SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 1, pp. 1-17, January 1989

001

SPECTRAL AND INNER-OUTER FACTORIZATIONS OF RATIONAL MATRICES*

TONGWEN CHEN† AND BRUCE A. FRANCIS†

Abstract. Spectral factorization and inner-outer factorization are basic techniques in treating many problems in electrical engineering. In this paper, the problems of doing spectral and inner-outer factorizations via state-space methods are studied when the matrix to be factored is real-rational and surjective on the extended imaginary axis. It is shown that our factorization problems can be reduced to solving a certain constrained Riccati equation, and that by examining some invariant subspace of the associated Hamiltonian matrix there exists a unique solution to this equation. Finally, a state-space procedure to perform the factorization is proposed.

Key words. spectral factorization, inner-outer factorization, algebraic Riccati equation, state-space methods

AMS(MOS) subject classifications. G5E05, 15A24

1. Introduction.

It has been well established in complex analysis that every function in the space $\mathbf{H}_p$ $(0 < p \leqq \infty)$ can be written as the product of an inner function and an outer function [15]. Within the class $\mathbf{RH}_\infty$ (prefix $\mathbf{R}$ means real-rational) of proper real-rational functions analytic in the right half-plane Re $s \geqq 0$, a function is *inner* if it has unit modulus on the imaginary axis and *outer* if it has no zeros in Re $s > 0$. Thus every function in $\mathbf{RH}_\infty$ can be represented as the product of an inner function and an outer function, and the factorization is unique up to sign.

These concepts generalize to the class, also denoted $\mathbf{RH}_\infty$, of proper real-rational matrices analytic in Re $s \geqq 0$. Such a matrix $G(s)$ is *inner* if $G^{\sim}(s)G(s) = I$ ($G^{\sim}(s)$ is defined as $G(-s)^T$), and *outer* if it has full row rank for every Re $s > 0$. Obviously, an inner matrix is tall (number of rows $\geqq$ number of columns) and an outer matrix is wide (number of rows $\leqq$ number of columns). Then every G in $\mathbf{RH}_\infty$ has an inner-outer factorization $G = G_iG_o$, G_i inner, G_o outer, the factors unique up to multiplication by an orthogonal matrix.

The basic idea in order to get an outer factor of $G(s)$ is to do spectral factorization of $G^{\sim}(s)G(s)$, and then get an inner factor by matrix inversion. Anderson [1] first studied the problem of doing spectral factorization by state-space methods. He showed that spectral factorization could be performed by solving an algebraic Riccati equation with a certain eigenvalue inequality, and he invoked Potter's result [14] connecting Riccati equations and Hamiltonian matrices. However, the procedure in [1] contains some gaps. More recently, Bart, Gohberg, and Kaashoek [2] and Bart et al. [3] developed a geometric factorization theory based on state-space models. Their work yields spectral factorization as a special case. Finally, Doyle [6] worked out a state-space procedure for spectral and inner-outer factorizations when $G(s)$ is injective on the extended imaginary axis. This procedure is essentially Anderson's with the gaps filled in.

Spectral factorization and inner-outer factorization are basic techniques in treating many problems in electrical engineering. In its matrix form, spectral factorization provides a tool for the solution of the optimal filtering problem [17], the impedance synthesis of n-port networks [13], among others. The need for inner-outer factorization arises in $\mathbf{H}_\infty$ optimal control (e.g., [7], [20]). Inner-outer factorization is also relevant to the robustness problem of feedback stability [16].

* Received by the editors August 17, 1987; accepted for publication (in revised form) April 27, 1988. This research was supported by the Natural Science and Engineering Research Council of Canada.

† Department of Electrical Engineering, University of Toronto, Toronto, Ontario, Canada, M5S 1A4.

Algorithms to do inner-outer factorization by polynomial methods were proposed in [5]. However, to our best knowledge, the problem of doing spectral factorization of $G^{\sim}(s)G(s)$ and inner-outer factorization of $G(s)$ by state-space methods remained unsolved when $G(s)$ in $\mathbf{RH}_\infty$ is surjective on the extended imaginary axis. In this case $G^{\sim}(s)G(s)$ is singular on the imaginary axis if $G(s)$ is strictly wide, and the spectral factorization of $G^{\sim}(s)G(s)$ is more delicate. The purpose of this paper is to study this problem and to derive a procedure for computing the factorizations starting from a state-space realization of the transfer matrix to be factored.

Let us now introduce some notation used in the following development. Assume $A \in \mathbf{R}^{n\times n}$. We identify the matrix A and its corresponding linear transformation $x \rightarrow Ax$ on $\mathbf{R}^n$. The matrix A is *stable* if the spectrum of A, denoted $\sigma(A)$, is in the left half-plane Re $s<0$. The image and kernel of the linear transformation are denoted Im A and Ker A, respectively. For a subspace $\mathbf{V} \subset \mathbf{R}^n$, $\langle A|\mathbf{V}\rangle$ is the *controllable subspace*

$$\langle A|\mathbf{V}\rangle = \mathbf{V} + A\mathbf{V} + \cdots + A^{n-1}\mathbf{V}.$$

The pair (A, B) is *controllable* if $\langle A|\operatorname{Im} B\rangle = \mathbf{R}^n$ and *stabilizable* if there exists a matrix F such that $A + BF$ is stable.

Let Λ and Γ be two disjoint sets such that $\sigma(A) = \Lambda \cup \Gamma$. Let $\alpha(s)$ be the characteristic polynomial of A and factor it as $\alpha = \alpha_\Lambda \alpha_\Gamma$, where α_Λ has zeros only in Λ, α_Γ only in Γ. Then we denote

$$\mathbf{X}_\Lambda(A) = \operatorname{Ker} \alpha_\Lambda(A), \qquad \mathbf{X}_\Gamma(A) = \operatorname{Ker} \alpha_\Gamma(A).$$

It is observed that $\mathbf{X}_\Lambda(A)$ is spanned by the generalized eigenvectors of A corresponding to eigenvalues in Λ, similarly for $\mathbf{X}_\Gamma(A)$. Moreover, if $\lambda \in \sigma(A)$, $\mathbf{X}_\lambda(A)$ denotes the generalized eigenspace of A corresponding to λ. If Λ is in Re $s<0$ and Γ in Re $s \geqq 0$, we call $\mathbf{X}_\Lambda(A)$, denoted $\mathbf{X}_-(A)$, the *stable modal subspace* relative to A and $\mathbf{X}_\Gamma(A)$, denoted $\mathbf{X}_+(A)$, the *unstable modal subspace*.

For a real matrix F, F^T is the transpose. If F is a complex matrix, F^* denotes the complex-conjugate transpose. For polynomial matrices $P_1(s)$ and $P_2(s)$, P_1 is *equivalent* to P_2, denoted by $P_1 \sim P_2$, if there exist unimodular matrices $M(s)$ and $N(s)$ such that $P_1 = MP_2N$.

It is convenient to let $[A, B, C, D]$ stand for the corresponding transfer matrix:

$$[A,B,C,D] := D + C(s-A)^{-1}B.$$

Some useful algebraic operations on transfer matrices using this data structure are collected in the appendix.

2. Problem formulation. Given $G(s)$ in $\mathbf{RH}_\infty$, bring in a minimal realization

$$G(s) = [A,B,C,D].$$

It follows from the operations in the appendix that a realization for $G^{\sim}(s)G(s)$ is

$$G^{\sim}(s)G(s) = [\bar{A},\bar{B},\bar{C},\bar{D}]$$

where

$$\bar{A} := \begin{bmatrix} A & 0 \\ -C^TC & -A^T \end{bmatrix}, \tag{1}$$

$$\bar{B} := \begin{bmatrix} B \\ -C^TD \end{bmatrix}, \tag{2}$$

$$\bar{C} := [D^TC \quad B^T], \tag{3}$$

$$\bar{D} := D^TD. \tag{4}$$

Let us first review the inner-outer factorization problem when $G(s)$ has full column rank on the extended imaginary axis. In this case, we see that $\bar{D}$ is nonsingular. Define

$$H := \bar{A} - \bar{B}(\bar{D})^{-1}\bar{C}.$$

We shall need the following lemma on spectral factorization.

LEMMA 1. (See e.g., [8].) *Assume $G(s)$ is injective on the extended imaginary axis. Then there exists a unique matrix X such that*

$$\mathbf{X}_-(H) = \operatorname{Im}\begin{bmatrix} I \\ X \end{bmatrix}.$$

Moreover,

$$G_-(s) := [A, B, (\bar{D})^{-1/2}(D^T C + B^T X), (\bar{D})^{1/2}]$$

is a spectral factor of $G^\sim(s)G(s)$.

From the lemma, we see that $G^\sim(s)G(s) = G_-^\sim(s)G_-(s)$ and G_-, $G_-^{-1} \in \mathbf{RH}_\infty$. Let

$$G_o := G_-, \qquad G_i := GG_-^{-1}.$$

Then we obtain a realization of G_i:

$$G_i = [A + BF, B(\bar{D})^{-1/2}, C + DF, D(\bar{D})^{-1/2}]$$

where

$$F := -(\bar{D})^{-1}(D^T C + B^T X).$$

Thus $G = G_i G_o$ is an inner-outer factorization of G with G_i inner and G_o outer.

Now suppose $G(s)$ is surjective on the extended imaginary axis, or equivalently, $G(s)$ has a right-inverse in $\mathbf{RL}_\infty$. Then $G^\sim(s)G(s)$ is not necessarily invertible in $\mathbf{RL}_\infty$. However, Lemma 1 says that the invertibility of $G^\sim G$ in $\mathbf{RL}_\infty$ guarantees the existence of an X in the lemma. This motivates us to perturb $G^\sim G$ into $G^\sim G + \varepsilon^2 I$, which is invertible in $\mathbf{RL}_\infty$, and to consider the spectral factor of $G^\sim G + \varepsilon^2 I$. Here ε is an arbitrary positive number.

It is easily seen that a realization of $G^\sim G + \varepsilon^2 I$ is given by the following:

$$G^\sim G + \varepsilon^2 I = [\bar{A}, \bar{B}, \bar{C}, \bar{D}_\varepsilon]$$

where $\bar{A}$, $\bar{B}$, $\bar{C}$ are defined in (1)–(3), and

$$\bar{D}_\varepsilon := D^T D + \varepsilon^2 I.$$

Let the matrix associated with the zeros of $G^\sim G + \varepsilon^2 I$ be

$$H_\varepsilon := \bar{A} - \bar{B}(\bar{D}_\varepsilon)^{-1}\bar{C}.$$

By the matrix inversion identity,

$$(I + AB)^{-1} = I - A(I + BA)^{-1}B,$$

we see that

$$\begin{aligned}(\bar{D}_\varepsilon)^{-1} = (\varepsilon^2 I + D^T D)^{-1} &= \frac{1}{\varepsilon^2}\left(I + \frac{1}{\varepsilon^2}D^T D\right)^{-1} \\ &= \frac{1}{\varepsilon^2}\left[I - \frac{1}{\varepsilon^2}D^T\left(I + \frac{1}{\varepsilon^2}DD^T\right)^{-1} D\right] \\ &= \frac{1}{\varepsilon^2}[I - D^T(\varepsilon^2 + DD^T)^{-1}D].\end{aligned}$$

The expansion

$$(\varepsilon^2 I + DD^T)^{-1} = (DD^T)^{-1} - \varepsilon^2 (DD^T)^{-2} + \varepsilon^4 (DD^T)^{-3} - \cdots$$

yields

$$(\bar{D}_\varepsilon)^{-1} = \frac{1}{\varepsilon^2}[I - D^T(DD^T)^{-1}D] + D^T(DD^T)^{-2}D - \varepsilon^2 D^T(DD^T)^{-3}D + \cdots .$$

To get an associated Riccati equation for our problem, the following derivation serves mainly as a conceptual tool. It will be justified later.

Define

$$E_0 := I - D^T(DD^T)^{-1}D, \tag{5}$$

$$E := D^T(DD^T)^{-2}D. \tag{6}$$

It is readily verified that $DED^T = I$. Hence the matrix E is the pseudo-inverse of D^TD. Suppose ε is small. We neglect the higher order terms. Thus

$$(\bar{D}_\varepsilon)^{-1} = \frac{1}{\varepsilon^2}E_0 + E.$$

Hence

$$H_\varepsilon = \begin{bmatrix} A & 0 \\ -C^TC & -A^T \end{bmatrix} - \begin{bmatrix} B \\ -C^TD \end{bmatrix} (\bar{D}_\varepsilon)^{-1} [D^TC \quad B^T]$$

$$= \begin{bmatrix} A - B\left(\frac{1}{\varepsilon^2}E_0 + E\right)D^TC & -B\left(\frac{1}{\varepsilon^2}E_0 + E\right)B^T \\ -C^TC + C^TD\left(\frac{1}{\varepsilon^2}E_0 + E\right)D^TC & -A^T + C^TD\left(\frac{1}{\varepsilon^2}E_0 + E\right)B^T \end{bmatrix}.$$

By the fact that E_0 is symmetric and $DE_0 = 0$, we have

$$H_\varepsilon = \begin{bmatrix} A - BED^TC & -\left(BEB^T + \frac{1}{\varepsilon^2}BE_0B^T\right) \\ 0 & -(A - BED^TC)^T \end{bmatrix}.$$

Assume that $A - BED^TC$ has no eigenvalues on the imaginary axis and that there exists a symmetric matrix X_ε such that

$$\mathbf{X}_-(H_\varepsilon) = \operatorname{Im}\begin{bmatrix} I \\ X_\varepsilon \end{bmatrix}.$$

Thus X_ε satisfies the algebraic Riccati equation

$$(A - BED^TC)^T X_\varepsilon + X_\varepsilon(A - BED^TC) - X_\varepsilon\left(BEB^T + \frac{1}{\varepsilon^2}BE_0B^T\right)X_\varepsilon = 0$$

and the matrix $(A - BED^TC - BEB^TX_\varepsilon - \varepsilon^{-2}\, BE_0B^TX_\varepsilon)$ is stable. It can be proved by a result in [18] that $X := \lim_{\varepsilon \to 0} X_\varepsilon$ exists. Assume

$$\lim_{\varepsilon \to 0} \frac{1}{\varepsilon^2} X_\varepsilon BE_0B^TX_\varepsilon = 0.$$

From (5), it is readily seen that $E_0^2 = E_0$. Hence E_0 is positive semidefinite. This together with the above assumptions implies that $XBE_0 = 0$. In this way we are led to solve the following constrained Riccati equation for X:

(7a) $$(A-BED^TC)^TX+X(A-BED^TC)-XBEB^TX=0,$$

(7b) $$X^T=X,$$

(7c) $$XBE_0=0,$$

(7d) $$(A-BED^TC-BEB^TX, BE_0) \text{ is stabilizable.}$$

3. Spectral and inner-outer factorizations. The existence and computation of the solution to the constrained Riccati equation are deferred to the next section. In this section we assume such X as satisfies (7) exists.

Define

(8) $$G_o := [A, B, C_o, D],$$

(9) $$C_o := C+DEB^TX.$$

We shall show that G_o is an outer matrix in $\mathbf{RH}_\infty$.

LEMMA 2. (See Minto [12].) *Assume $G \in \mathbf{RH}_\infty$, and $[A, B, C, D]$ is a minimal realization of G. Then the following conditions are equivalent*:

(i) *G has a right-inverse in $\mathbf{RH}_\infty$.*

(ii) *D is surjective and*

$$\mathbf{X}_+(A+BF)=\{0\}, \qquad C+DF=0$$

for some matrix F.

(iii) *D is surjective and*

$$\mathbf{X}_+(A-BED^TC) \subset \langle A-BED^TC \mid B \operatorname{Ker} D\rangle$$

where E is defined in (6).

Suppose (ii) is satisfied and a right-inverse of D is D^+. Then a right-inverse of $G(s)$ is given by

$$G^+ := [A+BF, BD^+, F, D^+].$$

From (5), the first fact to note is that

(10) $$\operatorname{Ker} D = \operatorname{Im} E_0.$$

By (7) and (9), the stabilizability of $(A - BED^TC - BEB^TX, BE_0)$ gives that of $(A - BED^TC_o, BE_0)$. Thus it follows from Theorem 2.3 of [19] and then (10) that

$$\mathbf{X}_+(A-BED^TC_o) \subset \langle A-BED^TC_o \mid \operatorname{Im} BE_0\rangle$$
$$= \langle A-BED^TC_o \mid B \operatorname{Ker} D\rangle.$$

Therefore, Lemma 2 says that the matrix G_o is right-invertible in $\mathbf{RH}_\infty$, and hence it is an outer matrix.

Let G_o^+ be a right-inverse of G_o in $\mathbf{RH}_\infty$. Define

(11) $$G_i := GG_o^+.$$

Then we have the following theorem.

THEOREM 1. *Assume G in $\mathbf{RH}_\infty$ is surjective on the extended imaginary axis and that there exists an X that solves* (7). *Then, with G_o, G_i defined by* (8), (11), *G_o is a spectral factor of $G^\sim G$ and G_i, G_o are inner and outer factors of G, respectively.*

Proof. From the previous discussion we know that G_o is right-invertible in $\mathbf{RH}_\infty$. To show that

$$G^\sim G = G_o^\sim G_o, \tag{12}$$

use (8) to get

$$G_o^\sim G_o = [\bar{A}_o, \bar{B}_o, \bar{C}_o, \bar{D}_o] \tag{13}$$

where

$$\bar{A}_o := \begin{bmatrix} A & 0 \\ -(C+DEB^TX)^T(C+DEB^TX) & -A^T \end{bmatrix},$$

$$\bar{B}_o := \begin{bmatrix} B \\ -(C+DEB^TX)^TD \end{bmatrix},$$

$$\bar{C}_o := [D^T(C+DEB^TX) \quad B^T],$$

$$\bar{D}_o := D^TD.$$

Define

$$T := \begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} \tag{14}$$

to get from (13) that

$$G_o^\sim G_o = [T^{-1}\bar{A}_oT, T^{-1}\bar{B}_o, \bar{C}_oT, \bar{D}_o]$$

and

$$T^{-1}\bar{A}_oT = \begin{bmatrix} A & 0 \\ (A-BED^TC)X + X(A-BED^TC) - XBED^TDEB^TX - C^TC & -A^T \end{bmatrix}$$

$$= \begin{bmatrix} A & 0 \\ -C^TC & -A^T \end{bmatrix} \text{ by (6) and (7a)},$$

$$T^{-1}\bar{B}_o = \begin{bmatrix} B \\ -C^TD - XBED^TD + XB \end{bmatrix}$$

$$= \begin{bmatrix} B \\ -C^TD + XBE_0 \end{bmatrix} \text{ by (5) and (6)}$$

$$= \begin{bmatrix} B \\ -C^TD \end{bmatrix} \text{ by (7c)}.$$

Similarly,

$$\bar{C}_oT = [D^TC \quad B^T].$$

Comparison of the above with the realization of $G^\sim G$ in (1)–(4) gives (12). Hence G_o is a spectral factor of $G^\sim G$.

To show that G_i is inner, we follow (11) and then (12):

$$G_i^\sim G_i = (G_o^+)^\sim (G^\sim G) G_o^+ = (G_o^+)^\sim (G_o^\sim G_o) G_o^+ = I.$$

Hence G_i is an inner matrix.

Finally, post-multiply (12) by G_o^+ to get

$$G^\sim G_i = G_o^\sim.$$

Post-multiply this by G_o and use (12) to get

$$G^{\sim}(G-G_iG_o)=0.$$

By our assumption, $G(j\omega)$ is surjective for $0 \leqq \omega \leqq \infty$. Then

$$G(j\omega)=G_i(j\omega)G_o(j\omega), \qquad 0 \leqq \omega \leqq \infty.$$

Hence $G = G_iG_o$. □

Now we find a realization for the inner factor G_i. Since a right-inverse of D is just ED^T, it follows from Lemma 2 that a right-inverse of G_o is

$$G_o^+=[A+BF, BED^T, F, ED^T]$$

where F is any matrix such that

(15) $$\mathbf{X}_+(A+BF)=\{0\} \quad \text{and} \quad C+DEB^TX+DF=0.$$

Since D is surjective, the solutions of (15) are all the matrices of the form

(16) $$F=-ED^T(C+DEB^TX)+F_1=-E(D^TC+B^TX)+F_1, \qquad DF_1=0.$$

Thus we compute F_1 such that

$$\mathbf{X}_+[A-BE(D^TC+B^TX)+BF_1]=\{0\} \quad \text{and} \quad DF_1=0.$$

(The existence of such F_1 is guaranteed by (7d).) Then define F as in (16). Hence

$$G_i=GG_o^+=:[A_i, B_i, C_i, D_i]$$

where

$$A_i=\begin{bmatrix} A & BF \\ 0 & A+BF \end{bmatrix}, \qquad B_i=\begin{bmatrix} BED^T \\ BED^T \end{bmatrix},$$
$$C_i=[C \quad DF], \qquad D_i=DED^T=I.$$

Define the similarity transformation

$$T:=\begin{bmatrix} I & I \\ 0 & I \end{bmatrix}$$

to get

(17) $$\begin{aligned} G_i &= [T^{-1}A_iT, T^{-1}B_i, C_iT, D_i] \\ &= \left[\begin{bmatrix} A & 0 \\ 0 & A+BF \end{bmatrix}, \begin{bmatrix} 0 \\ BED^T \end{bmatrix}, [C \quad C+DF], I\right] \\ &= [A+BF, BED^T, C+DF, I]. \end{aligned}$$

4. The constrained Riccati equation. We have assumed that $G(s)$ is surjective on the extended imaginary axis. It would be trivial if $G(s)$ is surjective for every Re $s > 0$, since in this case G is already outer. Hence we assume that $G(s)$ does not have full row rank on at least one point in Re $s > 0$.

Bring in the system matrix

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

DEFINITION. The zeros of $G(s)$ are the roots, counting multiplicities, of the invariant polynomials of $P(s)$.

Equations (18) and (6) yield

$$P(s) \sim \begin{bmatrix} A-BED^TC-s & B \\ C-DED^TC & D \end{bmatrix} = \begin{bmatrix} A-BED^TC-s & B \\ 0 & D \end{bmatrix}. \tag{19}$$

Let K be a base matrix of Ker D. Define

$$T = [D^T \quad K].$$

Then T is invertible. From (19), we have

$$\begin{aligned} P(s) &\sim \begin{bmatrix} A-BED^TC-s & BT \\ 0 & DT \end{bmatrix} = \begin{bmatrix} A-BED^TC-s & BD^T & BK \\ 0 & DD^T & 0 \end{bmatrix} \\ &\sim \begin{bmatrix} A-BED^TC-s & 0 & BK \\ 0 & DD^T & 0 \end{bmatrix} \\ &\sim \begin{bmatrix} I & 0 & 0 \\ 0 & A-BED^TC-s & BK \end{bmatrix}. \end{aligned} \tag{20}$$

We see from (20) that the nonconstant invariant polynomials of $P(s)$ are identical to those of the matrix $[A - BED^TC - s \quad BK]$. It follows then that if λ is a zero of $G(s)$, it is also an eigenvalue of $A - BED^TC$.

Let us denote the set of all rhp (right half-plane) zeros, counting multiplicities, of $G(s)$ by

$$\Lambda := \{\lambda_1, \lambda_2, \cdots, \lambda_m\}. \tag{21}$$

Then

$$\sigma(A-BED^TC) = \{\lambda_1, \cdots, \lambda_m, \lambda_{m+1}, \cdots, \lambda_n\} \tag{22}$$

for certain numbers $\lambda_{m+1}, \cdots, \lambda_n$, counting multiplicities. Define the set

$$\Gamma := \{\lambda_{m+1}, \cdots, \lambda_n\}. \tag{23}$$

We see that the sets Λ and Γ are symmetric with respect to the real axis in the complex plane because $G(s)$ is a real rational matrix.

Assumption 1. The sets Λ and Γ as defined are disjoint, i.e., $\lambda_i \neq \lambda_j$ for any $i = 1, \cdots, m$ and $j = m + 1, \cdots, n$.

This will be a standing assumption in the rest of the development.

PROPOSITION 1. *Under Assumption* 1, *suppose* $\lambda \in \sigma(A - BED^TC)$ *and* Re $\lambda > 0$. *Then* λ *is a rhp zero of* $G(s)$ *if and only if* $X_\lambda[(A - BED^TC)^T]$ *is orthogonal to* B Ker D.

Proof. Suppose λ is an eigenvalue of $A - BED^TC$ with multiplicity l. Do a Jordan decomposition of the matrix $(A - BED^TC)^T$, i.e., find a nonsingular matrix

$$Q := [q_1 q_2 \cdots q_l q_{l+1} \cdots q_n]$$

such that

$$Q^{-1}(A-BED^TC)^TQ = \text{diag}\,[J_\lambda, J] \tag{24}$$

where J_λ is an $l \times l$ matrix consisting of the Jordan blocks corresponding to λ,

$$J_\lambda = \begin{bmatrix} \lambda x & & & \\ & \cdot & & \\ & & \ddots & \\ & & & \cdot x \\ & & & \lambda \end{bmatrix}, \quad x \text{ can be either 1 or 0.}$$

It follows that $\{q_1, q_2, \cdots, q_l\}$ is a basis for $\mathbf{X}_\lambda[(A - BED^TC)^T]$.

Take complex-conjugate transpose of (24) to get

$$Q^*(A - BED^TC)(Q^*)^{-1} = \operatorname{diag}[J_\lambda^*, J^*].$$

Then

$$(25) \qquad \begin{aligned}[] [A - BED^TC - s \quad BK] &\sim [Q^*(A - BED^TC - s)(Q^*)^{-1} \quad Q^*BK] \\ &= [\operatorname{diag}(J_\lambda^* - s, J^* - s) \quad Q^*BK]. \end{aligned}$$

Necessity. Suppose further that λ is a rhp zero of $G(s)$. By Assumption 1, λ is an lth order zero. It follows then that $(s - \lambda)^l$ divides all the nth order minors of the matrix $[A - BED^TC - s \quad BK]$, and hence of the matrix in (25). Now since λ is not in $\sigma(J)$, we have

$$q_j^* BK = 0, \qquad j = 1, 2, \cdots, l.$$

Thus $\mathbf{X}_\lambda[(A - BED^TC)^T]$ is orthogonal to B Ker D.

Sufficiency. This follows by reversing the argument in the necessity part. □

Proposition 1 gives a characterization of rhp zeros of $G(s)$ in terms of the generalized eigenspace of $(A - BED^TC)^T$. Now let us turn to the constrained Riccati equation (7). The associated Hamiltonian matrix is defined as

$$(26) \qquad H := \begin{bmatrix} A - BED^TC & -BEB^T \\ 0 & -(A - BED^TC)^T \end{bmatrix}.$$

By (22) and (23), the spectrum of H consists of four parts:

$$\begin{aligned} \sigma(H) &= \sigma(A - BED^TC) \cup \sigma[-(A - BED^TC)^T] \\ &= \Lambda \cup \Gamma \cup -\Lambda \cup -\Gamma \end{aligned}$$

where

$$-\Lambda = \{-\lambda: \quad \lambda \in \Lambda\}, \qquad -\Gamma = \{-\gamma: \quad \gamma \in \Gamma\}.$$

The hypothesis that $\Lambda \cap \Gamma = \varnothing$ implies that $\mathbf{X}_\Gamma(A - BED^TC)$ is invariant under $A - BED^TC$. To establish the main theorem, we make the following assumption.

Assumption 2. There exists a subspace $\mathbf{V}$ in $\mathbf{R}^{2n}$ such that

(i) $\mathbf{V}$ is H-invariant.

(ii) $\sigma(H|\mathbf{V}) = -\Lambda$.

(iii) $\mathbf{V}$ is independent of $\mathbf{X}_\Gamma(A - BED^TC) \oplus \{0\}$.

A sufficient condition for the existence of such a $\mathbf{V}$ is that $-\Lambda \cap \Gamma = \varnothing$.

THEOREM 2. *Assume (A, B) is controllable. Then under Assumptions* 1 *and* 2, *there exists a unique matrix X such that*

$$(27) \qquad \mathbf{V} \oplus [\mathbf{X}_\Gamma(A - BED^TC) \oplus \{0\}] = \operatorname{Im} \begin{bmatrix} I \\ X \end{bmatrix}.$$

Moreover X solves the constrained Riccati equation (7).

Proof. Let a base matrix of $\mathbf{X}_\Gamma(A - BED^TC)$ be W_1. Then we have

$$(A-BED^TC)W_1 = W_1H_1 \tag{28}$$

for some matrix H_1 such that

$$\sigma(H_1) = \Gamma. \tag{29}$$

Let a base matrix of $\mathbf{V}$ be V. Then for some H_2,

$$HV = VH_2, \tag{30}$$

$$\sigma(H_2) = -\Lambda. \tag{31}$$

Partition V as

$$V = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}.$$

It follows from (30) that

$$(A-BED^TC)V_1 - BEB^TV_2 = V_1H_2 \tag{32}$$

and

$$-(A-BED^TC)^TV_2 = V_2H_2. \tag{33}$$

The above constructions yield

$$\mathbf{V} \oplus [\mathbf{X}_\Gamma(A-BED^TC) \oplus \{0\}] = \operatorname{Im} \begin{bmatrix} W_1 & V_1 \\ 0 & V_2 \end{bmatrix}. \tag{34}$$

Now we shall prove the results by proving a number of claims.

Claim 1. Ker $V_2 = \{0\}$.

Suppose that Ker $V_2 \neq \{0\}$. Equation (33) implies that Ker V_2 is H_2-invariant. Thus there exists a number λ in $\sigma(H_2)$ and a nonzero vector z such that $H_2z = \lambda z$, $z \in$ Ker V_2. Post-multiply (32) by z to get

$$(A-BED^TC)V_1z = \lambda V_1z. \tag{35}$$

If $V_1z = 0$, then $Vz = 0$. However V is a base matrix. Hence $V_1z \neq 0$. It follows from (35) that λ is an eigenvalue of $A - BED^TC$, specifically, $\lambda \in \Gamma$. Hence

$$V_1z \in \mathbf{X}_\Gamma(A-BED^TC) = \operatorname{Im} W_1.$$

Then

$$Vz \in \operatorname{Im} \begin{bmatrix} W_1 \\ 0 \end{bmatrix}.$$

This contradicts property (iii) of our Assumption 2. Thus Claim 1 follows.

Claim 2. Im W_1 *is orthogonal to* Im V_2.

Take transpose of (28), then post-multiply it by V_2 to get

$$W_1^T(A-BED^TC)^TV_2 = H_1^TW_1^TV_2.$$

Pre-multiply (33) by W_1^T:

$$-W_1^T(A-BED^TC)^TV_2 = W_1V_2H_2.$$

Add the above two equations to get

$$H_1^T W_1^T V_2 + W_1^T V_2 H_2 = 0. \tag{36}$$

By Assumption 1, $\sigma(H_1^T) \cap \sigma(-H_2) = \varnothing$. Then (36) has a unique solution $W_1^T V_2 = 0$ (Chapter 8 of [9]). Claim 2 follows then.

It follows from Claim 2 that

$$\operatorname{Im} V_2 \subset (\operatorname{Im} W_1)^\perp.$$

Claim 1 and the definitions of V_2 and W_1 imply that dim (Im V_2) + dim (Im W_1) = n. So

$$\operatorname{Im} V_2 = (\operatorname{Im} W_1)^\perp. \tag{37}$$

CLAIM 3. *The matrix $V_2^T V_1$ is symmetric.*

Take transpose of (33) and post-multiply it by V_1:

$$-V_2^T(A - BED^TC)V_1 = H_2^T V_2^T V_1.$$

Pre-multiply (32) by V_2^T to get

$$V_2^T(A - BED^TC)V_1 = V_2^T BEB^T V_2 + V_2^T V_1 H_2.$$

Then

$$H_2^T V_2^T V_1 + V_2^T V_1 H_2 = -V_2^T BEB^T V_2.$$

Since all the eigenvalues of H_2 are in Re $s > 0$, the above matrix equation has a unique solution, and it is symmetric.

To show that there exists a unique X such that (27) holds, by (34) it suffices to show that the matrix $[W_1 \quad V_1]$ is invertible.

Let us denote the subspace

$$\mathbf{Y} := (\operatorname{Im} [W_1 \quad V_1])^\perp.$$

Suppose, on the contrary, that $\mathbf{Y} \neq \{0\}$. Let a base matrix of $\mathbf{Y}$ be Y. Then

$$Y^T W_1 = 0, \tag{38}$$

$$Y^T V_1 = 0. \tag{39}$$

Equations (37) and (38) yield that $Y = V_2 M$, for some matrix M. Then from (39) $M^T V_2^T V_1 = 0$. Hence, by Claim 3,

$$V_2^T V_1 M = 0. \tag{40}$$

Pre-multiply (32) by Y^T and use (39) to get

$$Y^T(A - BED^TC)V_1 - Y^T BEB^T V_2 = 0,$$

i.e.,

$$M^T V_2^T(A - BED^TC)V_1 - M^T V_2^T BEB^T V_2 = 0. \tag{41}$$

Take transpose of (33) to get

$$V_2^T(A - BED^TC) = -H_2^T V_2^T.$$

Substitute the above into (41) to get

$$-M^T H_2^T V_2^T V_1 - M^T V_2^T BEB^T V_2 = 0.$$

Post-multiply this by M, noting $V_2^TV_1M = 0$:

$$M^TV_2^TBEB^TV_2M=0.$$

This together with the fact that E is positive semidefinite yields

$$Y^TBE=0 \tag{42}$$

By noting that Im E = Im D^T, we see that

$$Y^TB\,\mathrm{Im}\ D^T=0. \tag{43}$$

On the other hand, (33) and Claim 1 imply that Im $V_2 = \mathbf{X}_\Lambda[(A - BED^TC)^T]$. Proposition 1 says that $\mathbf{X}_\Lambda[(A - BED^TC)^T]$ is orthogonal to B Ker D, i.e.,

$$V_2^TB\,\mathrm{Ker}\ D=0. \tag{44}$$

Pre-multiply by M to get

$$Y^TB\,\mathrm{Ker}\ D=0. \tag{45}$$

Then (43) and (45) give $Y^TB = 0$. Hence

$$\mathrm{Im}\ B\subset\mathrm{Im}\,[W_1\quad V_1]. \tag{46}$$

Pre-multiply (28) by Y^T and use (38):

$$Y^T(A-BED^TC)W_1=0. \tag{47}$$

From (41) and (42), we see that

$$Y^T(A-BED^TC)V_1=0. \tag{48}$$

Thus (47), (48) together with the definition of Y yield that Im $[W_1\quad V_1]$ is invariant under $A - BED^TC$. This fact and (46) imply

$$\langle A-BED^TC\,|\,\mathrm{Im}\ B\rangle\subset\mathrm{Im}\,[W_1\quad V_1].$$

However, the hypothesis that (A, B) is controllable leads us to conclude that Im $[W_1\quad V_1] = \mathbf{R}^n$. Hence the matrix $[W_1\quad V_1]$ is invertible.

Define X by the equation

$$\mathrm{Im}\begin{bmatrix} W_1 & V_1 \\ 0 & V_2 \end{bmatrix}=\mathrm{Im}\begin{bmatrix} I \\ X \end{bmatrix}.$$

Then

$$X=[0\quad V_2][W_1\quad V_1]^{-1}.$$

To show X is symmetric, it suffices to show

$$\begin{bmatrix} W_1^T \\ V_1^T \end{bmatrix}[0\quad V_2]=\begin{bmatrix} 0 \\ V_2^T \end{bmatrix}[W_1\quad V_1]$$

which is equivalent to

$$\begin{bmatrix} 0 & W_1^TV_2 \\ 0 & V_1^TV_2 \end{bmatrix}=\begin{bmatrix} 0 & 0 \\ V_2^TW_1 & V_2^TV_1 \end{bmatrix}.$$

The latter follows from Claims 1 and 2.

Define

$$H_3:=A-BED^TC-BEB^TX$$

to get

$$H\begin{bmatrix} I \\ X \end{bmatrix} = \begin{bmatrix} I \\ X \end{bmatrix} H_3 \tag{49}$$

and

$$\sigma(H_3) = \Gamma \cup -\Lambda. \tag{50}$$

Pre-multiply (49) by $[-X \quad I]$ to get

$$(A - BED^TC)^TX + X(A - BED^TC) - XBEB^TX = 0. \tag{51}$$

Hence X solves (7a).

It follows from the fact that X is symmetric and then (44) that

$$XB \operatorname{Ker} D = X^TB \operatorname{Ker} D = ([W_1 \quad V_1]^{-1})^T \begin{bmatrix} 0 \\ V_2^T \end{bmatrix} B \operatorname{Ker} D = 0.$$

Hence X satisfies (7c).

To establish that (H_3, BE_0) is stabilizable, we first note that

$$X(A - BED^TC)^kB \operatorname{Ker} D = 0, \qquad k = 0, 1, 2, \cdots. \tag{52}$$

This can be proved by induction on k. First it is true for $k = 0$. Assume it is true for $k = i$. Then by Riccati equation (51),

$$\begin{aligned} &X(A - BED^TC)^{i+1}B \operatorname{Ker} D \\ &\qquad = [\mathrm{XBEB}^T\mathrm{X} - (A - BED^TC)^TX](A - BED^TC)^iB \operatorname{Ker} D \\ &\qquad = [XBEB^T - (A - BED^TC)^T]X(A - BED^TC)^iB \operatorname{Ker} D = 0. \end{aligned}$$

Thus (52) holds for all nonnegative integers.

This result immediately yields the following:

$$\langle H_3 | B \operatorname{Ker} D \rangle = \langle A - BED^TC | B \operatorname{Ker} D \rangle \tag{53}$$

Suppose, for a contradiction, that α is an uncontrollable mode of (H_3, BE_0) and $\operatorname{Re} \alpha > 0$. Then from (50),

$$\alpha \in \Gamma \subset \sigma(A - BED^TC).$$

From (52) and (53) we see that

$$H_3 | \langle H_3 | B \operatorname{Ker} D \rangle = (A - BED^TC) \, | \, \langle A - BED^TC | B \operatorname{Ker} D \rangle,$$

i.e., the controllable modes of (H_3, BE_0) and those of $(A - BED^TC, BE_0)$ are identical. It follows that α is an uncontrollable mode of $(A - BED^TC, BE_0)$. Thus by definition, e.g., [11],

$$\operatorname{rank} [A - BED^TC - \alpha \quad BE_0] < n.$$

This implies immediately that α is a zero of some invariant polynomial of the matrix $[A - BED^TC - s \quad BE_0]$; hence α is a rhp zero of $G(s)$. This is a contradiction. Therefore (H_3, BE_0) is stabilizable, i.e., (7d) holds. □

It is worth noting that the proof of Theorem 2 is constructive. Thus we can compute a solution X of equation (7) by finding bases for the eigenspaces of $\mathbf{X}_\Gamma(H)$ and $\mathbf{X}_{-\Lambda}(H)$.

PROPOSITION 2. *The constrained Riccati equation* (7) *has at most one solution.*

Proof. Suppose there exist two solutions X_1 and X_2 to (7). It follows from (7d) that

$$P_i := A - BED^TC - BEB^TX_i + BE_0F_i, \qquad i = 1,2$$

are stable for some matrices F_i, $i = 1, 2$. Then

$$\begin{aligned} &P_1^T(X_1 - X_2) + (X_1 - X_2)P_2 \\ &\qquad = (A - BED^TC - BEB^TX_1)^T(X_1 - X_2) \\ &\qquad\quad + (X_1 - X_2)(A - BED^TC - BEB^TX_2) \quad \text{by (7c)} \\ &\qquad = [(A - BED^TC)^TX_1 + X_1(A - BED^TC) - X_1BEB^TX_1] \\ &\qquad\quad - [(A - BED^TC)^TX_2 + X_2(A - BED^TC) - X_2BEB^TX_2] \quad \text{by (7b)} \\ &\qquad = 0 \quad \text{by (7a)}. \end{aligned}$$

Since $\sigma(P_1^T) \cap \sigma(-P_2) = \varnothing$, it follows from Chapter 8 of [9] that the equation $P_1^TX + XP_2 = 0$ has a unique solution $X = 0$, hence $X_1 = X_2$. □

5. Algorithm. We notice that the way to calculate the inner factor $G_i(s)$ can be further simplified. From (15) and (16) in § 3 we have

$$\begin{aligned} G_i &= [A - BED^TC - BEB^TX + BK, BED^T, -DEB^TX, I] \\ &=: [A_i, B_i, C_i, D_i] \end{aligned} \tag{54}$$

where K is such that $(A - BED^TC - BEB^TX + BK)$ is stable and $DK = 0$.

However, it is a fact that the inner factor of G is unique up to right-multiplication by an orthogonal matrix. Thus, it is claimed that the minimal realization of G_i in (54) does not depend on K. To see this, let $\mathbf{N}$ denote the unobservable subspace of (C_i, A_i) [19]. Since X solves the Riccati equation (7), it is easily shown that

$$\langle A - BED^TC \mid B \operatorname{Ker} D\rangle \subset \mathbf{N}.$$

By (53),

$$\langle A - BED^TC \mid B \operatorname{Ker} D\rangle = \langle A - BED^TC - BEB^TX \mid B \operatorname{Ker} D\rangle.$$

Thus

$$\langle A - BED^TC - BEB^TX \mid B \operatorname{Ker} D\rangle \subset \mathbf{N}.$$

This means that the BK-affected modes of A_i are unobservable by C_i. Hence

$$G_i = [A - BED^TC - BEB^TX, BED^T, -DEB^TX, I]. \tag{55}$$

We summarize the algorithm to do inner-outer factorization as follows.

Given $G(s)$ in $\mathbf{RH}_\infty$ and a realization $[A, B, C, D]$, assume $G(j\omega)$ is surjective for $0 \leqq \omega \leqq \infty$, and that (A, B) is controllable.

Step 1. Compute the set Λ of rhp zeros of $G(s)$:

First find the eigenvalues and generalized eigenvectors of $(A - BED^TC)^T$. Then check the orthogonality of $\mathbf{X}_\lambda[(A - BED^TC)^T]$ to $B \operatorname{Ker} D$ to determine whether a rhp $\lambda \in \Lambda$.

Step 2. Compute the solution X to the constrained Riccati equation:

Find a base matrix T for the subspace $\mathbf{X}_\Gamma(H) \oplus \mathbf{X}_{-\Lambda}(H)$. Partition T as

$$T = \begin{bmatrix} T_1 \\ T_2 \end{bmatrix}.$$

Then

$$X = T_2 T_1^{-1}.$$

Step 3. Compute inner factor G_i and outer factor G_o:

$$G_o = [A, B, C + DEB^T X, D],$$

$$G_i = [A - BED^T C - BEB^T X, BED^T, -DEB^T X, I].$$

Then get a minimal realization of G_i based on this realization.

The procedures to do spectral factorization of $G^{\sim}G$ are embedded in the above algorithm. The following is an illustrative example done on PC-MATLAB. Given

$$G(s) = \begin{bmatrix} \dfrac{s-3}{s+1} & \dfrac{7(s-3)}{s^2+10s+24} & \dfrac{4(s^2-5s+6)}{s^2+6s+9} \\ \dfrac{s^2+7s}{s^2+2s+1} & -\dfrac{s-5}{s+4} & -\dfrac{3s+1}{s+3} \end{bmatrix},$$

a controllable realization of $G(s)$ is the following:

$$A = \begin{bmatrix} -2 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -10 & -24 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -6 & -9 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

$$C = \begin{bmatrix} -4 & -4 & 7 & -21 & -44 & -12 \\ 5 & -1 & 9 & 54 & 8 & 24 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 4 \\ 1 & -1 & -3 \end{bmatrix}.$$

$G(s)$ has two rhp zeros, $\Lambda = \{3, 1\}$. The solution to the constrained Riccati equation is

$$X = \begin{bmatrix} 6 & 6 & 0 & 42 & 24 & -48 \\ 6 & 15 & -3 & 15 & 3 & -111 \\ 0 & -3 & 1 & 9 & 7 & 21 \\ 42 & 15 & 9 & 375 & 231 & -147 \\ 24 & 3 & 7 & 231 & 145 & -45 \\ -48 & -111 & 21 & -147 & -45 & 825 \end{bmatrix}.$$

Hence an inner factor and an outer factor of G are

$$G_i = \begin{bmatrix} \dfrac{s^2-3s}{s^2+4s+3} & \dfrac{s-3}{s^2+4s+3} \\ \dfrac{1}{s+1} & \dfrac{s}{s+1} \end{bmatrix},$$

$$G_o = \begin{bmatrix} \dfrac{s^2+4s}{s^2+2s+1} & \dfrac{2(4s+15)}{s^2+10s+24} & \dfrac{4s-1}{s+3} \\ \dfrac{s^2+7s+3}{s^2+2s+1} & -\dfrac{s^2+2s-21}{s^2+10s+24} & -\dfrac{3s+8}{s+3} \end{bmatrix}.$$

6. Conclusion. In this paper we have proposed a state-space procedure to perform spectral factorization of $G^{\sim}G$ and inner-outer factorization of G when $G(s)$ is surjective on the extended imaginary axis. It has been shown that our factorization problems are closely related to a certain constrained Riccati equation. To find the solution of this equation, we first calculate the set of rhp zeros of $G(s)$, then compute a special generalized eigenspace of the associated Hamiltonian matrix. Both steps amount to finding a basis for some invariant subspace of a real matrix. Employing orthogonal similarity transformations, the real ordered Schur decomposition provides a numerically reliable way to compute orthonormal bases for invariant subspaces (see, e.g., [10]). The algorithm derived in this paper has been coded in PC-MATLAB as *.m* files based on the real ordered Schur decomposition and satisfactory numerical results have been obtained. It is worth mentioning that the Hamiltonian–Schur decomposition can be employed to perform Step 2 in our algorithm. An algorithm to do this decomposition was described in [4].

Appendix. The transfer matrix corresponding to the state space realization (A, B, C, D) is

$$[A,B,C,D] := D + C(s-A)^{-1}B.$$

A collection of operations on transfer matrices in terms of this data structure follows:

$$[A,B,C,D] = [T^{-1}AT, T^{-1}B, CT, D],$$

$$[A,B,C,D]^{-1} = [A-BD^{-1}C, BD^{-1}, -D^{-1}C, D^{-1}] \quad (D \text{ nonsingular}),$$

$$[A,B,C,D]^{\sim} = [-A^T, -C^T, B^T, D^T],$$

$$[A_1,B_1,C_1,D_1][A_2,B_2,C_2,D_2] = \left[\begin{bmatrix} A_1 & B_1C_2 \\ 0 & A_2 \end{bmatrix}, \begin{bmatrix} B_1D_2 \\ B_2 \end{bmatrix}, [C_1 \quad D_1C_2], D_1D_2\right]$$

$$= \left[\begin{bmatrix} A_2 & 0 \\ B_1C_2 & A_1 \end{bmatrix}, \begin{bmatrix} B_2 \\ B_1D_2 \end{bmatrix}, [D_1C_2 \quad C_1], D_1D_2\right],$$

$$[A_1,B_1,C_1,D_1] + [A_2,B_2,C_2,D_2] = \left[\begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, [C_1 \quad C_2], D_1+D_2\right].$$

REFERENCES

[1] B. D. O. Anderson, *An algebraic solution to the spectral factorization problem*, IEEE Trans. Automat. Control, AC-12 (1976), pp. 410–414.

[2] H. Bart, I. Gohberg, and M. A. Kaashoek, *Minimal Factorization of Matrix and Operator Functions*, Birkhaüser, Basel, 1979.

[3] H. Bart, I. Gohberg, M. A. Kaashoek, and P. Van Dooren, *Factorizations of transfer matrices*, SIAM J. Control Optim., 18 (1980), pp. 675–696.

[4] R. Byers, *A Hamiltonian* QR *algorithm*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 212–229.

[5] B-C. Chang and J. B. Pearson, *Inner-outer factorization of rational matrices*, Tech. Rept. 8216, Dept. of Electrical Engineering, Rice University, Houston, TX, 1982.

[6] J. C. Doyle, *Lecture Notes in Advances in Multivariable Control*, ONR/Honeywell Workshop, Minneapolis, MN, 1984.

[7] B. A. Francis, J. W. Helton, and G. Zames, *H_∞-optimal feedback controllers for linear multivariable systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 888–900.

[8] B. A. Francis, *A Course in H_∞ Control Theory*, Springer-Verlag, New York, 1987.

[9] F. R. Gantmacher, *The Theory of Matrices*, Chelsea, New York, 1960.

[10] G. H. Golub and J. H. Wilkinson, *Ill-conditioned eigensystems and the computation of the Jordan canonical form*, SIAM Review, 18 (1976), pp. 578–619.

[11] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[12] K. D. MINTO, *Design of reliable control systems: theory and computations*, Ph.D. Thesis, Dept. Electrical Engineering, University of Waterloo, Waterloo, Canada, 1985.

[13] R. W. NEWCOMB, *Linear Multiport Synthesis*, McGraw-Hall, New York, 1966.

[14] J. E. POTTER, *Matrix quadratic solutions*, SIAM J. Appl. Math., 14 (1966), pp. 496–501.

[15] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.

[16] M. VIDYASAGAR, *Control System Synthesis*, MIT Press, Cambridge, MA, 1985.

[17] N. WIENER, *Extrapolation, Interpolation and Smoothing of Stationary Time Series*, John Wiley, New York, 1949.

[18] H. K. WIMMER, *Monotonicity of maximal solutions of algebraic Riccati equations*, Systems Control Lett., 5 (1985), pp. 317–319.

[19] W. M. WONHAM, *Linear Multivariable Control*, Springer-Verlag, New York, 1985.

[20] G. ZAMES AND B. A. FRANCIS, *Feedback, minimax sensitivity, and optimal robustness*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 585–601.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 1, pp. 18-29, January 1989

002

BLOCK KRONECKER PRODUCTS AND BLOCK NORM MATRICES IN LARGE-SCALE SYSTEMS ANALYSIS*

DAVID C. HYLAND† AND EMMANUEL G. COLLINS, JR.†

Abstract. Complex and large-scale systems are often viewed as collections of interacting subsystems. Properties of the overall system are then deduced from the properties of the individual subsystems and their interconnections. This analysis process for large-scale systems usually requires manipulating the matrix subblocks of block-partitioned matrices. Two tools that are useful in linear systems analysis are the Kronecker product and the matrix modulus ($|a_{ij}|$). However, these tools are designed for matrices partitioned into their scalar elements. Thus, this paper defines and presents properties of the block Kronecker product and block norm matrix, generalizations of the Kronecker product and matrix modulus to block-partitioned matrices. The utility of the results is illustrated by deriving in simplified fashion a recent result in robustness analysis.

Key words. block Kronecker product, block norm matrix, large-scale systems

AMS(MOS) subject classification. 15

1. Introduction.

In the analysis of complex and large-scale dynamic systems it is often advantageous to regard the overall system as a collection of interacting subsystems. Properties of the aggregate system can then be deduced from the properties of the individual subsystems and their interconnections. (See, e.g., [6], [7], [11], [14] for a small sample of the numerous published results that take this approach.) For linear systems, this type of analysis often involves manipulating the matrix subblocks of block-partitioned matrices.

Two tools that have been useful in systems analysis are the Kronecker product and the matrix modulus. The Kronecker product, for example, has been used to find the solution of linear matrix equations [1], [4], [10], [17], in the development of matrix calculus [4], [9], [13], [16], and in dynamic sensitivity analysis [3], [4]. However, the Kronecker product was designed for matrices partitioned into their scalar elements. For example if A, Q, and B are matrices, then the Kronecker product allows us to write

$$\text{vec}\,(AQB)=(B^T\otimes A)\,\text{vec}\,(Q) \tag{1.1}$$

where vec $(\cdot)$ is the vector-valued operator that stacks the columns of a matrix in a vector. However, suppose Q were partitioned into matrix subblocks. Then the operation (1.1) destroys this structure.

The matrix modulus of the matrix Q is the matrix $|q_{ij}|$ and has been used to develop robust stability conditions for dynamic linear systems [11], [18]. However, if Q is partitioned into matrix subblocks, the matrix modulus is too fine in that it is based on a property (the absolute value) of the scalar elements of a matrix. Conversely, a norm $\|Q\|_\theta$ of the matrix Q is too coarse in that it totally ignores the block-partitioned structure of the matrix.

Tools designed specifically for block-partitioned matrices are obviously needed. One such collection of tools has been based on matrix majorants and minorants [5]. This paper develops additional results based on the block Kronecker product and the block norm matrix generalizations, respectively, of the Kronecker product and modulus matrix to block-partitioned matrices.

The paper proceeds as follows. In § 2 the block Kronecker product is introduced and some of its algebraic properties are presented. Then, in § 3 the block norm matrix is defined and some related equalities and inequalities are given. Next, § 4 presents results

* Received by the editors September 28, 1987; accepted for publication (in revised form) May 9, 1988. This research was partly supported by Air Force Office of Scientific Research contract F49620-86-C-0038.

† Harris Corporation, Government Aerospace Systems Division, Melbourne, Florida 32902.

on block-diagonal and diagonal matrix structures. These results are useful in § 5, which uses results developed in the previous sections to derive in a simplified fashion the covariance block norm inequality of Proposition 4.2 in [7].

Before proceeding we present some notation. It is assumed that the matrices are, in general, complex.

I_p	$p \times p$ identity matrix
$\otimes$, $\oplus$	Kronecker product, Kronecker sum [4], [8]
$\mathrm{col}_i\,(Z)$	ith column of matrix Z
$\mathrm{vec}\,(Z)$	$\begin{bmatrix} \mathrm{col}_1\,(Z) \\ \mathrm{col}_2\,(Z) \\ \vdots \\ \mathrm{col}_q\,(Z) \end{bmatrix}, \quad Z \text{ is a } p \times q \text{ matrix}$
$\mathrm{vec}_{pq}^{-1}\,(z)$	$p \times q$ matrix defined such that $\mathrm{vec}\,[\mathrm{vec}_{pq}^{-1}\,(z)] = z$; z is a vector of dimension pq
Z^T	transpose of matrix Z
Z^H	conjugate transpose of matrix Z
z_{ij}	(i, j) element of matrix Z
$Y * Z$	Hadamard product of $p \times q$ matrices Y and Z $(Y * Z = [y_{ij} z_{ij}])$ [15]
$Y \leqq\leqq Z$	$y_{ij} \leqq z_{ij}$ for all i and j (Y and Z are real matrices of equal dimension.)
nonnegative matrix	matrix with nonnegative elements ($Z \geqq\geqq 0$)
$\mathrm{tr}\,(Z)$	trace of matrix Z
$\lambda_{\min}(Z)$, $\lambda_{\max}(Z)$	minimum and maximum eigenvalues of Hermitian matrix Z
$\sigma_i(Z)$	singular value of matrix Z
$\sigma_{\min}(Z)$, $\sigma_{\max}(Z)$	smallest and largest singular values of matrix Z
$\Vert Z \Vert_\theta$	any norm of matrix Z (not necessarily induced by a vector norm)
$\Vert Z \Vert_\phi$	any norm of matrix Z induced by a vector norm $\Vert \cdot \Vert_\alpha$ $(\Vert Z \Vert_\phi = \max_{\Vert y \Vert_\alpha = 1} \Vert Zy \Vert_\alpha)$
$\Vert y \Vert_2$	Euclidean norm of vector y
$\Vert Z \Vert_s$	Spectral norm of matrix Z, induced by the Euclidean norm $\Vert \cdot \Vert_2$
$\Vert Z \Vert_F$	Frobenius norm of matrix Z $(\Vert Z \Vert_F^2 = \sum_i \sum_j \lvert z_{ij}^2 \rvert)$

2. Block Kronecker products.

2. Block Kronecker products. This section introduces the block Kronecker product and a related vector-valued function vecb $(\cdot)$. The algebra associated with the block Kronecker product is also presented (see Table A). The reader familiar with the standard Kronecker algebra will quickly recognize that the block Kronecker algebra is almost identical. This is essentially due to property (A.3) of Table A.

It should be recognized that below the primary consideration is the special case of square matrices with square diagonal blocks. This restriction is to avoid notational complexity and confusion. However, most of the results extend to more general partitions. The extensions require a clear definition of how various matrices are partitioned (such as when multiplying rectangular matrices A and B).

Consider the $n \times n$ partitioned matrices

$$A = [A_{ij}]_{(i,j=1,\cdots,r)}, \tag{2.1a}$$

$$B = [B_{ij}]_{(i,j=1,\cdots,r)} \tag{2.1b}$$

where A_{ij} and B_{ij} are $n_i \times n_j$ and $\sum_{i=1}^{r} n_i = n$. The $n^2 \times 1$ vector vecb (A) is defined by

$$\text{(2.2)} \qquad \operatorname{vecb}(A) \triangleq \begin{bmatrix} \operatorname{vec}(A_{11}) \\ \vdots \\ \operatorname{vec}(A_{r1}) \\ \operatorname{vec}(A_{12}) \\ \vdots \\ \operatorname{vec}(A_{r2}) \\ \vdots \\ \operatorname{vec}(A_{1r}) \\ \vdots \\ \operatorname{vec}(A_{rr}) \end{bmatrix}.$$

Notice that vecb $(\cdot)$ is a linear operator.

We need to define an operation $A \bar{\otimes} B$ such that for an $n \times n$ matrix D partitioned identically to A and B

$$\text{(2.3)} \qquad \operatorname{vecb}(BDA^T) = (A \bar{\otimes} B) \operatorname{vecb}(D).$$

This motivates the definition of the *block Kronecker product* of A and B, denoted by $A \bar{\otimes} B$. $A \bar{\otimes} B$ is the $n^2 \times n^2$ matrix defined by

$$\text{(2.4)} \qquad A \bar{\otimes} B \triangleq \begin{bmatrix} A_{11} \circledast B & A_{12} \circledast B & \cdots & A_{1r} \circledast B \\ A_{21} \circledast B & A_{22} \circledast B & \cdots & A_{2r} \circledast B \\ \vdots & \vdots & \ddots & \vdots \\ A_{r1} \circledast B & A_{r2} \circledast B & \cdots & A_{rr} \circledast B \end{bmatrix}$$

where the $n_i n \times n_j n$ matrix product $A_{ij} \circledast B$ is defined by

$$\text{(2.5)} \qquad A_{ij} \circledast B \triangleq \begin{bmatrix} A_{ij} \otimes B_{11} & A_{ij} \otimes B_{12} & \cdots & A_{ij} \otimes B_{1r} \\ A_{ij} \otimes B_{21} & A_{ij} \otimes B_{22} & \cdots & A_{ij} \otimes B_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ A_{ij} \otimes B_{r1} & A_{ij} \otimes B_{r2} & \cdots & A_{ij} \otimes B_{rr} \end{bmatrix}.$$

The *block Kronecker sum* of A and B is denoted by $A \bar{\oplus} B$ and is defined by

$$\text{(2.6)} \qquad A \bar{\oplus} B = A \bar{\otimes} I_n + I_n \bar{\otimes} B.$$

Recognize that if the matrices A, B, and I_n are partitioned into their scalar elements (i.e., $r = n$) then $A \bar{\otimes} B = A \otimes B$ and $A \bar{\oplus} B = A \oplus B$, so that the block Kronecker product and block Kronecker sum reduce, respectively, to the Kronecker product and Kronecker sum.

Some of the basic algebraic properties of the block Kronecker product and block Kronecker sum are presented in Table A. In this table it is assumed that A and B are $n \times n$ matrices partitioned as in (2.1) and C and D are $n \times n$ matrices partitioned identically to A and B. Also, $f(\cdot)$ denotes an analytic function. The eigenvalues of A are denoted by $\lambda^{(i)}$ $(i = 1, \cdots, n)$ and $\alpha^{(i)}$ denotes any corresponding eigenvectors. Similarly, the eigenvalues of B are denoted by $\mu^{(i)}$, and $\beta^{(i)}$ denotes any corresponding eigenvectors. Recognize that if A or B have redundant eigenvalues, then it is possible to have $\alpha^{(j)} = \alpha^{(k)}$ or $\beta^{(j)} = \beta^{(k)}$ for $j \neq k$. Thus statements (A.16)–(A.18) in Table A are not redundant.

To understand (A.18) it is necessary to define the block Kronecker product of two vectors. Consider the n-dimensional partitioned vectors

$$\text{(2.7a)} \qquad x^T = [x_1^T, x_2^T, \cdots, x_r^T],$$

$$\text{(2.7b)} \qquad y^T = [y_1^T, y_2^T, \cdots, y_r^T]$$

TABLE A
Algebra of block Kronecker products.

(A.1)	$\operatorname{vecb}(ADB) = (B^T \bar{\otimes} A) \operatorname{vecb}(D)$
(A.2)	$\operatorname{vecb}(AD + DB) = (B^T \bar{\oplus} A) \operatorname{vecb}(D)$
(A.3)	$A \bar{\otimes} B = P^T(A \otimes B)P$ for some permutation matrix P
(A.4)	$(A + B) \bar{\otimes} C = A \bar{\otimes} C + B \bar{\otimes} C$
(A.5)	$A \bar{\otimes} (B + C) = A \bar{\otimes} B + A \bar{\otimes} C$
(A.6)	$(A \bar{\otimes} B)^T = A^T \bar{\otimes} B^T$
(A.7)	$(A \bar{\otimes} B)(C \bar{\otimes} D) = (AC) \bar{\otimes} (BD)$
(A.8)	$(A \bar{\otimes} B)^{-1} = A^{-1} \bar{\otimes} B^{-1}$
(A.9)	$B \bar{\otimes} A = Q(A \bar{\otimes} B)Q$ for some permutation matrix Q
(A.10)	$\det(A \bar{\otimes} B) = [\det(A) \det(B)]^n$
(A.11)	$\operatorname{tr}(A \bar{\otimes} B) = \operatorname{tr}(A) \operatorname{tr}(B)$
(A.12)	$(I_n \bar{\otimes} A)(B \bar{\otimes} I_n) = (B \bar{\otimes} I_n)(I_n \bar{\otimes} A)$
(A.13)	$f(I_n \bar{\otimes} A) = I_n \bar{\otimes} f(A)$
(A.14)	$f(A \bar{\otimes} I_n) = f(A) \bar{\otimes} I_n$
(A.15)	$\exp(A \bar{\oplus} B) = \exp(A) \bar{\otimes} \exp(B)$
(A.16)	The eigenvalues of $(A \bar{\otimes} B)$ are the n^2 numbers $\lambda^{(i)}\mu^{(j)}$ $(i, j = 1, 2, \cdots, n)$
(A.17)	The eigenvalues of $(A \bar{\oplus} B)$ are the n^2 numbers $\lambda^{(i)}\mu^{(j)}$ $(i, j = 1, 2, \cdots, n)$
(A.18)	$\alpha^{(i)} \bar{\otimes} \beta^{(j)}$ is an eigenvector of $A \bar{\otimes} B$ with eigenvalue $\lambda^{(i)}\mu^{(j)}$ and is also an eigenvector of $A \bar{\oplus} B$ with eigenvalue $\lambda^{(i)} + \mu^{(j)}$.

where x_i and y_i are n_i vectors and $\sum_{i=1}^{r} n_i = n$. Then, the $n^2 \times 1$ vector $x \bar{\otimes} y$ is defined by

$$x \bar{\otimes} y \triangleq \begin{bmatrix} x_1 \circledast y \\ x_2 \circledast y \\ \vdots \\ x_r \circledast y \end{bmatrix} \tag{2.8}$$

where

$$x_i \circledast y \triangleq \begin{bmatrix} x_i \otimes y_1 \\ x_i \otimes y_2 \\ \vdots \\ x_i \otimes y_r \end{bmatrix}. \tag{2.9}$$

The proofs of most of the properties presented in Table A are easy once the validity of (A.1) and (A.3) is established. Thus the proofs of these two statements are presented and then the proofs of the remaining results are discussed with the exception of property (A.18) whose proof is presented in detail.

Proof of (A.1). By definition

$$\operatorname{vecb}(ADB) = \begin{bmatrix} \operatorname{vec}((ADB)_{11}) \\ \vdots \\ \operatorname{vec}((ADB)_{r1}) \\ \operatorname{vec}((ADB)_{12}) \\ \vdots \\ \operatorname{vec}((ADB)_{r2}) \\ \vdots \\ \operatorname{vec}((ADB)_{1r}) \\ \vdots \\ \operatorname{vec}((ADB)_{rr}) \end{bmatrix}. \tag{2.10}$$

The (p, q) block of (ADB) is given by

$$(2.11)\qquad (ADB)_{pq} = \sum_{i=1}^{r} \sum_{j=1}^{r} A_{pj} D_{ji} B_{iq}.$$

Also,

$$(2.12)\qquad \operatorname{vec}(A_{pj} D_{ji} B_{iq}) = (B_{iq}^T \otimes A_{pj}) \operatorname{vec}(D_{ji}).$$

Substituting (2.11) and (2.12) into (2.10) shows that vecb (ADB) may be expressed as an $r \times r$ block matrix where the (p, q) block has dimension $n_p n_q \times n_p n_q$ and is given by

$$(2.13)\qquad [\operatorname{vecb}(ADB)]_{pq} = \begin{bmatrix} \sum_j (B_{qp}^T \otimes A_{1j}) \operatorname{vec}(D_{jq}) \\ \sum_j (B_{qp}^T \otimes A_{2j}) \operatorname{vec}(D_{jq}) \\ \vdots \\ \sum_j (B_{qp}^T \otimes A_{rj}) \operatorname{vec}(D_{jq}) \end{bmatrix}.$$

When we use the definition of $B_{qp}^T \circledast A$ (see (2.5)), it follows that (2.13) is equivalent to

$$(2.14)\qquad [\operatorname{vecb}(ADB)]_{pq} = (B_{qp}^T \circledast A) \begin{bmatrix} \operatorname{vec}(D_{1q}) \\ \operatorname{vec}(D_{2q}) \\ \vdots \\ \operatorname{vec}(D_{rq}) \end{bmatrix}.$$

Property (A.1) follows from (2.14). □

Proof of (A.3). Consider the equation

$$(2.15)\qquad ADB^T = C,$$

which is equivalent to

$$(2.16)\qquad (A \otimes B) \operatorname{vec}(D) = \operatorname{vec}(C).$$

Applying (A.1) to (2.15), we obtain

$$(2.17)\qquad (A \bar{\otimes} B) \operatorname{vecb}(D) = \operatorname{vecb}(C).$$

There exists a permutation matrix P such that

$$(2.18a)\qquad \operatorname{vecb}(C) = P \operatorname{vec}(C),$$

$$(2.18b)\qquad \operatorname{vecb}(D) = P \operatorname{vec}(D).$$

Substituting (2.18) into (2.17), premultiplying by P^T, and using $P^T P = I_{n^2}$ we obtain

$$(2.19)\qquad P^T (A \bar{\otimes} B) P \operatorname{vec}(D) = \operatorname{vec}(C).$$

Subtracting (2.16) from (2.19), we obtain

$$(2.20)\qquad [P^T (A \bar{\otimes} B) P - (A \otimes B)] \operatorname{vec}(D) = 0.$$

Since (2.20) is valid for *all* choices of D it follows that the expression in "[ ]" is identically zero. □

Property (A.2) now follows from (A.1). Property (A.3) implies that

$$(2.21)\qquad A \otimes B = P (A \bar{\otimes} B) P^T.$$

The proofs of (A.4)–(A.9) and (A.12) are then obtained by substituting (2.21) into the equivalent expressions for the standard Kronecker product and Kronecker sum [4], [8]. For example, substituting (2.21) into

$$(A\otimes B)(C\otimes D)=(AC)\otimes(BD) \tag{2.22}$$

we obtain

$$P(A\bar{\otimes}B)P^TP(C\bar{\otimes}D)P^T=P(AC)\bar{\otimes}(BD)P^T. \tag{2.23}$$

By pre- and post-multiplying (2.23) by P^T and P, respectively, we obtain (A.7).

