similar to a direct sum of unit multiples of 2×2 matrices of the form*

$$\begin{bmatrix} 1 & d \\ -\bar{d} & -1 \end{bmatrix},$$

with $0 < |d| < 1$, together with a diagonal unitary matrix.

(c) *A is unitarily similar to A' such that $|\det A'[\omega]| = r^{\wedge}_{n-1}(A)$ for all $\omega \in Q_{n-1,n}$, and for some positive number ν, the matrix $\nu A'$ is a direct sum of unit multiples of 2×2 matrices of the form*

$$\frac{1}{1-|d|^2}\begin{bmatrix} -1 & d \\ -\bar{d} & 1 \end{bmatrix},$$

with $0 < |d| < 1$, together with a diagonal unitary matrix.

Proof. Note that $r^{\wedge}_{n-1}(A) = r(C_{n-1}(A))$ and

$$E_{n-1}(\sigma_1(A), \cdots, \sigma_n(A)) = \sigma_1(C_{n-1}(A)) + \cdots + \sigma_n(C_{n-1}(A)).$$

Applying Theorem 3.1 to $C_{n-1}(A)$, we get the inequality.

Suppose the equality holds. Then A must be nonsingular, otherwise

$$nr(C_{n-1}(A)) > 2r(C_{n-1}(A)) \geq \sigma_1(C_{n-1}(A)) = E_{n-1}(\sigma_1(A), \cdots, \sigma_n(A)).$$

Now the condition (a) follows from Theorem 3.1.

Suppose condition (a) holds. Then, clearly, A is nonsingular. Since $C_{n-1}(A)^t$ is unitarily similar to $\operatorname{adj}(A) = \det(A)A^{-1}$, the matrix $r(C_{n-1}(A))^{-1}C_{n-1}(A)^t$ is unitarily similar to $r(\operatorname{adj}(A))^{-1}(\det(A)A^{-1})$, which is a unit multiple of $r(A^{-1})^{-1}A^{-1}$ as $r(\operatorname{adj}(A)) = r(\det(A)A^{-1}) = |\det(A)|r(A^{-1})$. Therefore A satisfies (b).

Suppose condition (b) holds and let $U \in \mathcal{U}_n$ be such that $U^*(r(A^{-1})^{-1}A^{-1})U$ is of the form described in (b). Then $A' = U^*AU$ will be of the form described in (c).

Suppose condition (c) holds. Then (e.g., see [15])

$$nr^{\wedge}_{n-1}(A) = \sum_{j=1}^{n} |C_{n-1}(A')_{jj}| \leq \sum_{j=1}^{n} \sigma_j(C_{n-1}(A')) = E_{n-1}(\sigma_1(A), \cdots, \sigma_n(A)).$$

Thus the equality holds. □

For $1 < m < n-1$, we have the following.

THEOREM 3.3. *Let $1 < m < n-1$, and let $A \in \mathbb{C}_{n\times n}$ have rank at least m. Then*

$$\binom{n}{m} r^{\wedge}_m(A) \geq E_m(\sigma_1(A), \cdots, \sigma_n(A)).$$

The equality holds if and only if νA is unitary for some nonzero $\nu \in \mathbb{C}$.

The long technical proof of Theorem 3.3 is postponed until the next section. We consider two consequences of the theorem in the following.

In [11], Marcus and Andresen attempted to prove the inequality

$$\binom{n}{m} r^{\wedge}_m(A) > \prod_{j=1}^{m} \sigma_j(A)$$

for $1 < m < n$ and obtained the result when $(m,n) \neq (2,4)$. We have the following corollary, which settles their unsolved case.

COROLLARY 3.4. *Let $1 < m < n$, and let $A \in \mathbb{C}_{n\times n}$ have rank at least m. Then*

$$\binom{n}{m} r_m^\wedge(A) > \prod_{j=1}^{m} \sigma_j(A).$$

Proof. By Theorems 3.2 and 3.3, we have

$$\binom{n}{m} r_m^\wedge(A) \geq E_m(\sigma_1(A), \cdots, \sigma_n(A)) \geq \prod_{j=1}^{m} \sigma_j(A).$$

Moreover, the first equality holds only if A is nonsingular. Clearly, the second equality holds if and only if A has rank m. Thus the two equalities cannot hold simultaneously. □

COROLLARY 3.5. *Let $1 < m < n-1$, and let $A \in \mathbb{C}_{n\times n}$ have rank at least m. Then*

$$\binom{n}{m} r(C_m(A)) \geq E_m(\sigma_1(A), \cdots, \sigma_n(A)).$$

The equality holds if and only if νA is unitary for some nonzero $\nu \in \mathbb{C}$.