The proofs of (A.10)–(A.11) and (A.16)–(A.17) follow from the equivalent results for the Kronecker product and Kronecker sum [4], [8], the property (A.3), and the fact that the determinant, trace, and eigenvalues of a matrix are invariant under similarity transformation.

Since $f(\cdot)$ is analytic, there exists a scalar sequence $\{f_i\}_{i=0}^{\infty}$ such that

$$f(\lambda)=\sum_{n=0}^{\infty} f_n\lambda^n. \tag{2.24}$$

Also, from (A.7) it follows that

$$(I_n\bar{\otimes}A)^i=I_n\bar{\otimes}A^i, \tag{2.25a}$$

$$(A\bar{\otimes}I_n)^i=A^i\bar{\otimes}I_n. \tag{2.25b}$$

The proofs of (A.13) and (A.14) follow from (2.24) and (2.25). Property (A.15) follows from (A.12), the fact that the exponential of the sum of commuting matrices is the product of exponentials (i.e., if $MN=NM$, $e^{(M+N)}=e^Me^N$) and (A.13), (A.14), and (A.7).

Finally, the proof of (A.18) is presented.

Proof of (A.18). Let $\mathrm{col}_1\,(M)$ denote the first column of the matrix M and let the $n\times n$ matrices E and F satisfy

$$\mathrm{col}_1\,(E)=\alpha^{(i)}, \tag{2.26a}$$

$$\mathrm{col}_1\,(F)=\beta^{(j)}. \tag{2.26b}$$

Then,

$$\mathrm{col}_1\,(AE)=\lambda^{(i)}\alpha^{(i)}, \tag{2.27a}$$

$$\mathrm{col}_1\,(BF)=\mu^{(j)}\beta^{(j)}. \tag{2.27b}$$

Using (A.7) we may write

$$(A\bar{\otimes}B)(E\bar{\otimes}F)=AE\bar{\otimes}BF, \tag{2.28}$$

and thus

$$(A\bar{\otimes}B)\,\mathrm{col}_1\,(E\bar{\otimes}F)=\mathrm{col}_1\,(AE\bar{\otimes}BF). \tag{2.29}$$

Recognize that for any $n\times n$ matrices M and N

$$\mathrm{col}_1\,(M\bar{\otimes}N)=\mathrm{col}_1\,(M)\bar{\otimes}\mathrm{col}_1\,(N). \tag{2.30}$$

The first part of the proof is completed by using (2.30) and substituting (2.26) and (2.27) into (2.29) to obtain

$$(A\bar{\otimes}B)(\alpha^{(i)}\bar{\otimes}\beta^{(j)})=\lambda^{(i)}\mu^{(j)}(\alpha^{(i)}\bar{\otimes}\beta^{(j)}). \tag{2.31}$$

From (2.6) and (A.7) it follows that

$$(A\bar{\oplus}B)(E\bar{\otimes}F)=AE\bar{\otimes}F+E\bar{\otimes}BF \tag{2.32}$$

and thus

$$(A\bar{\oplus}B)\operatorname{col}_1(E\bar{\otimes}F)=\operatorname{col}_1(AE\bar{\otimes}F)+\operatorname{col}_1(E\bar{\otimes}BF). \tag{2.33}$$

The latter part of the proof follows by using (2.30) and substituting (2.26) and (2.27) into (2.33) to obtain

$$(A\bar{\oplus}B)(\alpha^{(i)}\bar{\otimes}\beta^{(j)})=(\lambda^{(i)}+\mu^{(j)})(\alpha^{(i)}\bar{\otimes}\beta^{(j)}). \tag{2.34}$$ □

3. Block norm matrices. This section defines the block norm matrix and block comparison matrix of a given matrix. Then some basic properties of the block norm matrix are presented.

Consider the $p \times q$ partitioned matrix

$$N=[N_{ij}]_{(i=1,\cdots,u;j=1,\cdots,v)} \tag{3.1}$$

where N_{ij} is $p_i \times q_j$, $\sum_{i=1}^{u} p_i = p$ and $\sum_{i=1}^{v} q_i = q$. Then for any matrix norm $\|\cdot\|_\theta$ define the $u \times v$ *block norm matrix* $\bar{N}_\theta$ [12] by

$$\bar{N}_\theta=[\bar{N}]_\theta \triangleq [\,\|N_{ij}\|_\theta]_{(i=1,\cdots,p;j=1,\cdots,q)}. \tag{3.2}$$

The nonnegative matrix $\bar{N}_\theta$ is a generalization of the modulus matrix ($[\,|N_{ij}|\,]$) for scalar-partitioned matrices.

Also, consider the $p \times p$ partitioned matrix

$$M=[M_{ij}]_{(i,j=1,\cdots,u)} \tag{3.3}$$

where M_{ij} is $p_i \times p_j$. Let $\|\cdot\|_\phi$ denote a matrix norm induced by the vector norm $\|\cdot\|_\alpha$ and define the $u \times u$ *block comparison matrix* $\underline{M}_\phi$ [12] by

$$\underline{M}_\phi=[\underline{M}]_\phi \triangleq [\underline{m}_{ij}]_{(i,j=1,\cdots,u)} \tag{3.4a}$$

where

$$\underline{m}_{ii}=\|M_{ii}^{-1}\|_\phi^{-1}, \tag{3.4b}$$

$$\underline{m}_{ij}=-\|M_{ij}\|_\phi \quad \text{for } i \neq j. \tag{3.4c}$$

Here, by convention, if M_{ii} is singular, then $\|M_{ii}^{-1}\|_\theta^{-1} = 0$. $\underline{M}_\theta$ is a generalization of the comparison matrix [2] for scalar-partitioned matrices.

Some of the properties of block norm and block comparison matrices are presented in Table B. However, before discussing and proving these properties we state the following results on matrix norms.

PROPOSITION 3.1. *Let U be an $m \times n$ matrix and V an $n \times p$ matrix. Then*

$$\sigma_{\min}(U)\|V\|_F \leqq \|UV\|_F \leqq \sigma_{\max}(U)\|V\|_F, \tag{3.5a}$$

$$\|U\|_F\sigma_{\min}(V) \leqq \|UV\|_F \leqq \|U\|_F\sigma_{\max}(V). \tag{3.5b}$$

Proof. Express $\|UV\|_F^2$ as

$$\|UV\|_F^2=\operatorname{tr}(V^HU^HUV). \tag{3.6}$$

But U^HU has the modal composition

$$U^HU=E^H\Omega E \tag{3.7}$$

where E is unitary and

$$\Omega = \operatorname{diag}\{\sigma_i^2(U)\}_{i=1}^{n}. \tag{3.8}$$

Thus

$$\|UV\|_F^2 = \operatorname{tr}(V^H E^H \Omega E V) = \operatorname{tr}(\Omega E V V^H E^H). \tag{3.9}$$

It follows that

$$\sigma_{\min}^2(U)\operatorname{tr}(EVV^H E) \leqq \|UV\|_F^2 \leqq \sigma_{\max}^2(U)\operatorname{tr}(EVV^H E^H). \tag{3.10}$$

Inequality (3.5a) then follows since

$$\operatorname{tr}(EVV^H E) = \operatorname{tr}(VV^H) = \|V\|_F^2. \tag{3.11}$$

Inequality (3.5b) is proved similarly by using

$$\|UV\|_F^2 = \operatorname{tr}(UVV^H U^H). \tag{3.12}$$

□

PROPOSITION 3.2. *Let U and V be arbitrary matrices. Then,*

$$\sigma_{\max}(U \otimes V) = \sigma_{\max}(U)\sigma_{\max}(V). \tag{3.13}$$

Proof.

$$\sigma_{\max}^2(U \otimes V) = \lambda_{\max}((U \otimes V)(U \otimes V)^H). \tag{3.14}$$

Using known properties of the Kronecker product [4], [8] we find that

$$\begin{aligned} \sigma_{\max}^2(U \otimes V) &= \lambda_{\max}(UU^H \otimes VV^H) \\ &= \lambda_{\max}(UU^H)\lambda_{\max}(VV^H) \\ &= \sigma_{\max}^2(U)\sigma_{\max}^2(V). \end{aligned} \tag{3.15}$$

□

PROPOSITION 3.3. *Let U be an arbitrary matrix. Then,*

$$\|\operatorname{vec}(U)\|_F = \|U\|_F. \tag{3.16}$$

Proof. The result follows from the definition of the Frobenius norm $\|\cdot\|_F$. □

In Table B, c denotes a scalar, M is a $p \times p$ matrix partitioned as in (3.3), N and R are $p \times q$ matrices partitioned as in (3.1), and P is an $s \times p$ matrix partitioned compatibly with M and N. A, B, and D are $n \times n$ matrices partitioned identically in the form (2.1). The partitions of vecb (D) are assumed to be the vectors vec (D_{ij}) and the partitions of $(A \bar{\otimes} B)$ are chosen compatibly (i.e., the partitions are all of the form $A_{ij} \otimes B_{kl}$).

TABLE B
Block norm matrix properties.

(B.1)	$[\overline{cN}]_\theta = \lvert c \rvert \bar{N}_\theta$
(B.2)	$[\bar{N} + \bar{R}]_\theta \leqq\leqq \bar{N}_\theta R^{-}_\theta$
(B.3)	$[\overline{PN}]_\phi \leqq\leqq \bar{P}_\phi \bar{N}_\phi$
(B.4)	$[\overline{PN}]_F \leqq\leqq \bar{P}_s \bar{N}_F$
(B.5)	$[\overline{PN}]_F \leqq\leqq \bar{P}_F \bar{N}_s$
(B.6)	$[\overline{MN}]_F \geqq\geqq \underline{M}_s \bar{N}_F$
(B.7)	$[\overline{A \bar{\otimes} B}]_s = \bar{A}_s \otimes \bar{B}_s$
(B.8)	$[\overline{A \bar{\oplus} B}]_s \leqq\leqq \bar{A}_s \oplus \bar{B}_s$
(B.9)	$[\overline{\operatorname{vecb}(D)}]_F = \operatorname{vec}(\bar{D}_F)$

Properties (B.1) and (B.2) follow immediately from the norm properties $\|cN\|_\theta = |c|\ \|N\|_\theta$ and the triangle inequality, $\|N+R\|_\theta \leqq \|N\|_\theta + \|R\|_\theta$. Property (B.3) is a result of the triangle inequality and the induced norm property $\|PN\|_\phi \leqq \|P\|_\phi \|N\|_\phi$.

Before considering the remaining results, recognize that

$$\sigma_{\max}(N_{ij}) = \|N_{ij}\|_s, \tag{3.17a}$$

$$\sigma_{\min}(M_{ii}) = \|M_{ii}^{-1}\|_s^{-1}. \tag{3.17b}$$

Properties (B.4) and (B.5) then follow, respectively, from the triangle inequality and the right-hand side inequalities of (3.5a) and (3.5b).

Proof of (B.6).

$$\|(MN)_{ij}\|_F = \left\| M_{ii}N_{ij} + \sum_{\substack{k=1\\k\neq i}}^{u} M_{ik}N_{kj} \right\|_F . \tag{3.18}$$

It then follows from $\|N+R\| \geqq \|N\| - \|R\|$ and (3.5a) that

$$\|(MN)_{ij}\|_F \geqq \sigma_{\min}(M_{ii})\|N_{ij}\|_F - \left\| \sum_{\substack{k=1\\k\neq i}}^{u} M_{ik}N_{kj} \right\|_F . \tag{3.19}$$

But since

$$\left\| \sum_{\substack{k=1\\k\neq i}}^{u} M_{ik}N_{kj} \right\|_F \leqq \sum_{\substack{k=1\\k\neq i}}^{u} \|M_{ik}\|_F \|N_{kj}\|_F, \tag{3.20}$$

it follows that

$$\|(MN)_{ij}\|_F \geqq \sigma_{\min}(M_{ii})\|N_{ij}\|_F + \sum_{\substack{k=1\\k\neq i}}^{u} (-\|M_{ik}\|_F)\|N_{kj}\|_F, \tag{3.21}$$

which is equivalent to

$$\|(MN)_{ij}\|_F \geqq \sum_{k=1}^{u} (\underline{M}_s)_{ik}(\bar{N}_s)_{kj}. \tag{3.22}$$

Property (B.6) follows from (3.22). □

Proof of (B.7).

$$(A\bar{\otimes}B) = [A_{ij}\circledast B]_{(i,j=1,\cdots,r)} \tag{3.23}$$

where

$$A_{ij}\circledast B = \begin{bmatrix} A_{ij}\otimes B_{11} & A_{ij}\otimes B_{12} & \cdots & A_{ij}\otimes B_{1r} \\ A_{ij}\otimes B_{21} & A_{ij}\otimes B_{22} & \cdots & A_{ij}\otimes B_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ A_{ij}\otimes B_{r1} & A_{ij}\otimes B_{r2} & \cdots & A_{ij}\otimes B_{rr} \end{bmatrix} . \tag{3.24}$$

It follows that

$$\overline{(A\bar{\otimes}B)_s} = [\overline{(A_{ij}\circledast B)_s}]_{(i,j=1,\cdots,r)}. \tag{3.25}$$

Using (3.13), we write

$$\|A_{ij}\otimes B_{kl}\|_s = \|A_{ij}\|_s \|B_{kl}\|_s. \tag{3.26}$$

Substituting (3.26) into (3.24), we obtain

$$[\overline{A_{ij} \circledast B}]_s = [\bar{A}_s]_{ij} \bar{B}_s. \tag{3.27}$$

Property (B.7) follows from (3.25) and (3.27). □

Property (B.8) now follows from (B.2) and (B.7). Property (B.9) is a result of (3.16).

4. Diagonal structures.

4. Diagonal structures. In this section results concerning block-diagonal and diagonal matrices are presented. These results are used in the example of the next section.

Assume that A and B are $n \times n$ matrices of the following form:

$$A = \text{block-diag}\ \{A_i\}_{i=1}^{r}, \tag{4.1a}$$

$$B = \text{block-diag}\ \{B_i\}_{i=1}^{r} \tag{4.1b}$$

where A_i and B_i are $n_i \times n_i$ and $\sum_{i=1}^{r} n_i = n$. Then $A \bar{\otimes} B$ is the $n^2 \times n^2$ matrix

$$A \bar{\otimes} B = \text{block-diag}\ \{C_i\}_{i=1}^{r} \tag{4.2a}$$

where C_i is the $n_i n \times n_i n$ block-diagonal matrix

$$C_i = \begin{bmatrix} A_i \otimes B_1 & & 0 \\ & A_i \otimes B_2 & \\ & & \ddots \\ 0 & & A_i \otimes B_r \end{bmatrix}. \tag{4.2b}$$

Thus $A \bar{\otimes} B$ is block-diagonal with diagonal subblocks of the form $A_i \otimes B_j$.

It follows that $A \bar{\oplus} B$ is the $n^2 \times n^2$ matrix

$$A \bar{\oplus} B = \text{block-diag}\ \{D_i\}_{i=1}^{r} \tag{4.3a}$$

where D_i is the $n_i n \times n_i n$ block-diagonal matrix

$$D_i = \begin{bmatrix} A_i \oplus B_1 & & 0 \\ & A_i \oplus B_2 & \\ & & \ddots \\ 0 & & A_i \oplus B_r \end{bmatrix}. \tag{4.3b}$$

Thus $A \bar{\oplus} B$ is block-diagonal with diagonal subblocks of the form $A_i \oplus B_j$.

Now suppose v is an r^2 vector and E is an $r^2 \times r^2$ diagonal matrix. Express E as

$$E = \text{block-diag}\ \{E_i\}_{i=1}^{r} \tag{4.4a}$$

where the $r \times r$ diagonal matrix E_i is given by

$$E_i = \text{diag}\ \{\tilde{e}_{ij}\}_{j=1}^{r}. \tag{4.4b}$$

Then

$$\text{vec}_{rr}^{-1}\ (Ev) = \tilde{E}^T * \text{vec}_{rr}^{-1}\ (v) \tag{4.5}$$

where "$*$" denotes the Hadamard product and

$$\tilde{E} = [\tilde{e}_{ij}]_{(i,j=1,\cdots,r)}. \tag{4.6}$$

5. An illustrative example.

5. An illustrative example. We now use results from §§ 2–4 to derive the covariance block norm inequality found in Proposition 4.2 of [7].

Consider the nth order system

$$\dot{x}(t) = (A + G)x(t) + w(t) \tag{5.1}$$

where $w(t)$ is white noise with intensity V. It is assumed that the $n \times n$ matrix A is a stability matrix of the following form:

$$(5.2) \qquad A = \text{block-diag } \{A_i\}_{i=1}^{r}$$

where A_i is $n_i \times n_i$ $(\sum_{i=1}^{r} n_i = n)$ and represents the dynamics of the ith subsystem. G is an $n \times n$ matrix partitioned compatibly with A. The off-diagonal blocks of G represent the uncertain interactions among the various subsystems. It is assumed that for some nonnegative $r \times r$ matrix $\hat{G}$,

$$(5.3) \qquad \bar{G}_s \leqq\leqq \hat{G}.$$

Notice that $\hat{G}$ is a matrix majorant of G [5].

Assuming $(A + G)$ is a stability matrix, the asymptotic state covariance Q satisfies the Lyapunov equation

$$(5.4) \qquad 0 = (A+G)Q + Q(A+G)^T + V.$$

Assume that all matrices in (5.4) are partitioned compatibly. Then operating on (5.4) with vecb $(\cdot)$ and using (A.2), we obtain

$$(5.5) \qquad -(A \bar{\oplus} A) \operatorname{vecb}(Q) = (G \bar{\oplus} G) \operatorname{vecb}(Q) + \operatorname{vecb}(V),$$

and thus

$$(5.6) \qquad [\overline{-(A \bar{\oplus} A) \operatorname{vecb}(Q)}]_F = [\overline{(G \bar{\oplus} G) \operatorname{vecb}(Q) + \operatorname{vecb}(V)}]_F.$$

Considering the right-hand side of (5.6) and using (B.2), (B.4), (B.8), (B.9), and (5.3) consecutively, we obtain

$$(5.7) \qquad \begin{aligned} [\overline{(G \oplus G) \operatorname{vecb}(Q) + \operatorname{vecb}(V)}]_F &\leqq\leqq [\overline{(G \bar{\oplus} G) \operatorname{vecb}(Q)}]_F + [\overline{\operatorname{vecb}(V)}]_F \\ &\leqq\leqq [\overline{G \bar{\oplus} G}]_s [\overline{\operatorname{vecb}(Q)}]_F + [\overline{\operatorname{vecb}(V)}]_F \\ &\leqq\leqq (\bar{G}_s \oplus \bar{G}_s) \operatorname{vecb}(\bar{Q}_F) + \operatorname{vec}(\bar{V}_F) \\ &\leqq\leqq (\hat{G} \oplus \hat{G}) \operatorname{vec}(\bar{Q}_F) + \operatorname{vec}(\bar{V}_F). \end{aligned}$$

Similarly, considering the left-hand side of (5.6) and using (B.1), (B.6), and (B.9), we obtain

$$(5.8) \qquad [\overline{-(A \bar{\oplus} A) \operatorname{vecb}(Q)}]_F \geqq\geqq (\underline{A \bar{\oplus} A})_s \operatorname{vecb}(\bar{Q}_F).$$

Thus, from (5.6)–(5.8)

$$(5.9) \qquad (\underline{A \bar{\oplus} A})_s \operatorname{vec}(\bar{Q}_F) \leqq\leqq (\hat{G} \oplus \hat{G}) \operatorname{vec}(\bar{Q}_F) + \operatorname{vec}(\bar{V}_F).$$

It follows from (4.3) that $(\underline{A \bar{\oplus} A})_s$ is the $r^2 \times r^2$ diagonal matrix

$$(5.10\text{a}) \qquad (\underline{A \bar{\oplus} A})_s = \text{block-diag } \{[\underline{D}_i]_s\}_{i=1}^{r}$$

where $[\underline{D}_i]_s$ is the $r \times r$ diagonal matrix

$$(5.10\text{b}) \qquad [\underline{D}_i]_s = \operatorname{diag} \{\sigma_{\min}(A_i \oplus A_j)\}_{j=1}^{r}.$$

Define the $n \times n$ matrix $\tilde{A}$ by

$$(5.11) \qquad \tilde{A} = [\sigma_{\min}(A_i \oplus A_j)]_{(i,j=1,\cdots,r)}.$$

Then using (5.10) and (4.5) and operating with $\operatorname{vec}^{-1}(\cdot)$ on both sides of (5.9), we obtain

$$(5.12) \qquad \tilde{A}^T * \bar{Q}_F \leqq\leqq \hat{G}\bar{Q}_F + \bar{Q}_F\hat{G}^T + \bar{V}_F,$$

which is the covariance block norm inequality of Proposition 4.2 in [7].

REFERENCES

[1] S. BARNETT, *Matrix differential equations and Kronecker products*, SIAM J. Appl. Math., 24 (1973), pp. 1–5.

[2] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[3] J. W. BREWER, *Matrix calculus and the sensitivity analysis of linear dynamic systems*, IEEE Trans. Automat. Control, 23 (1978), pp. 748–751.

[4] ———, *Kronecker products and matrix calculus in systems theory*, IEEE Trans. Circuits Systems, 25 (1978), pp. 772–781; Erratum, IEEE Trans. Circuits Systems, 26 (1979), p. 360.

[5] G. DAHLQUIST, *On matrix majorants and minorants with applications to differential equations*, Linear Algebra Appl., 52/53 (1983), pp. 199–216.

[6] Y. S. HUNG AND D. J. N. LIMEBEER, *Robust stability of additively perturbed interconnected systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 1069–1075.

[7] D. C. HYLAND AND D. S. BERNSTEIN, *The majorant Lyapunov equation: A nonnegative matrix equation for robust stability and performance of large scale systems*, IEEE Trans. Automat. Control, 32 (1987), pp. 1005–1013.

[8] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, Academic Press, New York, 1985.

[9] H. NEUDECKER, *Some theorems on matrix differentiation with special reference to Kronecker matrix products*, J. Amer. Statist. Assoc., 64 (1969), pp. 953–963.

[10] ———, *A note on Kronecker matrix products and matrix equation systems*, SIAM J. Appl. Math., 17 (1969), pp. 603–606.

[11] O. D. I. NWOKAH, *The robust decentralized stabilization of complex feedback systems*, IEE Proc., 134 D (1987), pp. 43–47.

[12] A. M. OSTROWSKI, *On some metrical properties of operator matrices and matrices partitioned into blocks*, 2 (1961), pp. 161–209.

[13] D. S. G. POLLOCK, *Tensor products and matrix differential calculus*, Linear Algebra Appl., 67 (1985), pp. 169–193.

[14] D. D. SILJAK, *Large Scale Dynamic Systems: Stability and Structure*, Elsevier North-Holland, New York, 1978.

[15] G. P. H. STYAN, *Hadamard products and multivariate statistical analysis*, Linear Algebra Appl., 6 (1973), pp. 217–240.

[16] W. J. VETTER, *Matrix calculus operations and Taylor expansions*, SIAM Rev., 15 (1973), pp. 352–369.

[17] ———, *Vector structures and solutions of linear matrix equations*, Linear Algebra Appl., 10 (1975), pp. 181–188.

[18] R. K. YEDAVALLI, *Perturbation bounds for robust stability in linear state space models*, Internat. J. Control, 42 (1985), pp. 1507–1517.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 1, pp. 30–38, January 1989

A VARIANT OF KARMARKAR'S LINEAR PROGRAMMING ALGORITHM FOR PROBLEMS WITH SOME UNRESTRICTED VARIABLES*

JOHN E. MITCHELL† AND MICHAEL J. TODD‡

Abstract. A variant of Karmarkar's projective linear programming algorithm that can be used on problems with some unrestricted variables is considered. The variant is derived in two ways. One derivation involves eliminating the unconstrained variables, and the other involves solving a constrained least squares problem. The results of Gonzaga are used to show that our algorithm converges in $O(nq)$ iterations where n is the number of nonnegative variables and q is the precision required in the objective function value.

Key words. linear programming, Karmarkar's algorithm, unrestricted variables, constrained least squares problems

AMS(MOS) subject classifications. 90C05, 65F05, 65F25, 65F50

1. Introduction. In this paper we describe a variant of Karmarkar's algorithm [11] designed to solve linear programs where some of the variables are not constrained to be nonnegative. The linear programs are assumed to be in standard form. The variant we develop is a primal projective algorithm; we show that it takes $O(nq)$ steps, each requiring $O(n^3)$ work, where n is the number of variables constrained to be nonnegative and q is the precision required in the objective function value. We consider two approaches to finding a direction, one based on solving a constrained least squares problem and the other involving elimination of the unrestricted variables. We show that these two approaches lead to the same direction.

Vanderbei [15] has considered applying the affine variant of Karmarkar's algorithm (see, for example, [4] or [16]) to problems with unrestricted variables. He first assumes that the free variables have a lower bound and then derives a limiting direction as the lower bound approaches negative infinity.

Our original problem is written as follows:

$$(\mathrm{P})\qquad \begin{aligned} \text{minimize}\quad & c_A^T x_A + c_F^T x_F, \\ \text{subject to}\quad & A x_A + F x_F = 0, \\ & g_A^T x_A + g_F^T x_F = 1, \\ & x_A \geqq 0, \end{aligned}$$

where c_A, g_A, and x_A are n-vectors; c_F, g_F, and x_F are p-vectors; A is m by n of rank m; and F is m by p of rank p. We refer to the constraints

$$A x_A + F x_F = 0$$

* Received by the editors July 21, 1987; accepted for publication (in revised form) May 12, 1988.

† School of Operations Research and Industrial Engineering, College of Engineering, Cornell University, Ithaca, New York 14853-7501. Present address, Department of Mathematical Sciences, School of Science, Rensselaer Polytechnic Institute, Troy, New York 12180-3590 (mitchell@turing.cs.rpi.edu). The research of this author was partly supported by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University.

‡ School of Operations Research and Industrial Engineering, College of Engineering, Cornell University, Ithaca, New York 14853-7501(miketodd@cs.cornell.edu). The research of this author was partly supported by National Science Foundation grant ECS-8602534, by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University, and by Office of Naval Research contract N00014-87-K-0212.

(and any constraints equivalent to these) as the *subspace* constraints. We refer to the constraint

$$g_A^T x_A + g_F^T x_F = 1$$

(and any constraint equivalent to this) as the *normalizing* constraint.

The dual problem to (P) is

$$\text{(D)} \qquad \begin{aligned} &\text{maximize} && z, \\ &\text{subject to} && A^T y + z g_A \leqq c_A, \\ &&& F^T y + z g_F = g_A. \end{aligned}$$

We assume that (P) is feasible and that we have an initial feasible solution $(x_A^{0^T}, x_F^{0^T})^T$ with x_A^0 strictly positive. (Henceforth, we shall abuse notation and write (x_A^0, x_F^0), etc.) We also assume that the feasible region for (P) is compact. This is equivalent to saying that there does not exist a nonzero (x_A, x_F) satisfying

$$Ax_A + Fx_F = 0, \quad g_A^T x_A + g_F^T x_F = 0, \quad x_A \geqq 0.$$

Therefore, if we find a nonzero point (x_A, x_F) that satisfies all the constraints except the normalizing constraint, we automatically have

$$g_A^T x_A + g_F^T x_F > 0.$$

This observation is due to Gonzaga [9].

Note that if we have a general linear program in standard form with some unrestricted variables, we can transform it into a problem of the form (P). Assume we have a problem of the following form:

$$(\dot{\text{P}}) \qquad \begin{aligned} &\text{minimize} && \dot{c}_A^T x_A + \dot{c}_F^T x_F, \\ &\text{subject to} && \dot{A} x_A + \dot{F} x_F = b, \\ &&& x_A \geqq 0, \end{aligned}$$

where $\dot{c}_A$ and x_A are $(n - 1)$-vectors; $\dot{c}_F$ and x_F are p-vectors; b is an m-vector; $\dot{A}$ is m by $(n - 1)$; and $\dot{F}$ is m by p. We homogenize to obtain the following:

$$\begin{aligned} &\text{minimize} && \dot{c}_A^T x_A + \dot{c}_F^T x_F, \\ &\text{subject to} && [\dot{A} \mid -b] \begin{bmatrix} x_A \\ \xi \end{bmatrix} + \dot{F} x_F = 0, \\ &&& [0 \mid 1] \begin{bmatrix} x_A \\ \xi \end{bmatrix} = 1, \\ &&& x_A \geqq 0, \quad \xi \geqq 0. \end{aligned}$$

This formulation is equivalent to the original formulation and it is in the general form (P). Note that $g_F = 0$ and that if the feasible region of $(\dot{\text{P}})$ is compact, so is the feasible region of the homogenized version. Derivations similar to this (but working on problems without unrestricted variables) can be found in the work of, for example, Anstreicher [1], Gonzaga [9], Gay [5], Steger [13], and Ye and Kojima [17].

We propose to solve (P) using a variant of Karmarkar's algorithm. At iteration k of this algorithm we have a solution (x_A^k, x_F^k) to (P) and a lower bound z_k on the optimal value of (P). We define X_k to be the diagonal matrix with diagonal elements the entries of x_A^k, so $X^k e = x_A^k$, where e denotes a vector of ones of the appropriate dimension. Then

(e, x_F^k) is a feasible solution to the following rescaled problem:

$$(\bar{\mathrm{P}})\qquad \begin{array}{ll} \text{minimize} & \bar{c}_A^T\bar{x}_A + c_F^T x_F, \\ \text{subject to} & \bar{A}\bar{x}_A + F x_F = 0, \\ & \bar{g}_A^T\bar{x}_A + g_F^T x_F = 1, \\ & \bar{x}_A \geqq 0, \end{array}$$

where $\bar{A} := \bar{A}^k := AX_k$, $\bar{c}_A := \bar{c}_A^k := X_k c_A$, and $\bar{g}_A := \bar{g}_A^k := X_k g_A$.

Each iteration consists of three steps:

(1) Obtain a new lower bound $z = z_{k+1} \geqq z_k$.

(2) Find a suitable vector $\bar{d} = (\bar{d}_A, d_F)$ in the null space of $[\bar{A}\,|\,F]$. (We discuss how to choose $\bar{d}$ in later sections.) Choose a steplength α and set

$$\bar{x}_A \leftarrow e + \alpha\bar{d}_A, \qquad x_F \leftarrow x_F^k + \alpha d_F.$$

(Here $\alpha > 0$ is such that $\bar{x}_A > 0$. For example, α can be chosen by using a line search on an appropriate potential function, it can take a fixed value, or it can be chosen so that $\min\{\bar{x}_{A_i}\colon 1 \leqq i \leqq n\} = \gamma$ for some fixed γ such as 0.1 or 0.01. See references [11], [14], and [16].)

(3) Radially project so that the normalizing constraint is satisfied: Let

$$\zeta = \bar{g}_A^T\bar{x}_A + g_F^T x_F$$

and then rescale by setting

$$x_A^{k+1} \leftarrow X_k\bar{x}_A/\zeta, \qquad x_F^{k+1} \leftarrow x_F/\zeta.$$

In § 2 we derive a feasible direction $\bar{d}$ by solving a constrained least squares problem, in § 3 we derive a direction by eliminating the unrestricted variables x_F, and in § 4 we show that these two directions are equivalent.

2. Obtaining a direction using a least squares approach.

The problem is assumed to have the following form:

$$(\bar{\mathrm{P}})\qquad \begin{array}{ll} \text{minimize} & \bar{c}_A^T\bar{x}_A + c_F^T x_F, \\ \text{subject to} & \bar{A}\bar{x}_A + F x_F = 0, \\ & \bar{g}_A^T\bar{x}_A + g_F^T x_F = 1, \\ & \bar{x}_A \geqq 0, \end{array}$$

where $\bar{x}_A$, $\bar{c}_A$, and $\bar{g}_A$ are n-vectors; x_F, c_F, and g_F are p-vectors; $\bar{A} \in \Re^{m\times n}$; and $F \in \Re^{m\times p}$.

The dual to this problem is the following:

$$(\bar{\mathrm{D}})\qquad \begin{array}{ll} \text{maximize} & z, \\ \text{subject to} & \bar{A}^T y + z\bar{g}_A \leqq \bar{c}_A, \\ & F^T y + z g_F = c_F. \end{array}$$

(This is simply a rescaling of the dual (D) of (P) with the first constraints multiplied by the components of x_A^k.)

In this section we consider obtaining a direction by solving a constrained least squares problem. In the case when $p = 0$, that is, in a standard form linear program, the Karmarkar direction can be derived as the residual of the unconstrained least squares problem

$$(\mathrm{LS}(z)) \qquad \min_{y} \tfrac{1}{2}\|\bar{A}^T y-(\bar{c}_A-z\bar{g}_A)\|^2$$

where z is the current lower bound on the optimal value of the linear program. (See, for example, [6]. This reference augments the matrix $\bar{A}$ used in $(\mathrm{LS}(z))$ by a row of ones. However, as was shown by Gonzaga [9], the direction obtained when solving the least squares problem with this modified matrix is equivalent to the direction obtained by solving $(\mathrm{LS}(z))$.) Analogously to that derivation we find a direction for the problem (P) by solving the following constrained least squares problem:

$$(\mathrm{CLS}(z)) \qquad \begin{array}{ll} \min\limits_{y} & \tfrac{1}{2}\|\bar{A}^T y-(\bar{c}_A-z\bar{g}_A)\|_2^2 \\ \text{subject to} & F^T y=c_F-zg_F, \end{array}$$

where z is our current lower bound on the optimal value for (P). We choose this formulation because in $(\bar{\mathrm{D}})$ the dual constraints corresponding to the variables x_F are equality constraints and therefore any dual feasible solution (y, z) must satisfy the constraints in $(\mathrm{CLS}(z))$. Then a suitable direction for the constrained variables $\bar{x}_A$ in $(\bar{\mathrm{P}})$ is the optimal residual of the problem $(\mathrm{CLS}(z))$. We denote this direction by $\bar{d}_{\mathrm{CLS}_A}$.

We have assumed that F has full column rank and that A has full row rank. Therefore there exist feasible solutions to $(\mathrm{CLS}(z))$ and the optimal solution is unique. The Karush–Kuhn–Tucker optimality conditions [7] for y' to be an optimal solution to the problem $(\mathrm{CLS}(z))$ are as follows:

(1) There exists $v' \in \Re^p$ such that $\bar{A}(\bar{A}^T y' - (\bar{c}_A - z\bar{g}_A)) + Fv' = 0$.

(2) $F^T y' - (c_F - zg_F) = 0$.

It should be noted that v' is unique since we have assumed that F has full column rank.

Therefore, with $\bar{d}_{\mathrm{CLS}_A}$ as above, there exists a direction d_{CLS_F} $(:=v')$ such that $\bar{d}_{\mathrm{CLS}} := (\bar{d}_{\mathrm{CLS}_A}, d_{\mathrm{CLS}_F})$ is in the null space of $[\bar{A}\,|\,F]$. This is the direction we choose to move in at each iteration of our algorithm for solving (P). We then radially project in order to satisfy the normalizing constraint.

We now describe a straightforward way to update the lower bound z. Define $y(z)$ to be the optimal solution to $(\mathrm{CLS}(z))$. Then we can find a lower bound for (P) by solving

$$(\bar{\mathrm{D}}') \qquad \begin{array}{ll} \max\limits_{z} & z \\ \text{subject to} & \bar{A}^T y(z)+z\bar{g}_A \leqq \bar{c}_A. \end{array}$$

We will show later that $y(z)$ is a linear function of z (see (2.6)–(2.8)); it follows that $(\bar{\mathrm{D}}')$ can be solved by means of a ratio test. Let $\bar{z}'$ be the optimal value of $(\bar{\mathrm{D}}')$. Then $(y(\bar{z}'), \bar{z}')$ is feasible in $(\bar{\mathrm{D}})$ so $\bar{z}'$ is a valid lower bound for (P). If $\bar{z}' \leqq z_k$, we set $z_{k+1} \leftarrow z_k$ and $y_{k+1} \leftarrow y_k$; otherwise we set $z_{k+1} \leftarrow \bar{z}'$ and $y_{k+1} \leftarrow y(\bar{z}')$. When there are no unrestricted variables, this method of updating the lower bound z is equivalent to the method first given in Todd and Burrell [14].

A representation of the null space of F is necessary to solve the constrained least squares problem $(\mathrm{CLS}(z))$ and to find explicitly the direction $\bar{d}_{\mathrm{CLS}}$. The most numerically stable method of obtaining this representation is to form the QR-factorization of F. The standard procedure using this technique is given in Golub and Van Loan [8].

We first form the QR-factorization of F:

$$(2.1) \qquad F=[Q_1\,|\,Q_2]\left[\frac{R}{0}\right]$$

where $Q_1 \in \Re^{m \times p}$, $Q_2 \in \Re^{m \times (m-p)}$ and R is a $p \times p$ nonsingular upper triangular matrix. If we define

$$Q = [Q_1 \,|\, Q_2], \tag{2.2}$$

Q is an orthogonal matrix and Q_2 is an orthogonal basis for the null space of F.

For ease of notation we define several other quantities:

$$\check{y} := (\check{y}_1, \check{y}_2) := (Q_1^T y, Q_2^T y) \tag{2.3}$$

and

$$\check{A}_1 := A_1^k := Q_1^T \bar{A}, \qquad \check{A}_2 := A_2^k := Q_2^T \bar{A}. \tag{2.4}$$

Define $A_1 := Q_1^T A$ and $A_2 := Q_2^T A$. Then $\check{A}_1 = A_1 X_k$ and $\check{A}_2 = A_2 X_k$, so the products in (2.4) do not have to be computed at each iteration.

The problem (CLS(z)) can be rewritten as follows:

$$(\mathrm{CLS}'(z)) \qquad \begin{array}{ll} \displaystyle\min_{\check{y}_1, \check{y}_2} & \tfrac{1}{2} \| \check{A}_1^T \check{y}_1 + \check{A}_2^T \check{y}_2 - (\bar{c}_A - z \bar{g}_A) \|_2^2 \\ \text{subject to} & R^T \check{y}_1 = c_F - z g_F. \end{array}$$

Note that the constraints of (CLS$'(z)$) determine $\check{y}_1$ so (CLS$'(z)$) is equivalent to the following unconstrained least squares problem:

$$(\mathrm{CLS}''(z)) \qquad \min_{\check{y}_2} \; \tfrac{1}{2} \| \check{A}_2^T \check{y}_2 - (\bar{c}_A - z \bar{g}_A - \check{A}_1^T \check{y}_1) \|_2^2$$

where $\check{y}_1$ solves

$$R^T \check{y}_1 = c_F - z g_F. \tag{2.5}$$

Therefore the solution to (CLS$'(z)$) is

$$\check{y}_1(z) = R^{-T}(c_F - z g_F), \tag{2.6}$$

$$\check{y}_2(z) = [\check{A}_2 \check{A}_2^T]^{-1} \check{A}_2 (\bar{c}_A - z \bar{g}_A - \check{A}_1^T \check{y}_1(z)) \tag{2.7}$$

and the optimal solution to (CLS(z)) is

$$y_{\mathrm{CLS}}(z) := Q_1 \check{y}_1(z) + Q_2 \check{y}_2(z). \tag{2.8}$$

For any matrix M, let P_M denote the projection map onto the null space of M. The matrix operator corresponding to P_M is $I - M^T(MM^T)^{-1}M$. Then $\bar{d}_{\mathrm{CLS}_A}$, the optimal residual to the problem (CLS(z)), is given by

$$\bar{d}_{\mathrm{CLS}_A} = -P_{\check{A}_2}(\check{c}_A - z \check{g}_A), \tag{2.9}$$

where

$$\check{c}_A = (\bar{c}_A - \check{A}_1^T R^{-T} c_F) \tag{2.10}$$

and

$$\check{g}_A = (\bar{g}_A - \check{A}_1^T R^{-T} g_F). \tag{2.11}$$

The Karush–Kuhn–Tucker conditions given above yield

$$\begin{aligned} d_{\mathrm{CLS}_F} &= -R^{-1} \check{A}_1 \bar{d}_{\mathrm{CLS}_A} \\ &= R^{-1} \check{A}_1 P_{\check{A}_2}(\check{c}_A - z \check{g}_A). \end{aligned} \tag{2.12}$$

When we form the QR-factorization of F the constrained dual problem $(\bar{\mathrm{D}}')$ becomes

$$\begin{aligned} &\max_{z} && z \\ &\text{subject to} && \check{A}_2^T\check{y}_2(z)+z\check{g}_A \leqq \check{c}_A, \end{aligned}$$

from which $\bar{z}'$ can be obtained as above.

If A and F are sparse matrices, a different method should be used to solve the constrained least squares problem, since it is extremely unlikely that A_1 and A_2 will be sparse. Methods for solving sparse constrained least squares problems are discussed by Björck [2] and also by Coleman [3] and Heath [10]. One method puts a large weight on the equality constraints and includes them in the least squares objective. This approach is related to Vanderbei's algorithm [15] in that large weights are associated with variables that are far from their lower bounds. To satisfy the equality constraints in $(\mathrm{CLS}(z))$ exactly, it is necessary to find a basis W for the null space of F. In the dense case, one way to find such a basis is to perform the QR-factorization of F as in (2.1); the matrix Q_2 produced by this method is an orthogonal basis for the null space of F. When A and F are sparse, W should be sparse or represented in terms of sparse matrices, as in methods for large-scale, equality-constrained nonlinear programming. Then the major work at each iteration of our algorithm to solve the problem (P) is solving a system of the form

$$W^T\bar{A}\bar{A}^TWu=r$$

where r is a fixed vector. This can be solved indirectly provided we can efficiently form products involving W, $\bar{A}$, and their transposes.

3. Obtaining a direction by eliminating unrestricted variables. Consider the original problem

$$(\mathrm{P}) \qquad \begin{aligned} &\text{minimize} && c_A^Tx_A+c_F^Tx_F, \\ &\text{subject to} && Ax_A+Fx_F=0, \\ & && g_A^Tx_A+g_F^Tx_F=1, \\ & && x_A \geqq 0. \end{aligned}$$

In § 2 we obtained a direction by solving a constrained least squares problem. In this section we consider obtaining a direction by eliminating the unrestricted variables. This requires a representation of the null space of F; the resulting direction is independent of the basis we use for this null space. To facilitate comparison with the direction obtained in § 2 we define the QR-factorization of F as in (2.1), so Q_2 is a basis for the null space of F.

Substituting for x_F in (P), we obtain the following linear programming problem, which is a function of x_A only:

$$(\tilde{\mathrm{P}}) \qquad \begin{aligned} &\text{minimize} && \tilde{c}_A^Tx_A, \\ &\text{subject to} && \tilde{A}x_A=0, \\ & && \tilde{g}_A^Tx_A=1, \\ & && x_A \geqq 0, \end{aligned}$$

where

$$\tilde{c}_A=c_A-A^TQ_1R^{-T}c_F, \tag{3.1}$$

$$\tilde{g}_A=g_A-A^TQ_1R^{-T}g_F, \tag{3.2}$$

and

$$\tilde{A} = Q_2^T A. \tag{3.3}$$

For any x_A the corresponding x_F is given by

$$x_F = -R^{-1} Q_1^T A x_A. \tag{3.4}$$

If the feasible region for (P) is compact, so is that for ($\tilde{\text{P}}$).

($\tilde{\text{P}}$) is a standard form linear program so we can apply Karmarkar's algorithm to it directly. At iteration k we have a feasible solution x_A^k to ($\tilde{\text{P}}$) and a lower bound z_k on its value. Note that z_k is also a lower bound on the value of (P). We define X_k to be the diagonal matrix with elements the entries of x_A^k, so $X_k e = x_A^k$. Then e is a feasible solution to the rescaled problem

$$(\hat{\text{P}}) \qquad \begin{array}{ll} \text{minimize} & \hat{c}_A^T \hat{x}_A, \\ \text{subject to} & \hat{A}\hat{x}_A = 0, \\ & \hat{g}_A^T \hat{x}_A = 1, \\ & \hat{x}_A \geqq 0, \end{array}$$

where

$$\hat{c}_A := \hat{c}_A^k := X_k \tilde{c}_A, \quad \hat{g}_A := \hat{g}_A^k := X_k \tilde{g}_A, \quad \hat{A} := \hat{A}^k := \tilde{A} X_k. \tag{3.5}$$

To obtain the problem ($\hat{\text{P}}$), we first eliminated the unrestricted variables in (P) to give the problem ($\tilde{\text{P}}$), then scaled the remaining variables. Exactly the same problem ($\hat{\text{P}}$) would have resulted if we had first scaled (P) (giving the problem ($\bar{\text{P}}$)) and then eliminated the unrestricted variables.

All variables in ($\hat{\text{P}}$) are restricted to be nonnegative so the direction $\hat{d}_{\text{ELIM}_A}$ for the variables $\hat{x}_A$ is as defined in, for example, Mitchell and Todd [12]. Let z be our current lower bound on the optimal value of ($\hat{\text{P}}$) (and hence also on that of (P)). Then $\hat{d}_{\text{ELIM}_A}$ is given by

$$\hat{d}_{\text{ELIM}_A} = -P_{\hat{A}}(\hat{c}_A - z\hat{g}_A). \tag{3.6}$$

(We ignore the projection orthogonal to e, since Gonzaga [9] has shown that the final directions obtained are equivalent, whether or not this additional projection is performed.) After moving in the direction $\hat{d}_{\text{ELIM}_A}$, it is necessary to radially project the point obtained in order to satisfy the constraint

$$\hat{g}_A^T \hat{x}_A = 1. \tag{3.7}$$

Then we can find x_F by substituting for $\hat{x}_A$ (scaled by X_k) in (3.4). However, x_F is not necessary in order to solve the problem ($\tilde{\text{P}}$), and so we only need to solve for x_F using (3.4) after finding the optimal x_A.

Observe that instead of finding the new values of x_A by radially projecting to satisfy the constraint (3.7) and then using (3.4) to find x_F, it is possible to find a direction d_{ELIM_F} in the unrestricted variables. By (3.4), d_{ELIM_F} is given by

$$d_{\text{ELIM}_F} = -R^{-1} Q_1^T \hat{A} \hat{d}_{\text{ELIM}_A}. \tag{3.8}$$

Moving in the direction $\hat{d}_{\text{ELIM}} := (\hat{d}_{\text{ELIM}_A}, d_{\text{ELIM}_F})$ and then radially projecting to satisfy the appropriate normalizing constraint is exactly equivalent to the first approach.

Because there are no unrestricted variables in $(\hat{\mathrm{P}})$, we can update the lower bound z in the standard way [14] by solving the problem

$$(\hat{\mathrm{D}}') \qquad \begin{array}{ll} \max\limits_{z} & z, \\ \text{subject to} & \hat{A}^T y_{\mathrm{ELIM}}(z) + z\hat{g}_A \leqq \hat{c}_A, \end{array}$$

where

$$(3.9) \qquad y_{\mathrm{ELIM}}(z) = [\hat{A}\hat{A}^T]^{-1}\hat{A}(\hat{c}_A - z\hat{g}_A).$$

Let $\hat{z}'$ be the optimal value of $(\hat{\mathrm{D}}')$. If $\hat{z}' > z_k$, set $z_{k+1} \leftarrow \hat{z}'$; otherwise set $z_{k+1} \leftarrow z_k$.

4. Comparing the directions. In this section we show that the directions defined in the previous two sections are equivalent, and we use the results of Gonzaga [9] to show polynomial convergence of the algorithm when using either of these directions. We also discuss the benefits and disadvantages of using this approach as opposed to splitting the unrestricted variables.

THEOREM 1. *The solutions $\bar{z}'$ and $\hat{z}'$ to $(\bar{\mathrm{D}}')$ and $(\hat{\mathrm{D}}')$, respectively, are the same. For a given point (x_A^k, x_F^k) and lower bound z, the direction $\bar{d}_{\mathrm{CLS}_A}$ defined in § 2 (equation (2.9)) and the direction $\hat{d}_{\mathrm{ELIM}_A}$ defined in § 3 (equation (3.6)) are the same.*

Proof. Note that $\check{c}_A = \hat{c}_A$, $\check{g}_A = \hat{g}_A$, and $\check{A}_2 = \hat{A}$, where these quantities are defined in (2.10), (3.5), (2.11), (3.5), (2.4), and (3.5), respectively. Therefore for each z, $\check{y}_2(z)$ (defined in (2.7)) and $y_{\mathrm{ELIM}}(z)$ (defined in (3.9)) are the same. Hence $\bar{z}' = \hat{z}'$.

The equivalence of the directions follows directly from their definitions. □

It then follows from (2.12) and (3.8) that $d_{\mathrm{CLS}_F} = d_{\mathrm{ELIM}_F}$, and therefore the two procedures will generate the same sequence of iterates provided they start from the same point and use the same stepsizes. The results of Gonzaga [9] applied to $(\tilde{\mathrm{P}})$ then imply the following property of our algorithm.

THEOREM 2. *For suitable stepsizes α both of these algorithms take $O(nq)$ steps to converge.* □

Our algorithm requires the solution of a linear program of size m by n. In order to transform our original problem into this linear program, it was necessary to find a basis W for the null space of a matrix of size m by p, to premultiply our original constraint matrix A by the transpose of this basis, and also to perform several matrix-vector multiplications. These steps only have to be done once.

If instead we replaced each unrestricted variable by two nonnegative variables we would not have this extra cost at the beginning, but we would have to solve a linear program of size m by $(n + 2p)$. This linear program has an unbounded set of optimal solutions, so there is no guarantee that Karmarkar's algorithm would solve it in a polynomial number of iterations. One way to guarantee polynomial convergence is to place bounds on the split variables. However, this is not an elegant way to solve problems with unrestricted variables. (This contrasts with the simplex algorithm which includes unrestricted variables in the basis and never allows them to leave—only a minor alteration to the usual simplex algorithm.)

The main work at each iteration in our algorithm is the solution of a system of equations $W^T\bar{A}\bar{A}^TWu = r$, a system of equations of size $m - p$. The main work at each iteration in the alternative algorithm described in the previous paragraph is the solution of a system of equations $(\bar{A}\bar{A}^T + 2FF^T)y = v$, a system of equations of size m. In the dense case we would expect our algorithm to have a lower operation count per iteration than the alternative, regardless of the size of p. In the sparse case the balance is more delicate and depends on the representation and/or sparsity of the basis matrix W. For large p we would still expect our algorithm to outperform the alternative; for small p the comparison would have to be done on a case-by-case basis.

REFERENCES

[1] K. ANSTREICHER, *A monotonic projection algorithm for fractional linear programming*, Algorithmica, 1 (1986), pp. 483–498.

[2] A. BJÖRCK, *Least squares methods*, in Handbook of Numerical Analysis, P. G. Ciarlet and J. L. Lions, eds., Elsevier/North-Holland, Amsterdam, 1988.

[3] T. COLEMAN, *Large Sparse Numerical Optimization*, Springer-Verlag, Heidelberg, 1984.

[4] I. I. DIKIN, *Iterative solution of problems of linear and quadratic programming*, Dokl. Akad. Nauk SSSR, 174 (1967), pp. 747–748. (In Russian.) Soviet Math. Dokl., 8 (1967), pp. 674–675. (In English.)

[5] D. GAY, *A variant of Karmarkar's linear programming algorithm for problems in standard form*, Math. Programming, 37 (1987), pp. 81–90.

[6] P. GILL, W. MURRAY, M. SAUNDERS, J. TOMLIN, AND M. WRIGHT, *On projected Newton barrier methods for linear programming and an equivalence to Karmarkar's projective method*, Math. Programming, 36 (1986), pp. 183–209.

[7] P. GILL, W. MURRAY, AND M. WRIGHT, *Practical Optimization*, Academic Press, Orlando, FL, 1981.

[8] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.

[9] C. GONZAGA, *A conical projection algorithm for linear programming*, Math. Programming, 43 (1989), to appear.

[10] M. T. HEATH, *Numerical methods for large sparse linear least squares problems*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 497–513.

[11] N. KARMARKAR, *A new polynomial algorithm for linear programming*, Combinatorica, 4 (1984), pp. 373–395.

[12] J. E. MITCHELL AND M. TODD, *On the relationship between the search directions in the affine and projective variants of Karmarkar's linear programming algorithm*, Technical Report 725, School of Operations Research, Cornell University, Ithaca, NY, 1986.

[13] A. STEGER, *An extension of Karmarkar's algorithm for bounded linear programming problems*, M.Sc. Thesis, Dept. of Applied Mathematics and Statistics, State University of New York, Stony Brook, NY, August 1985.

[14] M. TODD AND B. BURRELL, *An extension of Karmarkar's algorithm for linear programming using dual variables*, Algorithmica, 1 (1986), pp. 409–424.

[15] R. VANDERBEI, *Affine scaling for problems with free variables*, Math. Programming, 43 (1989), to appear.

[16] R. VANDERBEI, M. MEKETON, AND B. FREEDMAN, *A modification of Karmarkar's linear programming algorithm*, Algorithmica, 1 (1986), pp. 395–407.

[17] Y. YE AND M. KOJIMA, *Recovering optimal dual solutions in Karmarkar's polynomial algorithm for linear programming*, Math. Programming, 39 (1987), pp. 305–317.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 1, pp. 39–64, January 1989

004

A MATRIX DECOMPOSITION METHOD FOR ORTHOTROPIC ELASTICITY PROBLEMS*

HSIN-CHU CHEN† AND AHMED H. SAMEH†

Abstract. The construction of an efficient numerical scheme for three-dimensional elasticity problems depends not only on understanding the nature of the physical problem involved, but also on exploiting special properties associated with its discretized system and incorporating these properties into the numerical algorithm. In this paper an efficient and parallelizable decomposition method is presented, referred to as the SAS domain decomposition method, for orthotropic elasticity problems with symmetrical domain and boundary conditions. Mathematically, this approach exploits important properties possessed by the special class of matrices A that satisfy the relation $A = PAP$, where P is some symmetrical signed permutation matrix. These matrices can be decomposed, via orthogonal transformations, into disjoint submatrices. Physically, the method takes advantage of the symmetry of a given problem and decomposes the whole domain of the original problem into independent subdomains. This method has potential for reducing the bandwidth of the stiffness (mass) matrix and lends itself to parallelism on three levels. Therefore, it is useful for sequential, vector, and multiprocessor computers.

Key words. SAS domain decomposition method, SAS property, SAS ordering, finite element method, reflection matrices, reflexive matrices, stiffness matrices, mass matrices, orthogonal similarity transformations, parallelism, speedup, orthotropic elasticity problems, eigenvalue problems

AMS(MOS) subject classifications. 15A04, 65F05, 65F15, 65N30

1. Introduction. The construction of an efficient numerical scheme depends on understanding the nature of the physical problem involved, exploiting special properties associated with its discretized system, and incorporating these properties into the numerical algorithm. The fast Fourier decomposition of the difference operators of the Poisson [BuGo70], [SaCh76] and the biharmonic equations [SaCh76], [Bjor83] is a typical example. The direct application of these fast solvers, however, is limited to separable problems [Schu77]. The separability of a physical problem depends not only on the differential equations, but also on the geometry of the boundary and on the form of the boundary conditions [Wein65]. The last two conditions do not often hold for problems in practice.

The SAS domain decomposition method (where the term SAS stands for "symmetrical and antisymmetrical") proposed in [ChSa87] and [Chen88] is a special decomposition method that decomposes certain classes of physical problems into independent subproblems by taking advantage of symmetry in these problems. This decomposition method has its origin in the idea of traditional symmetrical and antisymmetrical approaches [BlKa66], [Szil74], [WeJo87] and in generalized coordinate transformations [Rubi66]. A similar algorithm has been proposed independently in [NoPe87]. In our SAS decomposition method, we take advantage of the symmetry of the physical problems by exploiting important properties possessed by a special class of matrices A, $A \in \mathscr{C}^{n \times n}$, that satisfy the relation $A = PAP$ where P is some symmetrical signed permutation matrix. Unlike the fast Fourier decomposition method, the SAS approach directly applied to physical problems is constrained only by the conditions of the symmetry

* Received by the editors April 25, 1988; accepted for publication (in revised form) June 6, 1988. This work was partly supported by National Science Foundation grant NSF-MIP-8410110, U.S. Department of Energy grant DOE-DE-FG02-85ER25001, U.S. Air Force Office of Scientific Research grant AFOSR-85-0211, and an IBM donation.

† Center for Supercomputing Research and Development, University of Illinois at Urbana–Champaign, Urbana, Illinois 61801.

of domain, boundary conditions, and material properties. Therefore, this approach has much wider applications.

In this paper we present the SAS domain decomposition method for three-dimensional orthotropic elasticity problems with symmetrical domain and boundary conditions. In § 2, we introduce two special classes of vectors and the matrices mentioned above together with their important properties. In § 3, we present the SAS approach for decomposing algebraic linear systems and generalized eigenvalue problems whose coefficient matrices satisfy the desirable relation $A = PAP$. In § 4, the impact of such an approach on developing parallel algorithms for a variety of multiprocessors is illustrated and some possible parallel implementation strategies on existing and future supercomputers are provided. In § 5, we show how to decompose the element stiffness (mass) matrix of a rectangular hexahedral element [Melo63] into eight submatrices via orthogonal similarity transformations. These orthogonal transformations can be extended to decompose the system (mass) matrix if certain symmetry conditions exist. In § 6, numerical experiments on two isotropic prismatic bars are presented to demonstrate the applicability and usefulness of this domain decomposition method.

2. Special classes of vectors and matrices. Before presenting the SAS approach, we would like to introduce some basic definitions and fundamental properties regarding certain classes of vectors and matrices (see also [ChSa87], [Chen88]).

DEFINITION 2.1. Signed permutation and reflection matrices. A signed permutation matrix is a permutation matrix with its nonzero elements being either 1 or -1. A reflection matrix is a symmetric signed permutation matrix.

DEFINITION 2.2. Symmetrical and antisymmetrical vectors. Let P be some reflection matrix of order n. A vector $x \in \mathscr{C}^n$ is said to be symmetrical (or antisymmetrical) with respect to P if $x = Px$ (or if $x = -Px$).

Symmetrical and antisymmetrical subspaces. Let P be a reflection matrix. A vector subspace $S \subset \mathscr{C}^n$ is said to be symmetrical (or antisymmetrical) with respect to P if $x = Px$ (or if $x = -Px$) for any $x \in S$. Figure 2.1 shows geometrically the nonzero symmetrical subspace $\mathscr{R}^2_s(P)$ and antisymmetrical subspace $\mathscr{R}^2_a(P)$ of the $\mathscr{R}^2$ space with respect to some P.

DEFINITION 2.3. Reflexive matrices and subspaces. Let P be some reflection matrix of order n. A matrix $A \in \mathscr{C}^{n \times n}$ is said to be reflexive with respect to P if $A = PAP$. A subspace $S \subset \mathscr{C}^{n \times n}$ is said to be reflexive with respect to P if $A = PAP$ for any $A \in S$.

The SAS properties. A matrix $A \in \mathscr{C}^{n \times n}$ is said to possess the SAS property with respect to a reflection matrix P if A is reflexive with respect to P.

When a linear differential operator contains no odd derivative terms with domain and boundary condition symmetry, the corresponding matrix, say A, derived either from finite difference [Smit78], boundary element [LiLi83], or finite element discretization [Zien77], can often be arranged in such a way that A possesses the SAS property, namely,

$$A = PAP. \tag{2.1}$$

Let J_r, E_s, and F_s be symmetric matrices of order r, s, and s, respectively, and be defined by

$$J_r \equiv \begin{bmatrix} & 0 & 1 \\ & \cdot^{\cdot^{\cdot}} & \\ 1 & 0 & \end{bmatrix}, \qquad E_s = E_s^T \equiv \begin{bmatrix} & 0 & \pm 1 \\ & \cdot^{\cdot^{\cdot}} & \\ \pm 1 & 0 & \end{bmatrix}, \qquad F_s \equiv \begin{bmatrix} \pm 1 & & 0 \\ & \ddots & \\ 0 & & \pm 1 \end{bmatrix}. \tag{2.2}$$

Three of the most desirable forms of P are given by

$$P = J_r \otimes E_s, \quad P = J_r \otimes F_s, \quad P = I_r \otimes E_s \tag{2.3}$$

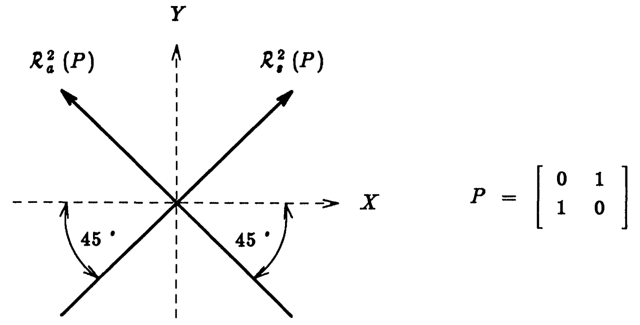


FIG. 2.1. *The nonzero symmetrical and antisymmetrical subspaces of $\mathscr{R}^2$ with respect to P.*

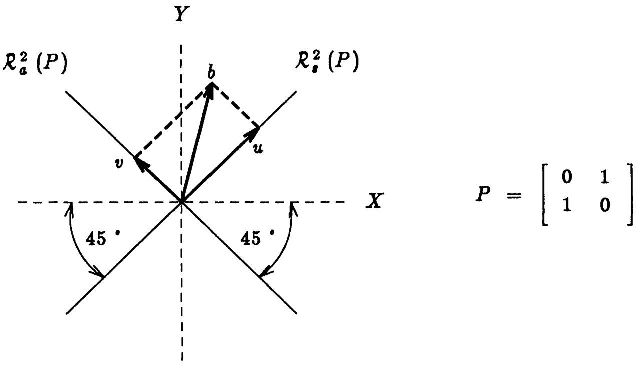


FIG. 2.2. *The decomposition of a vector b into its symmetrical part u and antisymmetrical part v.*

where $\otimes$ denotes the Kronecker product (or tensor product), $r \times s = n$, and I_r is the identity matrix of order r. Other forms of the matrix P are possible. It is worth noting that P is involutory.

Let P be some reflection matrix of order n and $\mathscr{C}_s^n(P)$ and $\mathscr{C}_a^n(P)$ be two subsets of the vector space $\mathscr{C}^n$ defined by

$$\mathscr{C}_s^n(P) \equiv \{x \mid x \in \mathscr{C}^n \text{ and } x = Px\}, \tag{2.4a}$$

$$\mathscr{C}_a^n(P) \equiv \{x \mid x \in \mathscr{C}^n \text{ and } x = -Px\}. \tag{2.4b}$$

THEOREM 2.4. *Given a reflection matrix P of order n, any vector $b \in \mathscr{C}^n$ can be decomposed into two parts, u and v, such that*

$$u + v = b \tag{2.5a}$$

where

$$u \in \mathscr{C}_s^n(P) \quad \textit{and} \quad v \in \mathscr{C}_a^n(P). \tag{2.5b}$$

The proof is readily established by taking $u = \frac{1}{2}(b + Pb)$ and $v = \frac{1}{2}(b - Pb)$. The geometrical decomposition of a vector b in $\mathscr{R}^2$ into its symmetrical and antisymmetrical parts with respect to some P is shown in Fig. 2.2.

THEOREM 2.5. *$\mathscr{C}_s^n(P)$ and $\mathscr{C}_a^n(P)$ are, respectively, symmetrical and antisymmetrical subspaces of $\mathscr{C}^n$ with respect to P over the field $\mathscr{C}$. Furthermore, $\mathscr{C}_s^n(P)$ and $\mathscr{C}_a^n(P)$ are mutually orthogonal.*

Proof. (1) $\mathscr{C}_s^n(P)$ and $\mathscr{C}_a^n(P)$ are subspaces. From Theorem 2.4, it is clear that $\mathscr{C}_s^n(P)$ is a nonempty subset of $\mathscr{C}^n$. Now let x and y be two arbitrary elements in $\mathscr{C}_s^n(P)$ and $\alpha \in \mathscr{C}$. The vector $(\alpha x + y)$ remains in $\mathscr{C}_s^n(P)$ since

$$(\alpha x+y)=P(\alpha x+y). \tag{2.6}$$

Therefore, $\mathscr{C}_s^n(P)$ is a subspace of $\mathscr{C}^n$ over the field $\mathscr{C}$. Similarly, $\mathscr{C}_a^n(P)$ is also a subspace of $\mathscr{C}^n$ over the field $\mathscr{C}$.

(2) $\mathscr{C}_s^n(P)$ and $\mathscr{C}_a^n(P)$ are orthogonal. For any $x \in \mathscr{C}_s^n(P)$ and any $y \in \mathscr{C}_a^n(P)$, we have

$$(x,y)=(Px,-Py)=-(x,y)=0 \tag{2.7}$$

where (., .) denotes an inner product. Hence $\mathscr{C}_s^n(P)$ and $\mathscr{C}_a^n(P)$ are mutually orthogonal.

(3) Since $\mathscr{C}_s^n(P)$ and $\mathscr{C}_a^n(P)$ are subspaces, we conclude from (2.4) and Definition 2.2 that $\mathscr{C}_s^n(P)$ and $\mathscr{C}_a^n(P)$ are, respectively, symmetrical and antisymmetrical subspaces of $\mathscr{C}^n$ with respect to P over the field $\mathscr{C}$. □

In the following, we present two more useful theorems. Theorem 2.6 indicates that the inverse of a reflexive matrix is also reflexive with respect to the same reflection matrix P, and therefore the solution of such a linear system will lie in the symmetrical (antisymmetrical) subspace if the right-hand side vector is symmetrical (antisymmetrical). Theorem 2.7 states that the addition and multiplication of two matrices that are reflexive with respect to the same reflection matrix P do not change the special SAS property.

THEOREM 2.6. *Given a linear system $Ax = f$, $A \in \mathscr{C}^{n\times n}$, and f, $x \in \mathscr{C}^n$, if A is nonsingular and reflexive with respect to some reflection matrix P, then*

$$A^{-1}=PA^{-1}P, \tag{2.8}$$

$$x\in \mathscr{C}_s^n(P) \quad \textit{iff}\; f\in \mathscr{C}_s^n(P), \tag{2.9a}$$

and

$$x\in \mathscr{C}_a^n(P) \quad \textit{iff}\; f\in \mathscr{C}_a^n(P). \tag{2.9b}$$

THEOREM 2.7. *Given two matrices A and B where A, $B \in \mathscr{C}^{n\times n}$, if A and B are both reflexive with respect to the same reflection matrix P, then*

$$(\alpha A\pm\beta B)=P(\alpha A\pm\beta B)P, \tag{2.10}$$

$$(\alpha A)(\beta B)=P(\alpha A)(\beta B)P \tag{2.11}$$

where α, $\beta \in \mathscr{C}$.

3. The SAS decomposition method.

3.1. Linear systems.

Consider the linear system

$$Ax=b \tag{3.1}$$

where A is reflexive with respect to some reflection matrix P and is nonsingular. The main idea of the SAS approach, based on the principle of superposition, is to decompose the right-hand side vector b into two parts, say u and v, such that (2.5a) and (2.5b) hold. This decomposition immediately enables us to handle (3.1) by solving two separate linear systems, say

$$Ay=u \quad \text{and} \quad Az=v \tag{3.2}$$

where $y + z = x$. At this point, it is still not clear that solving the two systems in (3.2) will lead to any advantage. The next important step is to decompose the matrix A into two submatrices, say A_1 of order k_1 and A_2 of order k_2, $k_1 + k_2 = n$, such that instead of solving (3.1) we can solve the following two smaller independent systems:

$$A_1\tilde{y} = \tilde{u} \quad \text{and} \quad A_2\tilde{z} = \tilde{v} \tag{3.3}$$

where $\tilde{u}$ and $\tilde{v}$ can be extracted from u and v and actually depend on the form of the matrix P. Theorem 2.6 yields information on how to decompose the matrix A. For example, if the matrix A satisfies (2.1) with P of the form

$$P = J_2 \otimes E_{n/2}, \qquad n \text{ even} \tag{3.4}$$

and the linear system in (3.1) is partitioned as

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \tag{3.5}$$

then by taking

$$\tilde{u} = (b_1 + E_{n/2}b_2), \qquad \tilde{v} = (b_2 - E_{n/2}b_1) \tag{3.6a}$$

and

$$\tilde{y} = (x_1 + E_{n/2}x_2), \qquad \tilde{z} = (x_2 - E_{n/2}x_1), \tag{3.6b}$$

we have

$$A_1 = A_{11} + A_{12}E_{n/2}, \qquad A_2 = A_{22} - A_{21}E_{n/2}. \tag{3.7}$$

Thus,

$$x_1 = \tfrac{1}{2}(\tilde{y} - E_{n/2}\tilde{z}) \quad \text{and} \quad x_2 = \tfrac{1}{2}(\tilde{z} + E_{n/2}\tilde{y}). \tag{3.8}$$

Note that the useful information about the relations between the components of the solution vector would have been wasted if (3.1) had been solved directly without exploiting the special property possessed by the matrix A.

In many cases, the decomposed submatrices A_1 and A_2 still have the desirable SAS property. The decomposition can then be further carried out to yield four independent subsystems with each submatrix of order approximately equal to one quarter of the order of A. This decomposition procedure can be applied recursively to those subsystems until no disjoint submatrices have the SAS property. For instance, the discrete biharmonic operator on a rectangular domain with boundary conditions symmetrical about its two centered axes can be decomposed into four independent subsystems [ChSa87], [Chen88]. The resulting subsystems can then be solved independently using either four processors as in the Cray X-MP/48 or four clusters as in the Cedar multiprocessor [DaKu86]. The final solution of the original system can be easily retrieved from the solution of the decomposed subsystems. The implementation of this approach on the Cedar machine takes full advantage of its three levels of parallelism: problem decomposition among clusters, parallelism within a cluster, and vectorization within each processor (compared to only two levels for the Cray X-MP).

3.2. Eigenvalue problems. When the matrix A possesses the SAS property, it can be shown (via similarity transformations) that the proposed decomposition approach can be used for solving eigenvalue problems much more efficiently. For instance, if P takes the form (3.4) and A is partitioned as in (3.5), then A is similar to $A_1 \oplus A_2$ through an orthogonal transformation X^TAX, where A_1 and A_2 are the same as in (3.7) and

$$X = \frac{1}{\sqrt{2}} \begin{bmatrix} I & -E_{n/2} \\ E_{n/2} & I \end{bmatrix}. \tag{3.9}$$

The high efficiency realized in handling the problem is due to three main features of the above technique. First, all the eigenvalues of the original matrix A can be obtained from the decomposed submatrices, which are of lower order. The amount of work is, therefore, greatly reduced even for sequential computations. Second, the extraction of the eigenvalues of the different submatrices can be performed independently, which implies that high parallelism can be achieved in addition to the computational savings. Third, the eigenvalues in each submatrix are in general better separated, which indicates faster convergence for schemes such as QR (e.g., see [Wilk65]).

To see how much effort can be saved, we consider the QR iterations for a real-valued full matrix (or order N). In using the QR iterations to solve for the eigenvalues, we usually reduce the original matrix to a Hessenberg form (a tridiagonal matrix for the symmetric problem). On a sequential machine, the reduction step takes about cN^3 flops [GoVa83] for some constant c. Suppose the matrix A satisfies $PAP = A$ and that it can be decomposed into four submatrices each of order $N/4$; then the amount of work required in the reduction step by using the proposed decomposition method is reduced to $cN^3/16$, i.e., one sixteenth of the original work. In addition, because of the fully independent nature of the subproblems we can further reduce the computing time by a factor between three and four by using, for example, the four central processing units (CPUs) of a Cray X-MP/48.

Depending on the form of the signed reflection matrix P, several similarity transformations can be derived for this special class of matrices $A = PAP$. Here we present two important and computationally attractive similarity transformations.

THEOREM 3.1. *Let* $A \in \mathscr{C}^{n \times n}$, *n even, be partitioned as*

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

with each submatrix being of order $n/2$. *Let P be of the form*

$$\begin{bmatrix} 0 & P_1^T \\ P_1 & 0 \end{bmatrix}$$

where P_1 *is some signed permutation matrix of order* $n/2$. *If A is reflexive with respect to P, then there exists an orthogonal matrix X such that A is similar to* $(A_{11} + A_{12}P_1) \oplus (A_{22} - A_{21}P_1^T)$.

Proof. Consider the orthogonal matrix

$$X = \frac{1}{\sqrt{2}} \begin{bmatrix} I & -P_1^T \\ P_1 & I \end{bmatrix}. \tag{3.10}$$

Then we have

(3.11)

$$X^{-1}AX = \frac{1}{2} \begin{bmatrix} I & P_1^T \\ -P_1 & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} I & -P_1^T \\ P_1 & I \end{bmatrix}$$

$$= \frac{1}{2} \begin{bmatrix} (A_{11}+A_{12}P_1)+(P_1^TA_{21}+P_1^TA_{22}P_1) & (A_{12}-P_1^TA_{21}P_1^T)+(P_1^TA_{22}-A_{11}P_1^T) \\ (A_{21}-P_1A_{12}P_1)+(A_{22}P_1-P_1A_{11}) & (A_{22}-A_{21}P_1^T)+(P_1A_{11}P_1^T-P_1A_{12}) \end{bmatrix}$$

$$= \begin{bmatrix} A_{11}+A_{12}P_1 & 0 \\ 0 & A_{22}-A_{22}P_1^T \end{bmatrix}.$$

The last expression is a direct consequence of the assumption that $A = PAP$.

THEOREM 3.2. *Let $A \in \mathscr{C}^{n\times n}$ be partitioned as (A_{ij}), $i, j = 1, 2$, and 3 with A_{11} and A_{33} of order r and A_{22} of order s, where $2r + s = n$. If $A = PAP$ where P is of the form*

$$P=\begin{bmatrix} 0 & 0 & P_1^T \\ 0 & I_s & 0 \\ P_1 & 0 & 0 \end{bmatrix}$$

in which P_1 is some signed permutation matrix of order r, then there exists an orthogonal matrix X such that A is similar to

$$\begin{bmatrix} A_{11}+A_{13}P_1 & \sqrt{2}A_{12} & 0 \\ \sqrt{2}A_{21} & A_{22} & 0 \\ 0 & 0 & A_{33}-A_{31}P_1^T \end{bmatrix}. \tag{3.12}$$

Proof. Consider the orthogonal matrix

$$X=\frac{1}{\sqrt{2}}\begin{bmatrix} I & 0 & -P_1^T \\ 0 & \sqrt{2}I_s & 0 \\ P_1 & 0 & I \end{bmatrix}. \tag{3.13}$$

Then the application of the similarity transformation $X^{-1}AX$ yields

(3.14)

$$X^{-1}AX=\frac{1}{2}\begin{bmatrix} I & 0 & P_1^T \\ 0 & \sqrt{2}I_s & 0 \\ -P_1 & 0 & I \end{bmatrix}\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}\begin{bmatrix} I & 0 & -P_1^T \\ 0 & \sqrt{2}I_s & 0 \\ P_1 & 0 & I \end{bmatrix}$$

$$=\frac{1}{2}\begin{bmatrix} (A_{11}+A_{13}P_1)+(P_1^TA_{31}+P_1^TA_{33}P_1) & \sqrt{2}(A_{12}+P_1^TA_{32}) & (A_{13}-A_{11}P_1^T)+(P_1^TA_{33}-P_1^TA_{31}P_1^T) \\ \sqrt{2}(A_{21}+A_{23}P_1) & \sqrt{2}A_{22} & \sqrt{2}(A_{23}-A_{21}P_1^T) \\ (A_{31}-P_1A_{13}P_1)+(A_{33}P_1-P_1A_{11}) & \sqrt{2}(A_{32}-P_1A_{12}) & (A_{33}-A_{31}P_1^T)-(P_1A_{13}-P_1A_{11}P_1^T) \end{bmatrix}$$

$$=\begin{bmatrix} A_{11}+A_{13}P_1 & \sqrt{2}A_{12} & 0 \\ \sqrt{2}A_{21} & A_{22} & 0 \\ 0 & 0 & A_{33}-A_{31}P_1^T \end{bmatrix}.$$

As in Theorem 3.1, the final equality expression in (3.14) is obtained by employing the assumption that $A = PAP$. It should be noted that if A is Hermitian, then both submatrices in (3.14) are still Hermitian. The same argument holds for the two submatrices in (3.11).

Note that the application of the above decomposition method can be extended to the generalized eigenvalue problem $Ax = \lambda Bx$ without difficulty if B also satisfies the same requirement, namely, $A = PAP$ and $B = PBP$.

4. Parallel implementations of the SAS approach. One of the important features of the SAS approach, whenever applicable, is the simple procedure involved in decomposing the problem into smaller independent subproblems. This approach is, therefore, very attractive for sequential, vector, and multiprocessor computers. Efficient implementations of the SAS scheme on parallel computers depends not only on the architecture of a given machine but also on how the compiler is designed to handle program statements that can be executed in parallel. If the matrix A is SAS-decomposable, then Table 4.1 shows the potential of parallelism inherent in solving linear systems $Ax = b$ via the SAS approach where *nsub* is the number of subdomains or submatrices decomposed by the

TABLE 4.1

Non-SAS Algorithm	SAS Algorithm
Form A Form b	Form $A_1, A_2, \cdots, A_{nsub}$ Form b
Solve $Ax = b$	Decompose b into $b_1, b_2, \cdots, b_{nsub}$ Solve $A_i x_i = b_i$ $(i = 1, \cdots, nsub)$ Retrieve x from $x_1, x_2, \cdots, x_{nsub}$

TABLE 4.2

Implementation	Quasi-Fortran statements	Execution mode
1	CVD\$L CNCALL DO $i = 1, nsub$ Solve $A_i x_i = b_i$ END DO	concurrent outer sequential/vector inner
2	CVD\$L NOCNCALL DO $i = 1, nsub$ Solve $A_i x_i = b_i$ END DO	sequential outer concurrent/vector inner
3	CVD\$L CLUSTERCALL DO $i = 1, nsub$ Solve $A_i x_i = b_i$ END DO	concurrent outer concurrent/vector inner

SAS approach. The first two implementations are currently available on the Alliant FX/8 parallel computer. The statements (CVD$L CNCALL) and (CVD$L NOCNCALL) are two Alliant optimization directives [Alli87] used to indicate whether the next loop is to be executed in concurrent mode. The third implementation in Table 4.2 is intended for the Cedar computer where all the A_i's can be formed simultaneously, one per cluster, and all the systems $A_i x_i = b_i$ are solved simultaneously, one per cluster.

In Table 4.1, the decomposition of the vector b into b_i and the retrieval of the solution x from x_i, $i = 1, \cdots, nsub$, involve only vector operations. The independence of the subsystems resulting from the SAS domain decomposition implies high-level parallelism. Hence, on machines such as the Alliant FX/8 or Cray X-MP/48 we can solve one subsystem per processor using a vectorized solver. On a machine like Cedar, however, we can solve each subsystem on each cluster using parallel solvers, thus taking full advantage of the three levels of parallelism of the Cedar architecture.

5. The application of the SAS approach to elasticity problems.

5.1. Orthotropic elasticity problems. To demonstrate the effectiveness of the SAS domain decomposition method, we consider the three-dimensional static analysis of orthotropic elasticity problems. The mathematical formulation for such an analysis [Lekh63] is briefly described below. Let Ω be the domain in R^3, Γ_1 the boundary surface where displacements are specified, and Γ_2 the boundary surface where tractions are known. Let the stress, displacement, body force, and surface traction vectors be denoted by σ, δ, $\mathbf{b}$, and $\mathbf{p}$, respectively, where $\sigma^T = [\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{13}, \sigma_{23}]$, $\delta^T = [\delta_1, \delta_2, \delta_3]$, $\mathbf{b}^T = [b_1, b_2, b_3]$, and $\mathbf{p}^T = [p_1, p_2, p_3]$. Here the superscript T stands for the transpose and the subscripts 1, 2, and 3 represent the three Cartesian directions x, y, and z, respectively.

The differential equations and boundary conditions for an orthotropic elasticity solid can be expressed in the following matrix form:

$$\mathscr{L}^T D \mathscr{L} \sigma + \mathbf{b} = \mathbf{0} \quad \text{in } \Omega \tag{5.1a}$$

subject to

$$\boldsymbol{\delta} = \boldsymbol{\delta}_b \qquad \text{on } \Gamma_1, \tag{5.1b}$$

$$\mathbf{p} \equiv C\sigma = \mathbf{p}_b \quad \text{on } \Gamma_2. \tag{5.1c}$$

Here $\mathscr{L}$ is the differential operator:

$$\mathscr{L}^T = \begin{bmatrix} \partial/\partial x_1 & 0 & 0 & \partial/\partial x_2 & \partial/\partial x_3 & 0 \\ 0 & \partial/\partial x_2 & 0 & \partial/\partial x_1 & 0 & \partial/\partial x_3 \\ 0 & 0 & \partial/\partial x_3 & 0 & \partial/\partial x_1 & \partial/\partial x_2 \end{bmatrix}; \tag{5.2}$$

C is the matrix of the direction cosines:

(5.3)

$$C = \begin{bmatrix} \cos(n,x_1) & 0 & 0 & \cos(n,x_2) & \cos(n,x_3) & 0 \\ 0 & \cos(n,x_2) & 0 & \cos(n,x_1) & 0 & \cos(n,x_3) \\ 0 & 0 & \cos(n,x_3) & 0 & \cos(n,x_1) & \cos(n,x_2) \end{bmatrix};$$

and D is the material property matrix:

$$D = \begin{bmatrix} d_{11} & d_{12} & d_{13} & & & \\ d_{12} & d_{22} & d_{23} & & & \\ d_{13} & d_{23} & d_{33} & & & \\ & & & d_{44} & & \\ & & & & d_{55} & \\ & & & & & d_{66} \end{bmatrix}. \tag{5.4a}$$

The symbols (n, x_i) in (5.3) denote the angles between x_i and the outward normal n to the surface Γ_2, and d_{ij} in (5.4a) represent the elastic constants for orthotropic material [Lekh63]. For isotropic material, the material property matrix is simplified to

(5.4b)

$$D = \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1-\nu & \nu & \nu & & & \\ \nu & 1-\nu & \nu & & & \\ \nu & \nu & 1-\nu & & & \\ & & & (1-2\nu)/2 & & \\ & & & & (1-2\nu)/2 & \\ & & & & & (1-2\nu)/2 \end{bmatrix}$$

where E and ν represent the material modulus and the Poisson's ratio, respectively. Solving (5.1) analytically is not always possible. Numerical approximation techniques are therefore necessary. In this paper, we employ the finite-element method using the basic 8-node rectangular hexahedral elements [Melo63], [Dawe84] for our numerical approximation. Figure 5.1 shows the node numbering and positive directions of the degrees of freedom of the element. At a given node the unknown in the direction of x_1 is always followed immediately by the unknowns in the direction of axes x_2 and x_3, respectively. We denote the dimensions of the element along x_1, x_2, and x_3 by $2l_1$, $2l_2$, and $2l_3$, respectively. The element stiffness matrix and mass matrix for an orthotropic element as shown in Fig. 5.1 are given [Melo63], [Prze68], [Chen88] in Appendices A and B, respectively.

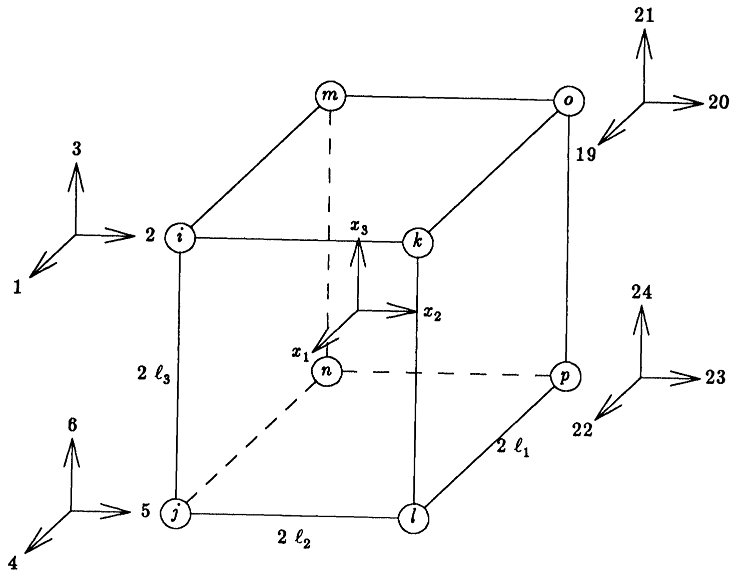


FIG. 5.1. *A basic rectangular hexahedral element.*

5.2. The SAS decomposition of the stiffness matrix. The SAS domain decomposition method is directly applicable to the system stiffness and mass matrices assembled from the elemental matrices when a three-dimensional orthotropic/isotropic elasticity problem with symmetrical domain and boundary conditions is symmetrically discretized by rectangular hexahedral elements. One way of showing this is to begin by proving that the element stiffness (mass) matrix possesses the SAS property with respect to some reflection matrix. In the following theorem we show how to recursively decompose the element stiffness matrix $K_{(e)}$ into eight submatrices.

THEOREM 5.1. *Let E_3, F_3, and G_3 be defined as*

$$(5.5)\qquad E_3=\begin{bmatrix}-1 & & \\ & 1 & \\ & & 1\end{bmatrix},\quad F_3=\begin{bmatrix}1 & & \\ & -1 & \\ & & 1\end{bmatrix},\quad G_3=\begin{bmatrix}1 & & \\ & 1 & \\ & & -1\end{bmatrix}$$

and I_k be the identity matrix of order k. If the material of the rectangular hexahedral element shown in Fig. 5.1 is either isotropic or orthotropic with its principal directions of orthotropy coinciding with the three coordinate axes, then the element stiffness matrix $K_{(e)}$ of this element, partitioned as (K_{ij}), $1 \leqq i, j \leqq 8$, with each submatrix being of order 3, is orthogonally similar to

$$(5.6)\qquad K_{sss}\oplus K_{ssa}\oplus K_{sas}\oplus K_{saa}\oplus K_{ass}\oplus K_{asa}\oplus K_{aas}\oplus K_{aaa}$$

where

$$K_{sss}=(A_s+C_sF_3)+(B_s+D_sF_3)G_3,$$

$$K_{ssa}=G_3(A_s+C_sF_3)G_3-G_3(B_s+D_sF_3),$$

$$K_{sas}=F_3(A_sF_3-C_s)+F_3(B_sF_3-D_s)G_3,$$

$$K_{saa}=G_3F_3(A_sF_3-C_s)G_3-G_3F_3(B_sF_3-D_s),$$

$$K_{ass}=(A_a+C_aF_3)+(B_a+D_aF_3)G_3,$$
$$K_{asa}=G_3(A_a+C_aF_3)G_3-G_3(B_a+D_aF_3),$$
$$K_{aas}=F_3(A_aF_3-C_a)+F_3(B_aF_3-D_a)G_3,$$
$$K_{aaa}=G_3F_3(A_aF_3-C_a)G_3-G_3F_3(B_aF_3-D_a), \tag{5.7}$$

with

$$\begin{aligned} A_s&=K_{11}+K_{15}E_3, & A_a&=E_3K_{11}E_3-E_3K_{15},\\ B_s&=K_{12}+K_{16}E_3, & B_a&=E_3K_{12}E_3-E_3K_{16},\\ C_s&=K_{13}+K_{17}E_3, & C_a&=E_3K_{13}E_3-E_3K_{17},\\ D_s&=K_{14}+K_{18}E_3, & D_a&=E_3K_{14}E_3-E_3K_{18}.\end{aligned}$$

Proof. From the explicit form of $K_{(e)}$ (see Appendix A), we observe the following three-level relations:

Level 1.

$$\begin{aligned}\begin{bmatrix} K_{15} & K_{16} & K_{17} & K_{18}\\ K_{25} & K_{26} & K_{27} & K_{28}\\ K_{35} & K_{36} & K_{37} & K_{38}\\ K_{45} & K_{46} & K_{47} & K_{48}\end{bmatrix} &= (I_4\otimes E_3)\begin{bmatrix} K_{51} & K_{52} & K_{53} & K_{54}\\ K_{61} & K_{62} & K_{63} & K_{64}\\ K_{71} & K_{72} & K_{73} & K_{74}\\ K_{81} & K_{82} & K_{83} & K_{84}\end{bmatrix}(I_4\otimes E_3),\\ \begin{bmatrix} K_{55} & K_{56} & K_{57} & K_{58}\\ K_{65} & K_{66} & K_{67} & K_{68}\\ K_{75} & K_{76} & K_{77} & K_{78}\\ K_{85} & K_{86} & K_{87} & K_{88}\end{bmatrix} &= (I_4\otimes E_3)\begin{bmatrix} K_{11} & K_{12} & K_{13} & K_{14}\\ K_{21} & K_{22} & K_{23} & K_{24}\\ K_{31} & K_{32} & K_{33} & K_{34}\\ K_{41} & K_{42} & K_{43} & K_{44}\end{bmatrix}(I_4\otimes E_3).\end{aligned} \tag{5.8}$$

Level 2.

$$\begin{aligned}\begin{bmatrix} K_{13} & K_{14}\\ K_{23} & K_{24}\end{bmatrix} &= (I_2\otimes F_3)\begin{bmatrix} K_{31} & K_{32}\\ K_{41} & K_{42}\end{bmatrix}(I_2\otimes F_3),\\ \begin{bmatrix} K_{33} & K_{34}\\ K_{43} & K_{44}\end{bmatrix} &= (I_2\otimes F_3)\begin{bmatrix} K_{11} & K_{12}\\ K_{21} & K_{22}\end{bmatrix}(I_2\otimes F_3),\\ \begin{bmatrix} K_{53} & K_{54}\\ K_{63} & K_{64}\end{bmatrix} &= (I_2\otimes F_3)\begin{bmatrix} K_{71} & K_{72}\\ K_{81} & K_{82}\end{bmatrix}(I_2\otimes F_3),\\ \begin{bmatrix} K_{73} & K_{74}\\ K_{83} & K_{84}\end{bmatrix} &= (I_2\otimes F_3)\begin{bmatrix} K_{51} & K_{52}\\ K_{61} & K_{62}\end{bmatrix}(I_2\otimes F_3).\end{aligned} \tag{5.9}$$

Level 3.

$$\begin{aligned} K_{12}&=G_3K_{21}G_3, & K_{52}&=G_3K_{61}G_3,\\ K_{22}&=G_3K_{11}G_3, & K_{62}&=G_3K_{51}G_3,\\ K_{32}&=G_3K_{41}G_3, & K_{72}&=G_3K_{81}G_3,\\ K_{42}&=G_3K_{31}G_3, & K_{82}&=G_3K_{71}G_3.\end{aligned} \tag{5.10}$$

Let P_1 and X_1, Q_1 and Y_1, and R_1 and Z_1 be defined as follows:

$$P_1=\begin{bmatrix} 0 & (I_4\otimes E_3) \\ (I_4\otimes E_3) & 0 \end{bmatrix} \quad\text{and}\quad X_1=\frac{1}{\sqrt{2}}\begin{bmatrix} I & -(I_4\otimes E_3) \\ (I_4\otimes E_3) & I \end{bmatrix},$$

$$Q_2=\begin{bmatrix} 0 & (I_2\otimes F_3) \\ (I_2\otimes F_3) & 0 \end{bmatrix} \quad\text{and}\quad Y_1=\frac{1}{\sqrt{2}}\begin{bmatrix} I & -(I_2\otimes F_3) \\ (I_2\otimes F_3) & I \end{bmatrix},$$

$$R_1=\begin{bmatrix} 0 & (I_1\otimes G_3) \\ (I_1\otimes G_3) & 0 \end{bmatrix} \quad\text{and}\quad Z_1=\frac{1}{\sqrt{2}}\begin{bmatrix} I & -(I_1\otimes G_3) \\ (I_1\otimes G_3) & I \end{bmatrix}.$$

From the relations of Level 1, it is seen that K is SAS-decomposable with respect to P_1. By applying the first orthogonal similarity transformation $X_1^TKX_1$, we have

$$X_1^TKX_1=K_s\oplus K_a \tag{5.11}$$

where

$$K_s=\begin{bmatrix} K_{11} & K_{12} & K_{13} & K_{14} \\ K_{21} & K_{22} & K_{23} & K_{24} \\ K_{31} & K_{32} & K_{33} & K_{34} \\ K_{41} & K_{42} & K_{43} & K_{44} \end{bmatrix}+\begin{bmatrix} K_{15} & K_{16} & K_{17} & K_{18} \\ K_{25} & K_{26} & K_{27} & K_{28} \\ K_{35} & K_{36} & K_{37} & K_{38} \\ K_{45} & K_{46} & K_{47} & K_{48} \end{bmatrix}\begin{bmatrix} E_3 & & & \\ & E_3 & & \\ & & E_3 & \\ & & & E_3 \end{bmatrix},$$

$$K_a=\begin{bmatrix} K_{55} & K_{56} & K_{57} & K_{58} \\ K_{65} & K_{66} & K_{67} & K_{68} \\ K_{75} & K_{76} & K_{77} & K_{78} \\ K_{85} & K_{86} & K_{87} & K_{88} \end{bmatrix}-\begin{bmatrix} K_{51} & K_{52} & K_{53} & K_{54} \\ K_{61} & K_{62} & K_{63} & K_{64} \\ K_{71} & K_{72} & K_{73} & K_{74} \\ K_{81} & K_{82} & K_{83} & K_{84} \end{bmatrix}\begin{bmatrix} E_3 & & & \\ & E_3 & & \\ & & E_3 & \\ & & & E_3 \end{bmatrix}. \tag{5.12}$$

From the relations of Levels 1 and 2 and the fact that E_3 and F_3 commute, it is not difficult to show that K_s and K_a are both SAS-decomposable with respect to Q_1. A second application of orthogonal similarity transformations yields

$$Y_1^TK_sY_1=K_{ss}\oplus K_{sa}, \qquad Y_1^TK_aY_1=K_{as}\oplus K_{aa} \tag{5.13}$$

where

(5.14)

$$K_{ss}=\begin{bmatrix} (K_{11}+K_{15}E_3) & (K_{12}+K_{16}E_3) \\ (K_{21}+K_{25}E_3) & (K_{22}+K_{26}E_3) \end{bmatrix}+\begin{bmatrix} (K_{13}+K_{17}E_3) & (K_{14}+K_{18}E_3) \\ (K_{23}+K_{27}E_3) & (K_{24}+K_{28}E_3) \end{bmatrix}\begin{bmatrix} F_3 & \\ & F_3 \end{bmatrix},$$

$$K_{sa}=\begin{bmatrix} (K_{33}+K_{37}E_3) & (K_{34}+K_{38}E_3) \\ (K_{43}+K_{47}E_3) & (K_{44}+K_{48}E_3) \end{bmatrix}-\begin{bmatrix} (K_{31}+K_{35}E_3) & (K_{32}+K_{36}E_3) \\ (K_{41}+K_{45}E_3) & (K_{42}+K_{46}E_3) \end{bmatrix}\begin{bmatrix} F_3 & \\ & F_3 \end{bmatrix},$$

$$K_{as}=\begin{bmatrix} (K_{55}-K_{51}E_3) & (K_{56}-K_{52}E_3) \\ (K_{65}-K_{61}E_3) & (K_{66}-K_{62}E_3) \end{bmatrix}+\begin{bmatrix} (K_{57}-K_{53}E_3) & (K_{58}-K_{54}E_3) \\ (K_{67}-K_{63}E_3) & (K_{68}-K_{64}E_3) \end{bmatrix}\begin{bmatrix} F_3 & \\ & F_3 \end{bmatrix},$$

$$K_{aa}=\begin{bmatrix} (K_{77}-K_{73}E_3) & (K_{78}-K_{74}E_3) \\ (K_{87}-K_{83}E_3) & (K_{88}-K_{84}E_3) \end{bmatrix}-\begin{bmatrix} (K_{75}-K_{71}E_3) & (K_{76}-K_{72}E_3) \\ (K_{85}-K_{81}E_3) & (K_{86}-K_{82}E_3) \end{bmatrix}\begin{bmatrix} F_3 & \\ & F_3 \end{bmatrix}.$$

Again from the relations of Levels 1, 2, and 3 and the fact that E_3, F_3, and G_3 commute, it can also be shown that K_{ss}, K_{sa}, K_{as}, and K_{aa} are all SAS-decomposable with respect to R_1. Applying the third orthogonal similarity transformation we obtain

$$\begin{aligned} &Z_1^TK_{ss}Z_1=K_{sss}\oplus K_{ssa}, && Z_1^TK_{sa}Z_1=K_{sas}\oplus K_{saa}, \\ &Z_1^TK_{as}Z_1=K_{ass}\oplus K_{asa}, && Z_1^TK_{aa}Z_1=K_{aas}\oplus K_{aaa}, \end{aligned} \tag{5.15}$$

where

(5.16)

$$
\begin{aligned}
K_{sss} &= ((K_{11}+K_{15}E_3)+(K_{13}+K_{17}E_3)F_3)+((K_{12}+K_{16}E_3)+(K_{14}+K_{18}E_3)F_3)G_3,\\
K_{ssa} &= ((K_{22}+K_{26}E_3)+(K_{24}+K_{28}E_3)F_3)-((K_{21}+K_{25}E_3)+(K_{23}+K_{27}E_3)F_3)G_3,\\
K_{sas} &= ((K_{33}+K_{37}E_3)-(K_{31}+K_{35}E_3)F_3)+((K_{34}+K_{38}E_3)-(K_{32}+K_{36}E_3)F_3)G_3,\\
K_{saa} &= ((K_{44}+K_{48}E_3)-(K_{42}+K_{46}E_3)F_3)-((K_{43}+K_{47}E_3)-(K_{41}+K_{45}E_3)F_3)G_3,\\
K_{ass} &= ((K_{55}-K_{51}E_3)+(K_{57}-K_{53}E_3)F_3)+((K_{56}-K_{52}E_3)+(K_{58}-K_{54}E_3)F_3)G_3,\\
K_{asa} &= ((K_{66}-K_{62}E_3)+(K_{68}-K_{64}E_3)F_3)-((K_{65}-K_{61}E_3)+(K_{67}-K_{63}E_3)F_3)G_3,\\
K_{aas} &= ((K_{77}-K_{73}E_3)-(K_{75}-K_{71}E_3)F_3)+((K_{78}-K_{74}E_3)-(K_{76}-K_{72}E_3)F_3)G_3,\\
K_{aaa} &= ((K_{88}-K_{84}E_3)-(K_{86}-K_{82}E_3)F_3)-((K_{87}-K_{83}E_3)-(K_{85}-K_{81}E_3)F_3)G_3.
\end{aligned}
$$

Using the three-level relations (5.8)–(5.10), we obtain (5.7) from (5.16). In summary, by combining these three levels of orthogonal similarity transformations (5.11), (5.13), and (5.15), we have the final expression

$$S^TKS = K_{sss} \oplus K_{ssa} \oplus K_{sas} \oplus K_{saa} \oplus K_{ass} \oplus K_{asa} \oplus K_{aas} \oplus K_{aaa}$$

where

$$S = XYZ$$

with

$$X = I_1 \otimes X_1, \quad Y = I_2 \otimes Y_1, \quad Z = I_4 \otimes Z_1.$$

Since X, Y, and Z are all orthogonal matrices, the above transformations are numerically stable. This decomposition can be represented graphically by a three-level binary tree as shown in Fig. 5.2. □

COROLLARY 5.2. *The element mass matrix* $M_{(e)}$ *of the rectangular hexahedral element* (*see Appendix* B) *shown in Fig.* 5.1 *is orthogonally similar to a matrix of the form* (5.6) *if the mass density* ρ *of the element is constant.*

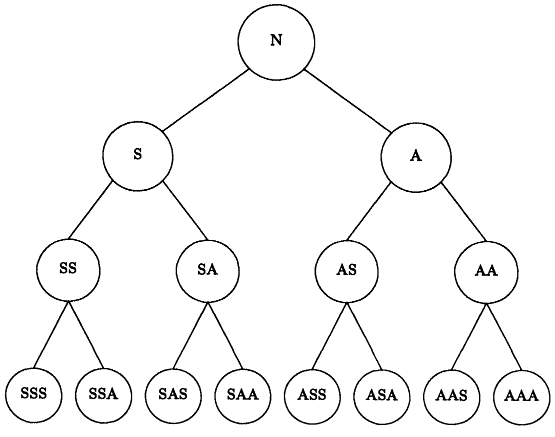


FIG. 5.2. *The binary tree representation of the* SAS *domain decomposition of three levels.*

Motivated by Theorem 5.1, we now present a more general theorem that allows us to recursively decompose a matrix A into eight submatrices using the same three levels of decompositions when certain relations exist. The proof of this theorem is analogous to Theorem 5.1 and therefore will be omitted.

THEOREM 5.3. *Let $A \in \mathbf{C}^{n\times n}$, $n = 8 \times s$, be partitioned as $[A_{ij}]$, $1 \leqq i, j \leqq 8$, with each submatrix being of order s. Let also E_s, F_s, and G_s be some reflection matrices each of order s and I_k the identity matrix of order k. Suppose that E_s, F_s, and G_s commute and the matrix A satisfies the same three-level relations* (5.8)–(5.10) *with E_3, F_3, and G_3 replaced by E_s, F_s, and G_s, respectively. Then A is orthogonally similar to*

$$A_{sss} \oplus A_{ssa} \oplus A_{sas} \oplus A_{saa} \oplus A_{ass} \oplus A_{asa} \oplus A_{aas} \oplus A_{aaa} \tag{5.17}$$

where

$$\begin{aligned}
A_{sss} &= (A_s + C_sF_s) + (B_s + D_sF_s)G_s,\\
A_{ssa} &= G_s(A_s + C_sF_s)G_s - G_s(B_s + D_sF_s),\\
A_{sas} &= F_s(A_sF_s - C_s) + F_s(B_sF_s - D_s)G_s,\\
A_{saa} &= G_sF_s(A_sF_s - C_s)G_s - G_sF_s(B_sF_s - D_s),\\
A_{ass} &= (A_a + C_aF_s) + (B_a + D_aF_s)G_s,\\
A_{asa} &= G_s(A_a + C_aF_s)G_s - G_s(B_a + D_aF_s),\\
A_{aas} &= F_s(A_aF_s - C_a) + F_s(B_aF_s - D_a)G_s,\\
A_{aaa} &= G_sF_s(A_aF_s - C_a)G_s - G_sF_s(B_aF_s - D_a),
\end{aligned} \tag{5.18}$$

with

$$\begin{aligned}
A_s &= A_{11} + A_{15}E_s, & A_a &= E_sA_{11}E_s - E_sA_{15},\\
B_s &= A_{12} + A_{16}E_s, & B_a &= E_sA_{12}E_s - E_sA_{16},\\
C_s &= A_{13} + A_{17}E_s, & C_a &= E_sA_{13}E_s - E_sA_{17},\\
D_s &= A_{14} + A_{18}E_s, & D_a &= E_sA_{14}E_s - E_sA_{18}.
\end{aligned}$$

THEOREM 5.4. *Let the domain of the three-dimensional linear isotropic or orthotropic elasticity problem* (5.1), *be a cube with its boundary conditions symmetrical about at least one principal plane. If the principal directions of orthotropy coincide with the three principal coordinate axes, then the problem can be discretized in such a way that the system stiffness matrix K assembled from $K_{(e)}$ and the system mass matrix M assembled from $M_{(e)}$ are both* SAS*-decomposable with respect to some reflection matrix P.*

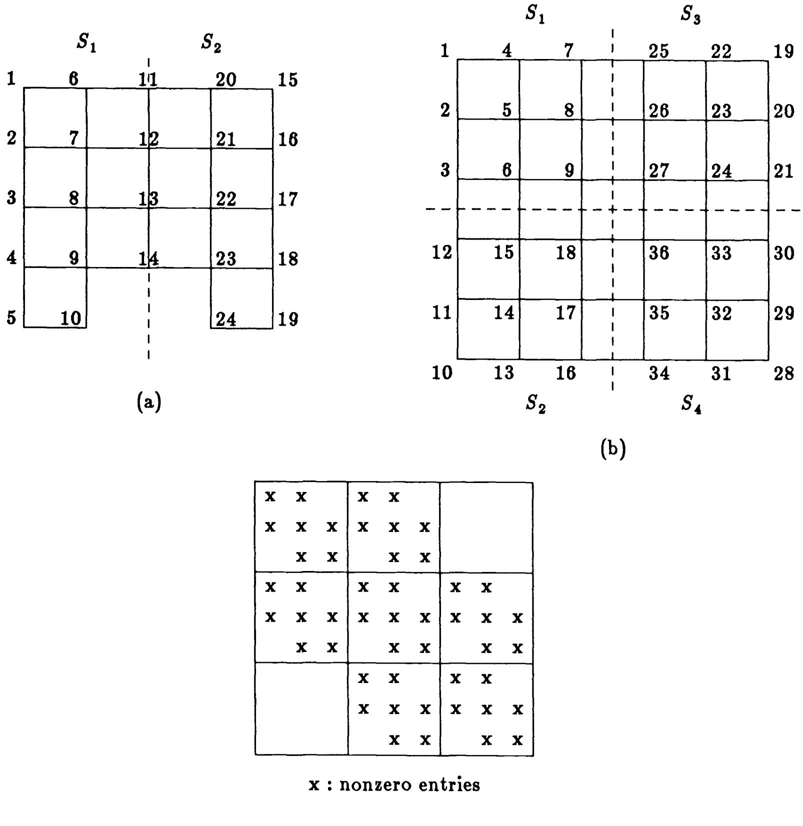
Proof. The proof of this theorem is rather lengthy and, therefore, will not be pursued here. Interested readers are referred to [Chen88] for a similar proof for orthotropic plate bending problems. □

DEFINITION 5.5. *The* SAS (*or reflexive*) *ordering.* An ordering is referred to as the SAS ordering if the following rules are satisfied.

(1) The whole domain is divided into two subdomains along a line (or a plane) of symmetry.

(2) Nodes in a subdomain are ordered such that any nodes on the line (or plane) of symmetry are ordered last.

(3) Nodes in the second subdomain are then numbered in the same order as their symmetrical counterparts in the first subdomain; see Fig. 5.3(a) for an ordering in a two-dimensional case. (SAS ordering in the three-dimensional case is treated similarly.)



(a)

(b)

x : nonzero entries

(c)

FIG. 5.3. *Two examples of the* SAS *ordering*. (*Dashed lines represent lines of symmetry*.)

If, however, the domain has more than one line (or plane) of symmetry the rules above are applied recursively on each of the subdomains created by the first line (or plane) of symmetry (see Fig. 5.3(b)). Although we have used the natural ordering within a subdomain as shown in Fig. 5.3, it should be mentioned that from the SAS ordering point of view, there is no restriction on what ordering should be used within a given subdomain. Different ordering, of course, results in different matrix structure. For the ordering shown in Fig. 5.3(b), the four decomposed submatrices all have the same structure, as shown in Fig. 5.3(c). We shall not be concerned with the structure of the undecomposed matrix because it need not be explicitly assembled.

In Theorem 5.4 we did not specify the form of the reflection matrix. Such a matrix depends on many factors, such as the ordering of unknowns, the plane(s) of symmetry, and the number of nodes on the plane(s) of symmetry. If, for example, we symmetrically discretize the problem into two subdomains and place no nodes on the plane of symmetry, then by employing the SAS ordering and retaining all dummy unknowns associated with the boundary conditions in the system (in other words each node retains all three unknowns in the same order as shown in Fig. 5.1 whether or not it has constraints), we obtain a system stiffness (or mass) matrix that is SAS-decomposable with respect to the following reflection matrix:

$$P=\begin{bmatrix} 0 & (I_L \otimes S_3) \\ (I_L \otimes S_3) & 0 \end{bmatrix} \tag{5.19}$$

where the subscript L represents the number of nodes in each of the two subdomains and

$$S_3 = \begin{cases} E_3 & \text{if } n_p \text{ is parallel to } x_1, \\ F_3 & \text{if } n_p \text{ is parallel to } x_2, \\ G_3 & \text{if } n_p \text{ is parallel to } x_3, \end{cases}$$

where E_3, F_3, and G_3 are defined in (5.5) and n_p stands for the normal of the plane of symmetry. If however there are K additional nodes on the plane of the symmetry then the reflection matrix may be given by

$$P = \begin{bmatrix} 0 & 0 & (I_L \otimes S_3) \\ 0 & (I_K \otimes S_3) & 0 \\ (I_L \otimes S_3) & 0 & 0 \end{bmatrix}. \qquad (5.20)$$

To close this section, we should bear in mind that the reflection matrix with respect to which a matrix is SAS-decomposable need not be unique.

6. Numerical experiments.

6.1. Physical problems. The physical problems we consider for our performance tests are two prismatic long bars, B1 and B2, as shown in Figs. 6.1 and 6.2, respectively. Both bars are assumed to be isotropic. Bar B1 [Wang53], [TiGo70] is fixed at the left end, i.e., the displacements in all three directions on the plane $x = 0$ are equal to zero. Bar B2, having the same uniform cross section, is fixed at both ends: $x = 0$ and $x = L$. The loading we consider for B1 is a simple bending moment M applied at its right end, while for B2 we apply a downward line load q at the two-thirds position of the bar from the left end. The dimensionless values for the constants L, M, $\cdots$, etc. are given in Table 6.1.

We use the basic 8-node rectangular hexahedral elements with several different finite-element discretization grids. These are $N_x \times N_y \times N_z$ grids where N_s ($s = x, y,$ or z) denotes the number of grid spacings in the direction of s. All discretized elements are identical in size. A 16 × 5 × 5 grid is shown in Fig. 6.3.

Since the domain and boundary conditions of bar B1 are symmetrical about planes xy and xz and the problem is symmetrically discretized, the system stiffness matrix can be decomposed into four disjoint submatrices. Similarly, we can decompose the system stiffness matrix of bar B2 into eight disjoint submatrices because it is symmetrical about three principal planes. In practice we do not actually decompose the assembled system

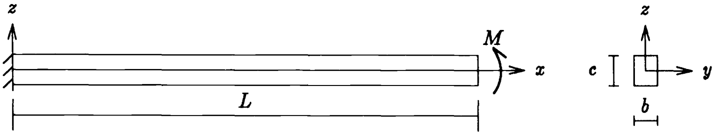


FIG. 6.1. *The prismatic bar* B1.

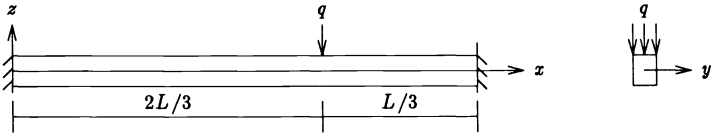


FIG. 6.2. *The prismatic bar* B2.

TABLE 6.1
Dimensionless constants.

Constants	L	b	c	E	ν	M	q
Values	30.0	1.5	2.0	1000.0	0.3	60.0	200.0

stiffness matrix. Instead, we need decompose only the element stiffness matrices involved before assembling them into the system.

To conveniently decompose and assemble the disjoint submatrices while taking advantage of the SAS property possessed by the system matrix, we use the SAS ordering to number the nodes between subdomains for all discretization. Within each subdomain we use the natural ordering, plane by plane, starting from the plane $x = L$, which has fewest points as compared with planes parallel to either $y = 0$ or $z = 0$, in order to minimize the bandwidth of the resulting matrix. On each plane the natural ordering is applied beginning with the direction which has fewer points. Note that the decomposability of the resulting linear system is independent of the right-hand side vector and therefore the symmetry of external loadings is not one of our main concerns.

6.2. Comparisons of results. The exact solutions of the three displacements for bar B1 are given [TiGo70], respectively, by

$$\delta_x = \left(\frac{M}{EI_y}\right)(-xz), \qquad \delta_y = \left(\frac{M}{EI_y}\right)(\nu xz),$$

$$\delta_z = \left(\frac{M}{2EI_y}\right)(x^2 - \nu y^2 + \nu z^2)$$

where I_y is the moment of inertia of the cross section of the bar with respect to the neutral axis parallel to the y axis. The comparison of displacements between the numerical approximation via the SAS decomposition technique and the exact solution for bar B1 at $(x, y, z) = (30.0, -0.75, 1.00)$ is shown in Table 6.2. The exact solution for bar B2 is not available. We therefore compare the numerical solution at $(20.0, -0.75, 1.00)$ with a solution obtained by using the isotropic parametric element L3DISOP [LoDo86] implemented in POLO-FINITE, a structural analysis software package developed at the Department of Civil Engineering, University of Illinois. The comparison is given in Table 6.3. For each discretization grid, identical numerical results (except round-off errors) were observed whether the problem was solved via the SAS approach or as a single domain without using decompositions.

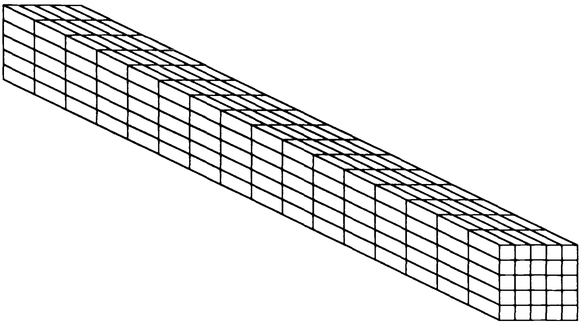


FIG. 6.3. *The* 16 × 5 × 5 *discretization grid* (*not in scale*).

TABLE 6.2
Solutions obtained via the SAS *approach.*
(*For bar* B1 *at* (30.0, −0.75, 1.00).)

Grid	16 × 5 × 5	32 × 5 × 5	48 × 5 × 5	64 × 5 × 5	Exact
$\delta_x/10$	−0.1337	−0.1652	−0.1727	−0.1756	−0.1800
$10\delta_y$	−0.1029	−0.1259	−0.1315	−0.1337	−0.1350
$\delta_z/100$	0.2001	0.2473	0.2588	0.2631	0.2700

TABLE 6.3
Solutions obtained via the SAS *approach.*
(*For bar* B2 at (20.0, −0.75, 1.00).)

Grid	15 × 5 × 5	45 × 5 × 5	75 × 5 × 5	15 × 5 × 5 (from POLO-FINITE)
$\delta_x/10$	−0.1561	−0.2085	−0.2145	−0.1561
δ_y	−0.1711	−0.2690	−0.2886	−0.1711
$\delta_z/100$	−0.2207	−0.2982	−0.3076	−0.2207

As far as the data storage and the computational efficiency are concerned, we compare the storage required and the CPU time (all in seconds) consumed on an Alliant FX/8 in solving the resulting linear systems with and without decompositions. Table 6.4 presents for the decomposed stiffness matrices the minimum half-bandwidth (a symmetric matrix $A = (a_{ij})$ is said to have half-bandwidth p if $a_{ij} = 0$ whenever $|i - j| \geqq p$), which can possibly be obtained through the SAS decomposition for both bars when the natural ordering is employed in each subdomain. The ordering we described earlier in this section gives this minimum half-bandwidth. It is clear that the SAS domain decomposition can greatly reduce the storage requirement for most cases. The following four algorithms are used to test the efficiency of the SAS approach. All computations are performed in double precision.

CHOLSE: An algorithm using Linpack solver DPBFA and DPBSL routines [DoMo79] on eight CEs (parallel implementation 2 (Table 4.2));

CHOLCN: An algorithm using Linpack solver DPBFA and DPBSL routines one per CE (parallel implementation 1 (Table 4.2));

GROWSE: An algorithm using Gaussian elimination where the matrix is stored by diagonals (rowwise) (parallel implementation 2 (Table 4.2)); and

GROWCN: An algorithm using Gaussian elimination where the matrix is stored by diagonals (rowwise) (parallel implementation 1 (Table 4.2)).

For this type of three-dimensional problems, the resulting banded system stiffness matrix is rather dense within the band. Direct banded solvers are much more efficient than iterative methods such as (preconditioned) conjugate gradient algorithms. The comparison of CPU time for non-SAS (NSUB = 1) and SAS (NSUB > 1) approach for bar B1 for a grid 64 × 5 × 5 grid is presented in Table 6.5. In Fig. 6.4 we plot the CPU time for all four discretization grids using the algorithm GROWSE with only one CE (FX/1). It is seen that the SAS approach is much more efficient than the classical one.

TABLE 6.4
Half-bandwidth for the decomposed submatrices.

Number of subdomains	1	2	4	8
Half-bandwidth	132	78	42	42

TABLE 6.5
CPU *time on the Alliant* FX/8 *in seconds for the* SAS *approach for bar* B1 *with grid* 64 × 5 × 5 *using direct methods.*

Algorithm	Number of subdomains	Number of processors			
		1	2	4	8
CHOLSE	1	95.43	67.09	52.11	46.47
	2	44.58	33.31	27.32	25.06
	4	15.16	13.08	11.97	11.55
CHOLCN	1	94.64	–	–	–
	2	44.30	22.80	–	–
	4	15.20	7.73	4.01	–
GROWSE	1	91.66	47.28	26.35	17.29
	2	37.55	19.55	10.89	7.06
	4	14.69	7.81	4.48	2.98
	4	(3.74)[1]	(2.00)[1]	(1.14)[1]	(0.75)[1]
GROWCN	1	90.71	–	–	–
	2	37.41	20.19	–	–
	4	14.65	7.60	4.22	–

[1] Numbers in parentheses indicate the CPU time actually consumed when the symmetry of the external loadings is taken into account by checking the right-hand side vectors of the decomposed subsystems.

This portion of savings in CPU time results mainly from the reduction of the bandwidth of the decomposed submatrices. Figure 6.5 shows the further reduction of CPU time contributed from the use of multiprocessors, namely, the effect of parallelism. Similar comparisons for bar B2 are given in Table 6.6 and Figs. 6.6 and 6.7.

If we define τ to be the ratio of the time required to solve the problem using the classical approach on one processor to the time consumed by the SAS approach using all eight CEs of the Alliant FX/8, then we can see from Tables 6.5 and 6.6 that the combination of the SAS domain decomposition and parallelism yields ratios τ ranging from 8.26 (CHOLSE) to 30.76 (GROWSE) for bar B1, and from 8.51 (CHOLSE) to 43.43 (CHOLCN) for bar B2. It should be pointed out that the above speedups did not take advantage of the symmetry of the external loadings. In other words, we solved all subsystems without checking their right-hand side vectors. The speedup in terms of the ratio of the CPU time using one CE to that using eight CEs ranges from 1.31 and 2.07 for the algorithm CHOLSE and from 4.93 to 5.48 for the algorithm GROWSE. Clearly if the concurrency is applied to solving a given linear subsystem, GROWSE (Gaussian elimination) has much more potential than CHOLSE (Cholesky decomposition). If, however, the concurrency can be applied to one level higher, i.e., solving several independent linear (sub)systems, then the Cholesky decomposition may still be competitive, depending on the number of independent (sub)systems and the number of processors available. For example, the algorithm CHOLCN, going from one CE to eight CEs, yielded a speedup of 6.75 for bar B2 when the domain is decomposed into eight independent

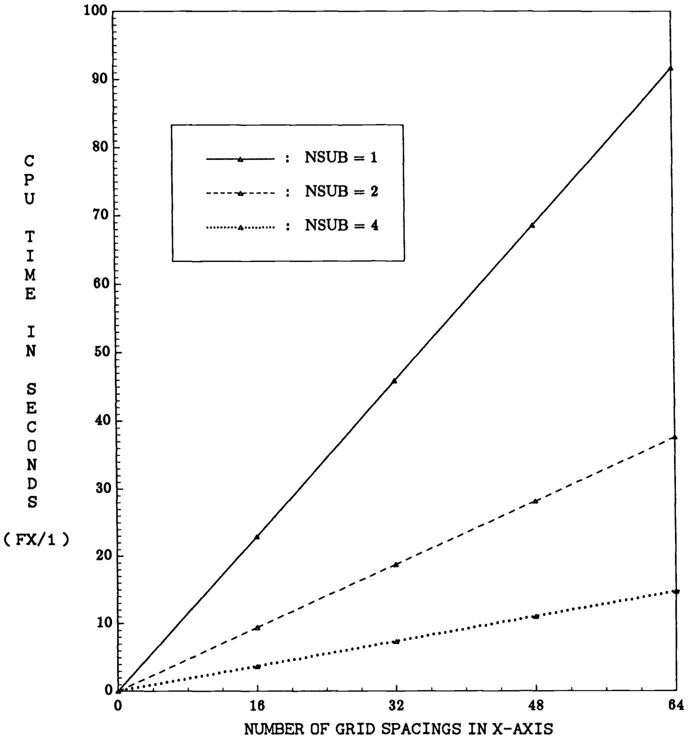


FIG. 6.4. CPU *time spent in solving linear systems from bar* B1 *via the* SAS *decompositions* (*Algorithm* GROWSE).

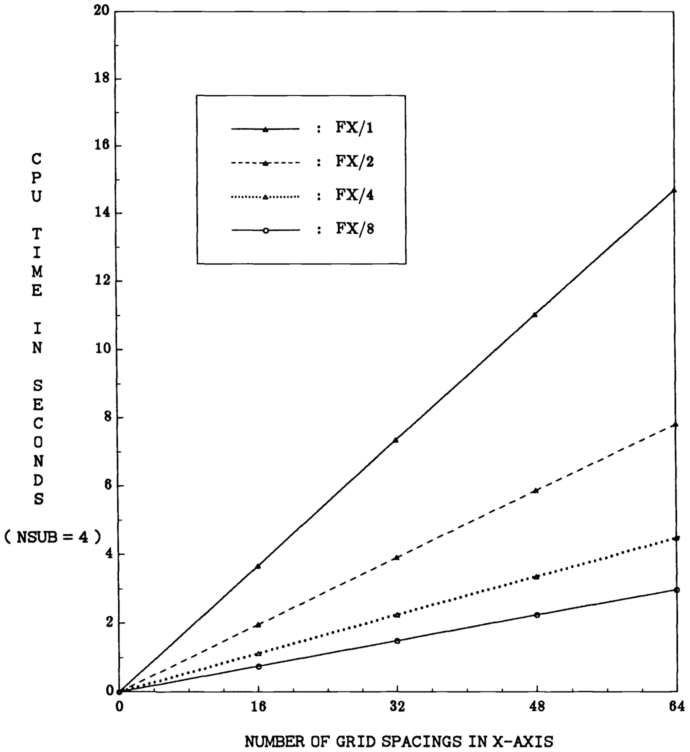


FIG. 6.5. CPU *time spent in solving linear systems from bar* B1 *using multiprocessors* (*Algorithm* GROWSE).

TABLE 6.6

CPU *time on the Alliant* FX/8 *in seconds for the* SAS *approach for bar* B2 *with grid* 75 × 5 × 5 *using direct methods.*

Algorithm	Number of subdomains	Number of processors			
		1	2	4	8
	1	114.56	78.86	62.29	55.30
CHOLSE	2	53.02	39.02	32.32	30.25
	4	18.10	15.37	15.28	13.66
	8	17.76	15.14	14.06	13.46
	1	114.65	–	–	–
CHOLCN	2	52.62	26.67	–	–
	4	18.13	9.12	4.79	–
	8	17.82	8.97	4.75	2.64
	1	112.95	55.21	32.34	20.59
GROWSE	2	45.30	22.91	13.11	8.88
	4	17.60	9.15	5.39	3.52
	8	17.53	9.06	5.29	3.50
	8	(8.67)[1]	(4.63)[1]	(2.64)[1]	(1.75)[1]
	1	114.60	–	–	–
GROWCN	2	45.55	23.88	–	–
	4	17.80	8.89	5.08	–
	8	17.60	8.71	5.09	3.48

[1] Numbers in parentheses indicate the CPU time actually consumed when the symmetry of the external loadings is taken into account by checking the right-hand side vectors of the decomposed subsystems.

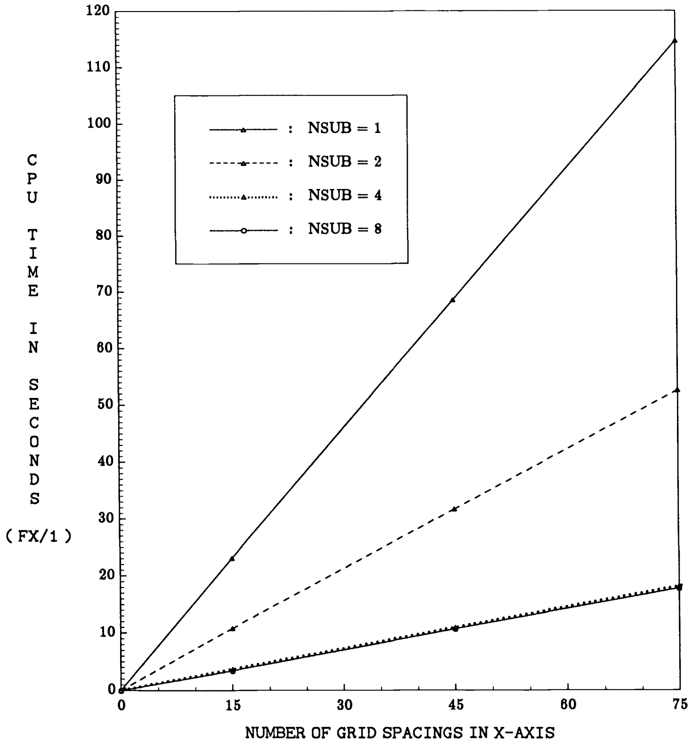


FIG. 6.6. CPU *time spent in solving linear systems from bar* B2 *via the* SAS *decompositions* (*Algorithm* CHOLCN).

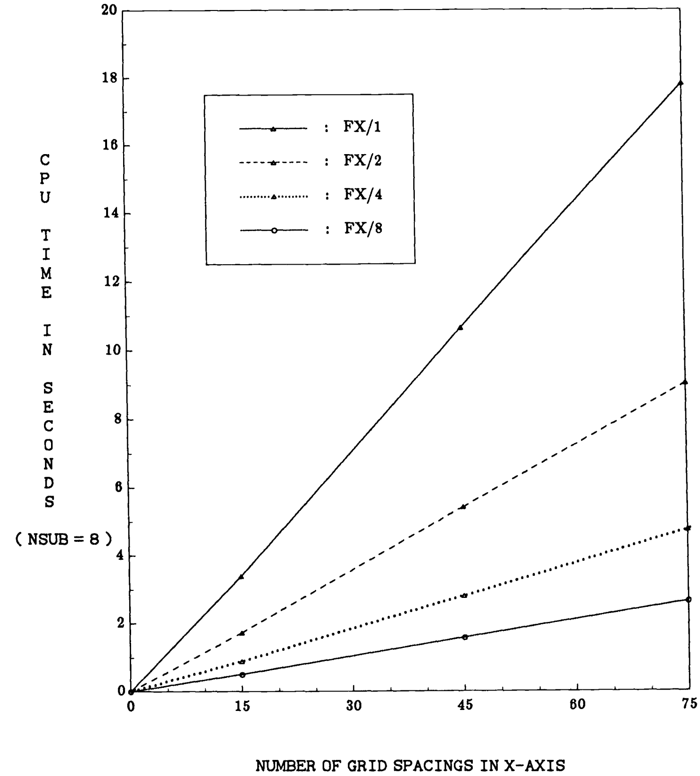


FIG. 6.7. CPU *time spent in solving linear systems from bar* B2 *using multiprocessors* (*Algorithm* CHOLCN).

subdomains by the SAS decomposition technique. This speedup would not have been possible had it not been for the exploitation of parallelism on a higher level.

7. Conclusions. The SAS domain decomposition method and its application to three-dimensional orthotropic/isotropic elasticity problems with domain and boundary condition symmetry have been presented. This decomposition method is an efficient and parallelizable approach for decomposing algebraic linear systems, eigenvalue problems, and generalized eigenvalue problems that possess the SAS property into smaller independent subsystems or subproblems. Mathematically, this approach exploits the important SAS property possessed by the special class of matrices $A = PAP$ where P is some reflection matrix (symmetrical signed permutation matrix). Using orthogonal (or unitary) transformations, we decompose the matrix into disjoint submatrices. Physically, the method takes advantage of the symmetry of a given problem and decomposes the whole domain of the original problem into independent subdomains. Unlike the fast Fourier decomposition method, this approach is constrained only by the SAS property and therefore has much wider applications.

From the outcome of the numerical experiments presented in this paper, it is clear that the SAS domain decomposition method is a very efficient approach for problems that are symmetrically discretizable. For problems that cannot be symmetrically discretized, the SAS domain decomposition method may still be promising, if used in conjunction with other domain decomposition techniques. Numerical experiments in this area will be reported later. Other advantages of the SAS approach that are worth mentioning are:

(1) This orthogonal decomposition of a matrix into disjoint submatrices is numerically stable.

(2) The approach lends itself to parallelism on three levels. It therefore is useful not only for supercomputers like the Cray X-MP series but for multiprocessors of the Cedar type [DaKu86], [ChSa87].

(3) It has potential for reducing the bandwidth of the matrix, thus reducing the storage requirements when the matrix is stored in a banded form.