Proof. Since $r(C_m(A)) \geq r_m^\wedge(A)$, the inequality follows from Theorem 3.3. If the equality holds, then

$$r(C_m(A)) = r_m^\wedge(A) = E_m(\sigma_1(A), \cdots, \sigma_n(A)).$$

By Theorem 3.3, νA is unitary for some nonzero $\nu \in \mathbb{C}$.

Conversely, if νA is unitary for some nonzero $\nu \in \mathbb{C}$, then $C_m(A)$ is a multiple of a unitary matrix and hence

$$\binom{n}{m} r(C_m(A)) = E_m(\sigma_1(A), \cdots, \sigma_n(A)) = \binom{n}{m} |\nu|^{-m}. \qquad \square$$

4. Proof of Theorem 3.3. This section is devoted to proving Theorem 3.3. We shall always assume that $1 < m < n-1$ and that $A \in \mathbb{C}_{n\times n}$ has rank at least m.

LEMMA 4.1. *There exists $U \in \mathcal{U}_n$ such that $U^*AU = DH$, where D is a diagonal unitary matrix and H is a positive-semidefinite hermitian matrix with eigenvalues $\sigma_1(A) \geq \cdots \geq \sigma_n(A)$.*

Proof. Let $A = VK$, where $V \in \mathcal{U}_n$ and K is a positive-semidefinite hermitian matrix with eigenvalues $\sigma_1(A) \geq \cdots \geq \sigma_n(A)$. Suppose $V = UDU^*$ where $U \in \mathcal{U}_n$ and D is a diagonal unitary matrix. Then $U^*AU = DU^*KU$ satisfies the conditions of the lemma. □

LEMMA 4.2. *Suppose $U \in \mathcal{U}_n$ satisfies the conditions of Lemma 4.1. Then*

$$\binom{n}{m} r_m^\wedge(A) \geq \sum_{\omega \in Q_{m,n}} |\det\, U^*AU[\omega]| = E_m(\sigma_1(A), \cdots, \sigma_n(A)).$$

The equality holds if and only if $r_m^\wedge(A) = |\det\, U^*AU[\omega]|$ for all $\omega \in Q_{m,n}$.

Proof. Note that

$$r_m^\wedge(A) \geq |C_m(U^*AU)_{jj}| = C_m(H)_{jj}$$

for $j = 1, \cdots, \binom{n}{m}$, and hence

$$\binom{n}{m} r_m^\wedge(A) \geq \sum_j |C_m(U^*AU)_{jj}| = \text{tr } C_m(H) = E_m(\sigma_1(A), \cdots, \sigma_n(A)).$$

Clearly, the equality holds if and only if $r_m^\wedge(A) = |\det U^*AU[\omega]|$ for all $\omega \in Q_{m,n}$. □

LEMMA 4.3. *If νA is unitary for some nonzero $\nu \in \mathbb{C}$, then*

$$\binom{n}{m} r_m^\wedge(A) = E_m(\sigma_1(A), \cdots, \sigma_n(A)).$$

Proof. Suppose νA is unitary for some nonzero $\nu \in \mathbb{C}$. Then $\nu^m C_m(A)$ is unitary and hence

$$|\nu|^{-m} = r(C_m(A)) \geq r_m^\wedge(A) \geq \binom{n}{m}^{-1} E_m(\sigma_1(A), \cdots, \sigma_n(A)) = |\nu|^{-m}. \qquad \square$$

We divide the proof of the converse of Lemma 4.3 into several lemmas. In the rest of the section we shall assume that A satisfies $\binom{n}{m} r_m^\wedge(A) = E_m(\sigma_1(A), \cdots, \sigma_n(A))$. Furthermore, after applying a suitable unitary similarity transform and multiplying A by a suitable constant, we assume that $A = DH$ such that D is a diagonal unitary matrix and H is a positive-semidefinite hermitian matrix with eigenvalues $\sigma_1(A) \geq \cdots \geq \sigma_n(A)$; and $|\det A[\omega]| = r_m^\wedge(A) = 1$ for all $\omega \in Q_{m,n}$.