Appendix A. The stiffness matrix for an orthotropic rectangular hexahedral element. (See Fig. 5.1 for notations and sign conventions.)

$$K_{(e)}=\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}=K_{(e)}^T,$$

$$S_{11}=\begin{vmatrix}
p_1 &&&&&&&&&&& \\
p_4 & p_2 &&&&&&&&&& \\
p_6 & p_5 & p_3 &&&&&&&&& \\
q_1 & q_4 & q_6 & p_1 &&&&& \text{sym.} &&& \\
q_4 & q_2 & q_5 & p_4 & p_2 &&&&&&& \\
-q_6 & -q_5 & q_3 & -p_6 & -p_5 & p_3 &&&&&& \\
r_1 & r_4 & r_6 & s_1 & s_4 & -s_6 & p_1 &&&&& \\
-r_4 & r_2 & r_5 & -s_4 & s_2 & -s_5 & -p_4 & p_2 &&&& \\
r_6 & -r_5 & r_3 & s_6 & -s_5 & s_3 & p_6 & -p_5 & p_3 &&& \\
s_1 & s_4 & s_6 & r_1 & r_4 & -r_6 & q_1 & -q_4 & q_6 & p_1 && \\
-s_4 & s_2 & s_5 & -r_4 & r_2 & -r_5 & -q_4 & q_2 & -q_5 & -p_4 & p_2 & \\
-s_6 & s_5 & s_3 & -r_6 & r_5 & r_3 & -q_6 & q_5 & q_3 & -p_6 & p_5 & p_3
\end{vmatrix},$$

$$S_{21}=\begin{vmatrix}
w_1 & w_4 & w_6 & x_1 & x_4 & -x_6 & y_1 & -y_4 & y_6 & z_1 & -z_4 & -z_6 \\
-w_4 & w_2 & w_5 & -x_4 & x_2 & -x_5 & -y_4 & y_2 & -y_5 & -z_4 & z_2 & z_5 \\
-w_6 & w_5 & w_3 & -x_6 & x_5 & x_3 & -y_6 & y_5 & y_3 & -z_6 & z_5 & z_3 \\
x_1 & x_4 & x_6 & w_1 & w_4 & -w_6 & z_1 & -z_4 & z_6 & y_1 & -y_4 & -y_6 \\
-x_4 & x_2 & x_5 & -w_4 & w_2 & -w_5 & -z_4 & z_2 & -z_5 & -y_4 & y_2 & y_5 \\
x_6 & -x_5 & x_3 & w_6 & -w_5 & w_3 & z_6 & -z_5 & z_3 & y_6 & -y_5 & y_3 \\
y_1 & y_4 & y_6 & z_1 & z_4 & -z_6 & w_1 & -w_4 & w_6 & x_1 & -x_4 & -x_6 \\
y_4 & y_2 & y_5 & z_4 & z_2 & -z_5 & w_4 & w_2 & -w_5 & x_4 & x_2 & x_5 \\
-y_6 & -y_5 & y_3 & -z_6 & -z_5 & z_3 & -w_6 & -w_5 & w_3 & -x_6 & -x_5 & x_3 \\
z_1 & z_4 & z_6 & y_1 & y_4 & -y_6 & x_1 & -x_4 & x_6 & w_1 & -w_4 & -w_6 \\
z_4 & z_2 & z_5 & y_4 & y_2 & -y_5 & x_4 & x_2 & -x_5 & w_4 & w_2 & w_5 \\
z_6 & z_5 & z_3 & y_6 & y_5 & y_3 & x_6 & x_5 & x_3 & w_6 & w_5 & w_3
\end{vmatrix},$$

$$S_{22}=\begin{vmatrix}
p_1 &&&&&&&&&&& \\
-p_4 & p_2 &&&&&&&&&& \\
-p_6 & p_5 & p_3 &&&&&&&&& \\
q_1 & -q_4 & -q_6 & p_1 &&&&& \text{sym.} &&& \\
-q_4 & q_2 & q_5 & -p_4 & p_2 &&&&&&& \\
q_6 & -q_5 & q_3 & p_6 & -p_5 & p_3 &&&&&& \\
r_1 & -r_4 & -r_6 & s_1 & -s_4 & s_6 & p_1 &&&&& \\
r_4 & r_2 & r_5 & s_4 & s_2 & -s_5 & p_4 & p_2 &&&& \\
-r_6 & -r_5 & r_3 & -s_6 & -s_5 & s_3 & -p_6 & -p_5 & p_3 &&& \\
s_1 & -s_4 & -s_6 & r_1 & -r_4 & r_6 & q_1 & q_4 & -q_6 & p_1 && \\
s_4 & s_2 & s_5 & r_4 & r_2 & -r_5 & q_4 & q_2 & -q_5 & p_4 & p_2 & \\
s_6 & s_5 & s_3 & r_6 & r_5 & r_3 & q_6 & q_5 & q_3 & p_6 & p_5 & p_3
\end{vmatrix},$$

$$\begin{aligned}
p_1 &= 4.0C_1 + 4.0C_2 + 4.0C_3, & w_1 &= -4.0C_1 + 2.0C_2 + 2.0C_3,\\
p_2 &= 4.0C_8 + 4.0C_9 + 4.0C_{10}, & w_2 &= 2.0C_8 - 4.0C_9 + 2.0C_{10},\\
p_3 &= 4.0C_{13} + 4.0C_{14} + 4.0C_{15}, & w_3 &= 2.0C_{13} - 4.0C_{14} + 2.0C_{15},\\
p_4 &= -2.0C_4 - 2.0C_5, & w_4 &= 2.0C_4 - 2.0C_5,\\
p_5 &= -2.0C_{11} - 2.0C_{12}, & w_5 &= -1.0C_{11} - 1.0C_{12},\\
p_6 &= 2.0C_6 + 2.0C_7, & w_6 &= -2.0C_6 + 2.0C_7,\\
q_1 &= 2.0C_1 + 2.0C_2 - 4.0C_3, & x_1 &= -2.0C_1 + 1.0C_2 - 2.0C_3,\\
q_2 &= 2.0C_8 + 2.0C_9 - 4.0C_{10}, & x_2 &= 1.0C_8 - 2.0C_9 - 2.0C_{10},\\
q_3 &= -4.0C_{13} + 2.0C_{14} + 2.0C_{15}, & x_3 &= -2.0C_{13} - 2.0C_{14} + 1.0C_{15},\\
q_4 &= -1.0C_4 - 1.0C_5, & x_4 &= 1.0C_4 - 1.0C_5,\\
q_5 &= -2.0C_{11} + 2.0C_{12}, & x_5 &= -1.0C_{11} + 1.0C_{12},\\
q_6 &= 2.0C_6 - 2.0C_7, & x_6 &= -2.0C_6 - 2.0C_7,\\
r_1 &= 2.0C_1 - 4.0C_2 + 2.0C_3, & y_1 &= -2.0C_1 - 2.0C_2 + 1.0C_3,\\
r_2 &= -4.0C_8 + 2.0C_9 + 2.0C_{10}, & y_2 &= -2.0C_8 - 2.0C_9 + 1.0C_{10},\\
r_3 &= 2.0C_{13} + 2.0C_{14} - 4.0C_{15}, & y_3 &= 1.0C_{13} - 2.0C_{14} - 2.0C_{15},\\
r_4 &= -2.0C_4 + 2.0C_5, & y_4 &= 2.0C_4 + 2.0C_5,\\
r_5 &= 2.0C_{11} - 2.0C_{12}, & y_5 &= 1.0C_{11} - 1.0C_{12},\\
r_6 &= 1.0C_6 + 1.0C_7, & y_6 &= -1.0C_6 + 1.0C_7,\\
s_1 &= 1.0C_1 - 2.0C_2 - 2.0C_3, & z_1 &= -1.0C_1 - 1.0C_2 - 1.0C_3,\\
s_2 &= -2.0C_8 + 1.0C_9 - 2.0C_{10}, & z_2 &= -1.0C_8 - 1.0C_9 - 1.0C_{10},\\
s_3 &= -2.0C_{13} + 1.0C_{14} - 2.0C_{15}, & z_3 &= -1.0C_{13} - 1.0C_{14} - 1.0C_{15},\\
s_4 &= -1.0C_4 + 1.0C_5, & z_4 &= 1.0C_4 + 1.0C_5,\\
s_5 &= 2.0C_{11} + 2.0C_{12}, & z_5 &= 1.0C_{11} + 1.0C_{12},\\
s_6 &= 1.0C_6 - 1.0C_7, & z_6 &= -1.0C_6 - 1.0C_7,
\end{aligned}$$

$$\begin{aligned}
C_1 &= d_{11}l_2^2l_3^2/V, & C_6 &= d_{13}l_2/12, & C_{11} &= d_{23}l_1/12,\\
C_2 &= d_{44}l_1^2l_3^2/V, & C_7 &= d_{55}l_2/12, & C_{12} &= d_{66}l_1/12,\\
C_3 &= d_{55}l_1^2l_2^2/V, & C_8 &= d_{22}l_1^2l_3^2/V, & C_{13} &= d_{33}l_1^2l_2^2/V,\\
C_4 &= d_{12}l_3/12, & C_9 &= d_{44}l_2^2l_3^2/V, & C_{14} &= d_{55}l_2^2l_3^2/V,\\
C_5 &= d_{44}l_3/12, & C_{10} &= d_{66}l_1^2l_2^2/V, & C_{15} &= d_{66}l_1^2l_3^2/V,
\end{aligned}$$

where d_{ij} are elastic constants; l_1, l_2, and l_3 are as shown in Fig. 5.1; and $V = 18l_1l_2l_3$.

Appendix B. The consistent mass matrix for a rectangular hexahedral element. (See Fig. 5.1 for notation and sign conventions.)

$$M_{(e)}=\left(\frac{\rho l_1 l_2 l_3}{216}\right) M_8 \otimes I_3,$$

$$M_8=\begin{vmatrix} 8 & & & & & & & \\ 4 & 8 & & & & & & \\ 4 & 2 & 8 & & & \text{sym.} & & \\ 2 & 4 & 4 & 8 & & & & \\ 4 & 2 & 2 & 1 & 8 & & & \\ 2 & 4 & 1 & 2 & 4 & 8 & & \\ 2 & 1 & 4 & 2 & 4 & 2 & 8 & \\ 1 & 2 & 2 & 4 & 2 & 4 & 4 & 8 \end{vmatrix}$$

where ρ is the density of the material and I_3 is the identity matrix of order 3.

REFERENCES

[Alli87] Alliant Computer Systems Corporation, FX/FORTRAN *Programmer's Handbook*, Alliant Computer Systems Corporation, Littleton, MA, 1987.

[Bjor83] P. Bjorstad, *Fast numerical solution of the biharmonic Dirichlet problem on rectangles*, SIAM J. Numer. Anal., 20 (1983), pp. 59–71.

[BlKa66] S. Blaszkowiak and Z. Kaczkowski, *Iterative Methods in Structural Analysis*, (A. Kacner and Z. Olesiak, transl.), Pergamon Press, Oxford, 1966.

[BuGo70] B. L. Buzbee, G. H. Golub, and C. W. Nielson, *On direct methods for solving Poisson's equations*, SIAM J. Numer. Anal., 7 (1970), pp. 627–657.

[Chen88] H. C. Chen, *The* SAS *domain decomposition method for structural analysis*, Ph.D. Thesis, Department of Computer Science, University of Illinois at Urbana–Champaign, Urbana, IL, 1988.

[ChSa87] H. C. Chen and A. Sameh, *Numerical linear algebra algorithms on the cedar system*, in Parallel Computations and Their Impact on Mechanics, A. K. Noor, ed., The American Society of Mechanical Engineers, AMD-Vol. 86, 1987, pp. 101–125.

[DaKu86] E. Davidson, D. Kuck, D. Lawrie, and A. Sameh, *Supercomputing tradeoffs and the cedar system*, Report CSRD-577, Center for Supercomputing Research and Development, University of Illinois at Urbana–Champaign, Urbana, IL, 1986.

[Dawe84] D. J. Dawe, *Matrix and Finite Element Displacement Analysis of Structures*, Clarendon Press, Oxford, 1984.

[DoMo79] J. J. Dongarra, C. B. Moler, J. R. Bunch, and G. W. Stewart, LINPACK *Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.

[GoVa83] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[Lekh63] S. G. Lekhnitskii, *Theory of Elasticity of an Anisotropic Elastic Body*, Holden-Day, San Francisco, 1963.

[LiLi83] J. A. Ligget and P. L-F. Liu, *The Boundary Integral Equation for Porous Media Flow*, George Allen & Unwin Ltd., London, 1983.

[LoDo86] L. A. Lopez, R. H. Dodds, D. R. Rehak, and R. J. Schmidt, POLO-FINITE, *a structural mechanics system for linear and nonlinear, static and dynamic analysis*, Engineering Systems Laboratory, University of Illinois at Urbana–Champaign, Urbana, IL; Department of Civil Engineering, University of Kansas, Lawrence, KS; Department of Civil Engineering, Carnegie-Mellon University, Pittsburgh, PA; and Department of Civil Engineering, University of Wyoming, Laramie, WY, 1986.

[Melo63] R. J. Melosh, *Structural analysis of solids*, ASCE J. Structural Div., 89 (1963), pp. 205–223.

[NoPe87] A. K. Noor and J. M. Peters, *Model-size reduction for the nonlinear dynamic analysis of quasi-symmetric structures*, Engrg. Comput., 4 (1987), pp. 178–189.

[Prze68] J. S. Przemieniecki, *Theory of Matrix Structural Analysis*, McGraw-Hill, New York, 1968.

[Rubi66] M. F. Rubinstein, *Matrix Computer Analysis of Structures*, Prentice-Hall, Englewood Cliffs, NJ, 1966.

[SaCh76] A. H. SAMEH, S. C. CHEN, AND D. J. KUCK, *Parallel Poisson and biharmonic solvers*, Computing, 17 (1976), pp. 219–230.

[Schu77] U. SCHUMANN, *On fast direct methods for the solution of discretized elliptic equations*, in Proc. GAMM-Workshop on Fast Solution Methods for the Discretized Poisson Equation, Karlsruhe, 1977.

[Smit78] G. D. SMITH, *Numerical Solutions of Partial Differential Equations: Finite Difference Methods*, 2nd ed., Clarendon Press, Oxford, 1978.

[Szil74] R. SZILARD, *Theory and Analysis of Plates: Classical and Numerical Methods*, Prentice Hall, Englewood Cliffs, NJ, 1974.

[TiGo70] S. P. TIMOSHENKO AND J. N. GOODIER, *Theory of Elasticity*, McGraw-Hill, New York, 1st ed. 1934; 2nd ed. 1951; 3rd ed. 1970.

[Wang53] C. T. WANG, *Applied Elasticity*, McGraw-Hill, New York, 1953.

[Wein65] H. F. WEINBERGER, *A First Course in Partial Differential Equations with Complex Variables and Transform Methods*, Blaisdell, New York, 1965.

[WeJo87] W. WEAVER, JR. AND P. R. JOHNSON, *Structural Dynamics by Finite Elements*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[Wilk65] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

[Zien77] O. C. ZIENKIEWICZ, *The Finite Element Method*, 3rd ed., McGraw-Hill, London, 1977.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 1, pp. 65-79, January 1989

005

LINEAR AGGREGATION OF INPUT-OUTPUT MODELS*

ERIC C. HOWE† AND CHARLES R. JOHNSON‡

Abstract. The aggregation of input-output models is analyzed. Three axioms are shown to characterize a simple functional form for aggregation; then the properties of the aggregated model are analyzed relative to the original model. Since an input-output model is driven by a square substochastic matrix, these results can also be viewed as facts about abstract mappings involving substochastic matrices.

Key words. input-output, aggregation, substochastic matrices, axiomatic aggregation theory, economics

AMS(MOS) subject classifications. 90, 15

1. Introduction. It is clear that some aggregation is necessary for input-output analysis. The array of goods and services produced by an economy is so vast that aggregation is a prerequisite to estimation of an input-output matrix; additional aggregation is usually required by promising confidentiality to survey respondents or by legislating confidentiality restrictions for statistical agencies. However, most analyses substantially exceed the minimum amount of aggregation. Few researchers utilize the largest, least aggregated tables that are available, perhaps due to the inconvenience of using large matrices, or a belief that the information lost by aggregation is not a significant source of error.

Methodological questions about aggregation have generated a lively debate in many areas of economics; this has not been true, however, in input-output analysis. Practically all theoretical analyses of input-output aggregation have utilized the same functional form; none has addressed the question of why that functional form was selected. Therefore, part of our purpose in this paper is to consider input-output aggregation more generally, and discover what can be deduced axiomatically about functional forms. Special aspects of the structure of an input-output model make the aggregation literature largely inapplicable. Two of the most general axiomatic analyses of aggregation, Wilson (1975) and Rubinstein and Fishburn (1986), are examples.

Our second aim is to prove several new results about input-output aggregation and link these results to the existing literature. Most of the known results were established in several papers in the mid-1950s. Economics has changed in the three decades since, so the original presentation of these results make them fairly inaccessible. As a result, we will also provide modern statements of the pioneering results, and brief proofs that stress their logical interconnection. We begin with a brief review of how input-output aggregation is generally performed.

Suppose that an n-by-n input-output matrix A is to be aggregated from n "micro commodities," denoted $N = \{1, \cdots, n\}$, to m "macro commodities", $M = \{1, \cdots, m\}$, with $m < n$. Let an m-by-n matrix S indicate which micro commodities are to be combined: for all $i \in M$ and $j \in N$, $s_{i,j}$ equals 1 if micro commodity j is to be included in macro commodity i and equals zero otherwise. Thus S is a 0, 1 matrix with exactly one 1 in every column and at least one 1 in every row. (Hence S is column stochastic.) Let an n-by-m matrix T indicate the proportional weights of each micro

* Received by the editors January 18, 1988; accepted for publication (in revised form) July 6, 1988.

† Department of Economics, University of Saskatchewan, Saskatoon, Saskatchewan, Canada S7N 0W0 (howe@sask).

‡ Department of Mathematics, College of William and Mary, Williamsburg, Virginia 23185 (crjoh2@wmmvs.bitnet). The research of this author was supported by National Science Foundation grant DMS-8713762 and Office of Naval Research contracts N00014-86-K-0693 and N00014-86-K-0012.

commodity in its macro aggregate. For all $i \in M$ and $j \in N$, $t_{j,i} \in [0, 1]$ if micro commodity j is included in macro commodity i and $t_{j,i}$ equals 0 otherwise, and the sum of the weights of the micro industries assigned to a given macro industry is assumed to be 1. Thus T is also column stochastic.

The input-output aggregator used in the literature is computed as the matrix product, SAT.

At least conceptually, however, there are other methods of matrix aggregation available. For example, McManus (1956) and Morimoto (1971) consider aggregation where S may contain any positive weights; Neudecker (1970) considers an aggregator chosen to optimize a particular objective function; and Leontief (1967) considers an alternative notion—aggregation through algebraic elimination of variables. Other possibilities include applying the above aggregation procedure to the $(I - A)^{-1}$ matrix to obtain $S(I - A)^{-1}T$ and then computing the aggregation of A as $(I - [S(I - A)^{-1}T]^{-1})$. Alternatively, we might use the micro commodities as data to estimate an aggregated matrix, using a variant of the econometric estimation techniques suggested by Gerking (1976) for estimating input-output models. Or we might first aggregate the columns to produce a rectangular model with m multiproduct industries, and then convert the rectangular model to a square m-by-m model using the approach outlined by Miller and Blair (1985) for converting rectangular models to square models.

What can be said about the functional form in input-output aggregation? We turn first to the axiomatic treatment of general linear aggregation.

2. Axiomatic analysis of linear aggregation. In this section we consider a general aggregator f, mapping n-by-n input-output matrices into m-by-m input-output matrices $m < n$. We suggest three axioms we might require of an input-output aggregator and show that these axioms are equivalent to a basic functional form of f.

Denote the set of all real n-by-m matrices by $M_{n,m}$; when $n = m$, $M_{n,n}$ is abbreviated to M_n. The usual vector space of real n-tuples will be denoted by R^n, and vectors from that space will be assumed to be column vectors.

We consider an (open) input-output model with n commodities given by

$$x = Ax + y.$$

Here $x \in R^n$ is an output vector, $y \in R^n$ is a final demand vector, and $A \in M_n$ is an input-output matrix. We measure amounts of commodities in a common unit (e.g., \$) and make the ordinary assumption of a positive rate of value added in the production of each commodity. For an input-output matrix $A = (a_{i,j})$, the entry $a_{i,j}$ may be interpreted as the amount (value) of commodity i necessary in the production of a unit (dollar's worth) of commodity j, given the technology represented by A. Thus $1 - \sum_i a_{i,j}$ is the value added per unit of production of commodity j, which is assumed positive. In this event, the matrix A has nonnegative entries and column sums all less than 1. Such a matrix is usually called (strictly) column substochastic (since a column stochastic matrix is a nonnegative one with column sums of 1).

We note that we make these assumptions primarily for simplicity, as broader situations may be accommodated with no substantive change in our results. In particular, we could assume that the input-output matrix has column sums that are less than or equal to 1, and that every irreducible component of A contains one column where this inequality is strict. Modifications would have to be made to some of the following theorems because, as we will see, aggregation can alter the irreducibility of a matrix.

DEFINITION 2.1. By an *input-output* matrix we simply mean a square column substochastic matrix.

The input-output equation $y = (I - A)x$ can be used to transform an output vector to a final demand vector. Conversely, since A is column substochastic, $(I - A)$ must be nonsingular, so that $x = (I - A)^{-1}y$ can be used to transform a final demand vector to an output vector. Since A is substochastic, the inverse matrix $(I - A)^{-1}$ is nonnegative and if A is irreducible, then $(I - A)^{-1}$ is strictly positive.

We wish to consider general functions mapping n-by-n input-output matrices into m-by-m input-output matrices.

DEFINITION 2.2. An *input-output matrix aggregator* is a function $f\colon M_n \twoheadrightarrow M_m$ that maps the n-by-n input-output matrices into m-by-m input-output matrices $m < n$. The k, l element of the matrix $f(A)$ will be denoted by $f(A)_{k,l}$. An input-output matrix B will be referred to as an *aggregation* of the input-output matrix A if it is the result of some aggregator applied to A.

Considering that input-output models are linear, one natural assumption is that the aggregator is linear. Additional reasons for considering linear aggregators will emerge in the sections that follow (most especially Corollary 4.1).

AXIOM 1 (Linearity). The aggregator $f\colon M_n \twoheadrightarrow M_m$ is a linear function.

A second axiom requires that the aggregator not distort the payments to primary factors in one particularly obvious situation. We use e to denote the column vector of 1's whose number of components is determined by the context.

AXIOM 2 (Value Added Neutrality). For each $0 < \alpha < 1$ and each input-output matrix A satisfying $e^TA = \alpha e^T$, the aggregator must satisfy $e^Tf(A) = \alpha e^T$.

This axiom asserts that if the proportion of value added is the same in the production of all micro commodities, then it should be the same for all of the macro commodities, and the micro and macro proportions of value added should be equal. Thus if all micro commodities require $(1 - \alpha)$ units of the primary factors per dollar of output, then a macro commodity, which is just an aggregation of micro commodities, should also require $(1 - \alpha)$ units of primary factors per dollar of output. We note that under the assumption of linearity on f, the above axiom could be stated to hold for one particular value of α, which would imply that it holds for all values of α.

Generally, in the process of aggregation we think of forming each macro commodity from one or more micro commodities. Our last axiom is a precise statement of this idea.

AXIOM 3 (Partitioning). There exist functions h_I mapping N onto M, called the input partition, and h_O mapping N onto M, called the output partition, which have the following three properties. Let $i, j \in N$; $k, l \in M$; and suppose that $\delta \in R$ is such that $(A + \delta E_{i,j})$ is an input-output matrix.

(a) Input partitioning: $f(A + \delta E_{i,j})_{k,l} = f(A)_{k,l}$ unless $k = h_I(i)$.

(b) Output partitioning: $f(A + \delta E_{i,j})_{k,l} = f(A)_{k,l}$ unless $l = h_O(j)$.

(c) Coincidental partitioning: The output partition and the input partition are equal, $h_O = h_I = h$.

Here $E_{i,j}$ denotes the matrix with a 1 in the i, j position and 0's elsewhere. In forming the aggregation, the function h_I indicates how the micro commodities, regarded as inputs, are to be assigned to the macro commodities, and h_O indicates how the assignment is to be made regarding the commodities as outputs. The partitioning axiom states that a perturbation of an entry of A should not change the input requirements (column elements) of any macro commodity other than the one to which the commodity is mapped. Similarly, a perturbation of an entry of A should not change the output (row elements) of any macro commodity other than the one to which the commodity is mapped. Finally, the input partition and the output partition should be the same. An immediate consequence of Axiom 3 is that

$$f(A+\delta E_{i,j})_{k,l}=f(A)_{k,l} \text{ unless } k=h(i) \text{ and } l=h(j)$$

so a perturbation in the i, j element of A does not effect the k, l element of $f(A)$ unless h maps i to k and j to l.

We will be principally interested in aggregators that satisfy the preceding axioms.

DEFINITION 2.3. We call an input-output aggregator f that satisfies Axioms 1, 2, and 3 a *standard aggregator*. For convenience, we also call an input-output matrix B a *standard aggregation* of the input-output matrix A if B is the result of some standard aggregator applied to A.

The following theorem characterizes the functional form of standard aggregators.

THEOREM 2.1. *An input-output aggregator f: $M_n \to M_m$ is standard if and only if f may be represented as*

$$f(A) = SAT$$

in which $S \in M_{m,n}$ is a 0, 1 *column stochastic matrix*, $T \in M_{n,m}$ *is column stochastic, and*

$$ST = I \in M_m.$$

Proof. For sufficiency, first note that $f(A) = SAT$ satisfies Axiom 1, since for two matrices $A^{(1)}, A^{(2)} \in M_n$ and $\alpha, \beta \in R$, we have

$$f(\alpha A^{(1)} + \beta A^{(2)}) = S(\alpha A^{(1)} + \beta A^{(2)})T = \alpha f(A^{(1)}) + \beta f(A^{(2)}).$$

Indeed, the matrix representation of f, relative to the "standard" basis, is

$$\text{vec } f(A) = (T^T \otimes S) \text{ vec } A.$$

(The function vec maps $M_{m,n}$ into R^{mn} by "stacking" the matrix columns, taken from left to right, and $\otimes$ denotes the Kronecker product.) The stated form also satisfies Axiom 2 since, if $e^TA = \alpha e^T$, then we may calculate:

$$e^TSAT = e^TAT = \alpha e^TT = \alpha e^T.$$

Let h: $N \to M$ be given by $h(i) = k$ if and only if $s_{k,i} = 1$. The column stochasticity of the 0, 1 matrix S implies that for each $i \in N$, $s_{k,i} = 1$ for exactly one $k \in M$, which ensures that h is single-valued. Moreover, since $ST = I$, S must have at least one 1 in every row, so h is onto. For S and T to be column stochastic, while ST $(=I)$ is also column stochastic, T can only have a nonzero entry in position i, j if S has a 1 in position j, i. (If T had a positive entry in position i, j and S had a 0 in position j, i, then the i, i element of the product ST would be less than 1, contrary to the hypothesis.) Computation verifies, then, that $SE_{i,j}T$ contains only zeros, except possibly in position $h(i), h(j)$, so the form SAT satisfies the third, and hence all three, of the axioms.

The proof of the necessity of the stated form proceeds in stages to indicate the functional forms that are possible as more axioms are applied.

If f is a linear aggregator, i.e., if Axiom 1 is assumed, then f has a matrix representation, G, in the "standard" basis, i.e.,

$$\text{vec } f(A) = G \text{ vec } A$$

in which G is m^2-by-n^2. Because an aggregator must map nonnegative matrices to nonnegative matrices, G must be componentwise nonnegative. Partition G as follows:

$$G = (G_{p,q})$$
$$= \begin{bmatrix} G_{1,1} & \cdots & G_{1,n} \\ \vdots & & \vdots \\ G_{m,1} & \cdots & G_{m,n} \end{bmatrix}$$

in which $G_{p,q} \in M_{m,n}$ for $p = 1, \cdots, m$ and $q = 1, \cdots, n$. Also, partition vec A and vec $f(A)$ as follows:

$$\text{vec } A = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}, \qquad \text{vec } f(A) = \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}$$

in which $u_q \in R^n$ for $q = 1, \cdots, n$, and $v_p \in R^m$ for $p = 1, \cdots, m$. Then

$$v_p = \sum_{q=1}^{n} G_{p,q} u_q.$$

Next we show that the value added neutrality axiom implies that there exist nonnegative numbers $\alpha_{p,q}$ for $p = 1, \cdots, m$ and $q = 1, \cdots, n$ such that

$$G_{p,q} = \alpha_{p,q} H_{p,q} \quad \text{for all } p \text{ and } q$$

in which $H_{p,q} \in M_{m,n}$ is column stochastic and

$$\sum_{q=1}^{n} \alpha_{p,q} = 1 \quad \text{for } p = 1, \cdots, m.$$

To show this, note that

$$e^T v_p = \sum_q e^T G_{p,q} u_q = \sum_q z_{p,q}^T u_q$$

in which $z_{p,q}^T$ is the column sum vector of $G_{p,q}$. Note that value added neutrality requires that the preceding expression equal α for all $p = 1, \cdots, m$ whenever all u_q, $q = 1, \cdots, n$ have the property that $e^T u_q = \alpha$. By varying the u_q among those with that property, it is clear that $\sum z_{p,q}^T u_q$ can take any value in the interval

$$\alpha[\sum (z_{p,q}^T)_{\min}, \sum (z_{p,q}^T)_{\max}]$$

in which $(\cdot)_{\min}$ denotes the minimum entry and $(\cdot)_{\max}$ denotes the maximum entry of the indicated vector. This degenerates into the trivial interval α if and only if

$$(z_{p,q}^T)_{\min} = (z_{p,q}^T)_{\max} = \alpha_{p,q} \quad \text{for } q = 1, \cdots, n \text{ and } p = 1, \cdots, m$$

and

$$\sum_{q=1}^{n} \alpha_{p,q} = 1 \quad \text{for } p = 1, \cdots, m.$$

This verifies the assertion.

Finally, input partitioning implies that

$$(G_{l,j})_{k,i} = 0$$

unless $k = h_I(i)$. This, plus the preceding assertion, implies that $G_{l,j} = \alpha_{l,j} S$, in which S is a 0, 1 column stochastic matrix that is independent of l, j. Thus $G = (\alpha_{l,j}) \otimes S$ and $f(A) = SA(\alpha_{l,j})^T$ and $(\alpha_{l,j})^T$ is column stochastic. Coincidental partitioning implies that $(\alpha_{l,j})$ has a zero entry wherever S does. Denote $(\alpha_{l,j})^T$ by T and the zero pattern plus column stochasticity of T implies $ST = I$, to complete the proof. □

In the context of the theorem, the statement $ST = I$ simply means that the nonzero entries of T are contained among the positions indicated by the 1's of S^T.

As noted in the Introduction, SAT is the most common functional form used for input-output aggregation. Usually the columns of T measure the proportions of the micro commodities in their macro aggregates in a base-period output vector x_o. That is, most applications set

$$T = \operatorname{diag}(x_o) S^T \{\operatorname{diag}(Sx_o)\}^{-1}$$

(where diag (·) denotes the square diagonal matrix whose elements are the components, in natural order, of the indicated vector). According to Charnes and Cooper, "The main justification for this mode of consolidation is that it conforms to the way data would be synthesized ab initio if [SAT] rather than A were the objective" (1961). Theorem 2.1 has established another significant reason to favor this particular functional form, although not necessarily this particular choice of T.

The form SAZ where Z is the Moore–Penrose generalized inverse of S is another special case of standard aggregation. The Moore–Penrose generalized inverse of S is the unique matrix $Z \in M_{n,m}$ such that $SZS = S$; $ZSZ = Z$; and SZ and ZS are symmetric. (See, for example, Horn and Johnson (1985).) The fact that $ST = I$ shows that $Z = T$ must satisfy the first two of these conditions for any choice of T (of the form given by Theorem 2.1), but the third need not be satisfied since TS need not be symmetric in general. However, suppose Z is obtained as the result of dividing the columns of S^T by their sums, that is

$$Z = S^T \{\operatorname{diag}(Se)\}^{-1}.$$

Then Z is column stochastic, and $SZ = I$, so $T = Z$ satisfies the requirements of Theorem 2.1. Moreover, $(ZS)^T = (S^T\{\operatorname{diag}(Se)\}^{-1}S)^T = ZS$, so Z satisfies the final symmetry requirement, and hence is the generalized inverse of S. The choice of T has been studied by Ijiri (1968), Kymn (1977), and others.

Theorem 2.1 shows that the three axioms are necessary and sufficient to guarantee that an aggregator has form SAT for S and T of the specified type. The following examples show that none of the axioms are superfluous. If the linearity axiom is dropped, then numerous different aggregators become possible, for example, the aggregator $f(A) = SAT$ in which the entries of S and T depend explicitly on the entries of A. (We note that the nonlinear function $f(A) = I - (S(I - A)^{-1}T)^{-1}$ mentioned in § 1 does not qualify as an aggregator because it does not always produce a nonnegative matrix, as shown by simple examples.) If value added neutrality is dropped (and the other two axioms retained) then S need not be column stochastic. This would correspond to "weighted aggregation" studied by McManus (1956) and Morimoto (1971). If the partitioning axiom is relaxed to permit the perturbation of one element of A to affect more than one element of $f(A)$, then S can have more than one positive entry per column; it need not be a 0, 1 matrix, but must still be column stochastic. This would correspond to an aggregation scheme where each micro commodity could be proportioned among several macro commodities.

Some additional terminology will prove useful. The proof of Theorem 2.1 makes it clear that the partition function h and the matrix S are equivalent. We introduce the following terminology to take account of that fact.

DEFINITION 2.4. If h maps N onto M then the 0, 1 matrix $S \in M_{m,n}$ given by

$$s_{i,j} = \begin{cases} 1 & \text{if } h(j) = i, \\ 0 & \text{otherwise} \end{cases}$$

will be called the *matrix representation* of h. Conversely, if $S \in M_{m,n}$ is a 0, 1 column stochastic matrix with no row containing only zeros, then S will be called a *partitioning matrix*, and the function h mapping N onto M given by $h(j) = i$ if $s_{i,j} = 1$ will be called the *function representation* of S.

Plainly, the matrix representation of h is column stochastic and has no row containing only zeros. The function representation of S is single-valued and onto. Also, the matrix S in Theorem 2.1 is a partitioning matrix.

3. Standard aggregation. This section discusses some features of the input-output matrix A that are, in general, preserved by a standard aggregator, $B = SAT$. The results in this section appear to be new. First, we prove the following result, which will prove useful throughout the remainder of the paper.

THEOREM 3.1. *Suppose $S \in M_{m,n}$ is a partitioning matrix, $T \in M_{n,m}$ is column stochastic, and $ST = I$. Then TS is a column stochastic, idempotent matrix of rank m.*

Proof. The matrix TS is idempotent because $(TS)(TS) = T(ST)S = TS$. Since T and S are column stochastic, TS is column stochastic. Moreover, S is a 0, 1 matrix with exactly one 1 in each column and at least one 1 in each row, so the columns of TS are each columns from T, and each column of T appears at least once in TS. The matrix T has full column rank equal to m because $ST = I \in M_m$. Hence TS also has rank m. □

Theorem 3.1 implies that the set of eigenvalues of TS includes 1 with multiplicity m and 0 with multiplicity $(n - m)$, since TS is idempotent.

The previous theorem can be used to establish a close relationship between standard aggregators and the notion of matrix similarity. Two matrices, $G, H \in M_n$ are said to be similar if there exists a nonsingular matrix Q such that $G = Q^{-1}HQ$. That is, there must exist a pair of matrices Q_1 and Q_2 such that $G = Q_1HQ_2$ with $Q_1Q_2 = Q_2Q_1 = I$. Note that a standard aggregator yields $B = SAT$ with $ST = I$. Although TS is not an idempotent matrix of rank n (and hence the identity), as would be the case if A and B were similar, it is an idempotent matrix of rank m. It is reasonable, then, to inquire about the relationship between the set of eigenvalues of SAT, $\sigma(SAT)$, and the set of eigenvalues of A, $\sigma(A)$. Although similar matrices have the same sets of eigenvalues, counting multiplicities, an additional condition must be satisfied in order for SAT to inherit an eigenvalue from A.

THEOREM 3.2. *Suppose that $A \in M_n$ is an input-output matrix, $S \in M_{m,n}$ is a partitioning matrix, $T \in M_{n,m}$ is column stochastic, and $ST = I$. Let $\lambda \in \sigma(A)$ with eigenvector x. If x is also an eigenvector of TS associated with the eigenvalue $1 \in \sigma(TS)$, then $\lambda \in \sigma(SAT)$ with eigenvector Sx.*

Proof. Suppose that $\lambda \in \sigma(A)$ with an associated eigenvector $x \neq 0$, and suppose that, in addition, $x = TSx$. Then $Ax = \lambda x$, so $A(TSx) = \lambda x$, and hence $SATSx = \lambda Sx$. Now, $Sx \neq 0$, since otherwise $0 = T(Sx) = x$ in violation of the hypothesis. Thus $\lambda \in \sigma(SAT)$ with associated eigenvector Sx. □

The hypothesis of Theorem 3.2 is highly restrictive because it would be coincidental for A to have an eigenvector x, which was also an eigenvector of TS associated with $1 \in \sigma(TS)$. However, a consequence of the theorem is to guarantee that for any irreducible input-output matrix A, and *any* partitioning matrix S, there always exists a standard aggregation of A that preserves the Perron eigenvalue of A. Moreover, the Perron eigenvector of the aggregated matrix is the aggregation of the Perron eigenvector of A.

COROLLARY 3.1. *Suppose that $A \in M_n$ is an irreducible input-output matrix with Perron eigenvalue λ and eigenvector x, and that $S \in M_{m,n}$ is a partitioning matrix. Then there exists a column stochastic matrix $T \in M_{n,m}$ such that SAT is a standard aggregation with Perron eigenvalue λ and eigenvector Sx.*

Proof. Since A is an irreducible square nonnegative matrix, A has a unique positive eigenvalue λ, called the Perron eigenvalue, with an associated positive eigenvector x. Let $T = \text{diag}\,(x)S^T\{\text{diag}\,(Sx)\}^{-1}$. The positivity of x implies that Sx is positive, which implies that $\{\text{diag}\,(Sx)\}^{-1}$ is a positive diagonal matrix. Thus T is nonnegative since it is the product of three nonnegative matrices. Moreover, T can be confirmed to be column stochastic:

$$e^T[\operatorname{diag}(x)S^T\{\operatorname{diag}(Sx)\}^{-1}]=x^TS^T\{\operatorname{diag}(Sx)\}^{-1}=e^T.$$

Finally, since T is the product of S^T and two diagonal matrices, $t_{i,j}$ can be positive only if $s_{j,i}$ is 1. Thus SAT is a standard aggregator.

The following computation shows that x is an eigenvector associated with a unit eigenvalue of TS:

$$TSx=[\operatorname{diag}(x)S^T\{\operatorname{diag}(Sx)\}^{-1}]Sx=\operatorname{diag}(x)S^Te=\operatorname{diag}(x)e=x.$$

Consequently, Theorem 3.2 guarantees that $\lambda\in\sigma(SAT)$ with eigenvector Sx. □

The significance of the Perron eigenvalue and eigenvector of an input-output matrix is discussed in many references, among them Takayama (1985) and Woods (1978). The spectral radius of a matrix A, denoted $\rho(A)$, equals $\max\{|\lambda|: \lambda\in\sigma(A)\}$. When A is nonnegative and irreducible, $\rho(A)$ is the Perron eigenvalue of A, so Corollary 3.1 could be restated in terms of the preservation of the spectral radius.

The previous corollary can be compared to the analyses of McManus (1956) and Morimoto (1971), whose results for weighted aggregation (mentioned in § 2) establish that any irreducible input-output matrix can be aggregated to a one-by-one scalar matrix so as to preserve the Perron eigenvalue. Corollary 3.1 establishes that any irreducible input-output matrix can be aggregated to an m-by-m matrix for any given partitioning matrix, so as to preserve the Perron eigenvalue and eigenvector.

Irreducibility is a technically useful concept in the analysis of input-output matrices. We next investigate how a standard aggregator can affect the reducibility of a matrix. It is straightforward to construct both examples in which an input-output matrix A is reducible but a standard aggregation of A is not, and also examples in which the reverse is true. For example, in the following expression A is reducible, but SAT is irreducible:

$$SAT=\begin{bmatrix}1&0&0\\0&1&1\end{bmatrix}\begin{bmatrix}.1&.1&.1\\.1&.1&.1\\0&0&.1\end{bmatrix}\begin{bmatrix}1&0\\0&.5\\0&.5\end{bmatrix}=\begin{bmatrix}.1&.1\\.1&.15\end{bmatrix}.$$

In the following, A is irreducible, but SAT is reducible:

$$SAT=\begin{bmatrix}1&1&0&0\\0&0&1&1\end{bmatrix}\begin{bmatrix}.1&.1&.1&.1\\.1&.1&.1&.1\\0&.1&.1&.1\\0&.1&.1&.1\end{bmatrix}\begin{bmatrix}1&0\\0&0\\0&.5\\0&.5\end{bmatrix}=\begin{bmatrix}.2&.2\\0&.2\end{bmatrix}.$$

Examination of the latter example shows that it is somewhat special, in that the null second row of T causes the 3, 2 and 4, 2 entries in A to have no weight in SAT. Indeed, we can prove the following theorem, which indicates that a standard aggregator maps irreducible matrices to irreducible matrices, except in extreme cases such as the previous example.

THEOREM 3.3. *Suppose that $A\in M_n$ is an input-output matrix, $S\in M_{m,n}$ is a partitioning matrix, $T\in M_{n,m}$ is column stochastic, and $ST=I$. Suppose for all $i\in N$ and $j\in M$ that $t_{i,j}$ is positive whenever $s_{j,i}$ is 1. If A is irreducible then SAT is irreducible.*

Proof. Let $h: N\rightarrow M$ be the function representation of S. Consider $S(E_{i,j})T$ for some $i,j\in N$. Since S is column stochastic, it must have a 1 somewhere in every column, so $SE_{i,j}$ cannot be zero. Because T has a positive entry in every position where S^T has a 1, T must have a positive entry somewhere in every row, so $SE_{i,j}T$ cannot be the zero matrix. As a consequence, since SAT is a linear function of A, we have that, if $a_{i,j} > 0$ then $(SAT)_{h(i),h(j)} > 0$.

Let $p, q \in M$. Pick $i \in h^{-1}(p)$ and $j \in h^{-1}(q)$. It is necessary and sufficient for A to be irreducible such that, for any $i, j \in N$, there must exist a sequence $i = k_1, k_2, \cdots, k_l = j$ in N such that $a_{k_t,k_{t+1}} > 0$ for all $t = 1, \cdots, l-1$. Consider such a sequence. The preceding paragraph demonstrates that $(SAT)_{h(k_t),h(k_{t+1})} > 0$, so $p = h(k_1), h(k_2), \cdots, h(k_l) = q$ is such a sequence in M for SAT. Thus SAT is irreducible. □

Thus a standard aggregator may create irreducibility, but it may not destroy irreducibility unless there is a micro commodity with a zero weight in T. It is possible to revise Theorem 3.3 to cover the case where A is reducible, with at least one positive number in every row, and with $l > 1$ irreducible subcomponents given by the partition of $N_1, \cdots, N_l$ of N. Consider a graph G with l vertices representing the irreducible subcomponents $N_1, \cdots, N_l$ and with an (undirected) arc connecting N_i and N_j if and only if there is a macro good that contains a micro good in N_i and a micro good in N_j. It can be shown that SAT is irreducible if and only if G is a connected graph.

We note that examples may easily be constructed to show that $\rho(SAT)$ may be larger or smaller than $\rho(A)$ for a standard aggregation SAT of A.

4. Aggregation error. In this section, we will (i) describe the errors that can result from aggregation, (ii) present bounds on the magnitude of the errors, (iii) discuss the special case when aggregation error is identically zero, and (iv) note that aggregation error will always be zero for some vectors though not necessarily all.

Input-output matrices are used in numerous different sorts of calculations, and aggregation can have different effects on the accuracy of each. The two most common uses involve (a) computing a final demand vector from a given output vector, and (b) computing an output vector from a given final demand vector. A final demand vector for the macro commodities can be obtained as $S(I-A)x$, the aggregation of the expression obtained from the unaggregated A matrix, or as $(I-B)Sx$, the output vector resulting from use of the standard aggregation $B = SAT$. Similarly, an output vector for the macro commodities can be obtained as $S(I-A)^{-1}y$ or as $(I-B)^{-1}Sy$. In general, the former member of each of these two pairs of expressions is not the same as the latter, and their differences may be regarded as errors resulting from the particular aggregation.

DEFINITION 4.1. Let $A \in M_n$ and $B \in M_m$ be input-output matrices, and suppose that $S \in M_{m,n}$ is a partitioning matrix. For $x \in R^m$ the m-dimensional vector

$$S(I-A)x-(I-B)Sx=[BS-SA]x$$

is called *type* a *error*, and, for $y \in R^n$, the m-dimensional vector

$$S(I-A)^{-1}y-(I-B)^{-1}Sy=[S(I-A)^{-1}-(I-B)^{-1}S]y$$

is called *type* b *error*. If both type a and type b aggregation errors are identically zero for all x and y, then B will be said to be a *zero error aggregation* of A associated with S.

Type a error results from the transformation of an output vector to a final demand vector and type b error results from the reverse transformation. Note that a Neumann expansion of type b error yields

$$S(I-A)^{-1}y-(I-B)^{-1}Sy=(BS-SA)y+(B^2S-SA^2)y+\cdots.$$

The usual terminology is to call type b error "total aggregation bias" and type a error, "first-order aggregation bias," which seems to imply that the transformation of output to final demand is inconsequential. We have adopted our more neutral terminology to reflect the fact that input-output analyses frequently involve both types of transformations.

What bounds can be established for relating the magnitudes of type a and b errors? Let $\|\cdot\|$ be any vector norm, and let $|||\cdot|||$ be the matrix norm induced by $\|\cdot\|$, i.e., for all $F \in M_n$

$$|||F||| = \max_{\|x\|=1} \|Fx\|.$$

(See, e.g., Horn and Johnson (1985).)

LEMMA 4.1. *Let $A \in M_n$ and $B \in M_m$ be input-output matrices, and suppose that $S \in M_{m,n}$ is a partitioning matrix. Then,*

$$\frac{\|(BS-SA)x\|}{\|x\|} \leqq |||BS-SA|||$$

and

$$\frac{\|[S(I-A)^{-1}-(I-B)^{-1}]y\|}{\|y\|} \leqq |||S(I-A)^{-1}-(I-B)^{-1}S|||.$$

The lemma follows immediately from the definition of an induced norm. The first expression is a bound on the size of type a error relative to the size of the output vector, and the second is a bound on the size of type b error relative to the size of the final demand vector. It is also clear from the definition of an induced norm that each bound is tight. The following theorem establishes a relationship between these two bounds.

THEOREM 4.1. *Let $A \in M_n$ and $B \in M_m$ be input-output matrices, and suppose that $S \in M_{m,n}$ is a partitioning matrix. Then,*

$$|||BS-SA||| \leqq |||I-B|||\;|||S(I-A)^{-1}-(I-B)^{-1}S|||\;|||I-A|||$$

and

$$|||S(I-A)^{-1}-(I-B)^{-1}S||| \leqq |||(I-B)^{-1}|||\;|||BS-SA|||\;|||(I-A)^{-1}|||.$$

Proof. To establish the first expression, we have

$$\begin{aligned} |||BS-SA||| &= |||(I-B)S-S(I-A)||| \\ &= |||(I-B)[S(I-A)^{-1}-(I-B)^{-1}S](I-A)||| \\ &\leqq |||I-B|||\;|||S(I-A)^{-1}-(I-B)^{-1}S|||\;|||I-A||| \end{aligned}$$

where the inequality follows from the submultiplicative property of induced matrix norms. To establish the second, we have

$$\begin{aligned} |||S(I-A)^{-1}-(I-B)^{-1}S||| &= |||(I-B)^{-1}[(I-B)S-S(I-A)](I-A)^{-1}||| \\ &= |||(I-B)^{-1}(-BS+SA)(I-A)^{-1}||| \\ &\leqq |||(I-B)^{-1}|||\;|||BS-SA|||\;|||(I-A)^{-1}|||. \end{aligned}$$ □

A simple example confirms that the bounds in the previous theorem are tight. Suppose $|||\cdot|||$ is the maximum absolute column sum norm, and let

$$A=\begin{bmatrix} .5 & 0 \\ 0 & 0 \end{bmatrix}, \quad S=[1 \quad 1], \quad T^{(1)}=\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad T^{(2)}=\begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

If the standard aggregation associated with S and $T^{(1)}$ is used, then both sides of the first inequality have a value of .5. If $T^{(2)}$ is used in place of $T^{(1)}$, then both sides of the second inequality have a value of 1. (The null column indicates that the second commodity is produced with only primary inputs, such as capital and labor. The example can also be regarded as a limiting case where the zeros are replaced by a small number ε.)

In addition to establishing a relationship between the magnitude of type a and type b errors, Theorem 4.1 has the following useful corollary.

COROLLARY 4.1. *Let $A \in M_n$ and $B \in M_m$ be input-output matrices, and suppose that $S \in M_{m,n}$ is a partitioning matrix. The following conditions are equivalent*:

(i) *B is a zero error aggregation of A associated with S.*

(ii) *Type* a *error is identically zero.*

(iii) *Type* b *error is identically zero.*

(iv) $BS = SA$.

(v) $S(I-A)^{-1} = (I-B)^{-1}S$.

Any of these conditions imply the following:

(vi) $(I-B)^{-1} = S(I-A)^{-1}T$.

(vii) *$B = SAT$ in which $T \in M_{n,m}$ is any column stochastic matrix with $ST = I$.*

Proof. (i) ⇒ (ii). This follows from Definition 4.1.

(ii) ⇒ (iii). If type a error is identically zero, i.e., $(BS - SA)x = 0$ for all x, then $|||BS - SA||| = 0$. Consequently, the second inequality in Theorem 4.1 implies that $|||S(I-A)^{-1} - (I-B)^{-1}S||| = 0$, so we have that $(S(I-A)^{-1} - (I-B)^{-1}S)y = 0$ for all y. That is, type b error is identically zero.

(iii) ⇒ (iv). Just as in the previous paragraph, if type b error is identically zero, then $|||S(I-A)^{-1} - (I-B)^{-1}S||| = 0$, so the first inequality in Theorem 4.1 implies that $|||BS - SA||| = 0$, which implies that $BS = SA$.

(iv) ⇒ (v). If $BS = SA$, then $|||BS - SA||| = 0$, so by Theorem 4.1

$$|||S(I-A)^{-1} - (I-B)^{-1}S||| = 0,$$

and hence $S(I-A)^{-1} = (I-B)^{-1}S$.

(v) ⇒ (i). If $S(I-A)^{-1} = (I-B)^{-1}S$ then type b error is identically zero, and hence type a error is identically zero, so B is a zero error aggregation of A associated with S.

(v) ⇒ (vi). Note that, postmultiplying (v) by T yields (vi), so that any of the five equivalent conditions implies (vi). Similarly, (iv) ⇒ (vii). □

Variants of the conditions in Corollary 4.1 have been used for some time. Hatanaka (1952), McManus (1956), and Charnes and Cooper (1961) all note the equivalence of (ii) and (iv). Morimoto (1970) and Ara (1959) prove that (iii) and (iv) are equivalent. Kossov (1970) proves that (ii) implies (iv). Miller and Blair (1985) and Bulmer-Thomas (1982) prove that (iv) implies (v).

Since (vii) follows from (i), if we wish to find aggregations of A that have zero aggregation error (if they exist), we need not look beyond the standard aggregations of A. In particular, if a correct S is known, then any column stochastic T with $ST = I$ will produce the zero error (standard) aggregation B.

In § 3, we have discussed the relationship between matrix similarity and standard aggregators, and presented cases in which a standard aggregator preserves part of the spectrum of A. The following theorem shows that if B is a zero error aggregation of A associated with S, then $\sigma(B) \subset \sigma(A)$. Thus the likeness between standard aggregators and matrix similarity is particularly compelling in the case of zero error aggregation. (Part (iii) was noted by Ara (1959).)

THEOREM 4.2. *Let $A \in M_n$ and $B \in M_m$ be input-output matrices, and suppose that $S \in M_{m,n}$ is a partitioning matrix. Suppose that B is a zero error aggregation of A associated with S. Then we have the following*:

(i) *If* $\lambda \in \sigma(A)$ *with right eigenvector* x, *then either* $Sx = 0$ *or* $\lambda \in \sigma(B)$ *with right eigenvector* Sx.

(ii) *If* $\lambda \in \sigma(B)$ *with left eigenvector* z, *then* $\lambda \in \sigma(A)$ *with left eigenvector* S^Tz.

(iii) $\rho(A) = \rho(B)$.

Proof. To prove statement (i), suppose that $\lambda \in \sigma(A)$ with right eigenvector $x \neq 0$. Then we have $Ax = \lambda x$, which implies that $SAx = \lambda Sx$. But, since B is a zero error aggregation of A associated with S, Corollary 4.1 states that $BS = SA$, so we have $BSx = \lambda Sx$. Therefore, either Sx must equal zero or $\lambda \in \sigma(B)$ with right eigenvector Sx.

For statement (ii), suppose that $\lambda \in \sigma(B)$ with left eigenvector $z \neq 0$. Then we have $z^TB = \lambda z^T$, which implies that $z^TBS = \lambda z^TS$. But, since $BS = SA$, we have that $z^TSA = \lambda z^TS$. The matrix S is a 0, 1 column stochastic matrix with at least one 1 in every row, so the entries of the vector z^TS are precisely the entries of z^T with every entry appearing at least once. Thus $z^TS \neq 0$. Therefore, $\lambda \in \sigma(A)$ with left eigenvector S^Tz.

Statement (iii) is an immediate consequence of statement (i). The Perron eigenvector x of A is nonnegative, so Sx cannot equal 0, so the Perron root of B is the same as that of A. □

Our final remarks on zero error aggregation demonstrate how to determine from an input-output matrix A whether there exists a partitioning matrix S such that there is a zero error aggregation of A associated with S. For given matrices A and S, we know that such an aggregation exists if and only if there is a matrix B that satisfies $BS = SA$. The following theorem gives an interpretation of the latter condition. Let $A[h^{-1}(k), h^{-1}(l)]$ denote the (not necessarily square) submatrix of A obtained by including the rows in the set $h^{-1}(k)$ and the columns in the set $h^{-1}(l)$.

THEOREM 4.3. *Suppose that* $A \in M_n$ *is an input-output matrix and* $S \in M_{m,n}$ *is a partitioning matrix, and let* $h: N \rightarrow M$ *be the function representation of* S. *There is a zero error aggregation of* A *associated with* S *if and only if there is a constant* $b_{k,l}$ *for each* $k, l \in M$ *such that*

$$e^TA[h^{-1}(k), h^{-1}(l)] = b_{k,l}e^T.$$

Furthermore, in this event, any standard aggregation of A *corresponding to* S *has zero aggregation error and results in the matrix* $B = (b_{k,l}) \in M_m$.

Proof. Let $B = (b_{k,l}) \in M_m$ be a zero error standard aggregation of A associated with S. By Corollary 4.1, $BS = SA$. But this is exactly the condition asserted in the theorem. Note that this condition is independent of the matrix T.

On the other hand, if there is a constant $b_{k,l}$ for each $k, l \in M$ such that

$$e^TA[h^{-1}(k), h^{-1}(l)] = b_{k,l}e^T,$$

then $BS = SA$ for the S given as the matrix representation of h and for B defined by $B = (b_{k,l})$. Then for any T satisfying $ST = I$ and this S, B is a standard aggregation of A. □

Theorem 4.3 suggests a simple algorithm to determine, for a given input-output matrix A, the smallest number of macro commodities m, for which there exists a partitioning matrix S and a standard aggregation B, in which B is a zero error aggregation of A associated with S. The algorithm operates by taking a sequence of successively finer refinements of a given partition, until the condition stated in Theorem 4.3 is satisfied. The algorithm proceeds as follows.

In step 0, partition $N = \{1, \cdots, n\}$ into one set, forming the partition $P^{(0)} = \{P_1^{(0)}\}$ with $P_1^{(0)} = N$.

In step r, use $P^{(r-1)}$ to partition the matrix A, and compute the column sums within every resulting submatrix of A. Refine $P^{(r-1)}$ to form $P^{(r)} = \{P_1^{(r)}, P_2^{(r)}, \cdots, P_{k_r}^{(r)}\}$ so that all $j \in P_i^{(r)}$ have common submatrix column sums. That is, obtain $P^{(r)}$ so that

$$\sum_{i \in P_l^{(r-1)}} a_{i,j} \quad \text{is constant for all } j \in P_s^{(r)}$$

for all $s = 1, \cdots, k_r$ and for all $l = 1, \cdots, k_{r-1}$. Continue to step $r + 1$ if $P^{(r)}$ differs from $P^{(r-1)}$ and, otherwise, stop.

Clearly in the usual case for an input-output matrix, the above algorithm would not stop until the partition contained n singleton sets, which would indicate that any standard aggregation of A would have (both type a and type b) aggregation error.

The algorithm must produce a partition that satisfies the requirements of the Theorem 4.3 because refinements are carried out until all of the submatrices of A have constant column sums. Note, however, that the refinement process only separates commodities if such a separation is necessary for zero error aggregation. Consequently, the algorithm produces the smallest possible number of macro commodities consistent with a zero error aggregation. There may, of course, be other zero error aggregations that have larger m.

To illustrate the algorithm, consider the following input-output matrix.

$$A = \begin{bmatrix} .2 & .1 & 0 & 0 & 0 \\ 0 & .3 & 0 & 0 & .2 \\ .2 & 0 & 0 & .3 & .2 \\ 0 & 0 & .3 & 0 & .2 \\ 0 & 0 & .1 & .1 & .2 \end{bmatrix}.$$

Step 0. Let $P^{(0)}$ be the partition containing only one set, $\{1, 2, 3, 4, 5\}$.

Step 1. The column sums over A are equal for the first four columns, but not for the fifth, so $P^{(1)}$ consists of two sets $P_1^{(1)} = \{1, 2, 3, 4\}$ and $P_2^{(1)} = \{5\}$.

Step 2. Partition A into submatrices according to $P^{(1)}$, to obtain the following matrix.

$$A = \left[\begin{array}{cccc|c} .2 & .1 & 0 & 0 & 0 \\ 0 & .3 & 0 & 0 & .2 \\ .2 & 0 & 0 & .3 & .2 \\ 0 & 0 & .3 & 0 & .2 \\ \hline 0 & 0 & .1 & .1 & .2 \end{array}\right].$$

All of the submatrices have constant column sums except the ones in the 1, 1 and 2, 1 positions. Thus $P^{(2)}$ consists of three sets $P_1^{(2)} = \{1, 2\}$, $P_2^{(2)} = \{5\}$, $P_3^{(2)} = \{3, 4\}$.

Step 3. Partition A into submatrices according to $P^{(2)}$ to obtain the following matrix:

$$A = \left[\begin{array}{cc|cc|c} .2 & .1 & 0 & 0 & 0 \\ 0 & .3 & 0 & 0 & .2 \\ \hline .2 & 0 & 0 & .3 & .2 \\ 0 & 0 & .3 & 0 & .2 \\ \hline 0 & 0 & .1 & .1 & .2 \end{array}\right].$$

All of the submatrices have constant column sums except the ones in the 1, 1 and the 2, 1 positions. Thus $P^{(3)}$ consists of the four sets $P_1^{(3)} = \{1\}$, $P_2^{(3)} = \{5\}$, $P_3^{(3)} = \{3, 4\}$, $P_4^{(3)} = \{2\}$.

Step 4. Using $P^{(3)}$ to partition A yields submatrices with constant column sums, so no additional refinements are required. Thus

$$B=\begin{bmatrix} .2 & .1 & 0 & 0 \\ 0 & .3 & 0 & .2 \\ .2 & 0 & .3 & .4 \\ 0 & 0 & .1 & .2 \end{bmatrix}, \qquad S=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

yield a zero error aggregation of A for S.

The preceding discussion gives conditions for aggregation error to be zero, regardless of the choice of x and y. Those conditions were restrictive and most input-output matrices would not satisfy them for any choice of S. However, for any standard aggregation, there are always subspaces within which type a and type b errors are zero.

THEOREM 4.4. *Suppose that* $A \in M_n$ *is an input-output matrix*, $S \in M_{m,n}$ *is a partitioning matrix*, $T \in M_{n,m}$ *is column stochastic, and* $ST = I$. *There exists a subspace of* R^n, *of dimension at least* m, *such that type* a *error is zero for all* x *in the subspace. There exists a subspace of* R^n, *of dimension at least* m, *such that type* b *error is zero for all* y *in the subspace.*

Proof. The matrix TS is idempotent, with rank m, so 1 occurs in $\sigma(TS)$ and is associated with an m-dimensional subspace, X, of eigenvectors. Let $x \in X$, so $TSx = x$. Note that type a error is

$$(BS-SA)x = SATSx - SAx = SAx - SAx = 0.$$

Thus type a error is zero for x taken from the m-dimensional subspace X.

Now, suppose that y is taken from the m-dimensional subspace given by $(I-A)X$, the image of X under multiplication by $(I-A)$. Let $x=(I-A)^{-1}y$. Then type b error equals

$$\begin{aligned}[S(I-A)^{-1}-(I-B)^{-1}S]y &= S(I-A)^{-1}y-(I-SAT)^{-1}Sy \\ &= (I-SAT)^{-1}[(I-SAT)STSx-S(I-A)x] \\ &= 0\end{aligned}$$

where the third equality follows because $TSTSx = x$. □

The proof of Theorem 4.4 shows that type a error is zero for all x that satisfy $TSx = x$. Which x are these? Let $e_i \in R^m$ denote the ith vector in the "standard" basis of R^m, so $Te_i = T_i$, the ith column of the matrix T. Note that

$$TS(T_i) = TS(Te_i) = T(ST)e_i = Te_i = T_i.$$

Thus type a error is zero for all $x \in R^n$ in the subspace spanned by $T_1, \cdots, T_m$. That is, type a error is zero for all output vectors whose micro commodities are output in the proportions given by the columns of the T matrix. Ordinarily negative output of a commodity is meaningless so only the nonnegative polyhedral cone generated by $T_1, \cdots,$ T_m would be of interest.

Similarly, type b error is zero for all y in the subspace spanned by $(I-A)T_1, \cdots,$ $(I-A)T_m$. This latter subspace will usually include final demand vectors that have mixed sign, but that are readily interpretable in a model that is open with respect to international trade, since negative final demand for a commodity simply indicates that the commodity is predominantly imported.

Balderston and Whitin (1954) and others have demonstrated that type a error is zero on a one-dimensional subspace. Morimoto (1970) has noted that type a error is zero on an m-dimensional subspace.

REFERENCES

K. ARA (1959), *The aggregation problem in input-output analysis*, Econometrica, 27, pp. 257–262.

J. B. BALDERSTON AND T. M. WHITIN (1954), *Aggregation in the input-output model*, in Economic Activity Analysis, Oskar Morgenstern, ed., John Wiley, New York, pp. 79–128.

V. BULMER-THOMAS (1982), *Input-Output Analysis in Developing Countries: Sources, Methods, and Applications*, John Wiley, New York.

A. CHARNES AND W. W. COOPER (1961), *Management Models and Industrial Applications of Linear Programming*, Vol I, John Wiley, New York.

J. S. CHIPMAN (1976), *Estimation and aggregation in economics: an application of the theory of generalized inverses*, Generalized Inverses and Applications, in M. Zuhair Nashed, ed., Academic Press, New York.

W. D. FISHER (1958), *Criteria for aggregation in input-output analysis*, Rev. Econom. Statist., 40, pp. 250–260.

——— (1969), *Clustering and Aggregation in Economics*, Johns Hopkins Press, Baltimore, MD.

S. GERKING (1976), *Input-output as a simple econometric model*, Rev. Econom. Statist., 58, pp. 274–282.

M. HATANAKA (1952), *Note on consolidation within a Leontief system*, Econometrica, 20, pp. 301–303.

R. A. HORN AND C. R. JOHNSON (1985), *Matrix Analysis*, Cambridge University Press, Cambridge.

Y. IJIRI (1968), *The linear aggregation coefficient as the dual of the linear correlation coefficient*, Econometrica, 36, pp. 252–259.

——— (1971), *Fundamental queries in aggregation theory*, J. Amer. Statist. Assoc., 66, pp. 766–782.

L. JOHANSEN (1961), *A note on "aggregation in Leontief matrices and the labour theory of value"*, Econometrica, 29, pp. 221–222.

V. KOSSOV (1970), *The theory of aggregation in input-output models*, in Contributions to Input-Output Analysis, A. P. Carter and A. Brody, eds., North-Holland, Amsterdam.

K. O. KYMN (1977), *Interindustry energy demand and aggregation of input-output tables*, Rev. Econom. Statist., 59, pp. 371–374.

W. LEONTIEF (1967), *An alternative to aggregation in input-output analysis and national accounts*, Rev. Econom. Statist., 49, pp. 412–419.

E. MALINVAUD (1956), *Aggregation problems in input-output models*, in The Structural Interdependence of the Economy, Tibor Barna, ed., John Wiley, New York, pp. 189–202.

M. MCMANUS (1956), *General consistent aggregation in Leontief models*, Yorkshire Bull. Econom. Social Res., 8, pp. 28–48.

R. E. MILLER AND P. D. BLAIR (1985), *Input-Output Analysis: Foundations and Extensions*, Prentice-Hall, Englewood Cliffs, NJ.

Y. MORIMOTO (1970), *On aggregation problems in input-output analysis*, Rev. Econom. Stud., 37, pp. 119–126.

——— (1971), *A note on weighted aggregation in input-output matrices*, Internat. Econom. Rev., 12, pp. 138–143.

M. MORISHIMA AND F. SETON (1961), *Aggregation in Leontief matrices and the labour theory of value*, Econometrica, 29, pp. 203–220.

H. NEUDECKER (1970), *Aggregation in input-output analysis: an extension of Fisher's method*, Econometrica, 38, pp. 921–926.