LEMMA 4.4. *If B is a principal $k \times k$ submatrix of A with $k > m$, then*

$$1 = |\det B[\delta]| = r_m^\wedge(B) = \binom{k}{m}^{-1} E_m(\sigma_1(B), \cdots, \sigma_k(B))$$

for all $\delta \in Q_{m,k}$.

Proof. Suppose $B = A[\gamma] = D[\gamma]H[\gamma]$ where $\gamma \in Q_{k,n}$. Then

$$1 = r_m^\wedge(A) \geq r_m^\wedge(B) \geq |\det B[\delta]| = 1$$

for all $\delta \in Q_{m,k}$, and the result follows from Lemma 4.2. □

LEMMA 4.5. *Suppose P is a permutation matrix with $P^tDP = D_0 \oplus D_1 \oplus \cdots \oplus D_k$ such that D_0 and $-D_0$ have no common eigenvalues, $D_i = \theta_i I_{p_i} \oplus (-\theta_i) I_{q_i}$ for $i = 1, \cdots, k$, and D_i and D_j have no common eigenvalues for $0 \leq i < j \leq k$. Then*

$$P^tAP = A_0 \oplus A_1 \oplus \cdots \oplus A_k$$

such that A_i and D_i have the same size for $i = 0, 1, \cdots, k$, and A_0 is a digonal matrix.

Proof. To prove the lemma, we show that $A_{ij} = 0$ if (i) $D_{ii} \neq D_{jj}$ and $D_{ii} \neq -D_{jj}$; or (ii) $D_{ii} = D_{jj} \neq -D_{tt}$ for all t, as follows. Let (i, j) satisfy (i) or (ii) and let $B = A[\gamma]$ with $\gamma \in Q_{m+1,n}$ such that $i, j \in \{\gamma(1), \cdots, \gamma(m+1)\}$. For simplicity, we assume $\gamma(t) = t$ for $t = 1, \cdots, m+1$, and $i < j$. Then $(m+1) r_m^\wedge(B) = E_m(\sigma_1(B), \cdots, \sigma_{m+1}(B))$ by Lemma 4.4. Note that $E_m(\sigma_1(B), \cdots, \sigma_{m+1}(B)) = \sum_{t=1}^{m+1} \sigma_t(C_m(B))$ and $r_m^\wedge(B) = r(C_m(B))$. It follows that $(m+1) r(C_m(B)) = \sum_{t=1}^{m+1} \sigma_t(C_m(B))$. Also, note that $C_m(B) = C_m(D[\gamma]) C_m(H[\gamma])$, and $C_m(D[\gamma])_{tt} = \eta_t$ with $\eta_{m-t+2} = \det(D[\gamma])/D_{tt}$ for $t = 1, \cdots, m+1$. By our assumption on D_{ii} and D_{jj}, we see that (i) $\eta_{m-i+2} \neq \eta_{m-j+2}$ and $\eta_{m-i+2} \neq -\eta_{m-j+2}$, or (ii)

$\eta_{m-i+2} = \eta_{m-j+2} \neq -\eta_t$ for all t. As a result, either (i) the $(m-j+2)$nd and the $(m-i+2)$nd diagonal entries of $C_m(B)$ are neither the same nor the negative to each other, or (ii) the $(m-j+2)$nd and the $(m-i+2)$nd diagonal entries of $C_m(B)$ are the same but different from the negative of any other diagonal entry. By Lemma 6 (see also the proof of Lemma 7) in [13], we see that $C_m(B)$ is nonsingular and is the direct sum of matrices of smaller sizes. In particular, we can find a positive integer p satisfying $m-j+2 \leq p < m-i+2$ such that $C_m(B) = B_1 \oplus B_2$ with $B_1 \in \mathbb{C}_{p\times p}$. (Note that B_1 and B_2 may be the direct sum of matrices of smaller sizes.) Since $B^{-1} = (\det\ B)^{-1}(\mathrm{adj}(B)) = (-1)^{m+2}(\det\ B)^{-1}(QC_m(B)Q)^t$ where

$$Q_{st} = \begin{cases} (-1)^s & \text{if } s+t = m+2, \\ 0 & \text{otherwise.} \end{cases}$$

It follows that $B = (-1)^{m+2}(\det\ B)(Q^t)^{-1}(C_m(B)^t)^{-1}(Q^t)^{-1}$. Since $(C_m(B)^t)^{-1} = (B_1^t)^{-1} \oplus (B_2^t)^{-1}$, one easily checks that $B = B_1' \oplus B_2'$ with $B_2' \in \mathbb{C}_{p\times p}$. Thus $A_{ij} = 0$. □

Note that $A = A_0 \oplus A_1 \oplus \cdots \oplus A_k$ implies $H = H_0 \oplus H_1 \oplus \cdots \oplus H_k$, accordingly. We shall prove that $H = I$ in order to establish the converse, Lemma 4.3. The proof will be by induction on m. We first obtain the following two lemmas, which are useful for the induction step.

LEMMA 4.6. *Suppose $n = m+2$. There exists an $n\times n$ unitary matrix V satisfying $VD = DV$ such that V^*AV is a direct sum of matrices of smaller sizes.*

Proof. Suppose A is not the direct sum of matrices of smaller sizes. By Lemma 4.5, we may assume that $D = \theta I_p \oplus (-\theta)I_{n-p}$ for some $\theta \in \mathbb{C}$ with $|\theta| = 1$ and $0 < p < n$. For simplicity, we assume that $\theta = 1$. Let A' be obtained from A by deleting its kth row and kth column, where $k = 1$ or n, so that the number of positive diagonal entries and the number of negative diagonal entries are both nonzero and are different. For simplicity, we assume $k = n$ and the number of positive diagonal entries of A' equals $p > n-1-p$, which is the number of negative diagonal entries of A'. Let $D' = D[1,\cdots,n-1]$ and $H' = H[1,\cdots,n-1]$. Then $C_m(A') = C_m(D')C_m(H')$ satisfies

$$(m+1)r(C_m(A')) = \sigma_1(C_m(A')) + \cdots + \sigma_{m+1}(C_m(A')).$$

By Theorem 3.2, A' is nonsingular. Since $C_m(D') = \delta I_{n-1-p} \oplus (-\delta)I_p$ with $\delta = 1$ or -1, and $\det\ H'[\gamma] = 1$ for all $\gamma \in Q_{m,m+1}$, by Lemma 6 (see also the proof of Lemma 7) in [13],

$$C_m(A') = \begin{bmatrix} \delta I_{n-1-p} & Y \\ -Y^* & -\delta I_p \end{bmatrix}.$$

Let $W_1 \in \mathcal{U}_{n-p-1}$ and $W_2 \in \mathcal{U}_p$ be such that $(W_1^*YW_2)_{ii} = \sigma_i(Y)$ for $i = 1,\cdots,n-p-1$. Then for $W = W_1 \oplus W_2$ the matrix $W^*C_m(A')W$ is unitarily similar to a direct sum of unit multiples of 2×2 matrices of the form

$$\begin{bmatrix} 1 & d \\ -\overline{d} & -1 \end{bmatrix}$$

with $0 \leq |d| < 1$, together with a diagonal unitary matrix. Note that

$$A' = (-1)^{m+2}(\det\ A')Q(C_m(A')^t)^{-1}Q,$$

where

$$Q_{st} = \begin{cases} (-1)^s & \text{if } s+t = m+2, \\ 0 & \text{otherwise.} \end{cases}$$

If $U = Q\overline{W}Q \in \mathcal{U}_{n-1}$, then $U = U_1 \oplus U_2$ with $U_1 \in \mathcal{U}_p$ and $U_2 \in \mathcal{U}_{n-1-p}$. Moreover, $U^*A'U$ is a nonzero multiple of a matrix which is the direct sum of 2×2 matrices of the form

$$(1 - |d|^2)^{-1} \begin{bmatrix} -1 & d \\ -\bar{d} & 1 \end{bmatrix},$$

together with a diagonal unitary matrix. By the assumption that $p > n - 1 - p > 0$, we see that there is at least one 2×2 block and that the diagonal part is nontrivial. Let P be a permutation matrix such that

$$B = P^tU^*A'UP = [B_{11}] \oplus \begin{bmatrix} B_{22} & B_{23} \\ -\overline{B_{23}} & -B_{22} \end{bmatrix} \oplus B',$$

where B_{11}, $B_{22} > 0$. Suppose $V = UP \oplus [1]$. Then

$$V^*AV = \begin{bmatrix} B & X \\ Y^* & A_{nn} \end{bmatrix},$$

where X, $Y \in \mathbb{C}^{n-1}$. Note that $V^*AV = (V^*DV)(V^*HV)$, where V^*DV is still in diagonal form with $(V^*DV)_{11} = (V^*DV)_{22} = 1$ and $(V^*DV)_{33} = -1$. We shall prove that $(V^*HV)_{1n} = (V^*HV)_{n1} = 0$ and hence $V^*AV = [(V^*AV)_{11}] \oplus V^*AV[2, \cdots, n]$.