A. RUBINSTEIN AND P. C. FISHBURN (1986), *Algebraic aggregation theory*, J. Economic Theory, 38, pp. 63–77.

A. TAKAYAMA (1985), *Mathematical Economics*, 2nd ed., Cambridge University Press, Cambridge.

H. THEIL AND P. URIBE (1967), *The information approach to the aggregation of input-output tables*, Rev. Econom. Statist., 49, pp. 451–461.

R. WILSON (1975), *On the theory of aggregation*, J. Economic Theory, 10, pp. 89–99.

J. E. WOODS (1978), *Mathematical Economics: Topics in Multi-Sectoral Economics*, Longman, London.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 1, pp. 80-93, January 1989

006

LOCALIZATION CRITERIA AND CONTAINMENT FOR RAYLEIGH QUOTIENT ITERATION*

CHRISTOPHER BEATTIE† AND DAVID W. FOX‡

Abstract. Rayleigh quotient iteration can often yield an eigenvalue-eigenvector pair of a positive-definite Hermitian problem in a very short time. The primary hindrance associated with its use as a regular computational tool lies with the difficulty of identifying and selecting the final regions of convergence. In this paper rigorous, accessible criteria for localizing Rayleigh quotient iteration to prespecified intervals of the spectrum are provided, as well as extensions to situations where only partial spectral information is available. An application for finding partial eigensolutions of symmetric tridiagonal matrices is given with results that compare very favorably with the EISPACK routine TSTURM.

Key words. symmetric matrix, eigenvalues, Raleigh quotient iteration

AMS(MOS) subject classifications. 15A18, 65F15

1. Introduction.

The generalized matrix eigenvalue problem

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{B}\mathbf{x} \tag{1.1}$$

with **A** and **B** Hermitian and **B** positive definite occurs frequently in science and engineering. Typically the matrices **A** and **B** are sufficiently large and sparse that a full spectral decomposition of (1.1) is inconvenient or too expensive, and ultimately only a relatively few eigenpairs satisfying (1.1) are needed. In such circumstances a wide variety of iterative methods may be considered (e.g., see [14]).

In this article we focus on the use of a simple vector iteration defined recursively through the solution of

$$(\mathbf{A} - \sigma_s \mathbf{B})\mathbf{x}_{s+1} = \omega_{s+1}\mathbf{B}\mathbf{x}_s, \qquad s = 0, 1, 2, \cdots. \tag{1.2}$$

The positive scaling factor ω_{s+1} is determined so that $\mathbf{x}_{s+1}^*\mathbf{B}\mathbf{x}_{s+1} = 1$. Now, if the shift σ_s is set to a fixed value γ, then (1.2) defines the usual *inverse iteration*. Excluding certain unstable situations, this will yield vector iterates that converge linearly to an eigenvector of (1.1) associated with the eigenvalue closest to γ. If, on the other hand, the shift is reset at every step to the Rayleigh quotient

$$\mu_s \equiv \mathbf{x}_s^* \mathbf{A}\mathbf{x}_s,$$

then (1.2) defines *Rayleigh quotient iteration*. In striking contrast to the linear convergence of inverse iteration, Rayleigh quotient iteration has an asymptotically *cubic* rate of convergence, though the vector iterates generally do *not* converge to an eigenvector corresponding to the eigenvalue closest to the initial shift μ_0.

In spite of the remarkable potential advantage that its speed may afford, in practice Rayleigh quotient iteration as an independent vector iteration method is rarely encountered, although it does lie at the heart of QL-QR algorithms commonly used for small to middle-sized matrix eigenvalue problems. This general disuse may be justified for two reasons: erratic convergence behavior of Rayleigh quotient iteration, and the considerable

[*] Received by the editors November 2, 1987; accepted for publication (in revised form) June 29, 1988.

[†] Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061. This research was conducted while the author was a visiting fellow at the Minnesota Supercomputer Institute and a visiting faculty member at the Department of Computer Science at the University of Minnesota, Minneapolis, Minnesota 55455.

[‡] Department of Computer Science, University of Minnesota, Minneapolis, Minnesota 55455.

expense of matrix factorizations needed in solving (1.2) repeatedly for consecutive shifts. Although Rayleigh quotient iteration is globally convergent for Hermitian problems, there is no assurance that closure will occur to an eigenvalue in the region of the initial shift or even that the sequence of Rayleigh quotient iterates will be restricted a priori to any given interval of the spectrum [10]. This issue has an aggravating influence on the latter point, which otherwise should not be overstated. While circumstances exist where it is critical to minimize the total number of matrix factorizations (as in [6]), in many other circumstances this overhead is acceptable as we may deduce from the evident utility of algorithms that utilize Sturm sequences or spectrum slicing to drive bisection or "determinant search" strategies (cf. [1]). For these methods the complexity of a single iteration is often comparable with that needed for a Rayleigh quotient iteration, yet we are left with a far slower asymptotic convergence rate.

The apparently unpredictable behavior of Rayleigh quotient iteration reflects the unsuitability of the Rayleigh quotient by itself as an indicator of alignment of a trial vector with any particular reducing space. However, Rayleigh quotient iteration will converge predictably to an eigenvalue within a given interval, provided that some trial vector $\mathbf{x}_k$, becomes well aligned with the reducing space of the eigenvectors corresponding to the eigenvalues contained in that interval. Such a situation could evolve, for example, if we precede Rayleigh quotient iteration with inverse iteration. In particular, to locate eigenvalues in the interval $[\gamma - \eta, \gamma + \eta] \equiv \mathbf{I}_\gamma(\eta)$ we may do inverse iteration with $\sigma_s = \gamma$ until the component of the iterated vector lying in the reducing space associated with the spectral interval $\mathbf{I}_\gamma(\eta)$ has been sufficiently magnified that subsequent Rayleigh quotient iteration converges in $\mathbf{I}_\gamma(\eta)$. Since the eigenvector components are a priori unknown, the question becomes: "How can we use observable quantities in order to determine when this component has been sufficiently magnified?"

Various ad hoc strategies have been proposed that try to avoid this difficult question. Typically inverse iteration is done a fixed number of times, and then Rayleigh quotient iteration is attempted. If it goes astray, then inverse iteration is restarted. This strategy can be found as early as 1958 [5]. Its shortcomings are apparent: there is no way of knowing if the inverse iterations are more numerous than necessary, thus delaying the onset of cubic convergence, or too few, leading to a uselessly wandering sequence of iterates (cf. [1]).

In reviewing Ostrowski's treatment of the local convergence behavior of Rayleigh quotient iteration [9a], [9b], we may observe that the explicit estimates of convergence neighborhoods for Rayleigh quotient iteration that are given may also be used to determine thresholds for switching from inverse iteration to Rayleigh quotient iteration. Unfortunately, the bounds are in terms of quantities involving the unknown eigenvectors, and hence they do not lead ultimately to useful switching criteria in practice. Some measure of a priori spectral information does appear necessary to derive rigorously valid switching criteria, however, and clearly such requirements must be modest if usable criteria are to be obtained.

Progress toward resolving these difficulties was made by Szyld and Widlund [18] and Szyld [16], [17], who first gave a rigorous switching criterion based on the isolation distance from $\mathbf{I}_\gamma(\eta)$ to the rest of the spectrum. Additionally, Szyld in [16], [17] formulated an ad hoc criterion for circumstances in which no a priori spectral information is available. Though this ad hoc criterion does fail in certain circumstances (in the sense that reversion to inverse iteration is necessary), it appears to work reasonably well in practice.

In this paper, we extend some of the previous results of Szyld, obtaining improved switching criteria that, on the one hand, make minimal demands on a priori spectral information and that, on the other, are best possible in a certain sense. We make fundamental use of the Kato–Temple inequalities in our derivation. Our results are illustrated in an implementation for resolving tridiagonal matrices that exhibit dramatic speedups over the comparable EISPACK routine, TSTURM.

In the following section we establish some additional notation and collect results that are used in our analysis. In § 3 we state our primary containment theorem (Theorem 1) and show how it is best possible using the information available. Section 4 discusses the localization of Rayleigh quotient iteration when the number of eigenvalues in an interval is available. In the final section we discuss application of our switching criteria and give some results of computational experiments.

2. Preliminaries. For an arbitrary **B**-normalized vector $\mathbf{x}_s$ with corresponding Rayleigh quotient $\mu_s = \mathbf{x}_s^* \mathbf{A}\mathbf{x}_s$ define the Rayleigh quotient residual vector

$$\mathbf{r}_s = (\mathbf{A} - \mu_s \mathbf{B})\mathbf{x}_s \tag{2.1}$$

and the inverse iteration residual vector

$$\mathbf{q}_s = (\mathbf{A} - \gamma \mathbf{B})\mathbf{x}_s. \tag{2.2}$$

The appropriate measure of magnitude for these quantities is the norm generated by $\mathbf{B}^{-1/2}$ [10, § 15.9], which we designate by $\|\cdot\|_b$. It is given by

$$\|\mathbf{x}\|_b = [\mathbf{x}^* \mathbf{B}^{-1} \mathbf{x}]^{1/2},$$

where * denotes the conjugate transpose. Thus

$$\|\mathbf{q}_s\|_b = [\mathbf{x}_s^* (\mathbf{A} - \gamma \mathbf{B})\mathbf{B}^{-1}(\mathbf{A} - \gamma \mathbf{B})\mathbf{x}_s]^{1/2} \tag{2.3}$$

and

$$\|\mathbf{r}_s\|_b = [\mathbf{x}_s^* (\mathbf{A} - \mu_s \mathbf{B})\mathbf{B}^{-1}(\mathbf{A} - \mu_s \mathbf{B})\mathbf{x}_s]^{1/2}. \tag{2.4}$$

It is important to note that $\|\mathbf{q}_s\|_b$ and $\|\mathbf{r}_s\|_b$ are available from information already computed during the course of an iteration step (1.2) with no need to solve an additional system involving **B**. In fact, for inverse iteration ($\sigma_s = \gamma$), premultiplication of (1.2) by $\mathbf{B}^{-1/2}$ shows immediately that

$$\|\mathbf{q}_{s+1}\|_b = \omega_{s+1}.$$

For either Rayleigh quotient iteration or inverse iteration (1.2) also yields

$$\|\mathbf{r}_{s+1}\|_b = [\omega_{s+1}^2 - (\sigma_s - \mu_{s+1})^2]^{1/2}.$$

Thus the norms of both residuals are available directly from the scaling ω_{s+1} with no additional work.

We note further that when $\sigma_s = \gamma$, premultiplication of (1.2) by $\mathbf{x}_s^*$ and the Cauchy–Schwarz inequality give

$$\|\mathbf{q}_{s+1}\|_b = |\mathbf{q}_s^* \mathbf{x}_{s+1}| \leqq \|\mathbf{q}_s\|_b.$$

Hence inverse iteration always produces a monotone decreasing sequence of inverse iteration residuals (see also [17]). Similarly, Rayleigh quotient iteration (with $\sigma_s = \mu_s$) always produces a monotone decreasing sequence of Rayleigh quotient residuals [10, § 4.8],

$$\|\mathbf{r}_{s+1}\|_b \leqq \|\mathbf{r}_s\|_b.$$

The monotonicity of these residuals is key in their use as thresholds for switching from inverse iteration to Rayleigh quotient iteration.

Before we start our analysis, we close this section with a few more useful facts about Rayleigh quotient iteration (see also [17] and [12]), which we state as a lemma.

LEMMA. *For Rayleigh quotient iteration* ($\sigma_s = \mu_s$) *we have*

$$\mu_{s+1} - \mu_s = (\mathbf{r}_s^* \mathbf{x}_{s+1})(\mathbf{x}_{s+1}^* \mathbf{B}\mathbf{x}_s) \tag{2.5}$$

and

$$\|\mathbf{r}_{s+1}\|^2 + (\mu_{s+1} - \mu_s)^2 = \omega_{s+1}^2 = |\mathbf{r}_s^* \mathbf{x}_{s+1}|^2. \tag{2.6}$$

Proof. Rewrite (1.2) as

$$(\mathbf{A} - \mu_{s+1}\mathbf{B})\mathbf{x}_{s+1} + (\mu_{s+1} - \mu_s)\mathbf{B}\mathbf{x}_{s+1} = \omega_{s+1}\mathbf{B}\mathbf{x}_s.$$

Premultiplication by $\mathbf{x}_{s+1}^*$ provides (2.5). Premultiplication by $\mathbf{B}^{-1/2}$ and the Pythagorean Theorem give (2.6).

3. Containment of Rayleigh quotient iterates. In this section we establish our central result on the containment of Rayleigh quotients. We show that if a Rayleigh quotient lies on one side of the midpoint of a known gap in the spectrum, then, provided that the residual is no greater than half of the gap width, all subsequent Rayleigh quotient iterates must lie on the same side of the midpoint. In an appropriate sense, this result is then shown to be the best possible. In all that follows, we will refer to the eigenvalues of (1.1) as the eigenvalues of **A** relative to **B**, or more simply as the eigenvalues of (**A**, **B**).

THEOREM 1. *Suppose that* (α, β) *is known to be a gap in the spectrum of* (**A**, **B**). *Let* μ *and* **r** *be the Rayleigh quotient and residual of the* **B**-*normalized vector* **x**. *Suppose* $\|\mathbf{r}_1\|_b \leqq (\beta - \alpha)/2$; *then we have the following*:

If $\mu_1 < (\alpha + \beta)/2$, *then* $\mu_s < (\alpha + \beta)/2$ *for all* $s \geqq 1$, and

if $\mu_1 > (\alpha + \beta)/2$, *then* $\mu_s > (\alpha + \beta)/2$ *for all* $s \geqq 1$.

Proof. We proceed by contradiction. Suppose that $\mathbf{x} = \mathbf{x}_1$ is such that $\mu_1 < (\alpha + \beta)/2$ and that $\|\mathbf{r}_1\|_b \leqq (\beta - \alpha)/2$. Let t be the lowest index for which $\mu_t < (\alpha + \beta)/2$ and $\mu_{t+1} \geqq (\alpha + \beta)/2$. Since $\|\mathbf{r}_s\|_b$ is nonincreasing with s, for all s we have

$$\|\mathbf{r}_s\|_b \leqq (\beta - \alpha)/2. \tag{3.1}$$

Since $(\lambda - \alpha)(\lambda - \beta) \geqq 0$ at each point λ of the spectrum of (**A**, **B**), we have from the spectral theorem (see also Kato [7]), for any **r** and μ corresponding to the **B**-normalized vector **x**

$$\|\mathbf{r}\|_b^2 + (\mu - \alpha)(\mu - \beta) \geqq 0.$$

In particular, since $\mu_t < (\alpha + \beta)/2$ this gives

$$\mu_t \leqq \frac{\alpha + \beta}{2} - \sqrt{\left(\frac{\beta - \alpha}{2}\right)^2 - \|\mathbf{r}_t\|_b^2}$$

and since $\mu_{t+1} \geqq (\alpha + \beta)/2$

$$\mu_{t+1} \geqq \frac{\alpha + \beta}{2} + \sqrt{\left(\frac{\beta - \alpha}{2}\right)^2 - \|\mathbf{r}_{t+1}\|_b^2}.$$

We subtract, square, and use $\|\mathbf{r}_t\|_b \geqq \|\mathbf{r}_{t+1}\|_b$ to obtain

$$(\mu_{t+1} - \mu_t)^2 \geqq (\beta - \alpha)^2 - 3\|\mathbf{r}_t\|_b^2 - \|\mathbf{r}_{t+1}\|_b^2.$$

Since $(\mu_{t+1} - \mu_t)^2 + \|\mathbf{r}_{t+1}\|_b^2 = |\mathbf{r}_t^* \mathbf{x}_{t+1}|^2$, we find

$$(\beta - \alpha)^2 \leqq 3\|\mathbf{r}_t\|_b^2 + |\mathbf{r}_t^* \mathbf{x}_{t+1}|^2 \leqq 4\|\mathbf{r}_t\|_b^2 \leqq (\beta - \alpha)^2,$$

where the center and right-hand inequalities are the result of the Cauchy–Schwarz inequality and of the bounds given by (3.1), respectively. Thus equality prevails and $\mathbf{r}_t$ must be collinear with $\mathbf{Bx}_{t+1}$. Now (2.5) yields the contradiction

$$\mu_{t+1} - \mu_t = (\mathbf{r}_t^* \mathbf{x}_{t+1})(\mathbf{x}_{t+1}^* \mathbf{B} \mathbf{x}_t) = \mathbf{r}_t^* \mathbf{x}_t = 0.$$

By a parallel argument, if $\mu > (\alpha + \beta)/2$, every Rayleigh quotient remains above $(\alpha + \beta)/2$. □

The inequalities in this theorem are the best possible in the sense that neither alone can be weakened. In fact, if the bound on $\|\mathbf{r}\|_b$ is weakened to $\|\mathbf{r}_1\|_b \leqq \kappa(\beta - \alpha)/2$ for any $\kappa > 1$, then the following counterexample shows that the theorem is false.

Suppose the known gap (α, β) is not optimal and that $\mathbf{A}$ has adjacent eigenvalues λ_- and λ_+ relative to $\mathbf{I}$ such that $\lambda_- < \alpha < \beta = \lambda_+$ (see Fig. 1). Define the unit vector $\mathbf{x}$ by

$$\mathbf{x} = [1 + \kappa]^{-1/2} \{ \mathbf{u}_- + \kappa^{1/2} \mathbf{u}_+ \}.$$

It is easy to verify that μ satisfies

$$\mu < (\alpha + \beta)/2 \quad \text{and} \quad \mu > (\lambda_+ + \lambda_-)/2$$

and that $\|\mathbf{r}\|$ satisfies

$$\|\mathbf{r}\| \leqq \kappa(\beta - \alpha)/2 \quad \text{and} \quad \|\mathbf{r}\| \leqq (\lambda_+ - \lambda_-)/2.$$

Thus the Rayleigh quotient iterates all have quotients lying above $(\lambda_+ + \lambda_-)/2$, and, in fact, the quotients converge to λ_+.

On the other hand, if the restriction on $\|\mathbf{r}\|_b$ is maintained, but $\mu_1 = (\alpha + \beta)/2$, all of the subsequent μ_s can remain at $(\alpha + \beta)/2$ as the following example shows. Suppose that $\mathbf{A}$ has the adjacent eigenvalues $\lambda_- = \alpha < \beta = \lambda_+$ relative to $\mathbf{I}$. The initial vector

$$\mathbf{x}_1 = 2^{-1/2}(\mathbf{u}_+ + \mu_-)$$

satisfies

$$\mu_1 = (\alpha + \beta)/2 \quad \text{and} \quad \|\mathbf{r}_1\| = (\beta - \alpha)/2;$$

however $\mu_s = (\alpha + \beta)/2$ for all $s \geqq 1$.

These examples give us the following theorem.

THEOREM 2. *Theorem* 1 *is optimal in the sense that if either condition is weakened there exist matrices and starting vectors that satisfy the relaxed criteria but have subsequent Rayleigh quotients that do not lie strictly on one side of the center of the specified gap.*

Note that if $\|\mathbf{r}_1\|_b < (\beta - \alpha)/2$, then μ_1 must lie strictly on one side or the other of the center of the specified gap, for there must be an eigenvalue within $\|\mathbf{r}_1\|_b$ of μ_1.

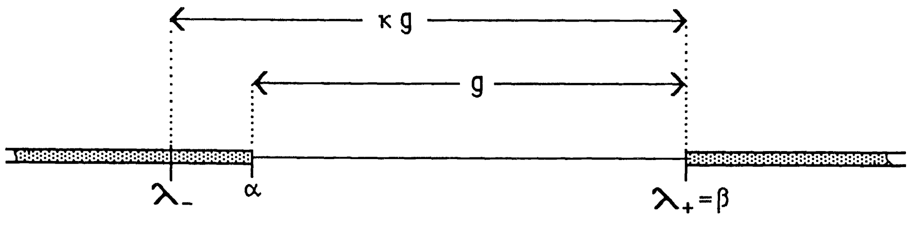


FIG. 1

4. Localization with partial spectral information. Now we are in a position to use results of the preceding section to establish conditions that ensure the convergence of Rayleigh quotient iteration to a point within a given interval. These results depend on determining, as part of the computation, gaps between eigenvalues which the sequence of Rayleigh quotients cannot cross.

Since our ultimate goal is to establish criteria for shifting from inverse iteration to Rayleigh quotient iteration as quickly as it is safe, we shall introduce notation here that is adapted to that purpose. If we have an interval $[a, b]$ containing eigenvalues of $(\mathbf{A}, \mathbf{B})$, we shall frequently use the center γ as the shift in the inverse iteration and denote by η the distance from the center to the endpoints, i.e., $\mathbf{I}_\gamma(\eta) = [a, b]$ with $\gamma = (a+b)/2$ and $\eta = (b-a)/2$.

To begin, suppose that the interval $[a, b]$ contains in its interior exactly one eigenvalue λ of $(\mathbf{A}, \mathbf{B})$ and that the Rayleigh quotient μ corresponding to the $\mathbf{B}$-normalized vector $\mathbf{x}$ lies in (a, b). Then the Kato–Temple inequalities (see [7] and [19]) imply

$$\mu - \|\mathbf{r}\|_b^2/(b-\mu) \equiv a' \leqq \lambda \leqq b' \equiv \mu + \|\mathbf{r}\|_b^2/(\mu - a),$$

which provides an interval $[a', b']$ that contains λ but is narrower than the original interval $[a, b]$ if $\|\mathbf{r}\|_b^2 < (b-\mu)(\mu - a)$.

When the Kato–Temple inequalities restrict the interval in this way, the improvement establishes a gap at each end, and Theorem 1 can be used to give conditions on $\|\mathbf{r}\|_b$ that will keep all of the subsequent Rayleigh quotients from crossing the gaps. This is embodied in Theorem 3.

THEOREM 3. *Suppose the following statements are true*:

(i) *The interval* (a, b) *contains exactly one eigenvalue* λ *of* $(\mathbf{A}, \mathbf{B})$.

(ii) *The Rayleigh quotient* μ *corresponding to the* $\mathbf{B}$*-normalized vector* $\mathbf{x}$ *lies in* (a, b).

(iii) *The corresponding residual vector* $\mathbf{r}$ *satisfies*

$$\|\mathbf{r}\|_b \leqq [(\mu - a)(b-a)]^{1/2} - (\mu - a);$$

then the subsequent Rayleigh quotient iterates all lie below $(b + b')/2$. *If* (i) *and* (ii) *hold and*

(iv) *The residual vector* $\mathbf{r}$ *satisfies*

$$\|\mathbf{r}\|_b \leqq [(b-\mu)(b-a)]^{1/2} - (b-\mu),$$

then the subsequent Rayleigh iterate quotients all lie above $(a' + a)/2$. *If* (i) *and* (ii) *hold and*

$$\|\mathbf{r}\|_b \leqq (\eta + |\mu - \gamma|)^{1/2}[(2\eta)^{1/2} - (\eta + |\mu - \gamma|)^{1/2}], \tag{4.1}$$

which is the smaller of the restrictions (iii) *and* (iv), *then the subsequent Rayleigh quotient iterates all lie in the interval* $([a + a']/2, [b + b']/2)$ *and converge to the eigenvalue* λ.

Proof. A quick computation shows that either of the restrictions on $\|\mathbf{r}\|_b$ implies that the Kato–Temple interval $[a', b']$ lies interior to $[a, b]$. Since $\mu < b' < (b + b')/2$, Theorem 1 implies that all of the subsequent Rayleigh quotients iterates also lie below $(b + b')/2$ provided that $\|\mathbf{r}\|_b \leqq (b - b')/2$, i.e., provided that $\|\mathbf{r}\|_b$ satisfies

$$2\|\mathbf{r}\|_b \leqq b - \mu - \|\mathbf{r}\|_b^2/(\mu - a).$$

But this is exactly what is implied by the first restriction on $\|\mathbf{r}\|_b$ in the hypothesis of the theorem. Similarly the second restriction of the hypothesis implies that the quotients remain above $(a + a')/2$. The convergence follows from the global convergence of Rayleigh quotient iteration. □

If an interval $[a, b]$ containing μ is known to contain n eigenvalues of $(\mathbf{A}, \mathbf{B})$, we can find bounds on $\|\mathbf{r}\|_b$ that will ensure that subsequent quotients remain in the interior of $[a, b]$. These conditions arise from determining the existence of a minimum gap in the spectrum within $[a, b]$ to the right and to the left of μ and are embodied in Theorem 4.

THEOREM 4. *Suppose n eigenvalues of $(\mathbf{A}, \mathbf{B})$ are known to be in $[a, b]$ and the Rayleigh quotient μ corresponding to a $\mathbf{B}$-normalized vector $\mathbf{x}$ lies in (a, b). Then if $\|\mathbf{r}\|_b \leqq (b-\mu)/(2n+1)$, the subsequent Rayleigh quotients satisfy $\mu_s < b-(b-\mu)/2n$. If $\|\mathbf{r}\|_b \leqq (\mu - a)/(2n+1)$, the subsequent Rayleigh quotients satisfy $a+(\mu-a)/2n < \mu_r$. Further, if*

$$\|\mathbf{r}\|_b \leqq [\eta - |\gamma-\mu|]/(2n+1), \tag{4.2}$$

which is the more restrictive of the two conditions on $\|\mathbf{r}\|_b$, then

$$a' \equiv a + [\eta - |\gamma-\mu|]/2n < \mu_r < b - [\eta - |\gamma - \mu|]/2n \equiv b',$$

and $\{\mu_s\}$ converges to a point in $[a', b']$.

Proof. We designate the n eigenvalues in $[a, b]$ by

$$a \leqq \lambda_{p+1} \leqq \lambda_{p+2} \leqq \cdots \leqq \lambda_{p+n} \leqq b.$$

Since $\|\mathbf{r}\|_b \leqq b-\mu$, there must be one of the n eigenvalues of $(\mathbf{A}, \mathbf{B})$ in $[a, b]$ within $b-\mu$ of μ. The width g^+ of the maximum gap in the spectrum of $(\mathbf{A}, \mathbf{B})$ having a right endpoint to the right of μ is given by

$$g^+ = \max_{j \geqq i} (\lambda_j - \lambda_{j-1}),$$

where λ_i is the first eigenvalue to the right of μ. If there are m eigenvalues in $[a, b]$ to the right of μ, $0 \leqq m \leqq n-1$, then g^+ can be bounded by

$$g^+ \geqq (\lambda_{p+n+1} - \lambda_{i-1})/(m+1) \geqq (b-\mu)/(m+1).$$

When all n eigenvalues in $[a, b]$ lie to the right of μ, then since λ_{p+1} must be within $\|\mathbf{r}\|_b$ of μ,

$$g^+ \geqq \max\{\|\mathbf{r}\|_b, (\lambda_{p+n+1} - \lambda_i)/n\} \geqq (b-\mu-\|\mathbf{r}\|_b)/n.$$

The last bound is the smallest of all and holds for a gap within $[a, b]$ lying entirely to the right of μ. To ensure that subsequent Rayleigh quotients cannot pass beyond the midpoint of this gap, by Theorem 1 it is enough to require that $\|\mathbf{r}\|_b$ satisfy

$$\|\mathbf{r}\|_b \leqq (b-\mu-\|\mathbf{r}\|_b)/2n,$$

which is the same as the first restriction on $\|\mathbf{r}\|_b$ in the hypotheses of this theorem. The location of the gap is known only to the extent that it lies entirely to the right of μ and within $[a, b]$, so that the best that can be said is that all subsequent quotients must lie to the left of $b-(b-\mu)/2n$.

The conditions to the left of μ are established by a parallel argument.

The interval $[a+(\eta - |\gamma-\mu|)/2n,\ b-(\eta-|\gamma-\mu|)/2n]$ must contain μ whenever both conditions on $\|\mathbf{r}\|_b$ are satisfied, and the convergence in the interval follows from the global convergence of Rayleigh quotient iteration. □

If no a priori information on the location of eigenvalues is available, it appears impossible to give rigorous criteria for switching from inverse iteration to Rayleigh quotient iteration that will guarantee localization of the subsequent Rayleigh quotient iterates. In [16] Szyld introduces the ad hoc criterion

$$\|\mathbf{q}\|_b < \eta, \tag{4.3}$$

which appears to perform well in most circumstances. Nonetheless he recognizes that in switching to Rayleigh quotient iteration based on this criterion, the subsequent Rayleigh quotients might leave the interval of interest, and his computational strategy involves reintroducing inverse iteration should that occur. The following simple example illustrates how that can happen.

The matrix

$$\mathbf{A}=\begin{bmatrix} 5/4 & 1 \\ 1 & 5/4 \end{bmatrix}$$

has eigenvalues 1/4 and 9/4 relative to $\mathbf{I}$. Suppose we perform inverse iteration with $\gamma = 0$ in order to locate the eigenvalues in the interval [2, 2] and obtain at step s

$$\mathbf{x}_s=\frac{1}{\sqrt{2}}\begin{bmatrix} 1+\sqrt{3} \\ -1+\sqrt{3} \end{bmatrix}.$$

We calculate $\|\mathbf{q}_s\| = [61/16]^{1/2} < \eta = 2$. In spite of this, subsequent Rayleigh quotient iteration converges to 9/4, which lies outside the interval of interest (an immediate consequence of Theorem 1).

Note that whenever either condition (4.1) of Theorem 3 or condition (4.2) of Theorem 4 holds, then Szyld's condition (4.3) will also be satisfied. In fact, we may compute from (4.2)

$$\begin{aligned}\|\mathbf{q}_S\|_b^2 = \|\mathbf{r}_S\|_b^2 + |\gamma-\mu_S|^2 &\leqq \frac{1}{(2n+1)^2}[\{(2n+1)^2+1\}\,|\gamma-\mu_S|^2 - 2\eta|\gamma-\mu_S| + \eta^2] \\ &< \eta^2\end{aligned}$$

for all values of $n \geqq 1$. A similar argument may be made for (4.1).

5. Computational issues. Our preceding analysis has had the goal of establishing efficient and accessible criteria for switching from inverse iteration to Rayleigh quotient iteration for reliable and rapid closure to those eigenpairs that are wanted. Recall that the results obtained in Theorems 3 and 4 depend only on the availability of eigenvalue counts in given intervals. This kind of information can be obtained from a variety of sources. Among them are Sturm sequences [2] and spectrum slicing [10], Gerschgorin discs [13], variational inequalities [10], and modification by a low rank matrix to obtain a matrix with known eigenvalues [3].

We have incorporated our ideas into a modification of the EISPACK procedure called TSTURM [15], [21]. TSTURM accepts a symmetric tridiagonal matrix and an interval within which eigenvalues are sought, and it returns the eigenvalues and the corresponding eigenvectors. For each irreducible submatrix in turn, TSTURM uses bisection with Sturm sequences until closure is obtained for each eigenvalue within the specified interval. These computed eigenvalues are then used with inverse iteration to obtain the corresponding eigenvectors. TSTURM is known to be a reliable and stable procedure and is the only EISPACK routine available for isolating eigenvalues interior to the spectrum. Although TSTURM has at best a linear rate of convergence, it can be expected to be faster than TQL2, the comparable EISPACK QL procedure for tridiagonal matrices, when fewer than 25 percent of the total matrix eigenvalues are contained within the specified interval.

Though there are some common elements in our work, Scott [11] refined the TSTURM strategy in a direction somewhat different from what we consider here. Scott's strategy involves maintaining eigenvector approximations throughout the iteration while using the associated sequence of Rayleigh quotients to direct subsequent bisection steps.

The method that he proposes has an asymptotic cubic rate of convergence (as does ours) and seems best suited to situations where a priori eigenvector information is available, although it is not restricted to this case. We note also that a variation of the TSTURM solution strategy has been implemented with substantial success in a multiprocessor environment by Lo, Philippe, and Sameh [8], and we expect that our modifications will provide further significant improvements in that environment as well.

In essence, our modification simply involves trading bisection steps for Rayleigh quotient iteration steps when it is safe to do so. Since evaluation of the Sturm sequence has the same order of complexity as solving a tridiagonal system (such as the system (1.2) in our case), we expect to reap rich benefits from the cubic convergence rate of Rayleigh quotient iteration, at least when the mantissa length of the computational word is sufficient to allow calculation in the asymptotic regime before closure is flagged. We avoid the expense of maintaining a vector iterate in the early stages of the process as well.

More specifically, let us begin by considering an interval containing a single eigenvalue λ that is isolated from adjacent eigenvalues by at least the gap width g. Suppose further that we begin Rayleigh quotient iteration in a neighborhood of this eigenvalue with an initial residual satisfying $\|\mathbf{r}_0\|_b \leqq g/2$. We discount rounding errors for the moment and assume that we are working with a mantissa of length T bits. We shall have approximately

$$\|\mathbf{r}_s\|_b \leqq g^{-2}\|\mathbf{r}_{s-1}\|_b^3 \tag{5.1}$$

for $s = 1, 2, \cdots$, (see, for example, [13]). Thus

$$|\lambda - \mu_s| \leqq \|\mathbf{r}_s\|_b \leqq g(\|\mathbf{r}_{s-1}\|_b/g)^3 \leqq g(\|\mathbf{r}_0\|_b/g)^{3^s} \leqq g \times 2^{-3^s}. \tag{5.2}$$

If g and λ have about the same order of magnitude, the number of Rayleigh quotient iterations to obtain full precision accuracy is approximately $\ln(T)/\ln 3$. On the other hand, if bisection is started on an interval of width g containing λ, then roughly $T-1$ bisection steps will be needed to achieve the same accuracy. Evaluation of a Sturm sequence requires roughly n floating point operations as opposed to about $8n$ for a single Rayleigh quotient iteration. Thus we expect a ratio of time in sequential computation given by the factor

$$\frac{\text{time of Rayleigh iteration}}{\text{time of bisection}} = \frac{8n \times \ln(T)/\ln 3}{n \times (T-1)} = \frac{8}{\ln 3} \cdot \frac{\ln(T)}{T-1}. \tag{5.3}$$

In *single precision* on a VAX 11/780, we have $T = 24$, and the value of the ratio is about 1.0, so no advantage can be seen; however for *double precision* on the same machine $T = 56$, and it appears that Rayleigh quotient iteration will require only about 53 percent of the operations needed for bisection! This casual analysis provides a basis for optimism. Many factors might affect the performance of Rayleigh iteration; rounding errors and delays in the onset of cubic convergence could affect our conclusion. However in the computations we have made thus far this anticipated advantage has been borne out.

The general strategy we follow can be viewed as a three-level iteration process, which for convenience we refer to as TLIP. For each irreducible submatrix in turn we use bisection with Sturm sequences until an eigenvalue is isolated from all others in a disjoint subinterval. Then inverse iteration is initiated using the midpoint of the subinterval as shift until the switching criterion of Theorem 3 (based on the known distance to adjacent subintervals) is satisfied. At that point Rayleigh quotient iteration is initiated and continued to closure for the eigenvalue in that interval.

Although the eigenvalues of any irreducible tridiagonal matrix are simple, they may appear multiple due to finite precision, and thus in unusual cases the initial bisection phase could go all the way to closure in attempting to isolate an eigenvalue. In normal circumstances, however, relatively little time is spent in bisection, and the procedure switches rapidly to inverse iteration and then to Rayleigh quotient iteration.

Our procedure was implemented as a straightforward modification of TSTURM with one exception. The starting vector for inverse iteration, which in TSTURM is taken as a multiple of $[1, 1, \cdots, 1]^*$ is replaced by a vector computed to enhance the growth of a solution vector in the sense of [4]. This change is significant in cases possessing substantial symmetry[1], for in TLIP the inverse iteration shift may not be very close to an exact eigenvalue so that amplification in the direction of the eigenvector is much less than in TSTURM, where agreement to within a small multiple of the machine precision can typically be ensured.

Our procedure has been tested on a variety of matrices. We include here detailed results from three such tests together with performance statistics accumulated over a variety of classes of random matrices. The numerical results presented in Tables 1–3 were all performed using double precision arithmetic on a VAX 11/780. Our results show an impressive improvement in speed in Examples 1 and 2. No improvement to speak of is seen in Example 3, which illustrates a situation where TLIP is unable to use the advantages of Rayleigh quotient iteration due to the near multiplicity of eigenvalues.

The first example we consider is the discrete one-dimensional Laplacian given by

$$a_{ij}=\begin{cases} 2, & i=j, \\ 1, & |i-j|=1, \\ 0, & \text{otherwise.} \end{cases}$$

This matrix has all of its eigenvalues in the range 0 to 4. The eigenvalues in the range 1.5 to 2.5 were sought for $n = 100$. Within this interval adjacent eigenvalues typically agree to two significant digits. Results are summarized in Table 1. Notice that the total number of iterations performed by TLIP in all three solution phases (i.e., bisection, inverse, Rayleigh iteration) is very small relative to the large number of bisections required by TSTURM. In all but the last eigenvalue (indexed 58) agreement with the values provided by TSTURM occurred to full precision. The last eigenvalue differed only in the last decimal digit. Thus, in effect, the final residual values that are listed may be considered as a measure of quality for the corresponding eigenvectors.

The second example is a severely graded matrix from [2], having a very broad spread in the spectrum. It is defined by

$$a_{ij}=\begin{cases} i^4, & i=j, \\ i-1, & i=j+1, \\ j-1, & i=j-1, \\ 0, & \text{otherwise.} \end{cases}$$

Eigenvalues in the range 2 to 20,000 were sought with $n = 100$. The results for this matrix are summarized in Table 2. On this problem when TSTURM was run with the internally generated convergence threshold, the final approximate eigenvectors had residual values that were unacceptably large (around 10^{-8}), although run times then became competitive with TLIP. When EPS1 was set to 10^{-12}, the final TSTURM residuals dropped to acceptable levels.

[1] For example, in our Example 1 the usual TSTURM starting vector is orthogonal to half of the eigenvectors and is an inappropriate choice as a starting vector for our inverse iteration phase.

TABLE 1
Example 1. *Discrete Laplacian.*

	TLIP				TSTURM	
Eigenindex	Bisections	Inverse iterations	Rayleigh iterations	$\|\mathbf{r}\|$ residual	Bisections	$\|\mathbf{q}\|$ residual
43	1	1	3	$6.02_{10^{-17}}$	50	$2.78_{10^{-16}}$
44	2	1	3	$2.34_{10^{-16}}$	51	$4.28_{10^{-16}}$
45	2	1	2	$1.86_{10^{-13}}$	50	$2.15_{10^{-16}}$
46	3	2	2	$5.64_{10^{-14}}$	52	$2.55_{10^{-16}}$
47	2	1	2	$4.24_{10^{-13}}$	50	$9.49_{10^{-17}}$
48	2	2	2	$2.50_{10^{-12}}$	51	$2.43_{10^{-16}}$
49	2	1	2	$2.18_{10^{-12}}$	50	$2.17_{10^{-16}}$
50	4	2	2	$2.08_{10^{-13}}$	53	$3.82_{10^{-16}}$
51	2	2	2	$2.47_{10^{-14}}$	50	$3.82_{10^{-16}}$
52	2	2	2	$2.14_{10^{-13}}$	51	$2.17_{10^{-16}}$
53	2	1	2	$4.59_{10^{-13}}$	50	$2.43_{10^{-16}}$
54	3	2	2	$7.12_{10^{-14}}$	52	$9.49_{10^{-17}}$
55	2	1	2	$2.18_{10^{-13}}$	50	$2.55_{10^{-16}}$
56	2	2	2	$6.30_{10^{-14}}$	51	$2.15_{10^{-16}}$
57	2	1	2	$7.53_{10^{-16}}$	50	$4.28_{10^{-16}}$
58	5	1	3	$1.96_{10^{-17}}$	54	$2.78_{10^{-16}}$

Relative solution time TLIP/TSTURM: 0.551.

The third example (see Table 3) is also taken from [2] (which is included in [21]) and is a variation of Wilkinson's test matrix $\mathbf{W}_{21}$ [20]. This matrix is defined by

$$a_{ij}=\begin{cases} 110-10i, & i=j=1,2,\cdots,11,\\ 10i-100, & i=j=12,13,\cdots,21,\\ 1, & |i-j|=1,\\ 0 \quad \text{otherwise.} \end{cases}$$

TABLE 2
Example 2. I4 *matrix.*

	TLIP				TSTURM[1]	
Eigenindex	Bisections	Inverse iterations	Rayleigh iterations	$\|\mathbf{r}\|$ residual	Bisections	$\|\mathbf{q}\|$ residual
2	2	2	3	$6.24_{10^{-16}}$	48	$5.74_{10^{-12}}$
3	2	2	2	$1.15_{10^{-13}}$	49	$8.18_{10^{-12}}$
4	2	2	2	$1.12_{10^{-15}}$	50	$2.81_{10^{-12}}$
5	2	5	2	$9.52_{10^{-15}}$	52	$1.22_{10^{-12}}$
6	2	2	2	$4.17_{10^{-13}}$	51	$1.41_{10^{-12}}$
7	3	2	2	$1.46_{10^{-13}}$	53	$1.99_{10^{-12}}$
8	2	2	2	$1.49_{10^{-13}}$	54	$1.87_{10^{-12}}$
9	2	1	1	$6.02_{10^{-12}}$	53	$1.32_{10^{-12}}$
10	3	4	2	$7.47_{10^{-14}}$	55	$1.49_{10^{-12}}$
11	2	2	3	$2.07_{10^{-13}}$	55	$1.98_{10^{-12}}$

Relative solution time TLIP/TSTURM: 0.617.

[1] EPS1 $= 10^{-12}$.

TABLE 3
Example 3. $\mathbf{W}_{21}$ *matrix.*

	TLIP				TSTURM	
Eigenindex	Bisections	Inverse iterations	Rayleigh iterations	$\lVert\mathbf{r}\rVert$ residual	Bisections	$\lVert\mathbf{q}\rVert$ residual
6	1	1	2	$2.13_{10^{-15}}$	33	$3.77_{10^{-16}}$
7	21	2	1	$1.05_{10^{-15}}$	53	$4.09_{10^{-15}}$
8	1	6	1	$2.22_{10^{-13}}$	22	$5.50_{10^{-15}}$
9	30	2	0	$7.36_{10^{-13}}$	52	$4.54_{10^{-16}}$
10	1	1	0	$1.34_{10^{-13}}$	10	$2.81_{10^{-15}}$
11	44	1	0	$2.64_{10^{-13}}$	54	$1.46_{10^{-15}}$
12	1	1	0	$8.88_{10^{-16}}$	1	$5.43_{10^{-16}}$
13	53	1	0	$2.66_{10^{-15}}$	53	$2.45_{10^{-14}}$
14	1	1	0	$3.55_{10^{-15}}$	1	$6.93_{10^{-16}}$
15	55	1	0	$5.32_{10^{-15}}$	55	$2.63_{10^{-14}}$

Relative solution time TLIP/TSTURM: 0.982.

Eigenvalues in the range 25 to 75 were sought. The first pair of eigenvalues (indexed 6 and 7) agree to eight decimal digits. The next two pairs agree to 11 and 14 decimal digits, respectively. The last two pairs of eigenvalues agree to 16 decimal digits—effectively the full precision used. Note that for the last six eigenvalues TLIP is operating essentially as TSTURM, since the near multiplicity of the eigenvalues drives the TLIP bisection phase to closure.

Table 4 summarizes comparisons performed on 700 random matrices. For these experiments TLIP and TSTURM were run in double precision on a SUN 3 workstation equipped with an MC68881 floating point coprocessor. For each matrix that was generated, roughly 25 percent of the eigenvalues and eigenvectors were found with each routine. Comparisons were made with respect to speed and accuracy in each case. The test matrices were separated into seven classes.

Class 1. These one hundred 100 × 100 matrices correspond to discretizations of one-dimensional Schrodinger operators with randomly generated potentials bounded by 1.0 in magnitude. Diagonal entries are uniformly distributed in the interval (1, 3) and off-diagonal entries are fixed at 1.0.

Class 2. These one hundred 50 × 50 matrices are gently graded from large at the upper left to small at the lower right. Consecutive diagonal and off-diagonal entries are in ratio $d_{i+1}/d_i = z^{1/2}$ and $e_{j+1}/e_j = z$, respectively, where z is uniformly distributed

TABLE 4

Matrix class	Relative solution time TLIP/TSTURM			Maximum deviation from orthogonality		Maximum scaled final residuals		Comparison of computed eigenvalues (number of digits in agreement)			
	avg	max	min	TSTURM	TLIP	TSTURM	TLIP	All	18	17	16
1	0.423	0.556	0.334	1.04E-13	5.41E-13	3.51E-14	2.99E-13	46%	52%	2%	0%
2	0.422	0.682	0.105	6.54E-12	1.03E-11	2.58E-14	1.69E-13	24%	11%	43%	22%
3	0.563	1.072	0.065	1.04E-12	3.52E-12	1.56E-14	2.97E-14	100%	0%	0%	0%
4	0.392	0.579	0.164	2.19E-14	3.45E-14	5.04E-15	4.94E-14	43%	53%	4%	0%
5	0.279	0.398	0.208	8.67E-12	8.72E-12	2.21E-14	2.24E-14	100%	0%	0%	0%
6	0.338	0.545	0.103	5.64E-13	8.78E-13	1.62E-14	1.71E-13	54%	44%	2%	0%
7	0.139	0.364	0.098	1.36E-12	1.34E-12	2.23E-14	2.19E-14	94%	4%	2%	0%

in (0, 1) and $d_1 = e_1 = 1.0$. Matrix entries typically will vary over 6 to 8 orders of magnitude.

Class 3. These one hundred 50×50 matrices are steeply graded from small at the upper left to large at the lower right. Consecutive diagonal entries are in ratio $d_{i+1}/d_i = z \times 10^r$, where z is uniform in (0, 1), r is an integer exponent uniformly distributed from 0 to 5, and $d_1 = 1.0$. Off-diagonals are generated in the same way. Matrix entries will vary up to 160 orders of magnitude.

Class 4. These one hundred 50×50 matrices are centrosymmetric and graded from large to small to large. For $i = 1$ to 25, consecutive diagonal entries are in ratio $d_i/d_{i+1} = z$ where z is uniform in (1, 2) and $d_1 = 1.0$. The remaining diagonal entries are defined by $d_i = d_{51-i}$. Off-diagonal entries are fixed at 1.0.

Class 5. These one hundred 50×50 matrices are also centrosymmetric but instead are graded from small to large to small. For $i = 1$ to 25, consecutive diagonal entries are in ratio $d_{i+1}/d_i = z$ where z is uniform in (1, 2) and $d_1 = 1.0$. The remaining diagonal entries are defined by $d_i = d_{51-i}$. Off-diagonal entries are fixed at 1.0.

Class 6. These one hundred 50×50 matrices have random diagonal and off-diagonal entries uniformly distributed in $(-1, 1)$.

Class 7. These one hundred 50×50 matrices have random entries in the diagonal and off-diagonal with the form $z \times 10^r$ where z is uniform in $(-1, 1)$ and r is an integer exponent uniformly distributed from 0 to 5.

For each matrix class, Table 4 lists the average, maximum, and minimum observed ratios of TLIP computation time to TSTURM computation time; the maximum observed deviation from orthonormality of the computed eigenvectors (as indicated by $\|\mathbf{I} - \mathbf{Z}^t\mathbf{Z}\|_\infty$ where $\mathbf{Z}$ is the computed matrix of eigenvectors); the maximum scaled residual norm; and the relative accuracy of eigenvalues returned by each routine. The percentages given under this last heading indicate to what extent TLIP reproduced values given by TSTURM. For example, within Class 1, 46 percent of all eigenvalues found (namely 4,605) agree to the last digit with the values produced by TSTURM, 52 percent differ only in the last digit, and the remaining 2 percent differ in the last two digits. The terminating residuals have been scaled by the norm of the matrix to permit relative comparisons among trials.

Notice that for the most part, the speedup of TLIP over TSTURM exceeds by a substantial margin the speedup predicted by (5.3), probably due to the pessimistic residual bound given by (5.2). The wide variation in performance for Class 3 appears to be due to the vast range of eigenvalue magnitudes within the search interval. In these circumstances, isolation of the eigenvalues is achieved comparatively quickly by bisection. But then the inverse iteration phase of TLIP may be entered while the isolating interval is still quite large, thus providing little information on the location of the contained eigenvalue. But while TSTURM continues on with bisection, there is the possibility that the inverse iteration phase of TLIP will converge much more slowly than bisection since the selected shift may still be quite far from any eigenvalue. This in turn suggests the need for a rational criterion for switching between bisection and inverse iteration. We do not consider this here but leave it as a question for further study.

The brief analysis that yielded (5.3) indicates that if we double our mantissa length (e.g., change from single to double precision), we might expect to see the TLIP/TSTURM timing ratios drop by at least 35 percent for usual word lengths and more for longer word lengths. On a Cray-2, we reran TSTURM and TLIP on Example 3, for which TLIP had shown little improvement, and obtained timing ratios of 98 percent and 58 percent, for default precision ($T = 48$) and double precision ($T = 112$), respectively—ultimately giving performance comparable to what we found with Examples 1 and 2. As we have noted earlier, bisection to obtain isolating intervals uses up a good part of the time; but once the eigenvalues are isolated, the speed of Rayleigh quotient iteration comes into play with a trebling of the number of significant digits each iteration.

Acknowledgment. The authors have had the benefit of discussions with Professor Daniel Szyld of Duke University, who found a flaw in the proof of an earlier version of Theorem 1 and to whom they wish to express their appreciation.

REFERENCES

[1] K.-J. BATHE AND E. L. WILSON, *Numerical Methods in Finite Element Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[2] W. BARTH, R. S. MARTIN, AND J. H. WILKINSON, *Calculation of the eigenvalues of a symmetric tridiagonal matrix by the method of bisection*, Numer. Math., 9 (1967), pp. 386–393.

[3] C. BEATTIE AND D. W. FOX, *Schur complements and the Weinstein-Aronszajn theory for modified matrix eigenvalue problems*. UMSI Technical Report 87/11, University of Minnesota Supercomputer Institute, Minneapolis, MN, February, 1987; Linear Algebra Appl., to appear.

[4] A. K. 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.

[5] W. L. FRANK, *Computing eigenvalues of complex matrices by determinant evaluation and by methods of Danilewski and Wielandt*, J. Soc. Indust. Appl. Math., 6 (1958), pp. 378–392.

[6] P. S. JENSEN, *The solution of large symmetric eigenproblems by sectioning*, SIAM J. Numer. Anal., 9 (1972), pp. 534–545.

[7] T. KATO, *On the upper and lower bounds of eigenvalues*, J. Phys. Soc. Japan, 4 (1949), pp. 334–339.

[8] S.-S. LO, B. PHILIPPE, AND A. SAMEH, *A multiprocessor algorithm for the symmetric tridiagonal eigenvalue problem*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. s155–s165.

[9a] A. M. OSTROWSKI, *On the convergence of the Rayleigh quotient iteration for the computation of characteristic roots and vectors*, I, Arch. Rational Mech. Anal., 1 (1958), pp. 233–241.

[9b] ———, *On the convergence of the Rayleigh quotient iteration for the computation of characteristic roots and vectors*, II, Arch. Rational Mech. Anal., 2 (1959), pp. 423–428.

[10] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[11] D. S. SCOTT, *Computing a few eigenvalues and eigenvectors of a symmetric band matrix*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 658–666.

[12] A. RUHE, *Computation of Eigenvalues and Eigenvectors*, in Sparse Matrix Techniques—Copenhagen 1976, V. A. Barker, ed., Lecture Notes in Mathematics 572, Springer-Verlag, Berlin, 1977.

[13] G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[14] ———, *A bibliographic tour of the large, sparse generalized eigenvalue problem*, in Sparse Matrix Computations, J. R. Bunch and D. J. Rose, eds., Academic Press, New York, 1976.

[15] B. T. SMITH, J. M. BOYLE, B. S. GARBOW, Y. IKEBE, V. C. KLEMA, AND C. B. MOLER, *Matrix Eigensystem Routines*—EISPACK *Guide*, Springer-Verlag, Heidelberg, 1974.

[16] D. SZYLD, *Criteria for combining inverse and Rayleigh quotient iteration to solve* $\mathbf{Ax} = \lambda\mathbf{Bx}$, in Innovative Numerical Methods in Engineering, R. D. Shaw, J. Periaux, A. Chaudouet, J. Wu, C. Marino, and C. A. Brebbia, eds., Springer-Verlag, New York, 1986.

[17] ———, *Criteria for combining inverse and Rayleigh quotient iteration*, SIAM J. Numer. Anal., to appear.

[18] D. SZYLD AND O. WIDLUND, *Applications of conjugate gradient type methods to eigenvalue calculations*, in Advances in Computer Methods for Partial Differential Equations—III, R. Vichnevetsky and R. Stapleman, eds., IMACS, 1979.

[19] G. TEMPLE, *The computation of characteristic numbers and characteristic functions*, Proc. London Math. Soc., 29 (1928), pp. 257–280.

[20] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, New York, 1965.

[21] J. H. WILKINSON AND C. H. REINSCH, *Handbook for Automatic Computation*, *Linear Algebra*, Vol. 2, Springer-Verlag, New York, 1971.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 1, pp. 94–104, January 1989

INHERITED MATRIX ENTRIES: *LU* FACTORIZATIONS*

CHARLES R. JOHNSON†, D. D. OLESKY‡, AND P. VAN DEN DRIESSCHE§

Abstract. For an n-by-n matrix $A = [a_{ij}]$ which has a unique unit LU factorization $A = LU$ with $U = [u_{ij}]$, combinatorial circumstances are determined under which $u_{ij} = a_{ij}$ for a given pair $i \leqq j$ or for all $i < j$ (or all $i \leqq j$). Analogous results are stated for other triangular factorizations and for the LU factorization of a principal submatrix of A. The relationship of the results to Gaussian elimination and sparse matrix analysis is discussed.

Key words. matrix factorization, Gaussian elimination, fill-in, directed graphs, sparse matrices

AMS(MOS) subject classifications. 05C50, 15A23, 65F50

1. Introduction. An n-by-n matrix $A = [a_{ij}]$ has an *LU factorization* if it may be written as a product $A = LU$ in which $L = [l_{ij}]$ and $U = [u_{ij}]$ are, respectively, lower and upper triangular n-by-n matrices. If there is such a factorization in which L is nonsingular, then there is one in which all diagonal entries of L are equal to 1; we call such a factorization in which L has unit diagonal a *unit LU factorization*. Our interest here is in a family of questions of the following type.

(1.1) Under what circumstances does $u_{ij} = a_{ij}$ for a *given* pair $i \leqq j$?
(1.2) Under what circumstances does $u_{ij} = a_{ij}$ for *all* pairs $i < j$ (or all $i \leqq j$)?

We refer to the equalities in (1.1) and (1.2) as *local* and *global inheritance*, respectively.

There are several familiar examples in which the phenomena requested by (1.2) occur. Perhaps the simplest is the case in which A itself is upper triangular. In this event $A = LU$ with $L = I$ and $U = A$, so that the upper triangular factor agrees with A above (and on) the diagonal. Another example is the following tridiagonal matrix A, which factors as $A = LU$ with $L = U^T$:

$$A = \begin{bmatrix} 1 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ 0 & & & -1 & 2 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & -1 & & & 0 \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & -1 \\ 0 & & & & 1 \end{bmatrix}.$$

Again, the upper triangular factor agrees with A above the diagonal, even though A is irreducible in this case. A very simple circumstance for the local question (1.1) is the case $i = 1$; it is well known and easy to check that the first row of A becomes the first row of U for any matrix A with unit LU factorization. It is the combinatorial basis for this sort of simplicity and sparsity preservation upon which we focus. If, for example,

* Received by the editors April 20, 1987; accepted for publication (in revised form) July 20, 1988.

† Department of Mathematics, College of William and Mary, Williamsburg, Virginia 23185. The work of this author was partly supported by National Science Foundation grant DMS-8713762 and Office of Naval Research contract N00014-87-K-0661 (crjoh2@wmmvs.bitnet).

‡ Department of Computer Science, University of Victoria, Victoria, British Columbia, Canada V8W 2Y2. The work of this author was partly supported by Natural Sciences and Engineering Research Council of Canada grant A-8214 (dolesky@uvunix.bitnet).

§ Department of Mathematics, University of Victoria, Victoria, British Columbia, Canada V8W 2Y2. The work of this author was partly supported by Natural Sciences and Engineering Research Council of Canada grant A-8965 (pvdd@uvvm.bitnet).

circumstances are such that the upper triangular factor agrees with A above the diagonal when the lower triangular factor has 1's on the diagonal, half of the assumed factorization may be written down immediately.

In order to ask our questions, we must assume that a unit LU factorization exists. To avoid possible ambiguities we shall also assume that it is unique, and, fortunately, this circumstance may be easily characterized. For index sets $\alpha, \beta \subseteq \{1, 2, \cdots, n\}$, we denote the submatrix of the n-by-n matrix A lying in the rows α and columns β by $A[\alpha \mid \beta]$. When $\beta = \alpha$, the submatrix is principal and we abbreviate $A[\alpha \mid \alpha]$ to $A[\alpha]$. We shall often be interested in the determinant of a leading principal submatrix of A and so adopt the notation $d_k(A) \equiv \det A[\{1, 2, \cdots, k\}]$. Our characterization slightly strengthens [10, Thm. 4]. Note that this result is well known when A is nonsingular (see, for example, [9, Cor. 3.5.5]).

THEOREM 1.1. *The n-by-n matrix A has a unique unit LU factorization if and only if*

$$d_k(A) \neq 0, \qquad k = 1, 2, \cdots, n-1. \tag{1.3}$$

Proof. Suppose that all the proper leading principal minors of A are nonzero, that is, condition (1.3) is met. Then by [9, Cor. 3.5.5] and the partitioning in [10], the required unit LU factorization exists and is unique. The converse is similar to that of [10], the only difference being that we take L (rather than U) as the normalized matrix. □

In view of Theorem 1.1 we shall generally assume that A satisfies condition (1.3).

In the spirit of sparse matrix analysis, we are not interested in circumstances such as (1.1) and (1.2) involving accidental numerical cancellation. We approach these questions from a combinatorial point of view, based upon the zero pattern of A rather than the values of the nonzero entries. For this purpose, recall that a directed graph D consists of a set of nodes and some directed edges; a subgraph is based upon the same set of nodes and a subset of the edges of D. See [1] for other graph-theoretic terms that we use. With a given n-by-n matrix $A = [a_{ij}]$ we associate a directed graph $D(A)$ on nodes $1, 2, \cdots, n$ by including the edge (i, j) from i to j if and only if $a_{ij} \neq 0$. This then precisely describes the zero pattern of A.

If, for example,

$$A = \begin{bmatrix} 1 & 1 & 1 & 2 \\ 1 & 2 & 1 & 1 \\ 2 & 6 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \qquad \text{then } U = \begin{bmatrix} 1 & 1 & 1 & 2 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix},$$

and $a_{34} = u_{34} = 0$ (see [1, ex. 4.1]). The inheritance of this zero entry is due to the numerical values of the entries, not to the structure of $D(A)$. To exclude such instances of "accidental cancellation," we give the following definition. Given a matrix A with digraph $D(A)$, we say that two values f and g computable from the entries of A are *equal generically* (written $f = g$ (generically)) if $f(\hat{A}) = g(\hat{A})$ for all $\hat{A}$ such that $D(A) = D(\hat{A})$.

We say that an n-by-n matrix A is *consistent* with a given directed graph D if $D(A)$ is a subgraph of D, and let $\mathscr{A}_D$ denote the set of all n-by-n matrices A that satisfy (1.3) and are consistent with D. The combinatorial phrasings of the matrix questions (1.1) and (1.2) which we address are then as follows.

(1.1′) For which directed graphs D does $A \in \mathscr{A}_D$ imply that $u_{ij} = a_{ij}$ for a given pair $i \leqq j$?

(1.2′) For which directed graphs D does $A \in \mathscr{A}_D$ imply that $u_{ij} = a_{ij}$ for all pairs $i < j$ (or all $i \leqq j$)?

We first answer questions (1.1) and (1.1′) with a simple graph-theoretic condition in § 2, and then use this answer to address questions (1.2) and (1.2′) in § 3. In § 4 we indicate the analogous solutions for the dual problems regarding L when U is normalized. Also in § 4 we consider the situation in which both $u_{ij} = a_{ij}$ for $i \leqq j$, and $l_{ij} = a_{ij}$ for $i \geqq j$. Some analogues for UL factorizations are stated in § 5, and the occurrence of $U(A[\beta])_{ij} = U(A)_{ij}$ is considered in § 6. The relationship with known results about Gaussian elimination is discussed in § 2 and further in § 7.

2. Local inheritance. In this section we address the local questions and begin with a sufficient condition for (1.1).

THEOREM 2.1. *Let A be an n-by-n matrix satisfying* (1.3). *If there is no path from i to $j \geqq i$ through $\{1, 2, \cdots, i-1\}$ in $D(A)$, then $u_{ij} = a_{ij}$ in the unit LU factorization of A.*

Proof. By [7, p. 26],

$$u_{ij} = \frac{\det A[\{1, \cdots, i-1, i\} \mid \{1, \cdots, i-1, j\}]}{d_{i-1}(A)} \quad \text{for } j \geqq i. \tag{2.1}$$

Expanding about the ith row,

$$\det A[\{1, \cdots, i-1, i \mid 1, \cdots, i-1, j\}] = a_{ij} d_{i-1}(A), \tag{2.2}$$

as there are no paths from i to j through $\{1, \cdots, i-1\}$. Equations (2.1) and (2.2) imply that $u_{ij} = a_{ij}$. □

The above path condition and an analogous one play an important part in our work, so we introduce some related terminology. For $1 \leqq i, j \leqq n$, A is (i, j) *lower restricted* if there is no path (of length $\geqq 2$) in $D(A)$ from i to j such that all intermediate nodes on this path are $<\min\{i, j\}$. Note that A is always $(1, j)$ and $(i, 1)$ lower restricted. Letting $i, j \in \{1, 2, \cdots, n\}$ and $S \subseteq \{1, 2, \cdots, n\}$, *$j$ is reachable from i through S* (see [1]) if there is a (simple) path in D (of length $\geqq 2$) from i to j such that all intermediate nodes on this path are in S. Thus for $1 \leqq i, j \leqq n$, A is (i, j) lower restricted if and only if j is not reachable from i through $S = \{1, 2, \cdots, \min\{i, j\}\}$. In the case that $i < j$ and A is (i, j) lower restricted, then $a_{ij} = 0$ implies that the submatrix $A[\{1, 2, \cdots, i, j\}]$ is reducible, whereas for $a_{ij} \neq 0$ this submatrix may be irreducible. If A is (i, i) lower restricted, then $A[\{1, 2, \cdots, i\}]$ is always reducible.

Provided A satisfies (1.3), Theorem 2.1 can be restated as follows. If A is (i, j) lower restricted, then $u_{ij} = a_{ij}$ for given $i \leqq j$ in the unit LU factorization of A. For example, if there is a positive integer p such that, for all i and j, $a_{ij} = 0$ when $j \geqq p + i$, then A is (i, j) lower restricted when $j \geqq p + i - 1$. Thus $u_{ij} = a_{ij}$ for all pairs (i, j) with $j \geqq p + i - 1$; and in particular $u_{ij} = a_{ij} = 0$ for all such pairs with $j \geqq p + i$. Note that this situation includes matrices of bandwidth $p - 1$ and lower Hessenberg matrices ($p = 2$).

Although our methodology is different, Theorem 2.1 is closely related to Theorem 1 of [12], which (for $i \leqq j$) can be interpreted as saying that if $a_{ij} = 0$ and node j is not reachable from node i through $\{1, 2, \cdots, i-1\}$, then $u_{ij} = 0$. In this circumstance, i.e., when node j is not reachable from node i through $\{1, 2, \cdots, i-1\}$, Theorem 2.1 shows that $u_{ij} = a_{ij}$ regardless of the value of a_{ij}. However, as examples (2.4) and (2.5) below illustrate, Theorem 1 of [12] does not characterize the combinatorial circumstances under which $u_{ij} = a_{ij}$ even when $a_{ij} = 0$; i.e., if $u_{ij} = a_{ij} = 0$ and the equality is due to the combinatorial structure of A, then it is not necessarily true that node j is not reachable from node i through $\{1, 2, \cdots, i-1\}$. We now give a characterization of the combinatorial circumstances under which $u_{ij} = a_{ij}$ (for arbitrary a_{ij}), thus answering (1.1′).