Let $A'' = V^*AV$, $H'' = V^*HV$, and $\lambda = \det\ H''[4, \cdots, n]$. Suppose $H''_{1n} \neq 0$. Since

$$1 = r_m^\wedge(A) = |\det\ A''[1, 4, \cdots, n]| = \det\ H''[1, 4, \cdots, n],$$

the matrix $H''[1, 4, \cdots, n]$ is nonsingular and hence is positive definite. Thus the matrix $G = H''[4, \cdots, n]$ is also nonsingular. Let μ be the last diagonal entry of G^{-1}. Then

$$1 = |\det\ A''[1, 4, \cdots, n]| = \det\ H''[1, 4, \cdots, n] = \lambda(H''_{11} - \mu|H''_{1n}|^2)$$

and

$$1 = |\det\ A''[2, 4, \cdots, n]| = \det\ H''[2, 4, \cdots, n] = \lambda(H''_{22} - \mu|H''_{2n}|^2).$$

Since $|A''_{11}/A''_{22}| = 1 - |d|^2$ for some $d \in \mathbb{C}$, we have $|H''_{2n}| \geq |H''_{1n}| > 0$. For $\phi \in [0, 2\pi)$, let $A''_\phi = Z_\phi^* A'' Z_\phi$ with

$$Z_\phi = \begin{bmatrix} \alpha \cos\ \phi & \alpha \sin\ \phi \\ \beta \sin\ \phi & -\beta \cos\ \phi \end{bmatrix} \oplus I_m,$$

where α and β satisfy $\overline{\alpha}H''_{1n} = |H''_{1n}|$ and $\overline{\beta}H''_{2n} = |H''_{2n}|$, respectively. Then A''_ϕ is still the product of a diagonal unitary matrix and a positive-semidefinite matrix. Thus

$$1 = |\det\ A''_\phi[1, 4, \cdots, n]| = \lambda(H''_{11} \cos^2 \phi + H''_{22} \sin^2 \phi - \mu(|H''_{1n}| \cos\ \phi + |H''_{2n}| \sin\ \phi)^2).$$

Putting $\phi = \pi/4$ and $\phi = 3\pi/4$, we see that at least one of H''_{1n} or H''_{2n} is zero, which is a contradiction. Thus the assumption of $H''_{1n} \neq 0$ cannot hold and the result follows. □

LEMMA 4.7. *Suppose $n = m + 2$ and $A = A_1 \oplus A_2$ such that $A_2 \in \mathbb{C}_{p\times p}$ with $p > 2$. Then A_2 satisfies*

$$\binom{p}{p-2} r_{p-2}^\wedge(A_2) = E_{p-2}(\sigma_1(A_2), \cdots, \sigma_p(A_2)).$$

Proof. Suppose $|\det\ A_1| = \eta$. Then $\eta^{-1} = |\det\ A_2[\gamma]| = r^{\wedge}_{p-2}(A_2)$; otherwise there exists $U \in \mathcal{U}_p$ such that $|\det\ U^*A_2U[1,\cdots,p-2]| > \eta^{-1}$ and hence

$$r^{\wedge}_m(A) \geq |\det\ A_1||\det\ U^*A_2U[1,\cdots,p-2]| > 1.$$

Applying Lemma 4.2 to A_2, we get the conclusion. □

The next two lemmas take care of the initial cases of the induction steps.

LEMMA 4.8. *If $n = m+2$ with $m = 2$ or 3, then $H = I$.*

Proof. By Lemma 4.6, we may assume that A is a direct sum of matrices. First we assume $m = 2$ and consider two cases.