THEOREM 2.2. *Let D be a directed graph on n nodes and let i, j be a given pair, $i \leqq j \leqq n$. Then for all $A \in \mathscr{A}_D$, $u_{ij} = a_{ij}$ in the unit LU factorization of A if and only if*

(i) *j is not reachable from i through $\{1, 2, \cdots, i-1\}$, or*

(ii) *if j is reachable from i through nodes $p_1, p_2, \cdots, p_t \in \{1, 2, \cdots, i-1\}$, then*

$$\det A[\{1,2, \cdots, i-1\} - \{p_1, p_2, \cdots, p_t\}] = 0 \quad \textit{for all } A \in \mathscr{A}_D.$$

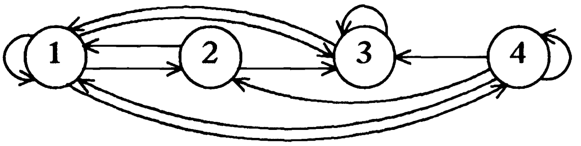
Proof. Expanding the numerator of (2.1) about the ith row (cf. (2.2)),

$$\begin{aligned} &\det A[\{1, \cdots, i-1, i \,|\, 1, \cdots, i-1, j\}] \\ (2.3)\qquad &= a_{ij}d_{i-1}(A) + \sum \pm a_{ip_1}a_{p_1p_2}\cdots a_{p_tj} \det A[\{1,2, \cdots, i-1\} - \{p_1, p_2, \cdots, p_t\}], \end{aligned}$$

where the summation is over all simple paths from i to j through $p_1, p_2, \cdots, p_t \in \{1, \cdots, i-1\}$, $t \geqq 1$, and the ± sign depends on i, j, and t, (see [1]). If $u_{ij} = a_{ij}$ for all $A \in \mathscr{A}_D$, then each term in this summation must be zero, so either there is no such path (condition (i)) or the complementary principal minor must be zero (condition (ii)). Conversely, if (i) is true then Theorem 2.1 gives $u_{ij} = a_{ij}$, while if (ii) is true (2.1) and (2.3) give this equality. □

Theorem 2.2 generalizes the well-known results concerning fill-in Gaussian elimination (see § 7), which are an attempt to determine when a nonzero entry will occur in the (i, j) position during the entire process of Gaussian elimination, given that $a_{ij} = 0$. To illustrate this point, consider the following digraph D (cf. the transpose of the example in [3, p. 944]):

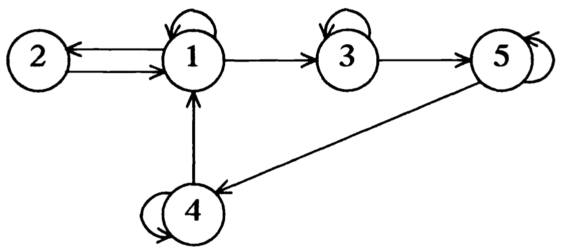
(2.4)



Subject only to condition (1.3), $u_{34} = a_{34} = 0$ in the unit LU factorization of any matrix $A \in \mathscr{A}_D$. However, by [12, Thm. 1] there is fill-in at the (3, 4) position of any such matrix $A \in \mathscr{A}_D$ if $a_{31}a_{14} \neq 0$. We note (using Theorem 1 of [12] and Theorem 2.2) that if there is fill-in at the (i, j) position for some $i < j$, and if $u_{ij} = a_{ij} = 0$ in the unit LU factorization of A, then $a_{kk} = 0$ for some $k < i$ (see also [3]).

The following example illustrates that it is possible to have fill-in at the (i, j) position even though $a_{ij} = 0$ and the (i, j) entry of the reduced matrix of Gaussian elimination is zero throughout the elimination. Let A be any matrix consistent with the digraph of (2.5) and satisfying (1.3).

(2.5)



Then, e.g., by [12, Thm. 1], there is fill-in at the (4, 5) position (provided that $a_{41}a_{13}a_{35} \neq 0$), although $u_{45} = a_{45} = 0$ and the (4, 5) entry is zero (in the absence of rounding error) at all stages of the elimination.

The following observation follows from the inheritance in U of a subset of a row of entries of A (cf. [5, Lemma 2]).

COROLLARY 2.3. *Let D be a directed graph on n nodes with a self loop at each node. Given $i \in \{1, 2, \cdots, n-1\}$, if $u_{ij} = 0$ for all $j > i$ in the unit LU factorization of all $A \in \mathscr{A}_D$ having $a_{kk} \neq 0$ for $k = 1, 2, \cdots, n-1$, then D is not strongly connected (that is, all such $A \in \mathscr{A}_D$ are reducible).*

Proof. Since $u_{ij} = 0$ for all such $A \in \mathscr{A}_D$, a_{ij} must equal zero and so D cannot have an edge (i, j) for all $j > i$. Since, for all such A, $a_{kk} \neq 0$ for $k = 1, 2, \cdots, n-1$, condition (ii) of Theorem 2.2 is vacuous. By condition (i), all such A are (i, j) lower restricted for all $j > i$. Thus there is no path from i to j in D, so the result follows. □

Note that this corollary is, in general, false without the condition that $a_{kk} \neq 0$ for $k = 1, 2, \cdots, n-1$. For example, with D as in (2.4), $u_{34} = 0$ (generically) in the unit LU factorizations of all $A \in \mathscr{A}_D$, although A may be irreducible. A result similar to Corollary 2.3 is given in [5, Lemma 2]. However, the results of [5] are based on a different understanding of when a matrix entry is zero/nonzero for combinatorial reasons. In the above example, u_{34} would be considered to be nonzero under the assumptions of [5] because it is computed from two nonzero values. However, u_{34} is in fact identically zero because of the combinatorial structure of A (and not because of accidental numerical cancellation). Thus our result is more precise in this regard.

When A is a nonsingular M-matrix [2, Chap. 6], then all principal minors of A are positive, so A has a unique unit LU factorization (with L and U also M-matrices) and condition (ii) of Theorem 2.2 is vacuous. In fact this last statement also holds for a singular, irreducible M-matrix A as every principal submatrix of a matrix in this set (other than A itself) is a nonsingular M-matrix [2, Thm. 6.4.16]. If A is a singular, reducible M-matrix, then it need not have an LU factorization. However, if in this case we impose our usual hypothesis (1.3), then $A[1, 2, \cdots, n-1]$ is a nonsingular M-matrix, and the same statement holds. Similarly, if A is a positive semidefinite matrix satisfying (1.3), then condition (ii) of Theorem 2.2 is also vacuous because $d_{n-1} \neq 0$ implies that $A[1, 2, \cdots, n-1]$ is positive definite. Thus we have the following corollary.

COROLLARY 2.4. *Let A satisfy* (1.3), *and be either an n-by-n M-matrix or positive semidefinite matrix, and let i, j be a given pair with $i \leqq j \leqq n$. Then $u_{ij} = a_{ij}$ in the unit LU factorization of A if and only if A is (i, j) lower restricted.* □

The analogous result regarding fill-in during Gaussian elimination when the coefficient matrix is a nonsingular M-matrix is given in [6, p. 290]. Similarly, Theorem 5.1.2 of [8] is an analogue of the positive semidefinite case. Note that, if A is any M-matrix that has a unique unit LU factorization, then the sign pattern of A insures that $u_{ij} \leqq a_{ij}$ for all $i \leqq j$. Moreover, if $u_{ij} < a_{ij}$, then the (i, j) entry of A monotonically decreases to u_{ij} during the Gaussian elimination process, implying that an equality $u_{ij} = a_{ij}$ can never be due to accidental numerical cancellation.

3. Global inheritance.

We now address questions (1.2), (1.2′) and first answer the more general graph question (1.2′) for all $i < j$. In contrast to the result for local inheritance, it turns out that the complementary minor condition (Theorem 2.2 (ii)) disappears from the characterization of global inheritance.

THEOREM 3.1. *Let D be a directed graph on n nodes. Then, for all $A \in \mathscr{A}_D$ and for all pairs i, j with $i < j \leqq n$, $u_{ij} = a_{ij}$ in the unit LU factorization of A if and only if j is not reachable from i in D through* $\{1, 2, \cdots, i-1\}$.

Proof. Assume that for all $i < j$, j is not reachable from i in D through $\{1, 2, \cdots, i-1\}$, that is, there is no path from i to j through nodes $< i$. By condition (i) of Theorem 2.2 this implies that $u_{ij} = a_{ij}$ for all $i < j$.

For the converse, assume that $u_{ij} = a_{ij}$ for all pairs $i < j$ and all $A \in \mathscr{A}_D$. Using Theorem 2.2, if condition (i) is true for all such pairs and all $A \in \mathscr{A}_D$, then our theorem is proved. Otherwise let $A \in \mathscr{A}_D$ and let i be the smallest node for which there is a path in $D(A)$ from i to $j > i$ through nodes $p_1, p_2, \cdots, p_t \in \{1, 2, \cdots, i-1\}$ such that $\det A[\alpha] = 0$, where $\alpha \equiv \{1, 2, \cdots, i-1\} - \{p_1, p_2, \cdots, p_t\}$. Then there exists a node $m \in \alpha$ such that

(a) $a_{mm} = 0$, since $\det A[\alpha] = 0$ for all $A \in \mathscr{A}_D$; and

(b) m lies on a cycle $m \to q_1 \to \cdots \to q_r \to m$ in $D(A[1, 2, \cdots, m])$ (since $d_m(A) \neq 0$) with some q_s, $1 \leqq s \leqq r$, not in α (since if each node in α having no 1-cycle lies on a cycle entirely in α, then there exists $A \in \mathscr{A}_D$ such that $\det A[\alpha] \neq 0$).

Thus $q_s \in \{p_1, p_2, \cdots, p_t\}$, so there exists a path $m \to q_1 \to \cdots \to (q_s = p_v) \to \cdots \to p_w$ where $p_v, \cdots, p_w \in \{p_1, p_2, \cdots, p_t, j\}$ and $p_w = \min\{j$, first node $> m$ on the path from i to $j\}$. All intermediate nodes on the path from m to p_w are $<m$, and, by the choice of i, all principal minors of $A[1, 2, \cdots, m-1]$ are nonzero. Thus by Theorem 2.2 $u_{mp_w} \neq a_{mp_w}$, which contradicts our assumption. So condition (ii) of Theorem 2.2 cannot hold in this case. □

Matrices with digraphs satisfying the condition of Theorem 3.1 are, in the terminology of § 2, (i, j) lower restricted for all pairs $i < j$; we call such matrices *forward lower restricted*. With this terminology and the result of Theorem 3.1, question (1.2) may now be answered succinctly as follows.

COROLLARY 3.2. *Let A be an n-by-n matrix satisfying* (1.3). *If A is forward lower restricted, then $u_{ij} = a_{ij}$ for all $i < j$ in the unit LU factorization of A. Conversely, if $u_{ij} = a_{ij}$ (generically) for all $i < j$, then A is forward lower restricted.* □

For example, a lower Hessenberg matrix is forward lower restricted, and thus satisfies the condition of the corollary; see the example after Theorem 2.1. If A is an M-matrix, then Corollary 2.4 shows that the conditions of Corollary 3.2 are necessary and sufficient without generic equality.

We now restrict A to be combinatorially symmetric and reconsider our global inheritance question (1.2) in this special case. When the undirected graph of A is a forest, we define A to be *invariantly ordered* if it has at most one nonzero entry in each column below the diagonal; for example, any tridiagonal matrix is invariantly ordered. We note that if the undirected graph of A is a minimum degree ordered forest (see, e.g., [8]), then A is invariantly ordered; however the converse is not necessarily true (unless the graph is a tree). This definition leads to the following characterization; we omit the proof, which is straightforward.

THEOREM 3.3. *Let A be an n-by-n combinatorially symmetric matrix satisfying* (1.3). *If the undirected graph of A is a forest and A is invariantly ordered, then $u_{ij} = a_{ij}$ for all $i < j$ in the unit LU factorization of A. Conversely, if $u_{ij} = a_{ij}$ (generically) for all $i < j$, then the undirected graph of A is a forest and A is invariantly ordered.* □

We now consider the more restrictive version of our question (1.2), in which we characterize inheritance of all entries for $i \leqq j$. Note that for the tridiagonal matrix given in the introduction, $u_{ij} = a_{ij}$ for all $i < j$ but not for all $i = j$. By analogy with Corollary 3.2 we have the following result.

COROLLARY 3.4. *Let A be an n-by-n matrix satisfying* (1.3). *If A is (i, j) lower restricted for all pairs $i \leqq j$, then $u_{ij} = a_{ij}$ for all $i \leqq j$ in the unit LU factorization of A. Conversely, if $u_{ij} = a_{ij}$ (generically) for all $i \leqq j$, then A is (i, j) lower restricted for all pairs $i \leqq j$.* □

Matrices which satisfy these conditions can have no p-cycle for $p \geqq 2$, and thus must be reducible; in fact, they must be essentially triangular (although not all essentially triangular matrices satisfy these conditions). In addition, if A is (i, j) lower restricted for all pairs $i \leqq j$, then $\det A = \prod_{i=1}^{n} a_{ii}$. Some examples of matrices satisfying these conditions are given below, where the entries a_{ij} are arbitrary (subject to (1.3)).

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22} & 0 & 0 \\ 0 & a_{32} & a_{33} & a_{34} \\ 0 & a_{42} & 0 & a_{44} \end{bmatrix}, \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & 0 \\ 0 & 0 & a_{43} & a_{44} \end{bmatrix}, \quad \begin{bmatrix} a_{11} & 0 & a_{13} & 0 \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & 0 \\ a_{41} & 0 & a_{43} & a_{44} \end{bmatrix}.$$

4. Inheritance in the right unit *LU* factorization. Results in the previous two sections characterize inheritance in the upper triangular factor when the main diagonal entries of the lower triangular factor are all unity. There are obviously analogous characterizations for inheritance in the lower triangular factor when the diagonal entries of the upper triangular factor are all unity. We call this the *right unit LU factorization* of a matrix A. By taking transposes, $A^T = U^TL^T$ and it is easy to see that Theorems 2.1 and 2.2 with $\{1, 2, \cdots, i-1\}$ replaced by $\{1, 2, \cdots, j-1\}$ in each case give conditions for $l_{ij} = a_{ij}$ for particular $i, j \leqq i$ in the right unit LU factorization of matrix A. Thus we have the result that if A satisfying (1.3) is (i, j) lower restricted, then $l_{ij} = a_{ij}$ for a given pair $i \geqq j$ in the right unit LU factorization of A.

To characterize global inheritance in this factorization, we introduce another definition. If a matrix A is (i, j) lower restricted for all pairs $i > j$, then we call A *backward lower restricted*. If this condition is true for a matrix A satisfying (1.3), then $l_{ij} = a_{ij}$ for all $i > j$ in its right unit LU factorization. The converse holds if the equality is generic (cf. Corollary 3.2). Assuming, in addition to the principal minor condition, that A is combinatorially symmetric, if A is invariantly ordered then $l_{ij} = a_{ij}$. The converse holds if the equality is generic (cf. Theorem 3.3). This follows because if A is combinatorially symmetric with a forest graph, then A is invariantly ordered if and only if A^T is. Thus, for the class of matrices specified above, inheritance in one unit LU factorization implies inheritance in the other unit LU factorization. Tridiagonal matrices are in this class (see the example in the introduction).

In the following result we characterize the simultaneous inheritance of entries (including those on the diagonal) in both LU factorizations.

THEOREM 4.1. *Let A be an n-by-n matrix satisfying* (1.3). *If for each $p \in \{1, 2, \cdots, n-1\}$ either $a_{pq} = 0$ for all q with $n \geqq q > p$ or else $a_{qp} = 0$ for all q with $n \geqq q > p$, then $u_{ij} = a_{ij}$ for all $i \leqq j$ in the unit LU factorization and $l_{ij} = a_{ij}$ for all $i \geqq j$ in the right unit LU factorization. The converse holds if the equalities are generic.*

Proof. Let $U = [u_{ij}]$ and $L = [l_{ij}]$, respectively, denote the upper and lower factors in the unit LU and right unit LU factorizations. Let $1 \leqq p \leqq n-1$ and assume first that either $a_{pq} = 0$ or else that $a_{qp} = 0$ for all $q > p$. Thus, in the digraph of A there cannot exist a path of the form $q_1 \rightarrow p \rightarrow q_2$ for any $q_1, q_2 > p$. That is, A is (i, j) lower restricted for all pairs (i, j), and therefore $u_{ij} = a_{ij}$ and $l_{ij} = a_{ij}$ for all $i \leqq j$ and $i \geqq j$, respectively.

To prove the converse, assume that $u_{ij} = a_{ij}$ (generically) and $l_{ij} = a_{ij}$ (generically) for $i \leqq j$ and $i \geqq j$, respectively. Let $1 \leqq p \leqq n-1$ and $p < r \leqq n$, and suppose that $a_{pr} \neq 0$. If $a_{qp} \neq 0$ for any $q > p$, then $a_{qp}a_{pr} \neq 0$ implies either that $u_{qr} \neq a_{qr}$ (when $q \leqq r$) or $l_{qr} \neq a_{qr}$ (when $q \geqq r$). In either case the assumption is contradicted and thus $a_{qp} = 0$ for all $q > p$. Similarly, it can be shown that $a_{rp} \neq 0$ for $r > p$ implies that a_{pq} must equal zero for all $q > p$. □

We call a matrix which satisfies the zero/nonzero pattern of Theorem 4.1 a *sawtooth* matrix, and note that it can be displayed as

$$A = \left[\begin{array}{c|c|c} a_{11} & \multicolumn{2}{c}{A_{12}} \\ \cline{1-1}\cline{2-3} A_{21} & a_{22} & A_{23} \\ \cline{2-2}\cline{3-3} & A_{32} & a_{33} \\ & & \ddots \end{array}\right],$$

where a_{ii} are arbitrary and one of $A_{i,i+1}$, $A_{i+1,i}$ is 0 while the other is arbitrary. Note that the first two examples at the end of § 3 are sawtooth matrices, but the third example is not.

5. *UL* factorizations. Whereas we have concentrated thus far on *LU* factorizations, we now state results for the analogous problems for *UL* factorizations of A. Again we consider the two normalizations; when all diagonal entries of U (respectively, L) are equal to 1, we call this a left (right) unit *UL* factorization. Each of these factorizations is unique (cf. Theorem 1.1) if and only if all proper trailing principal minors are nonzero, that is, if and only if

$$(5.1) \qquad \det A[\{n-k+1, \cdots, n\}] \neq 0 \quad \text{for } k=1,2,\cdots,n-1.$$

To consider inheritance of entries, we state a definition which is the analogue of (i,j) lower restricted. For $1 \leqq i, j \leqq n$, A is (i,j) *upper restricted* if there is no path (of length $\geqq 2$) in $D(A)$ from i to j such that all intermediate nodes on this path are $>\max\{i,j\}$. Note that A is always (i,n) and (n,j) upper restricted. A sufficient condition for local inheritance then parallels Theorem 2.1. If A satisfies (5.1) and is (i,j) upper restricted, then $l_{ij} = a_{ij}$ $(u_{ij} = a_{ij})$ for given $i \geqq j$ $(i \leqq j)$ in the left (right) unit *UL* factorization. Necessary and sufficient conditions for local inheritance can thus be given by analogy with Theorem 2.2.

For global inheritance we need definitions analogous to those for lower restricted. *A is backward (forward) upper restricted* if A is (i,j) upper restricted for all pairs $i > j$ $(i < j)$. The result analogous to Corollary 3.2 then takes the following form.

> Given A satisfying (5.1), if A is backward (forward) upper restricted, then $l_{ij} = a_{ij}$ $(u_{ij} = a_{ij})$ for all $i > j$ $(i < j)$ in the left (right) unit *UL* factorization of A.

The following result corresponds to Theorem 3.3, and contains an analogue of the condition that A is invariantly ordered.

> Given a combinatorially symmetric matrix A satisfying (5.1), if the undirected graph of A is a forest and A has at most one nonzero entry in each column above the main diagonal, then $l_{ij} = a_{ij}$ $(u_{ij} = a_{ij})$ for all $i > j$ $(i < j)$ in the left (right) unit *UL* factorization of A.

On requiring that all off-diagonal entries of A be inherited in L and U, we obtain the following (with the appropriate normalizations for L and U).

> Given an irreducible, combinatorially symmetric matrix satisfying (1.3) and (5.1), if A is tridiagonal then $u_{ij} = a_{ij}$ for all $i < j$ in the unit *LU* factorization of A and $l_{ij} = a_{ij}$ for all $i > j$ in the left unit *UL* factorization of A.

The converses of the three statements above all hold if the equalities are generic.

6. Submatrix inheritance. Suppose that A has a unique unit *LU* factorization, and that some submatrix $A[\beta]$ also has such a factorization $A[\beta] = L(A[\beta])U(A[\beta])$; we now characterize when the i,j entries of $U \equiv U(A)$ and $U(A[\beta])$ are equal. For $\beta = \{1, 2, \cdots, p\}$, $1 \leqq p \leqq n$, this equality obviously holds for any A satisfying (1.3), and for all $i, j \in \beta$. But for more general β this is not necessarily true, and we seek to determine combinatorial circumstances under which it does hold.

In the case that the i,j entry of A is inherited by U we have the following result. Given a directed graph D with a self loop at each node, we let $\mathscr{A}'_D$ denote the set of all n-by-n matrices A which have all principal minors nonzero and which are consistent with D. (Note that $\mathscr{A}'_D \subseteq \mathscr{A}_D$.)

THEOREM 6.1. *Let D be a directed graph with a self loop at each node; let $\beta \subseteq \{1, 2, \cdots, n\}$ and $i, j \in \beta$ with $i \leqq j$. If for all $A \in \mathscr{A}'_D$, $u_{ij} = a_{ij}$ in the unit LU factorization of A, then it is also the case that $U(A[\beta])_{ij} = a_{ij}$ in the corresponding factorization of $A[\beta]$.*

Proof. As all $A \in \mathscr{A}'_D$ have nonzero principal minors, we can apply Theorem 2.2 with condition (ii) vacuous. Thus $u_{ij} = a_{ij}$ for all $A \in \mathscr{A}'_D$ implies that j is not reachable from i through $\{1, 2, \cdots, i-1\}$ in $D(A)$. But if this condition holds on $D(A)$ it necessarily holds (with respect to $\beta \cap \{1, 2, \cdots, i-1\}$) on the subgraph of $D(A)$ induced by any set β containing i and j. Thus, using Theorem 2.2 again, the i, j entry is also inherited by $U(A[\beta])$, and so $U(A[\beta])_{ij} = u_{ij} = a_{ij}$. □

If the conditions of this theorem are relaxed to allow some $a_{ii} = 0$, then the result is not necessarily true. It is also easy to give an example that shows the converse of the theorem need not be true.

We now give a characterization of inheritance of any given entry in a certain submatrix.

THEOREM 6.2. *Let D be a directed graph with a self loop at each node and let $A \in \mathscr{A}'_D$. Then for given $i, j \in \beta \subseteq \{1, 2, \cdots, n\}$ with $i \leqq j$, $U(A[\beta])_{ij} = u_{ij}$ for all $A \in \mathscr{A}'_D$ (in the unit LU factorizations of $A[\beta]$ and A) if and only if every path from i to j in D through $\gamma = \{1, 2, \cdots, i-1\}$ passes through nodes only in β.*

Proof. Let $v_{ij} = U(A[\beta])_{ij}$. Then (cf. (2.1))

$$v_{ij} = \frac{\det A[(\gamma \cap \beta) \cup \{i\} \mid (\gamma \cap \beta) \cup \{j\}]}{\det A[\gamma \cap \beta]}. \tag{6.1}$$

This numerator can be expanded about the ith row (cf. (2.3)), giving

$$\begin{aligned} &\det A[(\gamma \cap \beta) \cup \{i\} \mid (\gamma \cap \beta) \cup \{j\}] \\ &\quad = a_{ij} \det A[\gamma \cap \beta] + \sum \pm a_{iq_1} a_{q_1 q_2} \cdots a_{q_s j} \det A[(\gamma \cap \beta) - \{q_1, q_2, \cdots, q_s\}] \end{aligned} \tag{6.2}$$

where the summation is now over all paths from i to j through nodes $q_1, q_2, \cdots, q_s \in \gamma \cap \beta$, and $s \geqq 1$.

Assume first that $v_{ij} = u_{ij}$ for all $A \in \mathscr{A}'_D$. Then the terms involved in the summations of (2.3) and (6.2) must be equal for all such A. If for some A there is a path from i to j in γ which includes a node not in β, then this path product multiplied by its complementary determinant will occur in u_{ij} (from (2.3)) but not in v_{ij} (from (6.2)). As A has all principal minors nonzero, in particular this complementary determinant is nonzero, so we have a contradiction.

For the converse, assume that every path from i to j in D through γ passes through nodes only in β. Then the summations in (2.3) and (6.2) are over an identical set of paths from i to j. Thus $u_{ij} = v_{ij}$ for all $A \in \mathscr{A}'_D$ if and only if we have for all such matrices

$$\frac{\det A[\gamma - \{p_1, p_2, \cdots, p_t\}]}{\det A[\gamma]} = \frac{\det A[(\gamma \cap \beta) - \{p_1, p_2, \cdots, p_t\}]}{\det A[\gamma \cap \beta]}, \tag{6.3}$$

for all $\{p_1, p_2, \cdots, p_t\} \subseteq \gamma \cap \beta$ such that there exists a path in D from i to j through $p_1, p_2, \cdots, p_t$. Now, as each A is assumed to have nonzero principal minors, all submatrices in (6.3) are nonsingular, so that Schur complements exist, and we can use ideas developed in [1]. Let ε be the set of nodes in $\gamma \cap \beta$ which are on no path from i to j. For any given path as above, let δ denote the set of all nodes in $\gamma \cap \beta$ that are on some path from i to j that contains at least one of the nodes $p_1, p_2, \cdots, p_t$. Note that $\{p_1, p_2, \cdots, p_t\} \subseteq \delta$. The nodes can be ordered so that A is given in partitioned form as

$$A[\gamma \cap \beta] = \begin{bmatrix} A[\varepsilon] & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \quad \text{where } A_{22} = \begin{bmatrix} A[\delta] & A_{23} \\ A_{32} & A[(\gamma \cap \beta) - \varepsilon - \delta] \end{bmatrix}.$$

On taking the Schur complement of $A[\varepsilon]$ in $A[\gamma \cap \beta]$:

$$\det A[\gamma \cap \beta] = \det A[\varepsilon] \det (A_{22} - A_{21} A[\varepsilon]^{-1} A_{12}).$$

But $A_{21} A[\varepsilon]^{-1} A_{12}$ must be zero for every matrix A satisfying our assumptions, as otherwise there would exist a path from i to j through $\gamma \cap \beta$ including nodes of ε, which contradicts the definition of ε. Consequently, $\det A[\gamma \cap \beta] = \det A[\varepsilon] \det A_{22}$. Similarly, taking the Schur complement of $A[\delta]$ in A_{22} gives

$$\det A[\gamma \cap \beta] = \det A[\varepsilon] \det A[\delta] \det A[(\gamma \cap \beta) - \varepsilon - \delta].$$

But $\{p_1, p_2, \cdots, p_t\} \subseteq \delta$, so the right side of (6.3) reduces to

$$(6.4) \qquad \frac{\det A[\delta - \{p_1, p_2, \cdots, p_t\}]}{\det A[\delta]}.$$

Using the same reasoning on the left side of (6.3), it is also equal to (6.4) and the result follows. □

Note that if there is no path from i to j in γ (that is, j is not reachable from i through $\{1, 2, \cdots, i-1\}$), then we conclude from Theorems 6.2 and 2.2 that $U(A[\beta])_{ij} = u_{ij} = a_{ij}$ for all $A \in \mathcal{A}'_D$, giving inheritance of this entry. Obvious analogous results hold for the other three unit factorizations described in §§ 4 and 5.

7. Relations with Gaussian elimination.

The relationship between Gaussian elimination and LU factorization is well known, especially in the numerical analysis literature. There also is a relationship between Gaussian elimination and Schur complements, so our results concerning inheritance of entries in U are related to results in [1].

Consider an n-by-n matrix A that satisfies (1.3) or, equivalently, for which all Gaussian elimination pivots are nonzero. Inheritance in the matrices L and U of an LU factorization of A is closely related to the concept of fill-in. When $a_{ij} = 0$ and node j is reachable from node i through $\{1, 2, \cdots, \min(i,j) - 1\}$, there is said to be *fill-in* at the (i,j) position. This idea is particularly important for large, sparse matrices, where it is desirable to minimize the fill-in; and has been discussed by many authors (see, e.g., [3], [4], [8], [11], [12]). Much of the emphasis in the literature concerning fill-in in sparse matrices concerns the determination of permutation matrices P, Q so that either PAP^T or PAQ has less fill-in than A; we have not discussed this important practical problem. The relationship between our results and those of [12] was given in § 2. In [8], consideration is restricted to symmetric positive definite matrices using the undirected graph of A and the Cholesky factorization $A = LL^T$. The following result is an immediate consequence of our Theorem 2.2 (and its analogue characterizing inheritance in L), and is an extension of Theorem 1 of [12] and Theorem 5.1.2 of [8].

COROLLARY 7.1. *Let A be an n-by-n matrix satisfying* (1.3) *and suppose that $a_{ij} = 0$. Then, ignoring accidental numerical cancellation, $u_{ij} \neq 0$ (if $i \leqq j$) or $l_{ij} \neq 0$ (if $i \geqq j$) in any LU factorization of A if and only if there exists a path $i \to k_1 \to k_2 \to \cdots \to k_t \to j$ in $D(A)$ with $k_p \in \{1, 2, \cdots, \min(i,j) - 1\}$, $1 \leqq p \leqq t$, and*

$$\det A[\{1, 2, \cdots, \min(i,j) - 1\} - \{k_1, k_2, \cdots, k_t\}] \neq 0. \qquad \square$$

Because of condition (ii) of Theorem 2.2, our results are more general and more precise than those contained in the literature concerning fill-in in sparse matrices. In addition, we require only that A has a unique unit LU factorization and we characterize the inheritance of both zero and nonzero entries. As shown by the example following Theorem 2.2, (see (2.4)), the use of condition (ii) of that theorem to deduce that $u_{ij} = a_{ij}$ (generically) is interesting in that the (i,j) entry of U may indeed change during the process of determining the LU factorization; however, the equality is guaranteed by the combinatorial structure.

REFERENCES

[1] W. W. BARRETT, C. R. JOHNSON, D. D. OLESKY, AND P. VAN DEN DRIESSCHE, *Inherited matrix entries: principal submatrices of the inverse*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 313–322.

[2] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[3] R. K. BRAYTON, F. G. GUSTAVSON, AND R. A. WILLOUGHBY, *Some results on sparse matrices*, Math. Comp., 24 (1970), pp. 937–954.

[4] I. S. DUFF, *Research directions in sparse matrix computations*, MAA Studies in Math., Studies in Numerical Analysis, G. H. Golub, ed., 24, 1984.

[5] I. S. DUFF, A. M. ERISMAN, C. W. GEAR, AND J. K. REID, *Some remarks on inverses of sparse matrices*, Argonne National Laboratory Technical Memorandum, 51 (1985), pp. 1–10.

[6] M. FIEDLER, *Special Matrices and their Applications in Numerical Mathematics*, Martinus Nijhorff, Dordrecht, The Netherlands, 1986.

[7] F. R. GANTMACHER, *Matrix Theory*, Vol. I, Chelsea, New York, 1959.

[8] J. A. GEORGE AND J. W. H. LIU, *Computer Solutions of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[9] R. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[10] C. M. LAU AND T. L. MARKHAM, *LU factorizations*, Czech. Math. J., 29 (1979), pp. 546–550.

[11] S. PISSANETSKY, *Sparse Matrix Technology*, Academic Press, New York, 1984.

[12] D. J. ROSE AND R. E. TARJAN, *Algorithmic aspects of vertex elimination on directed graphs*, SIAM J. Appl. Math., 34 (1978), pp. 176–197.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 1, pp. 105–117, January 1989

008

VARIATIONAL PRINCIPLES FOR EIGENVALUES OF NONSYMMETRIC MATRICES*

GILES AUCHMUTY†

Abstract. Some novel variational principles for finding real and complex eigenvalues of the generalized eigenproblem $Ax = \lambda Dx$ are formulated and analyzed. A is a general matrix, D is assumed to be real symmetric and positive definite. One class of principles is based on constrained minimization on the set where $\|Dx\| = 1$. The other class involves the minimization of smooth functions on the complement of certain closed convex sets. The minima of these functions occur either at certain eigenvectors, or certain singular vectors, of the problem. The eigenvalue is determined as a certain functional of the minimizer. A numerical implementation of these principles is described.

Key words. eigenvalues, matrices, variational principles

AMS(MOS) subject classifications. primary 49G05; secondary 15A18.

1. Introduction. This paper describes and analyzes some variational principles for finding "real and complex" eigenvalues and eigenvectors of general $n \times n$ matrices. The principles involve minimizing certain smooth functions on specific subsets of $\mathbb{R}^n$ (for real eigenvalues) or $\mathbb{R}^{2n} \times [0, 2\pi]$ (for complex eigenvalues). The minimizers occur at specific eigenvectors of the matrix and their norms give the eigenvalue—or its modulus. The functions are defined so that the optimal value is zero at eigenvectors, and is independent of the eigenvalue. There are critical points that are not minima of these problems and they are generalized singular vectors of associated matrices.

These variational principles do not require any symmetry or nonnegativity assumptions. They are quite different from Rayleigh's principle and related power methods for finding real eigenvalues of symmetric matrices. Also, they appear to be quite different to the variational characterizations of the largest eigenvalues and corresponding eigenvectors of stochastic matrices. In particular, the eigenvalue equations are not obtained by a direct application of the Lagrange multiplier rule, as in Rayleigh's principle. The methods provide information on all the eigenvalues of a nonsingular matrix, not just the largest or smallest ones.

Partial motivation for this work has been the questions arising in bifurcation and stability theory. Very often we are interested in studying how the eigenvalues and eigenvectors of a family of matrices depend on a parameter. For families of real symmetric matrices, Rayleigh's principle provides quick and simple methods for tracking various eigenvalues. The variational principles to be described here can be used in a similar manner. Also they provide estimates on the spectral radius of a matrix, and its inverse when it is nonsingular. In Theorem 8 and its corollaries we also describe various localization theorems on the eigenvalues. At present, it is not at all clear how the numerical solution of these variational principles compares with other numerical methods of eigenvalue estimation or localization. In § 6, however, we describe some results on implementing one of these principles for some simple families of matrices.

In § 2, we introduce some notation and background material for this problem. This problem is an example of nonstandard variational principles of the type described in [1]. The motivation for many of these comes from the systematic use of basic ideas in convex analysis, notably the use of conjugate convex functions and their extremality conditions.

* Received by the editors March 10, 1988; accepted for publication (in revised form) July 25, 1988. This research was partially supported by National Science Foundation grant DMS-8701886.

† Department of Mathematics, University of Houston, Houston, Texas 77004.

These are described in § 2. In § 3, we describe an elementary variational principle of least squares type for finding real eigenvalues of the generalized eigenvalue problem. In § 4, a very different principle based on a generalized Young's inequality is introduced and analyzed. The principle only finds positive eigenvalues and corresponding eigenvectors. Then § 5 describes how either of these variational principles may be extended to find complex eigenvalues of general matrices. In § 6, we conclude with a simple example of the numerical application of these principles.

2. Notation and background. Throughout this paper, $\mathbb{R}^n$ will be the usual n-dimensional real vector space with the inner product and norm defined by

$$\langle x,y\rangle=\sum_{j=1}^{n} x_j y_j, \qquad \|x\|=\left(\sum_{j=1}^{n} x_j^2\right)^{1/2}.$$

Let $A = (a_{ij})$ be a real $n \times n$ matrix, not necessarily symmetric. Its transpose is denoted A^T and the norm of A is

$$\|A\| = \sup_{\|x\|=1} \|Ax\|.$$

We shall only use the 2-norm and the corresponding induced matrix norm in this paper.

Our interest is in finding variational principles for the weighted, or generalized, eigenproblem of finding nontrivial solutions of

$$Ax=\lambda Dx \tag{2.1}$$

where D is a real symmetric, positive definite $n \times n$ matrix. The standard eigenproblem has $D = I_n$ being the $n \times n$ identity matrix, but since our methods work equally well in this weighted case the more general case will be analyzed.

A nonzero vector x in $\mathbb{R}^n$ is said to be normalized if $\|Dx\| = 1$ and $B_r = \{x \in \mathbb{R}^n\colon \|Dx\| \leqq r\}$, $S_r = \{x \in \mathbb{R}^n\colon \|Dx\| = r\}$ are the ball, respectively sphere, of radius r, center the origin in $\mathbb{R}^n$ with respect to D. When x, y are vectors in $\mathbb{R}^n$, then $x \wedge y = xy^T = (x_i y_j)$ is a rank 1 matrix. Other terms from linear algebra will be used in the sense defined in Horn and Johnson [2].

Let K be a closed subset of $\mathbb{R}^n$, $\bar{\mathbb{R}} = [-\infty, \infty]$ and $F\colon K \to \bar{\mathbb{R}}$ be a given function. Then $\alpha = \inf_{x\in K} F(x)$ is called the value of F on K. F is said to be bounded below on K is α is finite. A point $\hat{x}$ in K is a minimizer of F on K if $F(\hat{x}) = \alpha$ and we say that α is attained if F has a minimizer on K.

An interior point $\tilde{x}$ of K is a critical point of F if either (i) F is not differentiable at $\tilde{x}$, or (ii) $\nabla F(\tilde{x}) = 0$ where

$$\nabla F(\tilde{x})=\left(\frac{\partial F}{\partial x_1}(\tilde{x}), \cdots, \frac{\partial F}{\partial x_n}(\tilde{x})\right)^T$$

is the gradient of F at $\tilde{x}$.

When F is twice-continuously differentiable at x, then its Hessian at x is $D^2F(x) = (\partial^2 F(x)/\partial x_i \partial x_j)$ and is an $n \times n$ symmetric matrix.

When K is not bounded, we say that F is coercive on K provided

$$\operatorname*{Lim\ inf}_{\|x\|\to\infty} \frac{F(x)}{\|x\|} = \infty.$$

It is weakly coercive on K if $F(x) \to \infty$ as $\|x\| \to \infty$ for x in K.

We shall also use some terminology and results from convex analysis. When f: $\mathbb{R}^n \to \bar{\mathbb{R}}$ is a given function, then its conjugate convex (or polar) function f^*: $\mathbb{R}^n \to \bar{\mathbb{R}}$

is defined by

$$f^*(y) = \sup_{x \in \mathbb{R}^n} [\langle x,y\rangle - f(x)]. \tag{2.2}$$

This implies that

$$f(x) + f^*(y) \geqq \langle x,y\rangle \tag{2.3}$$

for all x, $y \in \mathbb{R}^n$. Equality holds in (2.3) if and only if

$$y \in \partial f(x) \tag{2.4}$$

where $\partial f(x) = \{w \in \mathbb{R}^n : f(y) - f(x) \geqq \langle w, y - x\rangle$ for all y in $\mathbb{R}^n\}$. $\partial f(x)$ is called the subdifferential of f at x and (2.3) is the generalized Young's inequality (see Zeidler [3, § 51.1]). Any other undefined terms in this paper should be taken in the same sense as in [3].

When $f(x) = (1/p)\|x\|^p$ with $1 \leqq p < \infty$, then

$$f^*(y) = \begin{cases} \chi_1(y) & \text{when } p = 1, \\ \dfrac{1}{q}\|y\|^q & \text{where } q = \dfrac{p}{p-1} \text{ and } 1 < p < \infty. \end{cases}$$

Here

$$\chi_1(y) = \begin{cases} 0 & \text{if } \|y\| \leqq 1, \\ \infty & \text{otherwise} \end{cases} \tag{2.5}$$

is the indicator function of the unit ball in $\mathbb{R}^n$.

For this f, inequality (2.3) becomes the classical Young's inequality

$$\frac{1}{p}\|x\|^p + \frac{1}{q}\|y\|^q \geqq \langle x,y\rangle \tag{2.6}$$

for all x, y in $\mathbb{R}^n$ and where $1/p + 1/q = 1$.

(2.7) Equality holds here if and only if $x = y = 0$ or else $y = \|x\|^{p-2}x$

3. Variational criterion for the existence of real eigenvalues. Our interest is in describing and analyzing some variational principles for the eigenvalues and eigenvectors of A with respect to D. We shall first describe some criteria of least squares type for finding real eigenvalues.

Consider the function $E\colon S_1 \to [0, \infty)$ defined by

$$\begin{aligned} E(x) &= \|Ax - \langle Ax, Dx\rangle Dx\|^2 \\ &= \|Ax\|^2 - \langle Ax, Dx\rangle^2. \end{aligned} \tag{3.1}$$

THEOREM 1. *Assume A, D, E, S_1 as above; then we have the following:*

(i) *$\alpha = \inf_{x \in S_1} E(x)$ is nonnegative, and this infimum is attained.*

(ii) *The eigenproblem (2.1) has a real eigenvalue if and only if α is zero. In this case, a minimizer of E on S_1 is a normalized eigenvector $\hat{x}$ corresponding to the real eigenvalue $\hat{\lambda} = \langle A\hat{x}, D\hat{x}\rangle$.*

Proof. Since D is positive definite, S_1 is a compact set. E is a continuous function, so Weierstrass' theorem implies α is finite and the infimum is attained. Nonnegativity follows from (3.1).

When $\alpha = 0$, then (i) implies there is an $\hat{x}$ in S_1 such that $E(\hat{x}) = 0$ and $A\hat{x} = \langle A\hat{x}, D\hat{x}\rangle D\hat{x}$. Hence $\hat{x}$ is an eigenvector of (A, D) corresponding to the real eigenvalue $\hat{\lambda} = \langle A\hat{x}, D\hat{x}\rangle$.

When (A, D) has a real eigenvalue $\tilde{\lambda}$, then there is a vector $\tilde{x}$ in S_1 obeying $A\tilde{x} = \tilde{\lambda}D\tilde{x}$. In this case $E(\tilde{x}) = 0$ and hence α is zero as claimed.

COROLLARY. *The eigenproblem* (2.1) *has no real eigenvalues if and only if A is nonsingular and*

$$\sup_{x \in S_1} \frac{|\langle Ax, Dx\rangle|}{\|Ax\|} = \beta < 1. \tag{3.2}$$

Proof. When A is singular then 0 is an eigenvalue of (2.1). When A is nonsingular, then $\|Ax\| \neq 0$ for x in S_1. From the theorem, (2.1) has a real eigenvalue if and only if α is zero which is equivalent to $\beta = 1$.

The function E also attains a finite maximum on S_1, and it may well have other critical points. Thus it has local extrema on S_1 which are not eigenvectors of (A, D). The following result shows that they are generalized singular vectors of a shifted matrix $A - \mu D$.

LEMMA 2.1. *Let $\tilde{x}$ be a constrained critical point of E on S_1. Then there exists real numbers μ_1, μ_2 such that $\tilde{x}$ obeys*

$$(A^T - \mu_1 D)(A - \mu_1 D)x = \mu_2 D^2 x. \tag{3.3}$$

Proof. From (3.1), we have

$$\tfrac{1}{2}\nabla E(x) = A^T Ax - \langle Ax, Dx\rangle(DA + A^T D)x. \tag{3.4}$$

The constrained critical points of E on S_1 obey

$$\nabla E(x) = \mu D^2 x$$

from the Lagrange multiplier rule, so when $\tilde{x}$ is a constrained critical point, we have (3.3) with

$$\mu_1 = \langle A\tilde{x}, D\tilde{x}\rangle \quad \text{and} \quad \mu_2 = \mu_1^2 + \mu.$$

In particular, when $D = I$, the critical points of E on S_1 are singular vectors of the shifted matrix $(A - \mu_1 I)$.

Computationally it is often preferable to work with unconstrained optimization problems. For this problem, consider $E_1 \colon \mathbb{R}^n \to \mathbb{R}$ defined by

$$E_1(x) = \|Ax - \langle Ax, Dx\rangle Dx\|^2 + (\|Dx\| - 1)^2. \tag{3.5}$$

E_1 differs from E only by the addition of a penalty term. Its properties may be summarized as follows.

THEOREM 2. *Assume A, D as before and E_1 is defined by* (3.5). *Then we have the following*:

(i) *E_1 is coercive and continuously differentiable on* $\mathbb{R}^n - \{0\}$.

(ii) *$\alpha_1 = \inf_{x \in \mathbb{R}^n} E_1(x)$ is nonnegative and attained.*

(iii) *(A, D) has a real eigenvalue if and only if $\alpha_1 = 0$. When $\alpha_1 = 0$, a minimizer $\hat{x}$ of E_1 on $\mathbb{R}^n$ is a normalized eigenvector of (A, D) corresponding to the real eigenvalue $\hat{\lambda} = \langle A\hat{x}, D\hat{x}\rangle$.*

(iv) *If $\tilde{x}$ is a nonzero critical point of E_1 on $\mathbb{R}^n$, then there exists constants μ_1, μ_2 such that $\tilde{x}$ obeys*

$$(A^T - \mu_1 D)(A - \mu_1 D)x = \mu_2 D^2 x. \tag{3.3}$$

Proof. (i) From (3.5) we have

$$E_1(x) \geqq \|Dx\|^2 - 2\|Dx\| + 1$$

since the first term is nonnegative. Since D is positive definite, there exists $d > 0$ such that $\|Dx\| \geqq d\|x\|$ for all x in $\mathbb{R}^n$.

Hence

$$\frac{E_1(x)}{\|x\|} \geqq d^2\|x\| - 2\|D\| + \frac{1}{\|x\|}, \tag{3.6}$$

and thus E_1 is coercive as claimed.

Upon expanding (3.5) we observe each term is differentiable when x is nonzero, and we find

$$\frac{1}{2}\nabla E_1(x) = \frac{1}{2}\nabla E(x) + (\|Dx\| - 1)\frac{D^2x}{\|Dx\|}$$

with $\nabla E(x)$ given by (3.4).

(ii) Since E_1 is coercive and continuous on $\mathbb{R}^n$, it is bounded below and attains its infimum. From (3.5) we have that $E_1(x) \geqq 0$ for all x, as each term is nonnegative.

(iii) This follows from (ii) of Theorem 1 since $\alpha_1 = 0$ if and only if $\alpha = 0$.

(iv) Using (3.4) and the above expression for $\nabla E_1(x)$, we have that if $\tilde{x}$ is a critical point of E_1, then it is a solution of

$$A^TAx - \langle Ax, Dx\rangle(DA + A^TD)x = (1 - \|Dx\|)\frac{D^2x}{\|Dx\|}.$$

After factoring this right-hand side, we obtain (3.3) with $\mu_1 = \langle A\tilde{x}, D\tilde{x}\rangle$ as before and $\mu_2 = (1 - \|D\tilde{x}\|)/\|D\tilde{x}\| + \mu_1^2$.

Here again, when $D = I$, we see that the nonzero critical points of E_1 on $\mathbb{R}^n$ are singular vectors of $A - \mu_1 I$.

4. Variational principles for positive real eigenvalues. The variational principles described in the last section provide information on all the real eigenvalues of (A, D). Here we shall describe quite a different variational principle based on the extremality conditions for Young's inequality. It will only provide information on positive eigenvalues and corresponding eigenvectors of (A, D).

Again A is assumed to be a real $n \times n$ matrix and D is a symmetric, positive definite, $n \times n$ real matrix. Define $F_p\colon \mathbb{R}^n \to \mathbb{R}$ by

$$F_p(x) = \frac{1}{p}\|Dx\|^p + \frac{1}{q}\|Ax\|^q - \langle Dx, Ax\rangle \tag{4.1}$$

where $1 < p < \infty$ and $q = p/(p-1)$ is the conjugate index to p.

For $R \geqq 0$, let $C_R = \{x \in \mathbb{R}^n\colon \|Dx\| \geqq R\}$. Then $C_0 = \mathbb{R}^n$ and C_R is the complement of the open ball with respect to D of center 0 and of radius R. C_R is always a connected, unbounded, closed set.

Consider the optimization problem of minimizing F_p on C_R. Let

$$\alpha_p(R) = \inf_{x \in C_R} F_p(x). \tag{4.2}$$

The following theorem summarizes the properties of this variational principle. Essentially it says that (A, D) has a positive eigenvalue λ obeying $\lambda \geqq R^{p-2}$ if and only if $\alpha_p(R) = 0$. When this holds, the minimizers of F_p on C_R are eigenvectors of (A, D)

corresponding to positive eigenvalues. To prove this theorem, we need the following lemma.

LEMMA 4.1. *Suppose A, D, F_p, p and q as above. Then $F_p(x) \geqq 0$ for all x in $\mathbb{R}^n$ and $F_p(\tilde{x}) = 0$ if and only if $\tilde{x}$* obeys

$$Ax = \|Dx\|^{p-2}Dx. \tag{4.3}$$

Proof. Consider the problem of evaluating

$$g(y) = \sup_{x \in \mathbb{R}^n} \left[\langle Dx, y \rangle - \frac{1}{p} \|Dx\|^p \right]$$

with $1 < p < \infty$. Since D is nonsingular, g is well defined and finite for all y in $\mathbb{R}^n$. The supremum is attained when

$$y = \|Dx\|^{p-2}Dx. \tag{4.4}$$

This implies that $g(y) = \|y\|^2/q$ where $1/q = 1 - 1/p$ or q is the conjugate index to p. Thus

$$\frac{1}{p}\|Dx\|^p + \frac{1}{q}\|y\|^q \geqq \langle Dx, y \rangle \tag{4.5}$$

for all x, y in $\mathbb{R}^n$ and equality holds here if and only (4.4) holds.

Substituting Ax for y we have the lemma.

This may be regarded as a generalized Young's inequality. Note that when $p = 2$, $F_2(x) = \frac{1}{2}\|Ax - Dx\|^2$ is a purely quadratic function which is minimized at solutions of

$$Ax = Dx.$$

In the rest of this section we shall only look at cases where $p \neq 2$. Moreover $p = 2$ is a dividing point between different behavior of the variational principle.

THEOREM 3. *Suppose A, D, C_R, F_p, and α_p as above with $R > 0$. When $2 < p < \infty$, then we have the following*:

(i) *F_p is coercive on C_R and $\alpha_p(R)$ is finite, nonnegative and it is attained.*

(ii) *$F_p(\tilde{x}) = 0$ if and only if $\tilde{x}$ is solution of* (2.1) *with*

$$\tilde{\lambda} = \|D\tilde{x}\|^{p-2}. \tag{4.6}$$

(iii) *(A, D) has a positive real eigenvalue λ obeying $\lambda \geqq R^{p-2}$ if and only if $\alpha_p(R) = 0$.*

When $1 < p < 2$, and A is nonsingular then (i) *and* (ii) *hold and* (iii)′ (2.1) *has a positive real eigenvalue λ, obeying*

$$\lambda \leqq R^{p-2} \quad \textit{if and only if } \alpha_p(R) = 0.$$

Proof. From Lemma 4.1 one has $F_p(x) \geqq 0$ for all x in $\mathbb{R}^n$ and thus $\alpha_p(R) \geqq 0$.

When $p > 2$, one has

$$F_p(x) \geqq \frac{1}{p}\|Dx\|^p - \|A\| \, \|D\| \, \|x\|^2$$

so D nonsingular implies F_p is coercive. F_p is obviously continuous so it attains its infimum on C_R and (i) holds. Property (ii) holds from Lemma 4.1.

When (2.1) has an eigenvalue λ obeying $\lambda \geqq R^{p-2}$, then let w be a corresponding normalized eigenvector of (2.1).

Let $\tilde{x} = \lambda^{\nu} w$ with $\nu = (p-2)^{-1}$; then $\|D\tilde{x}\| = \lambda^{\nu} \geqq R$ so $\tilde{x}$ is in C_R and $\tilde{x}$ is a solution of (2.1). Thus $F_p(\tilde{x}) = 0 \geqq \alpha_p(R)$. And hence $\alpha_p(R) = 0$. Conversely if $\alpha_p(R) = 0$, then from (i) there is an $\tilde{x}$ in C_R such that $F_p(\tilde{x}) = 0$ and hence from (ii), (A, D) has a real eigenvalue obeying $\lambda \geqq R^{p-2}$ as required.

$$\|Ax\| \geqq c\|x\| \tag{4.7}$$

for all x in $\mathbb{R}^n$. Thus

$$F_p(x) \geqq \frac{1}{p}\|Dx\|^p + \frac{c^q}{q}\|x\|^q - \|A\|\,\|D\|\,\|x\|^2.$$

When $1 < p < 2$, then $q > 2$ and hence F_p will be coercive. Thus (i), (ii) follow as before. Now $\lambda \leqq R^{p-2}$ implies $\lambda^{\nu} \geqq R$ so (iii)′ follows.

For fixed p, let $\Lambda_p(R) = \{\lambda_j : j \in J(R)\}$ be the set of distinct eigenvalues of (2.1) obeying $\lambda_j \geqq R^{p-2}$, when $p > 2$ or $\lambda_j \leqq R^{p-2}$ when $p < 2$.

When $\lambda_j \in \Lambda_p(R)$, let

$$E_j = \{x \in \mathbb{R}^n : Ax = \lambda_j Dx \text{ and } \|Dx\| = \lambda_j^{1/(p-2)}\}.$$

Then E_j is the set of critical points of F_p corresponding to the eigenvalue λ_j. If $\dim\ker(A - \lambda_j D) = d$, then E_j is diffeomorphic to a sphere of dimension $d-1$ when $d > 1$. When $d = 1$, E_j is a pair of points.

The following result quantifies the solutions of the variational problem.

COROLLARY 1. *Assume the conditions of the theorem and that* $\alpha_p(R) = 0$. *Define* $\Lambda_p(R)$ *and* E_j *as above, then the set of minimizers of* F_p *on* C_R *is* $\mathcal{M} = \bigcup_{j \in J(R)} E_j$. $\mathcal{M}$ *has a finite number of bounded, connected components. When each eigenvalue* λ_j *of* (2.1) *is simple, then* $\mathcal{M}$ *consists of exactly* $2J(R)$ *points.*

Proof. Since $\alpha_p(R) = 0$, the minimizers of F_p on C_R must obey $F_p(\tilde{x}) = 0$. The result now follows from (iii) of Theorem 3 and the definitions above.

The unusual feature of this variational principle is that the eigenvalue $\tilde{\lambda}$ is determined from the solution by (4.6). It is a functional of the solution—not a multiplier arising from the constraints as in Rayleigh-type principles. When $p > 2$, the zeros of F_p of larger (D–) norms are eigenvectors of (A, D) corresponding to larger real eigenvalues. When $1 < p < 2$, however, they correspond to the smaller positive eigenvalues.

To obtain a variational principle for the negative eigenvalues of A, we substitute $-A$ for A in (4.1) to obtain

$$\tilde{F}_p(x) = \frac{1}{p}\|Dx\|^p + \frac{1}{q}\|Ax\|^q + \langle Dx, Ax\rangle. \tag{4.8}$$

We may ask about the critical points of F_p on $\mathbb{R}^n$ or the local minimizers of F_p on C_R. The following result shows that these points again arise at generalized singular vectors of (A, D).

THEOREM 4. *Assume* A, D, F_p *as above. When* $1 < p < 2$, F_p *is continuously differentiable on* $\mathbb{R}^n - \{0\}$ *and when* $p > 2$, *is continuously differentiable on* $\mathbb{R}^n - \ker A$ *with*

$$\nabla F_p(x) = \|Dx\|^{p-2}D^2x + \|Ax\|^{q-2}A^TAx - (A^TD + DA)x. \tag{4.9}$$

When $\hat{x}$ *is a critical point of* F_p *on* $\mathbb{R}^n$ *with* $A\hat{x} \neq 0$, *then there exist* μ_1 *positive and* μ_2 *real such that*

$$(A^T - \mu_1 D)(A - \mu_1 D)\hat{x} = \mu_2 D^2\hat{x}. \tag{4.10}$$

When $\hat{x}$ is a local minimizer of F_p on C_R, then either $A\hat{x} = 0$ or (4.10) *again holds. If $\mu_2 = 0$ in* (4.10) *then $\hat{x}$ is an eigenvector of* (2.1) *corresponding to a positive eigenvalue.*

Proof. When $Ax \neq 0$, (4.9) follows by standard calculus. If $1 < p < 2$ then $q > 2$ and (4.9) has a continuous extension to $\mathbb{R}^n - \{0\}$.

When $\hat{x}$ is a critical point of F_p with $A\hat{x} \neq 0$, it must obey

$$\|A\hat{x}\|^{q-2}A^TAx-(A^TD+DA)x=-\|D\hat{x}\|^{p-2}D^2x.$$

Thus (4.10) holds with $\mu_1 = \|Ax\|^{2-q} > 0$ and $\mu_2 = \mu_1^2 - \|Dx\|^{p-2}\|Ax\|^{2-q}$. Take inner products of (4.10) with $\hat{x}$; then

$$\|(A-\mu_1 D)\hat{x}\|^2=\mu_2\|D\hat{x}\|^2.$$

Thus $\mu_2 = 0$ implies $\hat{x}$ is an eigenvector of (A, D) corresponding to the eigenvalue μ_1.

When $\hat{x}$ is a local minimizer of F_p on C_R with $A\hat{x} \neq 0$, then from the extremality conditions, there is a $\mu \geqq 0$ such that

$$\nabla F_p(\hat{x})-\mu D^2\hat{x}=0, \quad \text{and} \quad \mu(\|D\hat{x}\|^2-R^2)=0.$$

Thus if $\|D\hat{x}\| > R$, we have $\mu = 0$ and $\hat{x}$ is a critical point of F_p. When $\|D\hat{x}\| = R$, then μ may be nonzero and we have (4.10) with

$$\mu_2=\mu_1^2+(\mu-\|D\hat{x}\|^{p-2})\|A\hat{x}\|^{2-q}.$$

It is worth noting that the Hessian, or second derivative, of F_p is defined on $\mathbb{R}^n - \ker A$. From (4.9) we find that

$$\begin{aligned}D^2F_p(x)=&\ \|Dx\|^{p-2}D^2+\|Ax\|^{q-2}A^TA-(A^TD+DA)\\ &+(q-2)\|Dx\|^{p-4}Dx\wedge Dx+(q-2)\|Ax\|^{q-4}A^TAx\wedge A^TAx.\end{aligned} \tag{4.11}$$

When $1 < p \leqq 4/3$ this may be continuously extended to $\mathbb{R}^n - \{0\}$.

When $\hat{x}$ is a nonzero critical point of F_p, then $\hat{x}$ is said to be nondegenerate if $D^2F_p(\hat{x})$ is defined and nonsingular. The Morse index of $\hat{x}$ is the number of negative eigenvalues of $D^2F_p(\hat{x})$. We know that if $\hat{x}$ is an eigenvector of (A, D) corresponding to a positive eigenvalue, the Morse index $i(\hat{x})$ will be zero since $\hat{x}$ is a local (in fact global) minimizer of F_p. It would be interesting to know if there are other relationships between the Morse indices of critical points and some ordering of the singular values of A with respect to D.

When p increases to infinity, we have for each x in $\mathbb{R}^n$,

$$\lim_{p\to\infty} F_p(x)=F_\infty(x)=\chi_1(Dx)+\|Ax\|-\langle x,Ax\rangle \tag{4.12}$$

where χ_1 is defined by (2.5) and x is in $\mathbb{R}^n$. The problem of minimizing F_∞ on $\mathbb{R}^n$ is equivalent to minimizing

$$f(x)=\|Ax\|-\langle Ax,Dx\rangle \tag{4.13}$$

on the closed unit ball B_1 on $\mathbb{R}^n$. This is a constrained optimization problem which is very similar to the problem studied in § 3. The main difference is that here the constraint is that x lie in B_1 while in § 3 we required x be in S_1. The properties of this problem may be summarized as follows.

THEOREM 5. *Assume A, D, f as above, then* $\inf_{x\in B_1} f(x) = 0$. *If $\tilde{x}$ is a nonzero minimizer of f on B_1 then either $A\tilde{x} = 0$ or else $\tilde{x}$ is a normalized eigenvector of (A, D) corresponding to a positive eigenvalue.*

Proof. From Schwarz' inequality

$$f(x) \geqq \|Ax\|(1 - \|Dx\|) \geqq 0$$

for any x in B_1. Since $f(0) = 0$ we have $\inf_{x \in B_1} f(x) = 0$.

If $f(\tilde{x}) = 0$ with $A\tilde{x} \neq 0$, then we have $\|D\tilde{x}\| = 1$ and $\|A\tilde{x}\| = \langle D\tilde{x}, A\tilde{x}\rangle$. This implies $A\tilde{x} = \lambda D\tilde{x}$ from the equality condition in Schwarz' inequality and λ must be positive.

An interesting, related, variational principle is to minimize the function H_p: $\mathbb{R}^n - \{0\} \to \mathbb{R}$ defined by

$$H_p(x) = \|x\|^{-2} F_p(x) \tag{4.14}$$

where F_p is given by (4.1): Let

$$\gamma_p = \inf_{x \neq 0} H_p(x).$$

THEOREM 6. *Suppose* $1 < p < \infty$, $p \neq 2$, A *is nonsingular and* H_p *is defined by* (4.14). *Then we have the following*:

(i) H_p *is continuous and weakly coercive on* $\mathbb{R}^n - \{0\}$.

(ii) $\gamma_p \geqq 0$ *is finite and it is attained on* $\mathbb{R}^n - \{0\}$.

(iii) $\gamma_p = 0$ *if and only if* (2.1) *has a positive eigenvalue* λ. *In this case* $\tilde{x}$ *minimizes* H_p *on* $\mathbb{R}^n - \{0\}$ *if and only if* $\tilde{x}$ *is an eigenvector of* (A, D) *corresponding to the eigenvalue* $\tilde{\lambda} = \|D\tilde{x}\|^{p-2}$.

Proof. We have $H_p(x) \geqq 0$ for all x in $\mathbb{R}^n - \{0\}$ as $F_p(x) \geqq 0$. Since F_p is continuous one has H_p continuous. Since A is nonsingular, (4.7) holds so that

$$H_p(x) \geqq \frac{1}{p} \frac{\|Dx\|^p}{\|x\|^2} + \frac{c^q}{q} \|x\|^{q-2} - \|A\| \, \|D\|.$$

When D is positive definite we have from (3.6) that

$$H_p(x) \geqq \frac{d^p}{p} \|x\|^{p-2} + \frac{c^q}{q} \|x\|^{q-2} - \|A\| \, \|D\|. \tag{4.15}$$

Now $p > 2$ implies that H_p is weakly coercive on $\mathbb{R}^n - \{0\}$ and that $q < 2$. Thus $\operatorname{Lim\,inf}_{\|x\| \to 0} H_p(x) = +\infty$.

Equivalently for all positive k, the set $\{x \in \mathbb{R}^n - \{0\}: H_p(x) \leqq k\}$ is closed and bounded. It will be nonempty for k large enough. Thus $\gamma_p \geqq 0$ will be attained.

We have $\gamma_p = 0$ if and only if there is an $\tilde{x} \in \mathbb{R}^n - \{0\}$ for which $F_p(\tilde{x}) = 0$. Thus (iii) follows from Lemma 4.1.

Similarly when $1 < p < 2$, then $q > 2$ and (ii)–(iii) follow from (4.15) just as in the case $p > 2$.

Other functionals besides F_p and H_p could be used here. Let h: $[0, \infty] \to \mathbb{R}$ be a continuously differentiable, convex function, $f(x) = h(\langle Dx, x\rangle)$, and f^* be the conjugate convex function of f. Define F: $\mathbb{R}^n \to \mathbb{R}$ by

$$F(x) = f(x) + f^*(Ax) - \langle x, Ax\rangle. \tag{4.16}$$

Then $F(x) \geqq 0$ and $F(x) = 0$ if and only if

$$Ax \in \partial f(x) = \{2h'(\langle Dx, x\rangle)Dx\}. \tag{4.17}$$

Thus $F(\tilde{x}) = 0$ if and only if $\tilde{x}$ is an eigenvector of (2.1) corresponding to the eigenvalue $\tilde{\lambda} = 2h'(\langle Dx, x\rangle)$. Note that the range of the eigenvalues depends on the range of h', where h' is the derivative of h. When $h(s) = s^p/p$, $1 < p < \infty$ we have a theory similar to that already developed here.

5. Variational principles for complex eigenvalues. The results of §§ 3 and 4 may be generalized to provide variational principles for complex eigenvalues of the system (2.1).

To do this we observe that $\lambda_1 + i\lambda_2 = |\lambda| e^{i\theta}$ is a complex eigenvalue corresponding to the eigenvector $x = u + iv$ of (2.1) with u, v in $\mathbb{R}^n$ if and only if

$$\mathscr{A}_0 w = \begin{pmatrix} \cos\theta A & \sin\theta A \\ -\sin\theta A & \cos\theta A \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = |\lambda| \mathscr{D} w \tag{5.1}$$

where

$$\mathscr{D} = \begin{pmatrix} D & 0 \\ 0 & D \end{pmatrix} \quad \text{and} \quad w = \begin{pmatrix} u \\ v \end{pmatrix}.$$

This reduces the complex eigenvalue problem to one of finding positive eigenvalues of a $2n \times 2n$ system involving the angular variable θ.

Let $W_1 = \{(u, v) \in \mathbb{R}^{2n}\colon \|Du\|^2 + \|Dv\|^2 = 1\}$ and consider the function $\mathscr{E}$: $W_1 \times [0, 2\pi] \to [0, \infty)$ defined by

$$\begin{aligned} \mathscr{E}(u,v;\theta) &= \|\mathscr{A}_\theta w - \langle \mathscr{A}_\theta w, \mathscr{D}w\rangle \mathscr{D}w\|^2 \\ &= \|\mathscr{A}_\theta w\|^2 - \langle \mathscr{A}_\theta w, \mathscr{D}w\rangle^2. \end{aligned} \tag{5.2}$$

This is obtained from (3.1) by substituting $\mathscr{A}_\theta$ for A and $\mathscr{D}$ for D. The analogue of Theorem 1 now provides criteria for (2.1) to have eigenvalues in a given sector.

THEOREM 7. *Assume* A, D, $\mathscr{E}$, W_1 *as above and* $0 \leqq \theta_1 \leqq \theta_2 \leqq 2\pi$. *Then we have the following*:

(i) $$\alpha(\theta_1, \theta_2) = \inf_{(u,v,\theta) \in W_1 \times [\theta_1, \theta_2]} \mathscr{E}(u,v;\theta)$$

is nonnegative and is attained.

(ii) *The eigenproblem* (2.1) *has an eigenvalue* λ *lying in the sector* $\theta_1 \leqq \arg\lambda \leqq \theta_2$ *if and only if* $\alpha(\theta_1, \theta_2) = 0$. *In this case* $(\hat{u}, \hat{v}, \hat{\theta})$ *is a minimizer of* $\mathscr{E}$ *on* $W_1 \times [\theta_1, \theta_2]$ *if and only if* $\hat{u} + i\hat{v}$ *is an eigenvector of* (2.1) *corresponding to the eigenvalue* $\lambda = |\lambda| e^{i\hat{\theta}}$ *where*

$$|\lambda| = \langle \mathscr{A}_\theta \hat{w}, \mathscr{D}\hat{w}\rangle \quad \textit{and} \quad \hat{w} = (\hat{u}, \hat{v})^T.$$

Proof. This theorem follows directly from Theorem 1 and the representation (5.1).

Similarly the analogue of (4.1) for this problem is the function $\mathscr{F}_p$: $\mathbb{R}^{2n} \times [0, 2\pi] \to [0, \infty)$ defined by

$$\mathscr{F}_p(w,\theta) = \frac{1}{p}\|\mathscr{D}w\|^p + \frac{1}{q}\|\mathscr{A}_\theta w\|^2 - \langle \mathscr{D}w, \mathscr{A}_\theta w\rangle \tag{5.3}$$

with $1 < p < \infty$ and $q = p/(p-1)$ as before. The analogue of C_R is

$$D_R = \{(u,v,\theta) \in \mathbb{R}^{2n} \times [0, 2\pi]\colon \|Du\|^2 + \|Dv\|^2 \geqq R^2\}$$

and the variational problem is to minimize $\mathscr{F}_p$ on D_R: Let

$$\nu_p(R) = \inf_{D_R} \mathscr{F}_p(u,v,\theta). \tag{5.4}$$

From Theorem 3 and the representation (5.1) we obtain the following result.

THEOREM 8. *Suppose A, D, D_R, $\mathscr{F}_p$, ν_p as above with $R > 0$.*

When $p > 2$, then we have the following:

(i) *$\mathscr{F}_p$ is coercive on D_R and $\nu_p(R)$ is nonnegative and is attained.*

(ii) *$\mathscr{F}_p(\tilde{u}, \tilde{v}, \tilde{\theta}) = 0$ if and only if $\tilde{x} = \tilde{u} + i\tilde{v}$ is a solution of* (2.1) *corresponding to the eigenvalue $\tilde{\lambda} = |\tilde{\lambda}| e^{i\theta}$ where*

$$|\tilde{\lambda}|^2 = (\|D\tilde{u}\|^2 + \|D\tilde{v}\|^2)^{p-2}. \tag{5.5}$$

(iii) *A has a complex eigenvalue $\tilde{\lambda}$ obeying $|\tilde{\lambda}| \geqq R^{p-2}$ if and only if $\nu_p(R) = 0$.*

When $1 < p < 2$ with A nonsingular, then (i)–(ii) *above hold and*

(iii)$'$ *A has a complex eigenvalue $\tilde{\lambda}$ obeying $|\tilde{\lambda}| \leqq R^{p-2}$ if and only if $\nu_p(R) = 0$.*

COROLLARY 1. *Under the assumptions of the theorem, if $D = I$ is the identity matrix, $\nu_p(R) > 0$ and $p > 2$ then the spectral radius of A is less than R^{p-2}. A is a contraction if and only if $\nu_p(1) > 0$.*

Proof. This corollary follows from Theorem 8(iii). The spectral radius of A is $\max \{ |\lambda| : \lambda \text{ is an eigenvalue of } A \}$.

In a similar manner when $1 < p < 2$, A is nonsingular, we can obtain bounds on the spectral radius of A^{-1}. For all values of p, $p \neq 2$, $\nu_p(R)$ will be zero for R small enough; let

$$R_p = \inf \{ R > 0 : \nu_p(R) > 0 \}. \tag{5.6}$$

When $p > 2$, R_p is related to the largest (in absolute value) eigenvalue $\hat{\lambda}$ of (A, D) by $|\hat{\lambda}| = R_p^{p-2}$; while when $p < 2$, R_p is related to the least eigenvalue $\tilde{\lambda}$ by the same expression.

This functional may also be used to localize the eigenvalues. Suppose we wish to know if (A, D) has any eigenvalues in the region

$$K = \{ |\lambda| e^{i\theta} : \theta_1 \leqq \theta \leqq \theta_2 \text{ and } 0 < r_1 \leqq |\lambda| \leqq r_2 \}. \tag{5.7}$$

Let $\mathscr{K} = \{ (w, \theta) \in \mathbb{R}^{2n} \times [\theta_1 \theta_2] : r_1 \leqq \|\mathscr{D} w\|^{p-2} \leqq r_2 \}$ and consider the problem of minimizing $\mathscr{F}_p$ on $\mathscr{K}$.

COROLLARY 2. *Under the assumptions of the theorem with $1 < p < \infty$, $p \neq 2$, then (A, D) has an eigenvalue λ in K if and only if*

$$\inf_{(u,v,\theta) \in \mathscr{K}} \mathscr{F}_p(u, v, \theta) = 0. \tag{5.8}$$

Proof. This follows directly from Theorem 8(ii).

There also are similar results for the function H_p applied to the problem (5.1).

These results may even be generalized to complex eigenproblems for complex matrices $C = A + iB$ where A, B are real $n \times n$ matrices. The equation

$$Cx = \lambda D x \tag{5.9}$$

is equivalent to

$$\mathscr{C}_\theta W = |\lambda| \mathscr{D} w \tag{5.10}$$

where $w = (u, v)^T$, $x = u + iv$, $\lambda = |\lambda| e^{i\theta}$ and

$$\mathscr{C}_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} A & -B \\ B & A \end{pmatrix}.$$

Equation (5.10) is of the form (5.1) or (2.1) with $\mathscr{C}_\theta$ being a real $2n \times 2n$ matrix.

6. Numerical implementation. The variational principles described in the preceding sections were developed to help analyze and numerically compute eigenvalues of families of matrices. We have implemented the variational principle for finding the positive real eigenvalues of A by minimizing F_p on C_R as described in § 4 and Theorem 3, with $D = I$.

This is a constrained optimization problem defined on the complement of a convex set. It was converted into an unconstrained problem by adding a penalty term and the resulting function was K_p: $\mathbb{R}^n \to \mathbb{R}$ where

$$K_p(x) = \frac{1}{p}\|x\|^p + \frac{1}{q}\|Ax\|^q - \langle x, Ax\rangle + Q(\varepsilon^{-1}(R - \|x\|)) \tag{6.1}$$

where

$$Q(s) = \begin{cases} 0 & \text{for } s \leqq 0, \\ e^s - s - 1 & \text{for } s \geqq 0, \end{cases} \qquad \varepsilon > 0 \text{ and } 1 < p < \infty, \quad p \neq 2.$$

This function is C^1 on $\mathbb{R}^n$. To minimize K_p, a conjugate gradient method with an exact line search was used. In all the following calculations we took $p = 3$ and $\varepsilon = 0.1$. Given an initial vector $x^{(0)}$ and a choice of R, we sought the minima of K_3. When the minimal value was approximately zero, we computed the approximate eigenvalue by

$$\tilde{\lambda} = \|\tilde{x}\| \tag{6.2}$$

(since $p = 3$ in Theorem 3(ii)) with $\tilde{x}$ being the approximate minimizer and then computed the normalized residual

$$\tilde{r} = \frac{A\tilde{x} - \tilde{\lambda}\tilde{x}}{\|\tilde{x}\|}. \tag{6.3}$$

If $\|\tilde{r}\|$ is small, we have $\tilde{\lambda}$ is an approximate eigenvalue of A. Sometimes the minimizing $\tilde{x}$ was on the boundary of the domain ($\|\tilde{x}\| \cong R$) or is a singular vector of A as described in Theorem 4. In each of these cases $\|\tilde{r}\|$ was not small.

We shall describe the numerical results for two simple families of matrices.

Example 1. Let

$$A(\mu) = \begin{bmatrix} 1 & \mu \\ -\mu & 2 \end{bmatrix} \tag{6.4}$$

with μ being real. We have that $A(\mu)$ is skew-symmetric and its eigenvalues are given by

$$\lambda_\pm = \tfrac{1}{2}(3 \pm \sqrt{1 - 4\mu^2}).$$

They are real when $|\mu| \leqq \frac{1}{2}$; otherwise $A(\mu)$ has a pair of complex conjugate eigenvalues lying on the line Re $\lambda = \frac{3}{2}$.

TABLE 1
Results of minimizing K_3 with $A(\mu)$ defined by (6.4).

μ	Initial vector	No. of iterations	Minimal value	$\tilde{\lambda}$	$\tilde{x}$	$\|\tilde{r}\|$
0.0	*a*	8	5.519×10^{-12}	0.999999	$(0.999999, 3.24 \times 10^{-6})$	3.4×10^{-6}
0.0	*b*	6	3.02×10^{-14}	2.000000	$(2.48 \times 10^{-8}, 2.0)$	2.1×10^{-7}
0.2	*a*	7	1.122×10^{-12}	1.041754	(1.01977, 021284)	2.06×10^{-6}
0.2	*b*	11	2.78×10^{-11}	1.95826	(0.40008, 1.91696)	6.17×10^{-6}
0.4	*a*	13	5.925×10^{-11}	1.200001	(1.07331, 0.53667)	1.18×10^{-5}
0.4	*b*	13	4.688×10^{-11}	1.800011	(0.80499, 160998)	1.02×10^{-5}
0.48	*a*	34	8.53×10^{-10}	1.360124	(1.08800, 0.81621)	3.55×10^{-5}
0.6	*a*	7	7.415×10^{-3}	1.472521	(1.05900, 1.02314)	1.01×10^{-1}
1.0	*b*	21	1.359×10^{-1}	0.79945	−(0.667937, 4.43710)	7.37×10^{-1}

TABLE 2
Results of minimizing K_3 with $A(\mu)$ defined by (6.6).

μ	Initial vector	No. of iterations	Minimal value	$\tilde{\lambda}$	$\Vert\tilde{r}\Vert$
-1.0	*a*	15	1.354×10^{-11}	0.78282	7.77×10^{-6}
-1.0	*b*	40	6.657×10^{-11}	0.78280	1.80×10^{-5}
0.0	*a*	17	6.148×10^{-11}	1.000015	1.4×10^{-5}
0.0	*b*	13	2.045×10^{-13}	2.000000	4.6×10^{-7}
1.0	*a*	50	6.367×10^{-10}	1.35702	3.14×10^{-5}
1.0	*b*	74	6.586×10^{-10}	1.69213	2.79×10^{-5}
2.0	*a*	19	5.68×10^{-3}	1.51058	8.70×10^{-2}
2.0	*b*	37	5.68×10^{-3}	1.51058	8.70×10^{-2}

Table 1 summarizes the results of computations with $p = 3$, $\varepsilon = 0.1$, and either (a) the initial vector $x^{(0)} = (1.0, 0.9)$ or (b) $x^{(0)} = (0.5, 1.1)$. For each μ and initial vector we tabulate the minimal value of K_3, the computed minimizer and the value $\tilde{\lambda}$ from (6.2). The convergence criterion was that

$$\Vert\nabla K_3(\tilde{x})\Vert \leqq 1.0 \times 10^{-5}. \tag{6.5}$$

It is particularly informative to observe the behavior when $\mu > 0.5$. In these cases the minimal value and $\Vert\tilde{r}\Vert$ were much larger than those obtained when $0 \leqq \mu < 0.5$. All calculations were done in double precision arithmetic on a VAX 11/780.

Example 2. Let

$$A(\mu) = \begin{bmatrix} 1 & \mu & 0 \\ 0 & 2 & 1 \\ -1 & 0 & -3 \end{bmatrix}. \tag{6.6}$$

This matrix has the characteristic polynomial

$$\det(\lambda I - A(\mu)) = \lambda^3 - 7\lambda + (6 + \mu).$$

This matrix has three real eigenvalues if and only if

$$|\mu + 6| < 7\sqrt{28/27} \simeq 7.12845108$$

and it has positive eigenvalues if and only if

$$\mu < 7\sqrt{28/27} - 6 = 1.12845108.$$

Table 2 summarizes the results of computations with $p = 3$, $\varepsilon = 0.1$, and $R = 0.5$. The initial vectors were either (a) (1.5, 0.5, 0.5) or (b) (0.2, 1.2, 0.2) and the convergence criteria was (6.5) again. For economy of space, we shall omit the minimizing vector.

In this case, with $\mu = 2$, for many different initial data, we obtained the same minimizer with the same minimal value of 5.68×10^{-3}. Since we could not improve on this, and the residual was not small, we conclude that when $\mu = 2.0$, $A(\mu)$ has no positive eigenvalue.

REFERENCES

[1] G. AUCHMUTY, *Variational principles for operator equations and initial value problems*, J. Nonlinear Anal. TMA, 12 (1988), pp. 531–564.

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

[3] E. ZEIDLER, *Nonlinear Functional Analysis and Its Applications*, III, *Variational Methods and Optimization*, Springer-Verlag, Berlin, New York, 1985.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 1, pp. 118–134, January 1989

009

GENERALIZED INVERSES IN DISCRETE TIME MARKOV DECISION PROCESSES*

BERNARD F. LAMOND† AND MARTIN L. PUTERMAN‡

Abstract. A new self-contained approach based on the Drazin generalized inverse is used to derive many basic results in discrete time, finite state Markov decision processes. A product form representation for the transition matrix of a stationary policy gives new derivations of the average reward evaluation equations, Laurent series expansions, as well as the finite test for Blackwell optimality. This representation also suggests new computational methods.

Key words. dynamic programming, Markov decision processes, policy evaluation, generalized inverses, Drazin inverse

AMS(MOS) subject classifications. 15A09, 60J05, 90C40

1. Introduction. The purpose of this paper is to provide a simple and unified treatment of undiscounted Markov reward and decision process theory using purely algebraic methods. The cornerstone of this development is a matrix factorization of a stochastic matrix based on the Jordan canonical representation. All subsequent results are obtained using elementary linear algebra. Although the results obtained are not new, the approach herein is self-contained and more direct than the analytic methods of Blackwell (1962) and Veinott (1969). Rothblum (1981) has obtained closely related results but his methods are based on a more involved algebraic foundation. Ohno (1985) has used the Jordan canonical form in a different way to obtain bounds in undiscounted Markov decision processes.

Fundamental is the Drazin (1958) generalized inverse and a product form representation for it suggested by Campbell and Meyer (1979). The role of the generalized inverse for analyzing Markov chains has been recognized by many authors; a summary appears in Meyer (1982). Also, Kemeny (1981) and Hunter (1982), (1988) discussed the advantages of using generalized inverses in lieu of the usual fundamental matrix of Kemeny and Snell (1960) for analyzing finite Markov chains.

As shown in Lamond (1985), (1987), the product form representation of the Drazin inverse can be computed numerically by adapting standard matrix factorization methods. Hence our mathematically convenient transformation also suggests some new, efficient algorithms for policy evaluation.

In this paper, we show that the average reward policy evaluation equations of Howard (1960) are a direct consequence of the matrix decomposition, as is Veinott's (1969) Laurent series expansion of the resolvent. Moreover, we show that the recurrence relationship for its terms is a special case of a more general singular system of equations which we solve using the Drazin inverse. Similar equations were also considered by Veinott (1969). We provide a simple proof of the equivalence of $(n-m)$-discount optimality and Blackwell optimality in the case that the Blackwell optimal policy has m recurrent classes (cf. Veinott (1969), (1974)).

We now give a brief overview of this paper. The heart of the paper consists of §§ 3 and 4, which contain the main results on stationary Markov chains and Markov

* Received by the editors June 22, 1987; accepted for publication (in revised form) August 1, 1988. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada, under grant A-5527.

† Department of Systems and Industrial Engineering, The University of Arizona, Tucson, Arizona 85721.

‡ Faculty of Commerce and Business Administration, The University of British Columbia, Vancouver, British Columbia, Canada V6T 1Y8.

reward processes. Section 5 expands on the algebraic aspects of the policy evaluation equations and § 6 applies the results to Markov decision processes. The Drazin inverse is defined in § 2, first using the axioms of Drazin (1958) and next using the product form representation of Campbell and Meyer (1979).

In § 3, this matrix decomposition approach is used in a new derivation of the properties of the *limiting matrix*, the *deviation matrix* and the *fundamental matrix* of a discrete time Markov chain. In § 4, the decomposition of the transition matrix is used to give a direct derivation of the Laurent expansion of the resolvent and of the discounted value function, when the interest rate is small.

In § 5, we formulate the problem of finding the terms of a truncated series as a finite system of linear equations (as in Veinott (1969)). We show that the terms of the truncated series can be solved using the Drazin inverse of a special matrix of coefficients constructed from the Markov matrix of the process.

Finally, the application of the Drazin inverse approach to policy improvement algorithms is explored in § 6, with a new proof of the finite criterion for Blackwell optimality.

Section 7 provides conclusions and extensions.

2. Definition of the Drazin generalized inverse. Let A be a square $n \times n$ complex matrix, and let k be the smallest nonnegative integer such that

$$\operatorname{rank}(A^{k+1}) = \operatorname{rank}(A^k).$$

Then k is said to be the *index* of A, and is denoted by $k = \operatorname{ind}(A)$. Following Campbell and Meyer (1979), an $n \times n$ matrix A^D such that

$$A^D A A^D = A^D,$$

$$A A^D = A^D A,$$

$$A^{k+1} A^D = A^k$$

is called a *Drazin generalized inverse* of A. Evidently, A is nonsingular if and only if $\operatorname{ind}(A) = 0$ and in this case $A^D = A^{-1}$. Also, when $\operatorname{ind}(A) = 1$, A^D is called the *group inverse* and is denoted $A^{\#}$.

The following lemma gives a representation of the Drazin inverse of singular matrices. Recall that the *algebraic multiplicity* of a complex number μ for a square complex matrix A is the multiplicity of μ as a root in the characteristic polynomial of A. See, e.g., Campbell and Meyer (1979, Thm. 7.2.1) for the proof.

LEMMA 2.1. *Suppose A is a singular $n \times n$ complex matrix. Then there exists an $n \times n$ nonsingular complex matrix S such that*

$$A = S^{-1} \begin{pmatrix} B & 0 \\ 0 & C \end{pmatrix} S \tag{2.1}$$

where B is nonsingular and C is nilpotent. Moreover, given any such decomposition of A, the order of C is the algebraic multiplicity of zero for A, $\operatorname{ind}(A) = \min\{p = 1, 2, \cdots : C^p = 0\}$ and

$$A^D = S^{-1} \begin{pmatrix} B^{-1} & 0 \\ 0 & 0 \end{pmatrix} S. \tag{2.2}$$

Further, if A is real one can choose S, B and C of (2.1) to be real.

Note that the Drazin inverse A^D is unique even though, for fixed A, there are many different matrices S and B satisfying (2.1). The Jordan canonical form is one decomposition of the form given by (2.1). Of course, it does not usually have S as a real matrix. When A is real, decompositions given by (2.1) for which S is real are easier to compute than the Jordan form. For example, they do not necessarily require the computation of all eigenvalues of A. Computational algorithms for such decompositions are given in Wilkinson (1982) and Campbell and Meyer (1979). Also, an algorithm for computing A^D directly, without decomposing A as in (2.1), is given in Anstreicher and Rothblum (1987).

The next corollary specializes Lemma 2.1 to the case where $\operatorname{ind}(A) = 1$.

COROLLARY 2.2. *Suppose A is an $n \times n$ complex matrix. Then* $\operatorname{ind}(A) = 1$ *if and only if for some nonsingular matrix S,*

$$A = S^{-1}\begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} S \tag{2.3}$$

where B is a square nonsingular matrix. Moreover, in this case the order of B equals $n - m$, where m is the algebraic multiplicity of zero for A, and the Drazin inverse is given by

$$A^D = A^{\#} = S^{-1}\begin{pmatrix} B^{-1} & 0 \\ 0 & 0 \end{pmatrix} S. \tag{2.4}$$

Further, if A is real one can choose S and B of (2.3) *to be real.*

This special case is of particular interest for studying Markov decision processes. Furthermore, decompositions of the form (2.3) and (2.4) can be computed efficiently using the algorithm of Lamond (1987). As shown in Hunter (1988), many computational techniques for Markov chains can be expressed in terms of various generalized inverses.

In this paper, we focus on the Drazin inverse, its factorization in Corollary 2.2 and its use in Markov decision processes.

3. Decomposition of a stochastic matrix.

Suppose we are given a discrete time Markov chain with n states and transition matrix P. We assume that its transition matrix is *stochastic*, that is, $p_{ij} \geq 0$ for $i = 1, \cdots, n$ and $j = 1, \cdots, n$, and $\sum_{j=1}^{n} p_{ij} = 1$ for $i = 1, \cdots, n$. We assume, for simplicity, that the matrix is in normal form, i.e., that the states have been ordered in such a way that the transient states (if any) are first, followed by the recurrent classes, one after the other (see, e.g., Gantmacher (1960, page 74)). For example, a chain with two recurrent classes and some transient states would have a transition matrix P which could be transformed to appear as

$$P = \begin{pmatrix} P_{00} & P_{01} & P_{02} \\ 0 & P_{11} & 0 \\ 0 & 0 & P_{22} \end{pmatrix}.$$

We remark that for $k \geq 1$, P_{kk}, the submatrix corresponding to states in the recurrent class k, is stochastic and irreducible.

In this section, we further the approach of Campbell and Meyer (1979) and use the Perron-Frobenius theory to obtain a factorization for $I - P$ as in Corollary 2.2. Then we use this representation to obtain a new derivation of many important properties of the deviation matrix, the limiting matrix and the fundamental matrix of the associated Markov chain. Let $\sigma(A)$ denote the *spectral radius* of a matrix A, i.e., the modulus of its largest eigenvalue. The following lemma is well known and the main ingredients of its proof are included for the sake of completeness.

LEMMA 3.1. *If P is a stochastic matrix with m recurrent classes, then there is a similarity transformation S such that*

$$P = S^{-1}\begin{pmatrix} Q & 0 \\ 0 & I \end{pmatrix} S \tag{3.1}$$

where I is the $m \times m$ identity matrix and Q is an $(n-m) \times (n-m)$ matrix with no eigenvalue equal to 1. Moreover, $\sigma(Q) \leq 1$ and if P is aperiodic, $\sigma(Q) < 1$.

Proof. It is a classical result (Karlin (1968, Thm. 4.2.1)) that P has the eigenvalue 1 with algebraic multiplicity m and with m linearly independent eigenvectors. This implies that there exists a nonsingular matrix S such that (3.1) holds. Applying Lemmas 2.3 and 2.5 of Varga (1962) to the irreducible submatrices P_{kk}, we have that $\sigma(P_{00}) < 1$ and $\sigma(P_{kk}) \leq 1$ for $k = 1, \cdots, m$. This implies that $\sigma(Q) \leq 1$. The corollary to Theorem 2.3 of Varga (1962) implies that $\sigma(Q) < 1$ if every P_{kk} is aperiodic, $k = 1, \cdots, m$. □

Let $A = I - P$. Then Lemma 3.1 implies that A satisfies (2.3) with $B = I - Q$. This matrix B is nonsingular because Q does not have 1 as an eigenvalue. (One can see this using the Jordan canonical form: Q is similar to a lower triangular matrix with all diagonal entries different than 1, so that $I - Q$ is similar to a lower triangular matrix with nonzero diagonal entries and cannot be singular). Hence the group inverse $A^{\#}$ of A exists and is given by

$$A^{\#} = S^{-1}\begin{pmatrix} (I-Q)^{-1} & 0 \\ 0 & 0 \end{pmatrix} S. \tag{3.2}$$

In this context, $A^{\#}$ is usually called the *deviation matrix* of the Markov chain (see Veinott (1974)).

Another interesting consequence of Lemma 3.1 is that it provides an alternate proof of the following result of Kemeny and Snell (1960) that was fundamental in Blackwell (1962). Here we establish the result directly from the representation of P found in Lemma 3.1.

THEOREM 3.2. *If P is a stochastic matrix with m recurrent classes, then there is a unique matrix P^* such that*

$$PP^* = P^*P = P^*P^* = P^* \tag{3.3}$$

and $\operatorname{rank}(P^*) = m$. *Moreover*

$$P^* = S^{-1}\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} S \tag{3.4}$$

and $Z = (I - P + P^*)^{-1}$ exists, where I is the $m \times m$ identity matrix.

Proof. Decompose P as in (3.1), and apply the same similarity transformation S to P^*, to obtain

$$P^* = S^{-1}\begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} S$$

where R_{11} is $(n-m) \times (n-m)$, R_{22} is $m \times m$ and R_{12} and R_{21} have the appropriate dimensions. If $m = n$, then $P = I = P^*$ and we are done. Otherwise, using (3.1) and (3.3), we have (because $B = I - Q$ is nonsingular):

$$\begin{array}{rclcl} PP^* = P^* & \Longrightarrow & QR_{11} = R_{11} & \Longrightarrow & R_{11} = 0 \\ & & QR_{12} = R_{12} & \Longrightarrow & R_{12} = 0 \\ P^*P = P^* & \Longrightarrow & R_{21}Q = R_{21} & \Longrightarrow & R_{21} = 0 \\ P^*P^* = P^* & \Longrightarrow & R_{22}R_{22} = R_{22} & \Longrightarrow & R_{22} \text{ is a projection.} \end{array}$$

Now because rank$(P^*) = m$, this implies that rank$(R_{22}) = m$, and hence $R_{22} = I$ since the only projection of full rank is the identity. So we have (3.4). Also, it trivially follows that

$$I - P + P^* = S^{-1} \begin{pmatrix} B & 0 \\ 0 & I \end{pmatrix} S \tag{3.5}$$

is indeed nonsingular. □

We recall that P^* is called the *limiting matrix* of the Markov chain (P^* is also called the *stationary matrix*), while Z is called the *fundamental matrix* (Kemeny and Snell (1960)). The deviation matrix $A^{\#}$ is the matrix H of Blackwell (1962). Also, P^* is called the *eigenprojection of A at $\lambda = 1$*, by Rothblum (1981). The following well-known properties of P^*, $A^{\#}$ and Z are useful and are easily derived using (3.1), (3.2), (3.4), and (3.5). We prove the first one to further illustrate the simplicity of our approach.

COROLLARY 3.3. *Let $A^{\#}$ and P^* be defined as above. Then*

$$\text{(i)} \qquad (I - P)A^{\#} = A^{\#}(I - P) = I - P^*,$$

$$\text{(ii)} \qquad A^{\#}P^* = P^*A^{\#} = 0,$$

$$\text{(iii)} \qquad (I - P^*)A^{\#} = A^{\#}(I - P^*) = A^{\#},$$

$$\text{(iv)} \qquad P^* = I - AA^{\#},$$

$$\text{(v)} \qquad Z = (A + P^*)^{-1} = A^{\#} + P^* \quad \textit{and}$$

$$\text{(vi)} \qquad ZP^* = P^*.$$

Proof. We show only (i), the rest follow in a similar manner and are omitted.

$$\begin{aligned} (I - P)A^{\#} &= S^{-1} \begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} SS^{-1} \begin{pmatrix} B^{-1} & 0 \\ 0 & 0 \end{pmatrix} S = S^{-1} \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} S \\ &= S^{-1} \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} S - S^{-1} \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} S = I - P^*. \end{aligned}$$ □

The remainder of this section examines some results from Veinott (1969), (1974) which provide the probabilistic interpretation for the limiting and deviation matrices. Consider a sequence $\{A_i : i = 0, 1, 2, \cdots\}$ of $n \times n$ matrices, and let $B_N = \sum_{i=0}^{N-1} A_i$, for $N = 0, 1, 2, \cdots$. If $\lim_{N\to\infty} N^{-1}B_N = A$ exists, then A is called the *Cesàro limit of order one* of $\{A_i\}$, and we write

$$\lim_{N\to\infty} A_N = A \ (C, 1).$$

Further, if $\lim_{N\to\infty} B_N = B \ (C, 1)$, then we write

$$\sum_{N=0}^{\infty} A_N = B \ (C, 1).$$

We investigate the limits of sums of powers of a matrix. The following lemma (Veinott (1974, eq. 14)) is well known and is included for the sake of completeness.

LEMMA 3.4. *Suppose Q^N is bounded, for $N = 0, 1, 2, \cdots$, and $I - Q$ is nonsingular. Then*

$$\lim_{N\to\infty} Q^N = 0\ (C,1)$$

and

$$\sum_{N=0}^{\infty} Q^N = (I-Q)^{-1}\ (C,1).$$

Proof. The first limit follows from the identity

$$N^{-1}\sum_{i=0}^{N-1} Q^i(I-Q) = N^{-1}(I-Q^N).$$

The second one follows from the identities

$$N^{-1}\left(\sum_{i=0}^{N-1}\sum_{j=0}^{i} Q^j\right)(I-Q)^2 = N^{-1}\sum_{i=0}^{N-1}(I-Q^{i+1})(I-Q)$$
$$= I - Q - N^{-1}(Q - Q^{N+1}). \qquad \square$$

The following expansions (equations (15) and (18) of Veinott (1974)) can now be obtained as a corollary to our Theorem 3.2. Our proof again uses the matrix decomposition directly.

COROLLARY 3.5. *Suppose P is a stochastic matrix, P^* its limiting matrix and $A^\#$ its deviation matrix. Then*

$$\lim_{N\to\infty} P^N = P^*\ (C,1)$$

and

$$\sum_{N=0}^{\infty}(P^N - P^*) = A^\#\ (C,1).$$

Further, if P is aperiodic, then the ordinary limit can be used instead of $(C,1)$.

Proof. Decompose P as in (3.1). Then

$$P^N = S^{-1}\begin{pmatrix} Q^N & 0 \\ 0 & I \end{pmatrix} S$$

and

$$P^N - P^* = S^{-1}\begin{pmatrix} Q^N & 0 \\ 0 & 0 \end{pmatrix} S.$$

Since P^N is a stochastic matrix, we have that P^N is bounded, and hence so is Q^N. Moreover, $I - Q$ is nonsingular, by Lemma 3.1. Hence the result follows from Lemma 3.4. If P is aperiodic, then $\sigma(Q) < 1$ so that

$$\lim_{N\to\infty} Q^N = 0,$$

directly. □

The (well-known) consequence of Corollary 3.5 is that, for a Markov chain with stochastic transition matrix P, the entry π_{ij} of the limiting matrix P^* is the expected fraction of the time spent in state j, given that the system started in state i. Also, the entry α_{ij} of the deviation matrix $A^\#$ is the expected difference between the expected number of visits to state j when the system starts in state i, and the expected number of visits to state j when the system starts with the stationary distribution.

4. Application to discrete time Markov reward processes. Consider a homogeneous Markov chain $\{X(t) : t = 0, 1, 2, \cdots\}$ with finite state space $N = \{1, \cdots, n\}$ and stochastic transition matrix P, where

$$p_{ij} = \text{Prob}\{X(t+1) = j \mid X(t) = i\}, \qquad i, j = 1, \cdots, n.$$

Suppose also that at the end of every period t, the system earns a reward r_i, where $i = X(t)$. We say that r is the vector of *immediate rewards*. Such a Markov chain with rewards is called a *Markov reward process*, and is denoted by the triplet (N, r, P). Now suppose an arbitrary random variable Y is defined on the Markov chain $\{X(t)\}$. We define the conditional expectation operator E_i as

$$E_i(Y) = E\{Y \mid X(0) = i\} \text{ for } i = 1, \cdots, n.$$

Let $\rho > 0$ be the interest rate for one period. Then $\beta = (1+\rho)^{-1}$ is the one period discount factor, with $0 < \beta < 1$. We are interested in the present value, at time $t = 0$, of the total reward earned over an infinite horizon. More precisely, we define the expected total discounted reward function $v(\rho)$ such that

$$v_i(\rho) = E_i\Big(\sum_{k=0}^{\infty} \beta^{k+1} r_{X(k)}\Big), \qquad i = 1, \cdots, n.$$

The expectation can be expressed directly using matrix notation, and we get

$$v(\rho) = \beta \sum_{k=0}^{\infty} \beta^k P^k r = \beta(I - \beta P)^{-1} r. \tag{4.1}$$

The infinite series in (4.1) converges because

$$\lim_{k \to \infty} \beta^k P^k = 0,$$

with $0 < \beta < 1$.

We are interested in the case when β is in the neighborhood of 1. Following Veinott (1969), (1974), we replace the discount factor β by the interest rate ρ, so that

$$\begin{aligned} v(\rho) &= (1+\rho)^{-1}[I - (1+\rho)^{-1} P]^{-1} r \\ &= [(1+\rho)I - P]^{-1} r \\ &= [\rho I + A]^{-1} r \qquad (4.2) \end{aligned}$$

where $A = I - P$. The matrix $R_\rho = (\rho I + A)^{-1}$ is called the *resolvent* of A. We now provide a new, direct derivation of the Laurent expansion for the resolvent (Theorem 2 of Veinott (1969) and Theorem 3 of Veinott (1974)) using the decomposition properties of the matrix A and its Drazin inverse $A^{\#}$.

THEOREM 4.1. *Let λ be the nonzero eigenvalue of A with smallest modulus. If $0 < \rho < |\lambda|$ then*

$$R_\rho = \rho^{-1} P^* + \sum_{i=0}^{\infty} (-\rho)^i (A^{\#})^{i+1}. \tag{4.3}$$

Proof. By definition,

$$\begin{aligned} R_\rho = (\rho I + A)^{-1} &= S^{-1}\begin{pmatrix} \rho I + B & 0 \\ 0 & \rho I \end{pmatrix}^{-1} S \\ &= S^{-1}\begin{pmatrix} (\rho I + B)^{-1} & 0 \\ 0 & \rho^{-1} I \end{pmatrix} S \\ &= \rho^{-1} S^{-1}\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} S + S^{-1}\begin{pmatrix} (\rho I + B)^{-1} & 0 \\ 0 & 0 \end{pmatrix} S. \end{aligned}$$

The first term is $\rho^{-1}P^*$. Now consider the second term.

$$\begin{aligned} [\rho I + B]^{-1} &= [B(\rho B^{-1} + I)]^{-1} \\ &= [I + \rho B^{-1}]^{-1} B^{-1}. \end{aligned}$$

By hypothesis, $\sigma(\rho B^{-1}) = \rho/|\lambda| < 1$, so that

$$[I + \rho B^{-1}]^{-1} = \sum_{i=0}^{\infty} (-\rho)^i (B^{-1})^i.$$

But

$$S^{-1}\begin{pmatrix} (B^{-1})^i & 0 \\ 0 & 0 \end{pmatrix} S = (A^{\#})^i$$

so the second term above equals

$$\sum_{i=0}^{\infty} (-\rho)^i (A^{\#})^{i+1}$$

which gives the desired result. □

Although we formulated Theorem 4.1 in the specific case $A = I - P$, we remark that it is valid for any matrix A such that $\operatorname{ind}(A) = 1$. It is an extension of the simple case when the matrix A is nonsingular (i.e., $\operatorname{ind}(A) = 0$). Then, (4.3) becomes

$$R_\rho = \sum_{i=0}^{\infty} (-\rho)^i (A^{-1})^{i+1}.$$

Theorem 4.1 is itself a special case of a result of Rothblum (1981, Thm. 3.1), in which a matrix A with $\operatorname{ind}(A) = k \geq 0$ has terms in ρ^{-i}, for $i = 0, \cdots, k$. An extension of Rothblum's theorem to the case of a general pair $(A + \rho B)^{-1}$, where A and B are arbitrary square matrices, is derived in Lamond (1989).

COROLLARY 4.2. *Let $g = P^* r$, $h = A^{\#} r$ and*

$$w_i = (-1)^i (A^{\#})^{i+1} r, \quad i = 1, 2, \cdots. \tag{4.4}$$

Then

$$v(\rho) = \rho^{-1} g + h + \sum_{i=1}^{\infty} \rho^i w_i. \tag{4.5}$$

Proof. The corollary is proved by substituting (4.3) into (4.2). □

This Laurent expansion of the discounted reward function $v(\rho)$ is fundamental for the notion of discount optimality (Veinott (1969)) that will be discussed in § 6. Observe that our derivation uses only the property that $\operatorname{ind}(A) = 1$, that is, that the matrix A can be decomposed as in (2.3).

The term g is called the *gain* of the process and h is its *bias* (e.g., Denardo (1973)). This terminology is justified because, by Corollary 3.5,

$$g = \lim_{k\to\infty} P^k r \ (C,1)$$

and

$$h = \sum_{k=0}^{\infty}(P^k r - g) \ (C,1)$$

(using ordinary limits if P is aperiodic).

We conclude this section with a very simple derivation of a well-known and important equation for computing the gain and bias (see Howard (1960) and Blackwell (1962)). This equation is usually derived using a partial series expansion, but here it follows directly from the generalized inverse representation for P^*.

PROPOSITION 4.3. *Let g and h be defined as in Corollary 4.2. Then g and h satisfy*

$$h = r - g + Ph. \tag{4.6}$$

Proof. We use the fact that $P^* = I - AA^\#$. Then

$$g = (I - AA^\#)r = r - A(A^\# r) = r - (I-P)h.$$

The result follows by rearranging terms. □

5. Singular systems of equations. The technique of Proposition 4.3 for obtaining the gain and the bias vectors as solutions of a system of singular linear equations was extended in Veinott (1969) and Miller and Veinott (1969) to obtain all the terms w_i of Corollary 4.2 by solving an augmented system of linear equations.

Let A be an $n \times n$ matrix of index 1 (i.e., decomposition (2.3) is valid), and suppose b is a given column vector. Recall that the projection matrix $W = I - AA^\#$ is given by (3.4). In the special case $A = I - P$ with P a (stochastic) transition matrix, we have $W = P^*$, the limiting matrix. The following lemma is well known (see, e.g., Corollary 3.1.1 of Hunter (1982)) and provides a basic mechanism for solving systems of singular equations using the group inverse. We give a new proof based on the decomposition property. While this proof is not the shortest possible, we include it because this factorization can be used to compute solution vectors without computing the group inverse matrix itself (see Lamond (1987)).

LEMMA 5.1. *The (singular) system of equations $Ax = b$ has a solution if and only if $Wb = 0$. Further, if $Wb = 0$, then for arbitrary y*

$$x = A^\# b + Wy \tag{5.1}$$

is a solution of $Ax = b$.

Proof. Let

$$x = S^{-1}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad b = S^{-1}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \quad \text{and} \quad y = S^{-1}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}.$$

Then

$$Ax = S^{-1}\begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix} SS^{-1}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = S^{-1}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

which is equivalent to

$$\begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

This system has a solution if and only if $b_2 = 0$, which is equivalent to $Wb = 0$, because

$$Wb = S^{-1}\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} SS^{-1}\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = S^{-1}\begin{pmatrix} 0 \\ b_2 \end{pmatrix}.$$

Now we assume that $b_2 = 0$ and we take $x = A^{\#}b + Wy$, for some arbitrary vector y. We have

$$\begin{aligned} Ax &= AA^{\#}b + AWy \\ &= S^{-1}\begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} B^{-1} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} b_1 \\ 0 \end{pmatrix} + S^{-1}\begin{pmatrix} B & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \\ &= S^{-1}\begin{pmatrix} b_1 \\ 0 \end{pmatrix} + 0 = b. \end{aligned}$$ □

The promised extension of Proposition 4.3 follows as a simple consequence of the lemma and we give it without proof. The result is from page 1650 of Veinott (1969). The terms g, h and w_i, $i = 1, 2, \cdots$, are obtained by applying the theorem with $A = I - P$, $b_2 = r$, where r is the vector of one-period rewards, and $b_i = 0$ for $i \neq 2$. Then $g = x_1$, $h = x_2$ and $w_i = x_{i+2}$.

THEOREM 5.2. *Let $\{b_i, i = 1, 2, \cdots\}$ be a sequence of n-vectors and $W = I - AA^{\#}$. Then the system of equations*

$$Ax_1 = b_1, \tag{5.2}$$

$$x_{i-1} + Ax_i = b_i, \qquad i = 2, 3, \cdots \tag{5.3}$$

has a solution if and only if

$$Wb_1 = 0. \tag{5.4}$$

Furthermore, if (5.4) *holds, the solution is unique and is given by*

$$x_i = A^{\#}(b_i - x_{i-1}) + Wb_{i+1}, \qquad i = 1, 2, \cdots \tag{5.5}$$

where $x_0 = 0$.

As pointed out in Veinott (1969) and Miller and Veinott (1969), and as will be discussed in the next section, it is sufficient to evaluate only a finite number of terms w_i. Hence we now consider truncated systems of equations. Let A be an $n \times n$ complex matrix such that $\operatorname{ind}(A) \leq 1$ (i.e., A is either nonsingular or else it satisfies (2.3)). For $k \geq 1$, we define the matrix $A(k)$ as the block Jordan matrix of order k that has A for its diagonal blocks. For example, $A(1) = A$ and for $k = 3$, we have

$$A(3) = \begin{pmatrix} A & 0 & 0 \\ I & A & 0 \\ 0 & I & A \end{pmatrix}.$$

We are interested in solving the system of equations:

$$A(k)x(k) = b(k) \tag{5.6}$$

where $x(k)$ and $b(k)$ are partitioned so that they be conformable with A. For example, the gain and bias terms of Proposition 4.3 satisfy the system

$$\begin{pmatrix} A & 0 \\ I & A \end{pmatrix} \begin{pmatrix} g \\ h \end{pmatrix} = \begin{pmatrix} 0 \\ r \end{pmatrix}.$$

In the case when the matrix A is nonsingular, it is well known (and easy to verify) that the solution can be expressed as $x(k) = A(k)^{-1}b(k)$ where $A(k)^{-1}$ is a block lower triangular matrix such that for $i \geq j$, the block (i, j) is given by

$$(-1)^{i-j}(A^{-1})^{i-j+1}. \tag{5.7}$$

We now obtain a similar expression when the matrix A is singular, in the case when $\operatorname{ind}(A) = 1$, so that $A^{\#}$ exists and with $W = I - AA^{\#}$ as before. The proof of the following lemma is given in Lamond (1985, Lemma 3.2.6).

LEMMA 5.3. *If A is singular but satisfies* (2.3), *then $A(k)$, $k \geq 1$, has index k and its Drazin inverse $A(k)^D$ is a block lower triangular matrix such that its block (i, j), $i \geq j$, is given by*

$$(-1)^{i-j}(A^{\#})^{i-j+1}. \tag{5.8}$$

It is straightforward to verify that $A(k)A(k)^D$ is a block diagonal matrix with all diagonal blocks equal to $AA^{\#}$. The consequence is that $y(k) = A(k)^D b(k)$ is not a solution of (5.6) in general, because $AA^{\#}b_i \neq b_i$ for $i \geq 2$. Nonetheless, we can use $A(k)^D$ to construct a solution. Define a new matrix $R(k)$ whose only nonzero blocks are on the superdiagonal and are all equal to $W = I - AA^{\#}$. For example, $R(1) = 0$ and with $k = 3$ we have

$$R(3) = \begin{pmatrix} 0 & W & 0 \\ 0 & 0 & W \\ 0 & 0 & 0 \end{pmatrix}.$$

THEOREM 5.4. *If the system $Ax_1 = b_1$ has a solution, then the system $A(k)x(k) = b(k)$ has a solution given by*

$$x(k) = [A(k)^D + R(k)]b(k). \tag{5.9}$$

Moreover, $x_1, x_2, \cdots, x_{k-1}$ are uniquely determined while $x_k + Wy$ is also a solution, for any vector y.

Proof. Expanding the matrix product in (5.9) gives precisely the solution of (5.5). Because the system was truncated, the kth equation has a nonunique solution, as prescribed by Lemma 5.1. □

Of course, (5.9) gives precisely the same solution vector as obtained in Veinott (1969), (1974). See Veinott (1969, p. 1651) for a computational algorithm that takes advantage of the special structure available when $A = I - P$ with P stochastic. See Lamond (1987) for an algorithm using matrix factorization (valid for any matrix A of index 1).

6. Application to finite state and action Markov decision processes. Now consider a system in which an action must be taken before a transition occurs. Let $N = \{1, \cdots, n\}$ be the (finite) state space, and let A_i be the finite set of actions that can be taken when the system is in state i, for $i = 1, \cdots, n$. The system evolves in time according to the process $\{X(t) : t = 0, 1, 2, \cdots\}$, where $X(t)$ is the state of the system at time t. Let $\{a(t) : t = 0, 1, 2, \cdots\}$ be the sequence of selected actions, with $a(t) \in A_{X(t)}$.

We assume that the system satisfies the Markov property and that the transition probabilities do not depend explicitly on time. Hence we define

$$p_{ij}^k = \text{Prob}\{X(t+1) = j \mid X(t) = i, a(t) = k\},$$

for $k \in A_i$ and $i, j = 1, \cdots, n$. The Markov property implies that

$$\text{Prob}\{X(t+1) = j \mid X(0), a(0), \cdots, X(t), a(t)\} = \text{Prob}\{X(t+1) = j \mid X(t), a(t)\}.$$

Immediately before a transition occurs, the system earns an (expected) reward r_i^k, where $k = a(t)$ and $i = X(t)$. The action $a(t)$ is selected according to a function π, called a *policy*, such that

$$a(t) = \pi\big(t, X(0), a(0), \cdots, X(t-1), a(t-1), X(t)\big) \text{ for } t = 0, 1, 2, \cdots.$$

Now let us denote by P the transition probability function, by r the reward function and by Π the set of all policies. Then the above model is called a *Markov decision process* and is denoted by the quadruple (N, r, P, Π).

For a given policy $\pi \in \Pi$ and interest rate $\rho > 0$, we define the expected total discounted reward function $v_i^\pi(\rho)$ such that

$$v_i^\pi(\rho) = E_i^\pi\Big(\sum_{k=0}^{\infty} \beta^{k+1} r_{X(k)}^{a(k)}\Big), \qquad i = 1, \cdots, n,$$

where $\beta = (1+\rho)^{-1}$ and E_i^π is the conditional expectation operator under policy π. That is, for an arbitrary random variable Y defined from $\{X(t)\}$ and $\{a(t)\}$, we have

$$E_i^\pi(Y) = E^\pi\{Y \mid X(0) = i\}, \qquad i = 1, \cdots, n,$$

where the expectation is taken under policy π.

The problem is to find a policy $\pi \in \Pi$ that maximizes $v_i^\pi(\rho)$ in some sense, for $i = 1, \cdots, n$. Now define a function δ to be a *decision rule* if $\delta_i \in A_i$, for $i = 1, \cdots, n$, and let $\Delta = A_1 \times A_2 \times \cdots \times A_n$ be the set of all decision rules. Then a policy π is said to be a *Markov policy* if

$$\pi\big(t, X(0), a(0), \cdots, X(t)\big) = \delta_i(t),$$

where $i = X(t)$ and $\delta(t) \in \Delta$. Moreover, a Markov policy π is said to be *stationary* if $\delta(t) = \delta(0)$ for all t.

For discounted problems with an infinite time horizon, there exists a stationary policy that is optimal (Blackwell (1962, corollary to Thm. 3). Hence there is no loss of generality in restricting the problem to stationary policies. For simplicity of notation, we will denote a stationary policy by its decision rule δ, and the set of stationary policies by Δ.

For a fixed, stationary policy $\delta \in \Delta$, the system is equivalent to the Markov reward process (N, r^δ, P^δ), as in § 4, with transition matrix $P = P^\delta$ and reward vector $r = r^\delta$. For a given interest rate $\rho > 0$ and discount factor $\beta = (1+\rho)^{-1}$, the expected total discounted reward $v^\delta(\rho)$ is given by

$$v^\delta(\rho) = \sum_{i=0}^{\infty} \beta^{i+1} (P^\delta)^i r^\delta$$

as in § 4. Following Miller and Veinott (1969), we say that a policy δ is *ρ-optimal* if

$$v^\delta(\rho) \geq v^\gamma(\rho) \qquad \forall \gamma \in \Delta. \tag{6.1}$$

It was shown by Blackwell (1962) that there exists a policy $\delta \in \Delta$ and a $\rho^* > 0$ such that δ is ρ-optimal for all ρ in $0 < \rho < \rho^*$. Such a policy is said to be *Blackwell optimal.* Motivated by the series (4.5), Veinott (1969), (1974) defined a policy δ to be *k-discount optimal* if

$$\lim_{\rho \searrow 0} \rho^{-k}[v^\delta(\rho) - v^\gamma(\rho)] \geq 0 \qquad \forall \gamma \in \Delta. \tag{6.2}$$

(Veinott's definition also extends to the case of $\rho < 0$, but we will stick to the positive case, which Veinott(1969) calls k^+ discount optimality.) As special cases, a -1-discount optimal policy is also said to be *gain optimal*, while a 0-discount optimal policy is also said to be *bias optimal* (Denardo (1970)) or *nearly optimal* (Blackwell (1962)).

The policy iteration method of Howard (1960) and Blackwell (1962) produces a policy δ that is gain optimal. This method has been extended by Veinott (1966) to find a bias optimal policy and Miller and Veinott (1969) to find a Blackwell optimal policy. Veinott (1969) modified it to find a k-discount optimal policy, for any $k \geq -1$.

More precisely, let w_i^δ, $i = -1, 0, 1, \cdots$, be the terms of the Laurent series (4.5), for a policy $\delta \in \Delta$ (here $w_{-1}^\delta = g^\delta$ is the gain and $w_0^\delta = h^\delta$ is the bias). Then (6.2) implies that δ is k-discount optimal if

$$\sum_{i=-1}^{k} \rho^i w_i^\delta \geq \sum_{i=-1}^{k} \rho^i w_i^\gamma \qquad \forall \gamma \in \Delta \tag{6.3}$$

for all small enough $\rho > 0$. The extended algorithm evaluates w_{-1}^δ, $w_0^\delta, \cdots, w_{k+1}^\delta$ to show that no other policy is lexicographically better, up to w_k.

Now let n be the number of states of the process. Miller and Veinott (1969) showed that a policy is Blackwell optimal if and only if it is n-discount optimal. This result was extended by Denardo (1971) who showed that an $(n-1)$-discount optimal policy is Blackwell optimal. Then Veinott (1974), using the fact that $\operatorname{rank}(A) = n - m$, where $A = I - P$ and m is the number of recurrent classes of P, showed that any $(n-m)$-discount optimal policy is Blackwell optimal. We give a new proof of the latter result, using our matrix decomposition approach. As Miller and Veinott (1969), we need the following lemma from Gantmacher (1959), which we state without proof.

LEMMA 6.1. *Let M be a $k \times k$ matrix and L a linear subspace of $\mathbb{R}^k$. If $M^i x \in L$ for $i = 0, 1, \cdots, k-1$, then $M^i x \in L$ for $i = k, k+1, \cdots$.*

(The idea is that for $i \geq k$, $M^i x$ is a linear combination of x, $Mx, \cdots$, $M^{k-1}x$). This implies the following theorem.

THEOREM 6.2. *Suppose that some policy $\delta \in \Delta$ has m recurrent classes (i.e., $1 \leq m \leq n$). Suppose also that some policy $\gamma \in \Delta$ is such that $w_i^\gamma = w_i^\delta$, for $i = -1, 0, \cdots, n-m$. Then $w_i^\gamma = w_i^\delta$ for $i > n-m$.*

Proof. First, observe that in the special case $m = n$, the result is trivially true because $g^\gamma = g^\delta$, $A^\delta = I - P^\delta = 0$, and $(A^\delta)^\# = 0$, so that $h^\delta = w_0^\delta = w_1^\delta = \ldots = 0$. This implies that $h^\gamma = 0$. Now by (4.4),

$$w_i^\gamma = (-1)^i \left((A^\gamma)^\#\right)^i h^\gamma = 0 = w_i^\delta, \qquad i \geq 0.$$

In the case $m < n$, both policies have the same gain and bias, i.e., $w_{-1}^\delta = w_{-1}^\gamma = g$ and $w_0^\delta = w_0^\gamma = h$. Equation (4.4) implies that

$$w_i^\delta = (-1)^i \left((A^\delta)^\#\right)^i h, \qquad i \geq 0. \tag{6.4}$$

Decompose A^δ as in (2.3), and define

$$y_i = \begin{pmatrix} y_{i,1} \\ y_{i,2} \end{pmatrix} = S w_i^\delta$$

where $y_{i,1}$ and $y_{i,2}$ have dimensions $(n-m)$ and m, respectively. Define also

$$z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = Sh = S(A^\delta)^\# r^\delta = \begin{pmatrix} B^{-1} & 0 \\ 0 & 0 \end{pmatrix} S r^\delta.$$

Hence $z_2 = 0$. Multiply (6.4) by S to get

$$\begin{pmatrix} y_{i,1} \\ y_{i,2} \end{pmatrix} = S w_i^\delta = (-1)^i S S^{-1} \begin{pmatrix} (B^{-1})^i & 0 \\ 0 & 0 \end{pmatrix} Sh$$
$$= (-1)^i \begin{pmatrix} (B^{-1})^i & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$$

so that

$$\begin{aligned} y_{i,1} &= (-1)^i (B^{-1})^i z_1, \\ y_{i,2} &= 0. \end{aligned} \qquad i \geq 0. \tag{6.5}$$

Now apply the same transformation S to A^γ, giving

$$A^\gamma = S^{-1} \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} S \tag{6.6}$$

and define

$$x_i = \begin{pmatrix} x_{i,1} \\ x_{i,2} \end{pmatrix} = S w_i^\gamma.$$

Equation (5.3), with i=i+1, $A = A^\gamma$ and $b_{i+1} = 0$, together with (6.6), gives

$$\begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} \begin{pmatrix} x_{i+1,1} \\ x_{i+1,2} \end{pmatrix} = - \begin{pmatrix} x_{i,1} \\ x_{i,2} \end{pmatrix}, \qquad i \geq 0. \tag{6.7}$$

By hypothesis, $x_{i,1} = y_{i,1}$ and $x_{i,2} = 0$, for $i = 0, \cdots, n-m$ so

$$\begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} \begin{pmatrix} y_{i+1,1} \\ 0 \end{pmatrix} = - \begin{pmatrix} y_{i,1} \\ 0 \end{pmatrix}, \qquad i = 0, \cdots, n-m-1.$$

Using (6.5), we get

$$(X_{11}B^{-1} - I)(B^{-1})^{i} z_1 = 0 \tag{6.8}$$

and

$$X_{21}(B^{-1})^{i+1} z_1 = 0. \tag{6.9}$$

Now recall that z_1 is a vector of dimension $(n-m)$. Define a linear subspace L of $\mathbb{R}^{n-m}$ such that $x \in L$ if and only if $(X_{11}B^{-1} - I)x = 0$. Then (6.8) implies that $(B^{-1})^{i} z_1 \in L$, for $i = 0, \cdots, n-m-1$. By Lemma 6.1, it is true also for $i \geq n-m$ and hence (6.8) is also valid for $i \geq n-m$. The same argument applies to (6.9) as well.

We still have to show that it implies that $x_{i,1} = y_{i,1}$ and $x_{i,2} = y_{i,2} = 0$, for $i > n-m$. By Theorem 5.2, the recurrence (6.7) has a unique solution. Equations (6.8) and (6.9) imply that $x_{i,1} = y_{i,1} = (-1)^{i}(B^{-1})^{i} z_1$ and $x_{i,2} = y_{i,2} = 0$ is that solution for every $i \geq 0$. □

As in Veinott (1969), Miller and Veinott (1969) and Veinott (1974), we now apply the above result to dynamic programming.

THEOREM 6.3. *A policy with m recurrent classes is Blackwell optimal if and only if it is $(n-m)$-discount optimal.*

Proof. The theorem is proved by applying Theorem 6.2 to (6.3). □

It implies, in general, that a policy is Blackwell optimal if and only if it is $(n-1)$-discount optimal (because $m \geq 1$, always). We will now use this result and the notation of the previous section to formulate a policy iteration algorithm in a unified way. We first recall that with a fixed interest rate ρ, the policy iteration method for finding a ρ-optimal policy can be formulated as follows, with $\beta = (1+\rho)^{-1}$.

```
Step 0. Select an arbitrary policy $\delta \in \Delta$.
Step 1. Policy evaluation.
        Compute $v^{\delta} = (A^{\delta})^{-1} r^{\delta}$, where $A^{\delta} = I - \beta P^{\delta}$.
Step 2. Policy Improvement.
        Find $\gamma \in \Delta$ subject to $r^{\gamma} + \beta P^{\gamma} v^{\delta} = \max_{\eta \in \Delta}\{r^{\eta} + \beta P^{\eta} v^{\delta}\}$.
        (Choose $\gamma = \delta$ if possible).
Step 3. If $\gamma = \delta$ then stop
                  else set $\delta = \gamma$ and go to Step 1.
```

Using the notation of § 5, we can now formulate the policy iteration method for finding a k discount optimal policy as follows. Let $\kappa = k+3$ (we need to evaluate κ terms $w_{-1}, w_0, \cdots, w_{k+1}$). We simply replace steps 1 and 2 above as follows. At step 1, compute

$$x^{\delta}(\kappa) = [A^{\delta}(\kappa)^{D} + R^{\delta}(\kappa)] b^{\delta}(\kappa), \tag{6.10}$$

where $b_2^{\delta} = r$ and $b_i^{\delta} = 0$ for $i \neq 2$. At step 2, find $\gamma \in \Delta$ s.t.

$$b^{\gamma}(\kappa) + P^{\gamma}(\kappa) x^{\delta}(\kappa) = \max_{\eta \in \Delta}\{b^{\eta}(\kappa) + P^{\eta}(\kappa) x^{\delta}(\kappa)\}, \tag{6.11}$$

where $P^{\gamma}(\kappa) = I(\kappa) - A^{\gamma}(\kappa)$ and the maximization is done lexicographically (see, e.g., Veinott (1974)).

It is to be seen whether the latter algorithm could be formulated as a specialization of the method of Newton for solving systems of nonlinear equations, as the former was by Puterman and Brumelle (1979) using the ordinary inverse $(A^{\delta})^{-1}$.

It is important to note, however, that the formulation of (6.10) is not the most appropriate for computations. Indeed, we saw in § 5 that the terms w_i^{δ}, for $i = -1, \cdots, k$, can be evaluated successively. Hence it is more efficient to combine steps 1 and 2 in such a way that only those terms which are really required be computed.

There are two ways to do it. One is to successively evaluate $x^{\delta}(i) = w_{i-3}$ and produce Δ_i such that

$$\Delta_i = \{\gamma \in \Delta_{i-1} : b^{\gamma}(i) + P^{\gamma}x^{\delta}(i) \geq b^{\eta}(i) + P^{\eta}x^{\delta}(i), \forall \eta \in \Delta_{i-1}\}$$

where $\Delta_0 = \Delta$, and stop at the smallest $i \leq k$ for which Δ_i has only one element. This element is the policy γ of step 2. (If Δ_k has more than one element, choose any $\gamma \in \Delta_k$).

Another way is to stop at the smallest $i \leq k$ for which Δ_i does not contain δ, and choose $\gamma \in \Delta_i$ arbitrarily. Then γ is a better policy than δ, although it is not necessarily the best improvement. This latter approach might require fewer terms to be evaluated than the former.

7. Conclusions and extensions. The following are the main results with respect to our approach of Markov decision processes. Lemma 3.1 gives the basic decomposition of a transition matrix, which is used to define the limiting and fundamental matrices algebraically in Theorem 3.2. Then Corollary 3.5 identifies the algebraic definitions with the probabilistic notions of limiting and deviation matrices.

The results of particular interest for decision processes are Theorem 4.1, giving the Laurent expansion of the resolvent, and Proposition 4.3, which provides a particularly immediate derivation of the standard policy evaluation equation. Theorem 6.2 gives a fairly convincing proof of the finite test for Blackwell optimality of a stationary policy. Note that more than half the space taken by this proof is just notation. The mathematical argument itself if brief.

The computational aspects of the matrix factorization are discussed in Lamond (1985), (1987). Extensions to continuous time Markov decision processes and to semi-Markov decision processes are discussed in Lamond (1985). It is an open question whether Drazin inverse theory can be applied to processes with countably infinite state space and generator driven processes such as diffusion processes.

Acknowledgments. We would like to thank the associate editor, Uriel G. Rothblum, and anonymous referees for their stimulating comments and careful reading of earlier versions of this paper.

REFERENCES

K.M. Anstreicher and U.G. Rothblum (1987), *Using Gauss-Jordan elimination to compute the index, generalized nullspace and Drazin inverse*, Linear Algebra Appl., 85, pp. 221–239.

D. Blackwell (1962), *Discrete dynamic programming*, Ann. Math. Statist., 33, pp. 719–726.

S.L. Campbell and C.D. Meyer, Jr. (1979), *Generalized Inverses of Linear Transformations*, Pitman, London.

E.V. Denardo (1970), *Computing a bias-optimal policy in a discrete-time Markov decision problem*, Oper. Res., 18, pp. 279–289.

——— (1971), *Markov renewal programs with small interest rates*, Ann. Math. Statist., 42, pp. 477–496.

——— (1973), *A Markov decison problem,* in Mathematical Programming, T.C. Hu and S.M. Robinson, eds., Academic Press, New York, pp. 33–68.

M.P. DRAZIN (1958), *Pseudo-inverses in associative rings and semigroups,* Amer. Math. Monthly, 65, pp. 506–514.

F.R. GANTMACHER (1959), *The Theory of Matrices,* Vol. 1, Chelsea, New York.

——— (1960), *The Theory of Matrices,* Vol. 2, Chelsea, New York.

R.A. HOWARD (1960), *Dynamic Programming and Markov Processes,* John Wiley, New York.

J.H. HUNTER (1982), *Generalized inverses and their application to applied probability problems,* Linear Algebra Appl., 45, pp. 157–198.

——— (1988), *Characterizations of generalized inverses associated with Markovian kernels,* Linear Algebra Appl., 102, pp. 121–142.

S. KARLIN (1968), *A First Course in Stochastic Processes,* Academic Press, New York.

J.G. KEMENY (1981), *Generalization of a fundamental matrix,* Linear Algebra Appl., 38, pp. 193–206.

J.G. KEMENY AND J.L. SNELL (1960), *Finite Markov Chains,* Van Nostrand, New York.

B.F. LAMOND (1985), *Matrix Methods in Queueing and Dynamic Programming,* Ph.D. dissertation, University of British Columbia, Vancouver, Canada.

——— (1987), *An efficient factorization for the group inverse,* SIAM J. Algebraic Discrete Methods, 8, pp. 797–808.

——— (1989), *A generalized inverse method for asymptotic linear programming,* Math. Programming, 43, pp. 71–86.

C.D. MEYER, JR. (1982), *Analysis of finite Markov chains by group inversion techniques,* in Recent Applications of Generalized Inverses, S.L. Campbell, ed., Pitman, Boston, pp. 50–81.

B.L. MILLER AND A.F. VEINOTT, JR. (1969), *Discrete dynamic programming with a small interest rate,* Ann. Math. Statist., 40, pp. 366–370.

K. OHNO (1985), *Modified policy iteration algorithm with non-optimality tests for undiscounted Markov decision processes,* Technical report, Konan University, Kyoto, Japan.

M.L. PUTERMAN AND S.L. BRUMELLE (1979), *On the convergence of policy iteration in stationary dynamic programming,* Math. Oper. Res., 4, pp. 60–69.

U.G. ROTHBLUM (1981), *Resolvent expansions of matrices and applications,* Linear Algebra Appl., 38, pp. 33–49.

R.S. VARGA (1962), *Matrix Iterative Analysis,* Prentice-Hall, Englewood-Cliffs, NJ.

A.F. VEINOTT, JR. (1966), *On finding optimal policies in discrete dynamic programming with no discounting,* Ann. Math. Statist., 37, pp. 1284–1294.

——— (1969), *Discrete dynamic programming with sensitive discount optimality criteria,* Ann. Math. Statist., 40, pp. 1635–1660.

——— (1974), *Markov decision chains,* in Studies in Optimization, 10, MAA Studies in Mathematics, G.B. Dantzing and B.C. Eaves, eds., pp. 124–159.

J.H. WILKINSON (1965), *The Algebraic Eigenvalue Problem,* Clarendon Press, Oxford.

——— (1982), *Note on the practical significance of the Drazin inverse,* in Recent Applications of Generalized Inverses, S.L. Campbell, ed., Pitman, Boston, pp. 82–99.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 2, pp. 135-146, April 1989

NUMERICAL SOLUTION OF THE EIGENVALUE PROBLEM FOR HERMITIAN TOEPLITZ MATRICES*

WILLIAM F. TRENCH†

Abstract. An iterative procedure is proposed for computing the eigenvalues and eigenvectors of Hermitian Toeplitz matrices. The computational cost per eigenvalue-eigenvector for a matrix of order n is $O(n^2)$ in serial mode. Results of numerical experiments on Kac–Murdock–Szegö matrices and randomly generated real symmetric Toeplitz matrices of orders 100, 150, 300, 500, and 1,000 are included.

Key words. Toeplitz, Hermitian, symmetric, eigenvalue, eigenvector, Levinson–Durbin Algorithm

AMS(MOS) subject classifications. 65F15, 15A18, 15A57

1. Introduction. Here we present a method for computing the eigenvalues and eigenvectors of Hermitian Toeplitz matrices, i.e., matrices of the form

$$T_n = (t_{i-j})_{i,j=1}^{n}$$

with $t_r = \bar{t}_{-r}$. The method rests specifically and crucially on the special structure of T_n. There are efficient algorithms that exploit this simple structure to invert such matrices, or to solve systems $T_nX = Y$. There is also an extensive literature on the asymptotic distribution of the eigenvalues of a family $\{T_n\}$ of Hermitian Toeplitz matrices as $n \to \infty$, in the case where the $\{t_m\}$ are the Fourier coefficients of a function f which satisfies suitable integrability conditions. However, the development of efficient methods designed specifically to compute the eigenvalues and eigenvectors of these matrices is still in its early stages.

Several recent papers [5], [7], [9], [18], [23] have dealt with the spectral structure of Hermitian Toeplitz matrices, and numerical methods aimed mainly at finding the smallest eigenvalue of a positive definite Hermitian Toeplitz matrix have appeared [8], [11], [14], [15]. Some of these use inverse iteration with Rayleigh quotient shifting, exploiting the Levinson algorithm [17] for solving Toeplitz systems. Trench [22] has proposed a method which, on the basis of preliminary numerical experiments, appears to provide an effective procedure for computing the eigenvalues of Hermitian Toeplitz matrices generated by rational functions, at a cost per eigenvalue essentially independent of the order of the matrix. (Autocorrelation matrices of ARMA (Autoregressive Moving Average) processes are of this kind.)