Case 1. Suppose H is the direct sum of a 1×1 matrix F and another matrix G. For simplicity, we assume $H = F \oplus G$. By Lemma 4.7, if $A' = A[2,3,4]$, then $3r(A') = \sigma_1(A') = \sigma_2(A') = \sigma_3(A')$. By Theorem 3.1, we may assume that A', and hence G, is a direct sum of matrices of smaller sizes, and $G_{11} = G_{22} = G_{33} = r(A')$. Since $G_{ii}G_{jj} - |G_{ij}|^2 = r^{\wedge}_2(A) = 1$ for all $1 \leq i < j \leq 3$, we have $|G_{12}| = |G_{13}| = |G_{23}|$. As G is the direct sum of matrices of smaller sizes, we have $|G_{12}| = |G_{13}| = |G_{23}| = 0$. It follows that $H = I_4$.

Case 2. Suppose H is the direct sum of two 2×2 matrices, say $H = F \oplus G$. It follows that $F_{ii}G_{jj} = r^{\wedge}_2(A) = 1$ for $i,j = 1,2$. Thus $F_{11} = F_{22}$ and $G_{11} = G_{22}$. Since $\det\ F = \det\ G = 1$, we have $F_{ii} \geq 1$ and $G_{ii} \geq 1$ for $i = 1,2$. It follows that $F_{ii} = G_{ii} = 1$ for $i = 1,2$, and $F_{12} = G_{12} = 0$.

Now suppose $m = 3$. Again, we consider two cases.

Case 1′. Suppose A is the direct sum of a 1×1 matrix A_1 and another matrix A_2, say $A = A_1 \oplus A_2$. Let $H = F \oplus G$ accordingly. By Lemma 4.7,

$$\binom{4}{2} r^{\wedge}_2(A_2) = E_2(\sigma_1(A_2),\cdots,\sigma_4(A_2)).$$

By the result, when $m = 2$ (proved above), $G = \eta I$ for some positive number η. Since $\det\ H[\gamma] = 1$ for all $\gamma \in Q_{3,5}$, we conclude that $H = I$.

Case 2′. Case 1′ does not hold. Then A must be the direct sum of a 2×2 matrix A_1 and a 3×3 matrix A_2, say $A = A_1 \oplus A_2$. Let $D = D_1 \oplus D_2$ and $H = F \oplus G$ accordingly. By Lemma 4.7, if $A' = A[3,4,5]$, then $3r(A') = \sigma_1(A') = \sigma_2(A') = \sigma_3(A')$. By Theorem 3.1, A' is unitarily similar to D_2G' such that G' is a direct sum of matrices of smaller sizes, and $G'_{11} = G'_{22} = G'_{33} = r(A')$. Since $F_{11}(G'_{ii}G'_{jj} - |G'_{ij}|^2) = r^{\wedge}_3(A) = 1$ for all $1 \leq i < j \leq 3$, we have $|G'_{12}| = |G'_{13}| = |G'_{23}|$. As G' is the direct sum of matrices of smaller sizes, we have $|G'_{12}| = |G'_{13}| = |G'_{23}| = 0$. It follows that G', and hence G, equals I_3, which is a contradiction. □

LEMMA 4.9. *If $(m,n) = (4,6)$ and A is a direct sum of two 3×3 matrices, say $A = A_1 \oplus A_2$, then $H = I$.*

Proof. Assume A satisfies the hypotheses of the lemma. Let $H = F \oplus G$ and $D = D_1 \oplus D_2$ accordingly. By Lemma 4.7, $3r(A_2) = \sigma_1(A_2) = \sigma_2(A_2) = \sigma_3(A_2)$. By Theorem 3.1, A_2 is unitarily similar to D_2G' such that G' is a direct sum of matrices of smaller sizes, and $G'_{11} = G'_{22} = G'_{33} = r(A_2)$. Since $(\det\ F[1,2])(G'_{ii}G'_{jj} - |G'_{ij}|^2) = r^{\wedge}_4(A) = 1$ for all $1 \leq i < j \leq 3$, we have $|G'_{12}| = |G'_{13}| = |G'_{23}|$. As G' is the direct sum of matrices of smaller sizes, we have $|G'_{12}| = |G'_{13}| = |G'_{23}| = 0$. It follows that G', and hence G, equals I_3. By similar arguments, we can show that $F = I_3$ and the result follows. □

LEMMA 4.10. *If $m \geq 2$, then $H = I$.*

Proof of Theorem 3.3. To show that $H = I$, it suffices to show that all $(m+2) \times (m+2)$ principal submatrices of H equal I_{m+2}. Note that by Lemma 4.4, all $(m+2) \times (m+2)$ principal submatrices B of A satisfy

$$1 = |\det B[\delta]| = r_m^{\wedge}(B) = \binom{m+2}{m}^{-1} E_m(\sigma_1(B), \cdots, \sigma_k(B))$$

for all $\delta \in Q_{m,m+2}$. So we may confine our attention to the case when $n = m+2$ in the following. We prove that

$$\text{if } n = m+2 \geq 4 \quad \text{then} \quad H = I$$

by induction on m. If $m = 2, 3$, the statement is true by Lemma 4.8. Now suppose $m > 3$ and assume that the statement is true for all lower cases. By Lemma 4.6, we may assume that A is the direct sum of matrices of smaller sizes, say $A_1 \oplus A_2$ such that A_2 is $p \times p$ with $p \geq n - p$. Let $D = D_1 \oplus D_2$ and $H = F \oplus G$ accordingly. If $m = 4$ and both A_1 and A_2 are 3×3, the result is true by Lemma 4.9. Therefore, when $m = 4$, we may assume $p > 3$. If $m > 4$, then $p \geq (m+2)/2$ implies $p > 3$. Thus we may apply Lemma 4.7 and conclude that

$$\binom{p}{p-2} r_{p-2}^{\wedge}(A_2) = E_{p-2}(\sigma_1(A_2), \cdots, \sigma_p(A_2)).$$

Since $p - 2 \geq 2$, we can apply the induction assumption to A_2 and conclude that G is a scalar matrix. Now let $H = H' \oplus [H_{nn}]$ and $A = A' \oplus [A_{nn}]$ accordingly. By Lemma 4.7 again, we conclude that

$$\binom{m+1}{m-1} r_{m-1}^{\wedge}(A') = E_{m-1}(\sigma_1(A'), \cdots, \sigma_{m+1}(A')).$$

Applying the induction assumption on A', we see that H' is a scalar matrix. Consequently, $H = I$. □

By Lemmas 4.1, 4.2, 4.3, 4.4, and 4.10, we get the conclusions of Theorem 3.3.

Remark. Our proof of Theorem 3.3 is computational; it would be nice to have a conceptual proof.

REFERENCES

[1] P. ANDRESEN AND M. MARCUS, *Weyl's inequality and quadratic forms on the Grassmannian*, Pacific J. Math., 67 (1976), pp. 277–289.

[2] P.A. FILLMORE AND J.P. WILLIAMS, *Some convexity theorems for matrices*, Glasgow. Math. J., 12 (1971), pp. 110–117.

[3] M. GOLDBERG AND E.G. STRAUS, *Elementary inclusion relations for generalized numerical ranges*, Linear Algebra Appl., 18 (1977), pp. 1–24.

[4] R.A. HORN AND C.R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, U.K., 1991.

[5] C.R. JOHNSON AND C.K. LI, *Inequalities relating unitarily invariant norms and the numerical radius*, Linear and Multilinear Algebra, 23 (1988), pp. 183–191.

[6] C.K. LI, *The decomposable numerical radius and numerical radius of a compound matrix*, Linear Algebra Appl., 76 (1986), pp. 45–58.

[7] ———, *On the higher numerical radius and spectral norm*, Linear Algebra Appl., 80 (1986), pp. 55–70.

[8] C.K. LI AND N.K. TSING, *Linear operators preserving the decomposable numerical radius*, Linear and Multilinear Algebra, 23 (1988), pp. 333–341.

[9] ———, *Norms that are invariant under unitary similarities and the C-numerical radii*, Linear and Multilinear Algebra, 24 (1989), pp. 209–222.

[10] M. MARCUS, *Finite Dimensional Multilinear Algebra*, Part I, Marcel Dekker, New York, 1973.

[11] M. MARCUS AND P. ANDRESEN, *The numerical radius of exterior powers*, Linear Algebra Appl., 16 (1977), pp. 131–151.

[12] M. MARCUS AND I. FILIPPENKO, *Linear operators preserving the decomposable numerical range*, Linear and Multilinear Algebra, 7 (1979), pp. 27–36.

[13] M. MARCUS AND M. SANDY, *Singular values and numerical radii*, Linear and Multilinear Algebra, 18 (1985), pp. 337–353.

[14] T.Y. TAM, *Linear operators on matrices: The invariance of the decomposable numerical radius*, Linear Algebra Appl., 87 (1987), pp. 147–153.

[15] R.C. THOMPSON, *Singular values, diagonal elements and convexity*, SIAM J. Appl. Math., 32 (1977), pp. 39–63.