The method presented here combines the Levinson–Durbin Algorithm [6] for the shifted matrices $T_m - \lambda I_m (1 \leqq m \leqq n-1)$ with an iterative root-finding procedure to locate the zeros of the rational function

$$q_n(\lambda) = p_n(\lambda)/p_{n-1}(\lambda), \tag{1}$$

where

$$p_m(\lambda) = \det[T_m - \lambda I_m], \qquad 1 \leqq m \leqq n. \tag{2}$$

The basic idea of this approach did not originate with us. Cybenko and Van Loan [8] used the Levinson–Durbin Algorithm and Newton's method to compute the smallest

* Received by the editors November 16, 1987; accepted for publication (in revised form) July 15, 1988. This work was partially supported by National Science Foundation grant DMS-8707080.

† Department of Mathematics, Trinity University, 715 Stadium Drive, San Antonio, Texas 78284 (wtrench@trinity.bitnet).

eigenvalue of a symmetric positive definite Toeplitz matrix, and our work should be considered to be a continuation of theirs. However, our method will determine any eigenvalue of T_n that is not also an eigenvalue of any of the nested submatrices $T_1, \cdots,$ T_{n-1} (an assumption also required by Cybenko and Van Loan). The corresponding eigenvectors are obtained as by-products.

Delsarte and Genin [9] have used arguments based on the Levinson–Durbin Algorithm as applied to the shifted matrices $T_m - \lambda I_m$ to obtain theoretical results concerning the spectra of Hermitian Toeplitz matrices. For a result related to their work, see also Wilkes and Hayes [23].

2. The theoretical basis for the method. Most of the results in this section are not new, although we believe that this presentation in specific reference to the eigenvalue problem is somewhat more explicit and complete than previous discussions. In any case, it seems appropriate to include it here for the reader's convenience.

Since the eigenvalues of T_n are real, we assume throughout that λ is real. Let

$$U_{n-1} = [t_1, t_2, \cdots, t_{n-1}]^t \qquad ({}^t = \text{transpose}).$$

If λ is not an eigenvalue of T_{n-1}, then let

$$X_{n-1}(\lambda) = [x_{1,n-1}(\lambda), \cdots, x_{n-1,n-1}(\lambda)]^t$$

be the solution of

$$(T_{n-1} - \lambda I_{n-1}) X_{n-1}(\lambda) = U_{n-1}, \tag{3}$$

and define

$$Y_n(\lambda) = \begin{bmatrix} -1 \\ X_{n-1}(\lambda) \end{bmatrix}. \tag{4}$$

Recall the definitions (1) and (2) of $q_n(\lambda)$ and $p_m(\lambda)$. In the following, $\| \quad \|$ is the Euclidean norm.

THEOREM 1. *If λ is not an eigenvalue of T_{n-1}, then*

$$q_n(\lambda) = t_0 - \lambda - \bar{U}^t_{n-1} X_{n-1}(\lambda) \tag{5}$$

and

$$q'_n(\lambda) = -1 - \|X_{n-1}(\lambda)\|^2. \tag{6}$$

If, in addition, λ is an eigenvalue of T_n, then $Y_n(\lambda)$ is an associated eigenvector.

Proof. We partition $T_n - \lambda I_n$ in the form

$$T_n - \lambda I_n = \begin{bmatrix} t_0 - \lambda & \bar{U}^t_{n-1} \\ U_{n-1} & T_{n-1} - \lambda I_{n-1} \end{bmatrix}. \tag{7}$$

Subtract $x_{j,n-1}(\lambda)$ times column $j + 1$ from the first column of (7) for $j = 1, \cdots, n - 1$, and invoke (3) to obtain

$$\begin{aligned} p_n(\lambda) &= \begin{vmatrix} t_0 - \lambda - \bar{U}^t_{n-1} X_{n-1}(\lambda) & \bar{U}^t_{n-1} \\ 0 & T_{n-1} - \lambda I_{n-1} \end{vmatrix} \\ &= (t_0 - \lambda - \bar{U}^t_{n-1} X_{n-1}(\lambda)) p_{n-1}(\lambda), \end{aligned}$$

which implies (5). From (3)–(5) and (7), we have

$$(T_n - \lambda I_n) Y_n(\lambda) = -q_n(\lambda)[1, 0, \cdots 0]^t; \tag{8}$$

hence, if λ is an eigenvalue of T_n, then $Y_n(\lambda)$ is an associated eigenvector.

To verify (6), we differentiate (5):

$$q_n'(\lambda)=-1-\bar{U}_{n-1}^t X_{n-1}'(\lambda)\tag{9}$$
$$=-1-\bar{X}_{n-1}^t(\lambda)(T_{n-1}-\lambda I_{n-1})X_{n-1}'(\lambda),$$

where the second equality follows from (3) and the Hermitian symmetry of $T_{n-1}-\lambda I_{n-1}$. Since differentiating (3) shows that

$$(T_{n-1}-\lambda I_{n-1})X_{n-1}'(\lambda)=X_{n-1}(\lambda),$$

(9) implies (6).

Formula (6) is due to Cybenko and Van Loan [8]; however, they did not explicitly identify $q_n(\lambda)$ as the ratio $p_n(\lambda)/p_{n-1}(\lambda)$.

Except for a missing minus sign on the right, (3) is the Yule–Walker equation for $T_{n-1}-\lambda I_{n-1}$ (cf. [6]). The following theorem is essentially a statement of the Levinson–Durbin Algorithm for solving (3), with minor changes to account for the fact that the diagonal element of the matrix in (3) is $t_0-\lambda$ rather than unity. We omit the proof.

THEOREM 2. *If $T_m-\lambda I_m$ is nonsingular for $1\leqq m\leqq n-1$, then (3) can be solved recursively as follows. Let*

$$x_{11}(\lambda)=t_1/(t_0-\lambda),\qquad \Delta_1(\lambda)=t_0-\lambda,\tag{10}$$

and, for $2\leqq m\leqq n-1$,

$$\Delta_m(\lambda)=[1-|x_{m-1,m-1}(\lambda)|^2]\Delta_{m-1}(\lambda),\tag{11}$$

$$x_{mm}(\lambda)=\Delta_m^{-1}(\lambda)\left[t_m-\sum_{j=1}^{m-1}t_{m-j}x_{j,m-1}(\lambda)\right],\tag{12}$$

and

$$x_{jm}(\lambda)=x_{j,m-1}(\lambda)-x_{mm}(\lambda)\bar{x}_{m-j,m-1}(\lambda),\qquad 1\leqq j\leqq m-1.\tag{13}$$

For convenience, we say that a real number λ is *nondefective with respect to T_n* if it is not an eigenvalue of any of the principal submatrices $T_1,\cdots,T_{n-1}$. Conversely, λ is *defective with respect to T_n* if it is an eigenvalue of any of these matrices. An eigenvalue of T_n which is not simultaneously an eigenvalue of any of the principal submatrices will be said to be a *nondefective eigenvalue* of T_n. From the Cauchy Separation Theorem, a nondefective eigenvalue must be of multiplicity one. (Note that these are nonstandard usages of *defective* and *nondefective*.)

Cybenko [6] has shown that if $m\geqq 2$ and

$$L_m(\lambda)=\begin{bmatrix}1 & 0 & \cdots & 0 & 0\\ -x_{1,m-1}(\lambda) & 1 & \cdots & 0 & 0\\ -x_{2,m-1}(\lambda) & -x_{1,m-2}(\lambda) & \cdots & 1 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ -x_{m-1,m-1}(\lambda) & -x_{m-2,m-2}(\lambda) & \cdots & -x_{1,1}(\lambda) & 1\end{bmatrix},$$

then

$$\bar{L}_m^T(\lambda)(T_m-\lambda I_m)L_m(\lambda)=\operatorname{diag}[\Delta_m(\lambda),\cdots,\Delta_1(\lambda)].\tag{14}$$

Because of Sylvester's law of inertia, this implies the following theorem, which has been used previously in the algorithms of Cybenko and Van Loan [8] and Hu and Kung [15] for computing the smallest eigenvalue of a positive definite Hermitian Toeplitz matrix, and is also crucial for the more general algorithm presented here.

THEOREM 3. *Let* $\mathrm{Neg}_m(\lambda)$ *be the number of eigenvalues of* T_m (*counting multiplicities*) *less than* λ. *Then* $\mathrm{Neg}_m(\lambda)$ *equals the number of negative values among* $\{\Delta_1(\lambda), \cdots, \Delta_m(\lambda)\}$, *provided that* λ *is nondefective with respect to* T_m.

Since $\det L_m(\lambda) = 1$, (14) implies that

$$(15) \qquad p_m(\lambda) = \prod_{j=1}^{m} \Delta_m(\lambda), \qquad 1 \leqq m \leqq n,$$

which is essentially equivalent to a formula obtained in [21] for the determinant of a Hermitian Toeplitz matrix. Setting $m = n$ in (15) shows that $p_n(\lambda) = \Delta_n(\lambda)p_{n-1}(\lambda)$; hence

$$(16) \qquad q_n(\lambda) = \Delta_n(\lambda).$$

Henceforth we will use $q_n(\lambda)$ and $\Delta_n(\lambda)$ interchangeably.

Note that it is not necessary to carry out the computations in (13) for $m = n - 1$ in order to compute $q_n(\lambda)$ from (16), as it would be if we wished to use (5) obtained earlier. (However, (16) requires that $T_m - \lambda I_m$ be nonsingular for $1 \leqq m \leqq n - 1$, while (5) requires only that $T_{n-1} - \lambda I_{n-1}$ be nonsingular.)

THEOREM 4. *Suppose that* α *and* β *are nondefective with respect to* T_n, *and that* (α, β) *contains exactly one eigenvalue* (*with multiplicity one*) *of* T_n. *Suppose also that neither* α *nor* β *is an eigenvalue of* T_n. *Then* (α, β) *contains no eigenvalues of* T_{n-1} *if and only if* $\Delta_n(\alpha) > 0$ *and* $\Delta_n(\beta) < 0$.

Proof. Since $\mathrm{Neg}_n(\beta) = 1 + \mathrm{Neg}_n(\alpha)$ by assumption, Theorem 3 implies that the set $\{\Delta_1(\beta), \cdots, \Delta_n(\beta)\}$ has exactly one more negative member than the set $\{\Delta_1(\alpha), \cdots, \Delta_n(\alpha)\}$. Therefore, if either $\Delta_n(\alpha) < 0$ or $\Delta_n(\beta) > 0$, the set $\{\Delta_1(\beta), \cdots, \Delta_{n-1}(\beta)\}$ must contain more negative members than the set $\{\Delta_1(\alpha), \cdots, \Delta_{n-1}(\alpha)\}$, and therefore (α, β) contains at least one eigenvalue of T_{n-1}, by Theorem 3. On the other hand, if $\Delta_n(\alpha) > 0$ and $\Delta_n(\beta) < 0$, then the two sets mentioned in the last sentence must contain the same number of negative elements, and Theorem 3 implies that $\mathrm{Neg}_{n-1}(\beta) = \mathrm{Neg}_{n-1}(\alpha)$, i.e., that T_{n-1} has no eigenvalues in (α, β).

As observed in [8], the idea of computing $p_n(\lambda)/p_{n-1}(\lambda)$ by partitioning a Hermitian matrix as in (7) and then locating its zeros by combining inertia computations with a root finding method has been used by other authors (see, e.g., [19] and [24]); however, this approach requires $O(n^3)$ operations for the general Hermitian matrix, rather than the $O(n^2)$ required for Toeplitz matrices.

In connection with his work on real centrosymmetric matrices, Andrew [1] has defined a vector $V = [v_1, \cdots, v_n]^t$ to be *symmetric* if

$$(17) \qquad v_j = v_{n-j+1}, \qquad 1 \leqq j \leqq n,$$

or *skew-symmetric* if

$$(18) \qquad v_j = -v_{n-j+1}, \qquad 1 \leqq j \leqq n.$$

Cantoni and Butler [5] have shown that if T is a real symmetric Toeplitz matrix of order n, then R has an orthonormal basis consisting of $n - [n/2]$ symmetric and $[n/2]$ skew-symmetric eigenvectors of T. (Here $[x]$ is the integer part of x.) For convenience, let us say that an eigenvalue of T_n is *even* or *odd* if it has an associated eigenvector satisfying (17) or (18), respectively. It is clear from (11) with $m = n - 1$ that λ is a nondefective eigenvalue of a real symmetric Toeplitz matrix if and only if $x_{n-1,n-1}(\lambda) = \pm 1$. From the form of the associated eigenvector $Y_n(\lambda)$ in (4), we can see more specifically that λ is a nondefective even eigenvalue of T_n if and only if $x_{n-1,n-1}(\lambda) = -1$, or a nondefective odd eigenvalue if and only if $x_{n-1,n-1}(\lambda) = 1$.

3. The iterative procedure. If λ is defective with respect to T_n, then $q_n(\lambda)$ cannot be computed by means of Theorem 3. For practical purposes it is more appropriate to observe that $q_n(\lambda)$ cannot be computed in this way if at least one of the quantities $\Delta_1(\lambda), \cdots, \Delta_{n-1}(\lambda)$ is so small as to cause overflow in (12) for some m in $\{1, \cdots, n-1\}$. We will discuss this further in § 5; however, for now it suffices to say that in the numerical experiments reported in § 4, which comprise the computation of thousands of eigenvalues, there was not a single instance in which computation was terminated for this reason. Therefore, we will assume in this section that the eigenvalues of T_n (or at least those that we are trying to compute) are nondefective, and that none of the approximants to the eigenvalues generated by the procedure that we are about to describe are sufficiently close to being defective so as to cause overflow in (12).

We use an iterative procedure to locate the eigenvalues of T_n as the zeros of $q_n = p_n/p_{n-1}$. The iteration terminates when the difference between successive iterates is sufficiently small. In the following description of the procedure, we assume that $\Delta_n(\lambda) \neq 0$ for every value of λ encountered during the iteration. This is for convenience only; obviously, if $\Delta_n(\lambda)$ "underflows" to zero, then λ is an acceptable approximation to an eigenvalue. (This did not occur in any of our computations.)

Let the eigenvalues of T_n be

$$\lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_n,$$

and suppose that we wish to find a single nondefective eigenvalue λ_i. Our first task is to find an interval (α, β) containing λ_i, but not containing any other eigenvalues of T_n or any eigenvalues of T_{n-1}. On such an interval $q_n(\lambda)$ is continuous. Obviously α and β satisfy the first requirement if and only if

$$\text{(19)} \qquad \mathrm{Neg}_n(\alpha) = i-1 \quad \text{and} \quad \mathrm{Neg}_n(\beta) = i,$$

and, given this, Theorem 4 implies that the second holds if and only if

$$\text{(20)} \qquad \Delta_n(\alpha) > 0 \quad \text{and} \quad \Delta_n(\beta) < 0.$$

In the following, $\mathrm{Neg}_n(\lambda)$ is computed by means of Theorem 3.

To start, we find α and β, by trial and error, such that

$$\text{(21)} \qquad \mathrm{Neg}_n(\alpha) \leqq i-1 \quad \text{and} \quad \mathrm{Neg}_n(\beta) \geqq i.$$

If (19) and (20) hold for this α and β, then this phase of the computation is finished. If not, let $\gamma = (\alpha + \beta)/2$. If $\mathrm{Neg}_n(\gamma) \leqq i-1$, replace α by γ; if $\mathrm{Neg}_n(\gamma) \geqq i$, replace β by γ. Repeat this until (19) and (20) both hold, which must occur after finitely many steps.

Since q_n is continuous on the interval (α, β) that we have just determined, we can now switch from bisection to a more efficient zero finding method to locate λ_i. The method of false position [20] was unacceptably slow, but the Pegasus modification of this method yielded consistently good results. Since this procedure is well described in the literature (see, e.g., [10], [20]), we will not describe it here, except to say that if $\{\mu_j\}$ is the sequence of iterates produced by the Pegasus computation, starting with $\mu_0 = \alpha$ and $\mu_1 = \beta$, then we terminate this phase of the computation at the first integer r such that

$$\text{(22)} \qquad |\mu_r - \mu_{r-1}| < .5(1+\mu_r)10^{-K},$$

where K is a suitable positive integer. We then compute $\Delta_n(\mu_r)$ from (10)–(13), and continue (13) with $m = n-1$ (which is not required to compute $\Delta_n(\mu_r)$, as mentioned earlier) to compute $x_{1,n-1}(\mu_r), \cdots, x_{n-2,n-1}(\mu_r)$. Then we use Newton's method to obtain a final approximation to λ_i:

$$\begin{aligned}\mu_{r+1} &= \mu_r - \Delta_n(\mu_r)/\Delta_n'(\mu_r)\\ &= \mu_r + \Delta_n(\mu_r)/(1 + \|X_{n-1}(\mu_r)\|^2)\end{aligned}$$

(cf. (6) and (16)).

This application of Newton's method is "for good measure," and can probably be omitted without great loss. We included it without rigorously evaluating its effect because in some cases it appeared to allow the use of a smaller integer K in (22) without degrading the results, and we report it here since it was used in most of the computations reported in § 4. We did not use Newton's method as our principal iterative technique (after determining α and β as in (19) and (20)), since it could produce a sequence of iterates that converges to another eigenvalue or that does not converge at all. The Pegasus method does not suffer from this defect, and it has a respectable order of convergence (approximately 1.642).

It may be of interest to note that Newton's method as applied to this problem is actually a form of Rayleigh quotient iteration. To see this, suppose that λ is an approximation to an eigenvalue of T_n. Then the vector $Y_n(\lambda)$ in (4) is an approximation to a corresponding eigenvector, and a new approximation $\hat{\lambda}$ to the eigenvalue can be obtained by computing the Rayleigh quotient

$$\hat{\lambda} = \frac{\bar{Y}_n^t(\lambda) T_n Y_n(\lambda)}{\|Y_n\|^2}. \tag{23}$$

However, from (8) and (16),

$$T_n Y_n(\lambda) = \lambda Y_n(\lambda) - \Delta_n(\lambda)[1,0,\cdots,0]^t,$$

so that

$$\bar{Y}_n^t(\lambda) T_n Y_n(\lambda) = \lambda \|Y_n(\lambda)\|^2 + \Delta_n(\lambda).$$

Since

$$\|Y_n(\lambda)\|^2 = 1 + \|X_{n-1}(\lambda)\|^2 = -\Delta_n'(\lambda)$$

(cf. (6) and (16)), it now follows that the Rayleigh quotient $\hat{\lambda}$ in (23) can be rewritten as

$$\hat{\lambda} = \lambda - \Delta_n(\lambda)/\Delta_n'(\lambda),$$

which establishes our point.

Now suppose that we wish to find eigenvalues $\lambda_p, \cdots, \lambda_q$, where $1 \leqq p < q \leqq n$. Since it would be wasteful to simply apply the procedure just described independently for $i = p, \cdots, q$, we will define a method for finding $\xi_{p-1}, \cdots, \xi_q$ such that

$$\xi_{i-1} < \lambda_i < \xi_i, \qquad p \leqq i \leqq q. \tag{24}$$

Having accomplished this, we then apply the above described procedure for $i = p, \cdots, q$, taking the initial points in the search for λ_i to be $\alpha = \xi_{i-1}$ and $\beta = \xi_i$. (Clearly, (24) implies (21) in this case.) It is to be understood that as each ξ_i is determined, $\Delta_n(\xi_i)$ is retained for subsequent use.

The inequalities (24) are equivalent to

$$\begin{aligned}&\mathrm{Neg}_n(\xi_{p-1}) \leqq p-1,\\ &\mathrm{Neg}_n(\xi_i) = i, \qquad p \leqq i \leqq q-1,\\ &\mathrm{Neg}_n(\xi_q) \geqq q.\end{aligned} \tag{25}$$

We specify the method for choosing $\xi_{p-1}, \cdots, \xi_q$ inductively. We start by choosing a and b, by trial and error, such that $\mathrm{Neg}_n(a) \leqq p-1$ and $\mathrm{Neg}_n(b) \geqq q$, and let $\xi_{p-1} = a$ and $\xi_q = b$. Now suppose that at some step of our inductive procedure ξ_{p-1} and ξ_q have been specified, but at least one of the intermediate points $\xi_p, \cdots, \xi_{q-1}$ has not. Let r and s be the smallest integers such that $p \leqq r < s \leqq q$ and ξ_r has not been selected, while ξ_s has. Define

$$\gamma = (\xi_{r-1} + \xi_s)/2 \tag{26}$$

and $k = \mathrm{Neg}_n(\gamma)$. If $r = p$ and $k < p-1$ (which can occur only if the inequality holds in (25)), then we replace ξ_{p-1} by γ. Similarly, if $s = q$ and $k > q$, then we replace ξ_q by γ. In all other cases, $r - 1 \leqq k \leqq s$, and we let $\xi_k = \gamma$.

This procedure merely replaces a previously selected ξ_k unless k satisfies the stronger inequalities $r \leqq k \leqq s-1$; however, the bisection (26) will obviously cause the selection process to be completed in a finite number of steps.

Since ξ_{i-1} is no longer needed after λ_i has been obtained, λ_i can be stored in the location previously occupied by ξ_{i-1}.

4. Computational results. We considered real symmetric matrices only. All computations reported here were performed in double precision (15+ decimal places) in Fortran 77. The computations for all matrices of order less than 1,000 were performed on an IBM PC AT. Those for matrices of order 1,000 were performed on an IBM PS/2 Model 60. Both machines are equipped with the 80287 coprocessor. Due to the limitations of available computing equipment, we made no attempt to use parallel processing to solve the Levinson–Durbin system (3). Therefore, the computation of each eigenvalue and its associated eigenvector with our implementation of the proposed method requires $O(n^2)$ steps, where the "constant" buried in the "O" depends, of course, on the number of iterations required for the given eigenvalue. Although this number depends on the eigenvalue itself and on the starting values (α and β), its average value over all eigenvalue-eigenvector pairs for a matrix of order n appears to be essentially independent of n. Of course, it depends on K in (22). In the computations reported here, we took $K = 10$.

We consider two kinds of matrices: the Kac-Murdock-Szegö (KMS) matrices

$$T_n = (\rho^{|i-j|})_{i,j=1}^{n} \qquad (0<\rho<1) \tag{27}$$

discussed in [13] and [16], and matrices

$$T_n = (t_{i-j})_{i,j=1}^{n},$$

in which the defining elements $t_0, \cdots, t_{n-1}$ were randomly generated with a uniform distribution in $[-10, 10]$.

The eigenvalues of the KMS matrices can be computed quite easily, even on a handheld calculator. It is shown in [13] that if

$$\sin(n+1)\gamma - 2\rho \sin n\gamma + \rho^2 \sin(n-1)\gamma = 0, \tag{28}$$

then the quantity

$$\lambda = (1-\rho^2)(1-2\rho\cos\gamma + \rho^2)^{-1} \tag{29}$$

is an eigenvalue of T_n in (27). Moreover, it is also shown in [13] that (28) has roots $\gamma_1, \cdots, \gamma_n$ satisfying the inequalities

$$0 < \gamma_1 < \frac{\pi}{n+1} < \gamma_2 < \frac{2\pi}{n+1} < \cdots < \gamma_n < \frac{n\pi}{n+1}.$$

TABLE 1

Distribution of fractional errors $\{f_i\}$ *in the eigenvalues of* KMS *matrices of order* $n = 100$.

$\rho =$	.2	.3	.4	.5	.6	.7	.8	.9	.95	.995
$[10^{-8}, 10^{-7})$	0	0	0	1	0	0	0	0	0	0
$[10^{-10}, 10^{-9})$	0	0	0	0	0	0	0	0	0	18
$[10^{-11}, 10^{-10})$	0	0	0	0	0	0	0	0	0	19
$[10^{-12}, 10^{-11})$	0	3	0	0	1	1	0	1	12	34
$[10^{-13}, 10^{-12})$	1	7	1	1	2	4	15	33	53	24
$[10^{-14}, 10^{-13})$	3	50	9	14	23	31	47	49	29	3
$[10^{-15}, 10^{-14})$	41	31	60	64	61	54	31	16	5	2
$[10^{-16}, 10^{-15})$	47	0	27	17	10	8	6	1	0	0
$[0, 10^{-16})$	8	9	3	3	3	2	1	0	1	0

Average number of function evaluations per eigenvalue = 10.33.
Average running time per matrix = 23.26 min.

Given such precise information on their locations, it is a simple matter to find $\gamma_1, \cdots, \gamma_n$ by standard root-finding methods, and then to compute the eigenvalues $\lambda_1 < \lambda_2 < \cdots < \lambda_n$ from

$$\lambda_i = (1-\rho^2)(1-2\rho\cos\gamma_{n-i+1}+\rho^2)^{-1}, \qquad 1 \leqq i \leqq n. \tag{30}$$

For a considerable extension of this idea, see [22].

We used the algorithm proposed here to compute all eigenvalues of KMS matrices of orders $n = 100$, 300, 500, and 1,000 for various values of ρ. We also computed the same eigenvalues by the "exact" method; that is, by solving (28) iteratively with the Pegasus procedure to obtain $\gamma_1, \cdots, \gamma_n$, and then computing $\lambda_1, \cdots, \lambda_n$ from (30). We terminated the iteration for each γ_i as soon as the difference between successive iterates was less than 10^{-14}. We then computed the fractional error

$$f_i = (\hat{\lambda}_i - \tilde{\lambda}_i)/\tilde{\lambda}_i, \tag{31}$$

where $\hat{\lambda}_i$ and $\tilde{\lambda}_i$ are the estimates of λ_i obtained from our general algorithm and the "exact" method, respectively. The distributions of these fractional errors are shown in Tables 1, 2, 3, and 4; e.g., Table 1 shows that for $n = 100$ and $\rho = .5$, 14 of the fractional errors were in the interval $[10^{-14}, 10^{-13})$.

Tables 5 and 6 summarize results obtained in computing all eigenvalues of 20 randomly generated matrices of order 100, 24 of order 150, 22 of order 300, five of order 500, and two of order 1,000. As mentioned above, the eigenvectors were obtained as byproducts. We attempted to assess the results as follows.

TABLE 2

Distribution of fractional errors $\{f_i\}$ *in the eigenvalues of* KMS *matrices of order* $n = 300$.

$\rho =$	.2	.3	.4	.5	.6	.7	.8	.9	.95
$[10^{-11}, 10^{-10})$	0	0	0	0	0	1	0	2	8
$[10^{-12}, 10^{-11})$	0	1	1	2	4	6	10	15	93
$[10^{-13}, 10^{-12})$	7	5	13	15	17	39	76	160	144
$[10^{-14}, 10^{-13})$	66	80	102	137	165	178	167	107	48
$[10^{-15}, 10^{-14})$	157	150	142	119	98	62	40	14	5
$[10^{-16}, 10^{-15})$	60	57	37	24	12	14	5	1	2
$[0, 10^{-16})$	10	7	5	3	4	0	2	1	0

Average number of function evaluations per eigenvalue = 10.21.
Average running time per matrix = 10.25 hrs.

TABLE 3
Distribution of fractional errors $\{f_i\}$ in eigenvalues of KMS *matrices of order n = 500.*

$\rho =$	.5	.95
$[10^{-10}, 10^{-9})$	0	1
$[10^{-11}, 10^{-10})$	0	19
$[10^{-12}, 10^{-11})$	5	211
$[10^{-13}, 10^{-12})$	47	214
$[10^{-14}, 10^{-13})$	260	47
$[10^{-15}, 10^{-14})$	159	6
$[10^{-16}, 10^{-15})$	27	1
$[0, 10^{-16})$	2	1

We computed

$$Q_i = |q_n(\hat{\lambda}_i)|, \qquad 1 \leqq i \leqq n \tag{32}$$

(or, equivalently, $Q_i = |\Delta_n(\hat{\lambda}_i)|$), where $\hat{\lambda}_i$ is the final estimate of λ_i. We also computed

$$R_i = \min\{|x_{n-1,n-1}(\hat{\lambda}_i) - 1|, |x_{n-1,n-1}(\lambda_i) + 1|\}. \tag{33}$$

It is obvious from (1) that $Q_i = 0$ if $\hat{\lambda}_i = \lambda_i$. Also, since λ_i is an eigenvalue of T_n if $x_{n-1,n-1}(\lambda_i) = \pm 1$, $R_i = 0$ if $\hat{\lambda}_i = \lambda_i$. Table 5 shows the percentage distributions of $\{Q_i\}$ and $\{R_i\}$. Here n is the order of the matrix and m is the number of matrices of that order for which the results are given. Under each value of n there are two columns, headed Q and R, which show the percentage distributions of $\{Q_i\}$ and $\{R_i\}$, respectively, for all m matrices of the given order n. For example, 34.58 percent of the $\{Q_i\}$ and 11.74 percent of the $\{R_i\}$ fell in the interval $[10^{-9}, 10^{-8})$ for $n = 300$.

After a considerable portion of the computations summarized in Table 5 had been completed, we decided that a more decisive measure of error should be calculated, even though it engendered a substantial increase in computation time, namely,

$$\sigma_i = \|T_n - \hat{\lambda}_i Y_n(\hat{\lambda}_i)\| / \|Y_n(\hat{\lambda}_i)\|, \tag{34}$$

since $Y_n(\hat{\lambda}_i)$ (as defined in (4) with $\lambda = \hat{\lambda}_i$) is an approximate λ_i-eigenvector. Table 6 shows the percentage distribution of $\{\sigma_i\}$ for a subclass of the matrices considered in Table 5; again, m is the number of matrices of the given order n included in this subclass.

The computations in (10)–(13) require approximately n^2 flops for each λ. With $K = 10$ in (22), these computations were performed on the average approximately eleven

TABLE 4
Distribution of fractional errors $\{f_i\}$ in the eigenvalues of the KMS *matrix.*

$$T_n = (.9^{|i-j|})_{i,j=1}^{1000}.$$

$[10^{-9}, 10^{-8})$	1
$[10^{-10}, 10^{-9})$	3
$[10^{-11}, 10^{-10})$	33
$[10^{-12}, 10^{-11})$	313
$[10^{-13}, 10^{-12})$	507
$[10^{-14}, 10^{-13})$	127
$[10^{-15}, 10^{-14})$	15
$[0, 10^{-15})$	1

TABLE 5
Percentage distributions of $\{Q_i\}$ and $\{R_i\}$ for randomly generated matrices.

$n =$	100		150		300		500		1000	
$m =$	20		24		22		5		2	
	Q	R	Q	R	Q	R	Q	R	Q	R
$[1, 10)$	00.00	00.00	00.00	00.00	00.02	00.00	00.00	00.00	00.00	00.00
$[10^{-1}, 1)$	00.00	00.00	00.00	00.00	00.03	00.02	00.08	00.00	00.15	00.00
$[10^{-2}, 10^{-1})$	00.00	00.00	00.06	00.00	00.03	00.02	00.04	00.04	00.20	00.00
$[10^{-3}, 10^{-2})$	00.00	00.00	00.03	00.00	00.14	00.00	00.16	00.04	00.90	00.05
$[10^{-4}, 10^{-3})$	00.15	00.00	00.17	00.06	00.30	00.06	00.56	00.04	02.15	00.15
$[10^{-5}, 10^{-4})$	00.45	00.05	00.39	00.06	00.97	00.15	02.60	00.08	07.95	00.55
$[10^{-6}, 10^{-5})$	00.50	00.10	01.17	00.19	03.23	00.33	07.92	00.88	19.25	02.40
$[10^{-7}, 10^{-6})$	02.50	00.35	03.33	00.53	09.55	01.21	20.20	02.36	30.90	07.55
$[10^{-8}, 10^{-7})$	05.15	00.95	10.50	01.39	24.29	04.94	32.12	09.40	27.50	16.50
$[10^{-9}, 10^{-8})$	15.90	03.50	23.50	05.11	34.58	11.74	25.80	18.92	08.65	25.75
$[10^{-10}, 10^{-9})$	28.60	09.25	35.83	14.14	20.47	24.80	08.64	28.96	01.80	27.35
$[10^{-11}, 10^{-10})$	30.20	20.40	19.47	25.47	05.27	28.64	01.80	24.52	00.40	13.65
$[10^{-12}, 10^{-11})$	13.45	27.60	04.75	29.58	00.98	19.53	00.04	11.24	00.10	04.75
$[10^{-13}, 10^{-12})$	02.60	22.70	00.78	16.31	00.14	06.52	00.00	02.88	00.05	00.90
$[10^{-14}, 10^{-13})$	00.45	11.60	00.03	05.56	00.02	01.64	00.04	00.56	00.00	00.35
$[0, 10^{-14})$	00.05	03.50	00.00	01.61	00.00	00.41	00.00	00.08	00.00	00.05

Average number of function evaluations per eigenvalue: 10.92, 10.89, 10.81, 10.84.

times per eigenvalue (and this was essentially independent of the particular matrix or its order). Let us extrapolate from these computations and assume that this method requires approximately $M(K)n^2$ flops per eigenvalue, where $M(10) \approx 11$. By comparison, standard QR requires approximately $2n^3/3$ flops for the preliminary tridiagonalization of T_n, after which all the eigenvalues can be computed with $O(n)$ flops [12, §8.2]. On the basis of this count only, it would seem that the method presented here has a clear advantage over standard QR if it is desired to compute N eigenvalues ($1 \leq N \leq n$) of T_n, provided that N is small compared to $(2n)/3M(K)$, while the advantage shifts to standard QR if this

TABLE 6
Percentage distribution of errors $\{\sigma_i\}$ for randomly generated matrices.

$n =$	100	150	300	500	1000
$m =$	10	6	5	2	2
$[10^{-4}, 10^{-3})$	00.00	00.00	00.07	00.00	00.05
$[10^{-5}, 10^{-4})$	00.00	00.00	00.00	00.10	00.35
$[10^{-6}, 10^{-5})$	00.00	00.00	00.07	00.20	00.90
$[10^{-7}, 10^{-6})$	00.00	00.33	00.47	00.10	07.00
$[10^{-8}, 10^{-7})$	00.30	00.67	03.20	09.20	28.25
$[10^{-9}, 10^{-8})$	01.40	03.33	17.93	35.40	45.85
$[10^{-10}, 10^{-9})$	12.50	19.11	46.47	43.00	14.80
$[10^{-11}, 10^{-10})$	36.10	51.89	26.87	09.70	01.90
$[10^{-12}, 10^{-11})$	40.40	20.67	03.93	01.20	00.85
$[10^{-13}, 10^{-12})$	07.60	03.33	01.00	00.20	00.05
$[10^{-14}, 10^{-13})$	01.70	00.67	00.00	00.00	00.00
$[0, 10^{-14})$	00.00	00.00	00.00	00.00	00.00

Average number of function evaluations per eigenvalue: 10.92, 10.89, 10.81, 10.84.

is not so. In the context of parallel processing, the determination of the crossover point is more complicated; since the computations for distinct eigenvalues are completely independent of each other with the present method, it is straightforward to distribute the labor of computing many eigenvectors among multiple processors. Moreover, the memory requirement for the present method is $O(n)$, compared to $O(n^2)$ for standard QR.

The approximate average running times required on the IBM PC AT to find all eigenvalues and eigenvectors of T_n, with $K = 10$ in (22) and without computing σ_i (cf. (34)), were 24 min., 81 min., 10.6 hrs., and 49 hrs. for $n = 100, 150, 300$, and 500, respectively. For those runs in which σ_i was computed, the average running times were approximately 27 min., 88 min., 11.6 hrs., and 54 hrs., respectively. For the matrices of order 1,000 the average running time per eigenvalue-eigenvector pair was approximately 15 min. on the PS/2 Model 60.

5. The effects of defectiveness. Our program included a command to terminate computation of $q_n(\lambda) = \Delta_n(\lambda)$ if

$$(35) \qquad |1 - x^2_{m-1,m-1}(\lambda)| < 10^{-J}$$

for some m in $\{1, \cdots, n-1\}$. The purpose of this test is to prevent overflow in (12) if λ is too close to an eigenvalue of one of the principal submatrices $T_1, \cdots, T_{n-1}$ of T_n. In the computations reported in § 4 we took $J = 9$ (recall that $K = 10$ in (22)), and termination for this reason never occurred. Thus, the practical effect of defectiveness is not that it is likely to cause overflow (although this can be forced to happen in contrived situations), rather, it affects the accuracy of the results.

In most cases where the error indicators f_i, Q_i, and R_i, were relatively large, we were able to ascertain that the eigenvalues in question were close to being defective. For example, it can be seen in Table 5 that Q_i was in the interval $[1, 10)$ for one of the 6,600 eigenvalues computed for randomly generated matrices of order 300. This was $\hat{\lambda} = 122.4186385510399$, with $q_{300}(\hat{\lambda}) \cong 8.45$. To test for defectiveness, we reduced J in (35) to four and attempted to compute $q_{300}(\hat{\lambda})$. The calculation terminated with $m = 298$. Subsequent calculation showed that $q_{298}(\hat{\lambda}) \cong .31 \times 10^{-5}$, indicating that $\hat{\lambda}$ was close to an eigenvalue of T_{298}. Examination of other cases in which the error indicators were unusually large yielded similar results.

The results in Table 1 for the KMS matrix with $\rho = .5$ and $n = 100$ show that the fractional error f_i for one eigenvalue is in the interval $[10^{-8}, 10^{-7})$, while all the others are less than 10^{-12}. The eigenvalue for which this occurred is $\lambda = 1$, which is defective; indeed, it is straightforward to verify that if $\rho = .5$ and $n = 3m + 1$ ($m = 0, 1, 2, \cdots$), then $\gamma = \pi/3$ satisfies (28) and therefore, from (29), $\lambda = 1$ is an eigenvalue of T_n. Hence, $\lambda = 1$ is an eigenvalue of 33 principal submatrices of T_{100}.

Although our results indicate that defectiveness in the sense that we have defined it is not a major problem for the matrices that we have considered, it would still be worthwhile to develop methods to overcome it. Clearly, defectiveness is a problem—theoretically—mainly because the Levinson–Durbin Algorithm for solving the Yule–Walker equation (3) requires that all the principal submatrices of $T_n - \lambda I_n$ be nonsingular, i.e., that $T_n - \lambda I_n$ be "strongly nonsingular." Alternative methods have been proposed for solving Toeplitz systems with matrices that are not strongly nonsingular; for discussions of such methods see [2]–[4]. A possible direction for future research would be to incorporate some of the ideas in these references into the present method.

Acknowledgment. The author thanks G. Cybenko and the referees for several suggestions for improving the manuscript.

REFERENCES

[1] A. L. Andrew, *Eigenvectors of certain matrices*, Linear Algebra Appl., 7 (1973), pp. 151–162.

[2] R. P. Brent, F. G. Gustavson, and D. Y. Y. Yun, *Fast solution of Toeplitz systems of equations and computation of Padé approximants*, J. Algorithms, 1 (1980), pp. 259–295.

[3] J. R. Bunch, *Stability of methods for solving Toeplitz systems of equations*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 349–364.

[4] ———, *The weak and strong stability of algorithms in numerical linear algebra*, Linear Algebra Appl., 88/89 (1987), pp. 49–66.

[5] A. Cantoni and F. Butler, *Eigenvalues and eigenvectors of symmetric centrosymmetric matrices*, Linear Algebra Appl., 13 (1976), pp. 275–288.

[6] G. Cybenko, *The numerical stability of the Levinson-Durbin Algorithm for Toeplitz systems of equations*, SIAM J. Sci. Statist. Comput., 1 (1980), pp. 303–319.

[7] ———, *On the eigenstructure of Toeplitz matrices*, IEEE Trans. Acoust. Speech Signal Process., 32 (1984), pp. 275–288.

[8] G. Cybenko and C. Van Loan, *Computing the minimum eigenvalue of a symmetric positive definite Toeplitz matrix*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 123–131.

[9] P. Delsarte and Y. Genin, *Spectral properties of finite Toeplitz matrices*, in Mathematical Theory of Networks and Systems, Proc. MTNS-83 International Symposium, Beer Sheva, Israel, 1983, pp. 194–213.

[10] M. Dowell and P. Jaratt, *The "Pegasus" method for computing the root of an equation*, BIT, 12 (1972), pp. 503–508.

[11] D. R. Fuhrmann and B. Liu, *Approximating the eigenvalues of a symmetric Toeplitz matrix*, in Proc. 21st Annual Allerton Conference on Communications, Control and Computing, Monticello, IL, 1983, pp. 1046–1055.

[12] G. H. Golub and C. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, 1983.

[13] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, University of California Press, Berkeley, 1958.

[14] Y. H. Hu and S. Y. Kung, *Highly concurrent Toeplitz eigensystem solver for high resolution spectral estimation*, Proc. ICASSP, 83 (1983), pp. 1422–1425.

[15] ———, *Toeplitz eigensystem solver*, IEEE Trans. Acoust. Speech Signal. Process., 33 (1985), pp. 1264–1271.

[16] M. Kac, W. L. Murdock, and G. Szegö, *On the eigenvalues of certain Hermitian forms*, J. Rat. Mech. Anal., 2 (1953), pp. 767–800.

[17] N. Levinson, *The Weiner* RMS (*root mean square*) *error criterion in filter design and prediction*, J. Math. Phys., 25 (1947), pp. 261–278.

[18] J. Makhoul, *On the eigenvectors of symmetric Toeplitz matrices*, IEEE Trans. Acoust. Speech Signal Process., 29 (1981), pp. 868–872.

[19] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[20] A. Ralston and P. Rabinowitz, *A First Course in Numerical Analysis*, 2nd ed., McGraw-Hill, New York, 1978.

[21] W. F. Trench, *An algorithm for the inversion of finite Toeplitz matrices*, SIAM J. Appl. Math., 12 (1964), pp. 515–522.

[22] ———, *Numerical solution of the eigenvalue problem for symmetric rationally generated Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 291–303.

[23] D. M. Wilkes and M. H. Hayes, *An eigenvalue recursion for Toeplitz matrices*, IEEE Trans. Acoust. Speech Signal Process., 35 (1987), pp. 907–909.

[24] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 2, pp. 147–154, April 1989

002

INVARIANT FACTOR ASSIGNMENT ON HIGHER-ORDER SYSTEMS USING STATE-FEEDBACK*

ION ZABALLA†

Abstract. Some of the known results concerning invariant factor assignment by means of state-feedback on first-order systems to systems with higher order are extended. Also, some problems concerning realizations of the transfer function matrix of such systems are studied.

Key words. invariant factors, state-feedback, linearization, realization

AMS(MOS) subject classifications. 15A18, 93B15, 93B55

1. Introduction. Consider the following first-order time invariant system:

$$\dot{x}(t)=Ax(t)+Bu(t) \tag{1}$$

where $A\in \mathbf{K}^{n\times n}$ and $B\in \mathbf{K}^{n\times m}$ ($\mathbf{K}=\mathbf{R}$ or $\mathbf{C}$). In [4], Lancaster and Maroulas have dealt with the problem of finding a feedback matrix $F\in \mathbf{K}^{m\times n}$ such that, on writing $u(t)=Fx(t)+v(t)$, the new system

$$\dot{x}(t)=(A+BF)x(t)+Bv(t)$$

has a fundamental matrix $A+BF$ with prescribed spectral properties, namely, prescribed Jordan canonical form. (They worked with complex matrices.)

It is well known that the Jordan form of a complex matrix A is completely determined, up to a permutation of its diagonal blocks, by the eigenvalues and the sizes of the Jordan blocks associated to them. (Occasionally these sizes are called the Segre characteristic of A.) In other words, the Jordan form of A is determined by its elementary divisors and then by its invariant factors.

If system (1) is completely controllable (i.e., the dimension of the controllable subspace of (A, B):

$$\mathscr{C}(A,B)=\sum_{j=0}^{n-1} I_m A^j B$$

is n), then a well-known result by Rosenbrock (see [5, p. 190] or [6]) states the following.

ROSENBROCK'S THEOREM. *If $\gamma_1,\cdots,\gamma_n$ are monic polynomials such that $\gamma_i :> \gamma_{i+1}$, $1\leqq i\leqq n-1$ (the symbol $:>$ is used to mean "divides") and*

$$\sum_{j=1}^{n} d(\gamma_j)=n$$

and system (1) is completely controllable, then there exists a feedback matrix $F\in\mathbf{K}^{m\times n}$ such that $A+BF$ has $\gamma_1,\cdots,\gamma_n$ as invariant factors, if and only if

$$\gamma_i=1,\qquad 1\leqq i\leqq n-m, \tag{2}$$

$$(k_1,\cdots,k_m)\prec(d(\gamma_n),\cdots,d(\gamma_{n-m+1})) \tag{3}$$

* Received by the editors July 1, 1988; accepted for publication August 24, 1988. This work was partly supported by 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, Williamsburg, Virginia.

† Escuela Universitaria de Magisterio, Universidad del País Vasco, 0100 6. Vitoria-Gasteiz, Spain. Present address, Universidad del País Vasco, Departamento de Matemáticas, Portal de Lasarte s/n, 01007. Vitoria-Gasteiz, Spain.

where $k_1 \geqq \cdots \geqq k_m$ are the controllability indices of (A, B), $d(\bullet)$ denotes degree, and $\prec$ is the symbol of majorization in the Hardy, Littlewood, and Polya sense.

(See [8] for a definition of controllability indices and $\prec$.)

So, when $\mathbf{K}$ is $\mathbf{C}$ and in spite of the possibility of prescribing arbitrarily the spectrum of $A + BF$, the Jordan form of this matrix must be constrained to verify (3).

On the other hand, if system (1) is not completely controllable (i.e., dim $\mathscr{C} < n$) then the spectrum of $A + BF$ can no longer be prescribed arbitrarily (see [3, p. 204], [4]), but still we can say how far its Jordan canonical form may be assigned.

THEOREM 1 [10]. *Under the same conditions as in Rosenbrock's Theorem, if $\mathbf{K} = \mathbf{F}$ is an arbitrary field and $\alpha_1 :> \cdots :> \alpha_n$ are the invariant factors of (A, B) (i.e., those of the polynomial matrix $[\lambda I_n - A, -B]$), then there exists a feedback matrix $F \in \mathbf{F}^{m \times n}$ such that $A + BF$ has $\gamma_1, \cdots, \gamma_n$ as invariant factors if and only if the following conditions hold*:

$$\gamma_{i-m} :> \alpha_i :> \gamma_i, \qquad 1 \leqq i \leqq n, \tag{4}$$

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

where $k_1 \geqq \cdots \geqq k_m$ are the controllability indices of (A, B),

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

$0 \leqq j \leqq m$ and $\alpha_i = \gamma_i = 1$ for $i < 1$.

Taking into account that the controllability of (A, B) is equivalent to the condition $\alpha_i = 1$ for $1 \leqq i \leqq n$ (see [8, p. 124]), it is easily seen that Theorem 1 contains that of Rosenbrock as a particular case (see [10]). Furthermore, Theorem 1 solves completely the problem of the possible canonical forms for the similarity available under state feedback, including, of course, pole assignability.

In § 2, and following the suggestion made in [4], we will extend this result to higher-order systems. Section 3 is devoted to analyzing the possible realization of the transfer function matrix of higher-order systems as defined in § 2.

2. Higher-order systems. Consider now a p-order system of the form

$$\frac{d^p x(t)}{dt^p} - \sum_{j=0}^{p-1} A_j \frac{d^j x(t)}{dt^j} = Bu(t) \tag{6}$$

where $A_j \in \mathbf{K}^{n \times n}$, $0 \leqq j \leqq p-1$, and $B \in \mathbf{K}^{n \times m}$. Put

$$L(\lambda) = I_n \lambda^p - \sum_{j=0}^{p-1} A_j \lambda^j$$

and define

$$y(t) = \begin{bmatrix} x(t) \\ \dfrac{dx(t)}{dt} \\ \vdots \\ \dfrac{d^{p-1}x(t)}{dt^{p-1}} \end{bmatrix}; \tag{7}$$

then, a first-order system equivalent to (6) is

$$\dot{y}(t) = Cy(t) + Du(t) \tag{8}$$

where $C \in \mathbf{K}^{np \times np}$ is the first companion matrix of $L(\lambda)$, i.e.,

$$C = \begin{bmatrix} 0 & I_n & 0 & \cdots & 0 \\ 0 & 0 & I_n & \cdots & 0 \\ 0 & 0 & 0 & \cdots & I_n \\ A_0 & A_1 & A_2 & \cdots & A_{p-1} \end{bmatrix} \tag{9}$$

$$D = \begin{bmatrix} 0 \\ B \end{bmatrix} \in \mathbf{K}^{np \times m}. \tag{10}$$

Following [4], state feedback is interpreted in the form

$$u(t) = \sum_{j=0}^{p-1} F_j \frac{d^j x(t)}{dt^j} + v(t) \tag{11}$$

where $F_j \in \mathbf{K}^{m \times n}$. The performance of a state feedback as that of (11) on system (6) yields the transformed system

$$\frac{d^p x(t)}{dt^p} - \sum_{j=0}^{p-1} (A_j + BF_j) \frac{d^j x(t)}{dt^j} = Bv(t)$$

and associated to it we have the matrix polynomial

$$\hat{L}(\lambda) = I_n \lambda^p - \sum_{j=0}^{p-1} (A_j + BF_j) \lambda^j.$$

By using the same substitution as in (7), we obtain the following equivalent first-order system:

$$\dot{y}(t) = (C + DF) y(t) + Dv(t)$$

where

$$F = [F_0, \cdots, F_{p-1}] \in \mathbf{K}^{m \times np}.$$

Our main result in this section is Theorem 2.

THEOREM 2. *Let $\mathbf{K} = \mathbf{F}$ be an arbitrary field, let $\alpha_1 :> \cdots :> \alpha_n$ be the invariant factors of the matrix polynomial $[L(\lambda), -B] \in \mathbf{F}[\lambda]^{n \times (n+m)}$, and let $\gamma_1 :> \cdots :> \gamma_n$ be monic polynomials such that*

$$\sum_{j=1}^{n} d(\gamma_j) = np.$$

Let $k_1 \geqq \cdots \geqq k_m$ be the controllability indices of (C, D). There exist matrices $F_j \in \mathbf{F}^{m \times n}$, $0 \leqq j \leqq p-1$ such that $\hat{L}(\lambda)$ has $\gamma_1, \cdots, \gamma_n$ as invariant factors if and only if conditions (4) and (5) hold.

Proof. Let us put $\varepsilon_i = \alpha_{i-(p-1)n}$, $1 \leqq i \leqq pn$. (Recall that $\alpha_i = 1$ for $i < 1$.) Write

$$Q(\lambda) = \begin{bmatrix} \lambda I_n & -I_n & 0 & \cdots & 0 & 0 \\ 0 & \lambda I_n & -I_n & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & \lambda I_n & -I_n \\ I_n & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \in \mathbf{F}[\lambda]^{np \times np}$$

$$P(\lambda) = \begin{bmatrix} I_n & 0 & \cdots & 0 & 0 \\ 0 & I_n & \cdots & 0 & 0 \\ 0 & 0 & \cdots & I_n & 0 \\ L_{p-1}(\lambda) & L_{p-2}(\lambda) & \cdots & L_1(\lambda) & L_0(\lambda) \end{bmatrix} \in \mathbf{F}[\lambda]^{pn \times pn}$$

where $L_0(\lambda) = I_n$ and $L_j(\lambda) = \lambda L_{j-1}(\lambda) - A_{p-j}$, $1 \leqq j \leqq p-1$. (See [2, p. 13].) Then $P(\lambda)$ and $Q(\lambda)$ are unimodular matrices and

$$P(\lambda)[\lambda I_{np} - C, -D] = \begin{bmatrix} I_{n(p-1)} & 0 & 0 \\ 0 & L(\lambda) & -B \end{bmatrix} \begin{bmatrix} Q(\lambda) & 0 \\ 0 & I_n \end{bmatrix}.$$

So,

$$[\lambda I_{np} - C, -D] \quad \text{and} \quad \begin{bmatrix} I_{np-1} & 0 & 0 \\ 0 & L(\lambda) & -B \end{bmatrix}$$

are equivalent polynomial matrices, and therefore they have the same invariant factors. Bearing in mind that $\alpha_1, \cdots, \alpha_n$ are the invariant factors of $[L(\lambda), -B]$, we can conclude that $\varepsilon_1, \cdots, \varepsilon_{pn}$ are the invariant factors of (C, D).

Let $\delta_i = \gamma_{i-(p-1)n}$, $1 \leqq i \leqq pn$ ($\gamma_i = 1$ for $i < 1$). By Theorem 1, there exists $F \in \mathbf{F}^{m \times np}$ such that $C + DF$ has $\delta_1, \cdots, \delta_{pn}$ as invariant factors if and only if

$$\delta_{i-m} :> \varepsilon_i :> \delta_i, \qquad 1 \leqq i \leqq pn, \tag{12}$$

$$(k_1, \cdots, k_m) \prec (d(\sigma_m), \cdots, d(\sigma_1)), \tag{13}$$

where

$$\sigma_j = \frac{\beta^j}{\beta^{j-1}}, \quad \beta^j = \beta_1^j \bullet \cdots \bullet \beta_{pn+j}^j, \quad \beta_i^j = \text{l.c.m.}\,(\varepsilon_{i-j}, \delta_{i-m}),$$
$$1 \leqq i \leqq pn+j, \quad 0 \leqq j \leqq m.$$

Let F_j be the submatrix of F formed by all its rows and the columns jnth, $\cdots$, $(j+1)n$th, $0 \leqq j \leqq p-1$. It is easily seen that $C + DF$ is the first companion matrix of $\hat{L}(\lambda)$. So [2, p. 13], $C + DF$ and diag $(I_{(p-1)n}, \hat{L}(\lambda))$ have the same invariant factors. That is to say, $\delta_1, \cdots, \delta_{pn}$ are the invariant factors of

$$\begin{bmatrix} I_{(p-1)n} & 0 \\ 0 & \hat{L}(\lambda) \end{bmatrix}.$$

But $\delta_i = 1$ for $1 \leqq i \leqq (p-1)n$ and $\delta_{i+(p-1)n} = \gamma_i$, $1 \leqq i \leqq n$. Hence $\gamma_1, \cdots, \gamma_n$ are the invariant factors of $\hat{L}(\lambda)$. So, to complete the proof of the theorem we have only to see that (12) and (13) are equivalent to (4) and (5), respectively, which is easily done taking into account the definition of the ε's and δ's. □

Since $[L(\lambda), B]$ and (C, D) have the same nontrivial invariant factors (i.e., those different from one) and, as we said before, a characterization of the controllability of (C, D) is that all its invariant factors are equal to one, we can say that system (6) is completely controllable if all the invariant factors of $[L(\lambda), B]$ are equal to one. In such a case, we get the following generalization of Rosenbrock's result to higher-order systems.

COROLLARY 1. *Under the same conditions as in Theorem* 2, *if* (6) *is a completely controllable system, then there exist matrices* $F_j \in \mathbf{F}^{m \times n}$, $0 \leqq j \leqq p-1$ *such that* $\hat{L}(\lambda)$ *has* $\gamma_1, \cdots, \gamma_n$ *as invariant factors if and only if conditions* (2) *and* (3) *hold.*

The proof of this Corollary 1 can be obtained from Theorem 2 in the same way as Rosenbrock's Theorem can be obtained from Theorem 1 (see [10]).

The pair (C, D) as defined by (9) and (10) summarizes all the information we need to solve the problem of the invariant factor assignment on higher-order systems. The same can be said about another pair (C_1, D_1) such that $C_1 = PCP^{-1}$ and $D_1 = PDQ$ for some nonsingular matrices P and Q. But all these pairs (C_1, D_1) provide more information than we need, because they are such that C_1 is a linearization [2, p. 12] of $L(\lambda)$ and the invariant factors of $L(\lambda)$ do not play a role in the solution of the problem. (Of course, the invariant factors of $L(\lambda)$ are related to those of $[L(\lambda), B]$ (see [9]), but this relation is not manifest in the solution. Whatever the invariant factors of $L(\lambda)$ are, the important matter is which are the invariant factors of $[L(\lambda), B]$.) Thus, an important problem to be solved is characterizing the pairs (C_1, D_1), which can provide all the structure information of $[L(\lambda), B]$ needed to solve the problem of invariant factor assignment. We will delve deeply into this point by dealing with the possible realizations of the transfer function matrix of system (6) (assuming that the outputs are the states), $L(\lambda)^{-1}B$.

3. Realizations of $L(\lambda)^{-1}B$. First of all, it should be noted that if

$$H:=[I_n,0,\cdots,0]\in\mathbf{F}^{n\times np}, \tag{14}$$

then $H(\lambda I_n - C)^{-1}D$ is a state-space realization [1] of $L(\lambda)^{-1}B$. (To see this, it is enough to apply in a suitable way transformations (i)–(iv) and Theorem 3.4 of [5, p. 59] to

$$\begin{bmatrix} \lambda I_{np}-C & -D \\ -H & 0 \end{bmatrix}.)$$

On the one hand, if $H_1(\lambda I - C_1)^{-1}D_1$ and $H_2(\lambda I - C_2)^{-1}D_2$ are two minimal (or irreducible) state-space realizations of $L(\lambda)^{-1}B$ then [1] for $i = 1, 2$, (C_i, D_i) is completely controllable (c.c.) and (C_i, H_i) is completely observable (c.o.). Or in other words, using the terminology of [5], the matrices in state-space form,

$$P_1(\lambda)=\begin{bmatrix} \lambda I-C_1 & -D_1 \\ -H_1 & 0 \end{bmatrix}, \qquad P_2(\lambda)=\begin{bmatrix} \lambda I-C_2 & -D_2 \\ -H_2 & 0 \end{bmatrix},$$

have no decoupling zeros. Hence, by the corollary of Theorem 3.1 of [5, p. 106], $P_1(\lambda)$ and $P_2(\lambda)$ are system similar, i.e. [5, p. 56] there exists a nonsingular matrix U such that

$$H_2=H_1U^{-1}, \quad C_2=UC_1U^{-1}, \quad D_2=UD_1.$$

Thus, C_1, C_2 are similar, and [9] (C_1, D_1) and (C_2, D_2) have the same controllability indices.

On the other hand, if $L(\lambda)$ and B are relatively left prime (i.e., all the invariant factors of $[L(\lambda), B]$ are equal to one), then $H(\lambda I_{np} - C)^{-1}D$ is a minimal state-space realization of $L(\lambda)^{-1}B$ (C, D and H will denote those matrices given by (9), (10), and (14), respectively). In fact, (C, H) is constructed to be c.o. and (C, D) is c.c. because (C, D) and $[L(\lambda), B]$ have the same nontrivial invariant factors. Thus, if $H_1(\lambda I - C_1)^{-1}D_1$ is another minimal state-space realization of $L(\lambda)^{-1}B$, then C and C_1 are similar and therefore [2, p. 15] C_1 is another linearization of $L(\lambda)$.

Hence, if $L(\lambda)$ and B are relatively left prime, all the necessary information needed to solve the problem of invariant factor assignment is provided by any pair (C_1, D_1) such that C_1 is a linearization of $L(\lambda)$, (C_1, D_1) is c.c., and there exists a matrix H_1 such that (C_1, H_1) is c.o. and $L(\lambda)^{-1}B = H_1(\lambda I - C_1)^{-1}D_1$. And since in this case C and C_1 are similar the size of C_1 is $np \times np$.

If $L(\lambda)$ and B are not left relatively prime, then we can still get a minimal state-space realization of $L(\lambda)^{-1}B$. Let $H_1(\lambda I - C_1)^{-1}D_1$ be such a minimal realization. Since (C_1, D_1) is c.c., this pair and $[L(\lambda), B]$ no longer have the same invariant factors. Thus, whenever we consider minimal realizations of $L(\lambda)^{-1}B$ we are losing part of the structure information of $[L(\lambda), B]$. Nevertheless, we can give the following result.

THEOREM 3. *For $i = 1, 2$ let t_i be positive integers and $H_i \in \mathbf{F}^{n\times t_i}$, $C_i \in \mathbf{F}^{t_i\times t_i}$, $D_i \in \mathbf{F}^{t_i\times m}$ matrices such that (C_i, H_i) is a c.o. pair. If for $i = 1, 2$, $H_i(\lambda I - C_i)^{-1}D_i$ are* *two state-space realizations of the same rational function matrix* $T(\lambda)$, *then* (C_1, D_1) *and* (C_2, D_2) *have the same controllability indices.*

Proof. For $i = 1, 2$ and by using Algorithm 7.2 of [5, p. 81],

$$Q_i(\lambda) = \begin{bmatrix} \lambda I_{t_i} - C_i & -D_i \\ -H_i & 0 \end{bmatrix}$$

can be brought by system similarity to the form (Kalman form)

$$P_i(\lambda) = \begin{bmatrix} \lambda I_{s_i} - C_{i1} & -C_{i2} & -D_{i1} \\ 0 & \lambda I_{t_i - s_i} - C_{i3} & 0 \\ -H_{i1} & -H_{i2} & 0 \end{bmatrix}$$

where (C_{i1}, D_{i1}) are c.c. and (C_{i1}, H_{i1}) are c.o. In other words, for $i = 1, 2$

$$P_{i1}(\lambda) = \begin{bmatrix} \lambda I_{s_i} - C_{i1} & -D_{i1} \\ -H_{i1} & 0 \end{bmatrix}$$

has no decoupling zeros. Since $P_i(\lambda)$ and $Q_i(\lambda)$ are system similar, they give rise to the same transfer function matrix [5, p. 59], and so

$$T(\lambda) = [H_{i1} H_{i2}] \begin{bmatrix} \lambda I_{s_i} - C_{i1} & -C_{i2} \\ 0 & \lambda I_{t_i - s_i} - C_{i3} \end{bmatrix}^{-1} \begin{bmatrix} D_{i1} \\ 0 \end{bmatrix}$$

$$= H_{i1}(\lambda I_{s_i} - C_{i1})^{-1} D_{i1}, \qquad i = 1, 2.$$

As $P_{11}(\lambda)$ and $P_{21}(\lambda)$ have no decoupling zeros, by Theorem 3.2 of [5, p. 108], we can conclude that $s_1 = s_2$ (say $s_1 = s_2 = s$) and again by the corollary of Theorem 3.1 of [5, p. 106], it turns out that $P_{11}(\lambda)$ and $P_{21}(\lambda)$ are system similar, and therefore there exists a nonsingular matrix $U \in \mathbf{F}^{s \times s}$ such that

$$C_{21} = U C_{11} U^{-1} \quad \text{and} \quad D_{21} = U D_{11}.$$

So, (C_{11}, D_{11}) and (C_{21}, D_{21}) have the same controllability indices and the theorem follows from the fact that the controllability indices of (C_i, D_i) are those of (C_{i1}, D_{i1}), $i = 1, 2$ [8, Lemma 2.8]. □

After this theorem the controllability indices of $[L(\lambda), B]$ are the controllability indices of the pair (C_1, D_1) in any state-space realization $H_1(\lambda I - C_1)^{-1} D_1$ of $L(\lambda)^{-1}B$ such that (C_1, H_1) is c.o., i.e., if $H_1(\lambda I - C_1)^{-1} D_1$ is any minimal realization of $L(\lambda)^{-1}B$ then the controllability indices of $[L(\lambda), B]$ are those of (C_1, D_1).

Since any state-space realization $H_1(\lambda I - C_1)^{-1} D_1$ of $L(\lambda)^{-1}B$ such that (C_1, H_1) is c.o. keeps the information about the controllability indices of $[L(\lambda), B]$ and $H(\lambda I_{np} - C)^{-1} D$ gives complete information concerning the invariant factors of $[L(\lambda), B]$, a new question arises in a natural way: What realizations of $L(\lambda)^{-1}B$ provide information concerning the invariant factors of $[L(\lambda), B]$?

THEOREM 4. *Let* $H_1(\lambda I - C_1)^{-1} D_1$ *be a realization of* $L(\lambda)^{-1}B$ *such that* (C_1, H_1) *is a c.o. pair. Let*

$$\begin{bmatrix} C_{11} & C_{12} & D_{11} \\ 0 & C_{13} & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} E_1 & C_2 & D_{12} \\ 0 & C_3 & 0 \end{bmatrix}$$

be the Kalman forms of (C_1, D_1) *and* (C, D), *respectively. Then* $[L(\lambda), B]$ *and* (C_1, D_1) *have the same controllability indices and nontrivial invariant factors if and only if we have the following*:

(i) C_{11} *and* E_1 *are similar*;

(ii) (C, D) *and* (C_1, D_1) *are* Γ*-equivalent.*

(See [8] for a definition of Γ-equivalence. This equivalence relation is called "block-similarity" in [3].)

Proof. We already proved the necessity of (i) in Theorem 3. Moreover, (C, D) and (C_1, D_1) are Γ-equivalent if and only if ([8, Thm. 2.12]) they have the same invariant factors and the same controllability indices. The theorem follows from the fact that the controllability indices and the nontrivial invariant factors of $[L(\lambda), B]$ are those of (C, D). □

One way to construct a pair (C_1, D_1) with the same controllability indices and the same nontrivial invariant factor as $[L(\lambda), B]$ is the following:

(1) Get a minimal state-space realization of $L(\lambda)^{-1}B$ by means of one of the several available procedures (for instance, see [5, p. 122]), and let $H_{11}(\lambda I_s - C_{11})^{-1}D_{11}$ be such a minimal realization. The dimension of this realization, s, is $np - q$ where $q = \sum_{j=1}^{n} d(\alpha_j)$, $\alpha_1 :> \cdots :> \alpha_n$ being the invariant factors of $[L(\lambda), B]$.

(2) Let $\alpha_{t+1}, \cdots, \alpha_n$ be the invariant factors of $[L(\lambda), B]$ different from one. Denote by N_i a companion matrix of α_{t+i}, $i = 1, \cdots, n-t$ put

$$C_{12} = \operatorname{diag}(N_1, \cdots, N_{n-t}) \in \mathbf{F}^{q\times q}.$$

(3) Let X be an $s \times q$ arbitrary matrix.

(4) Write

$$C_1 = \begin{bmatrix} C_{11} & X \\ 0 & C_{12} \end{bmatrix} \quad \text{and} \quad D_1 = \begin{bmatrix} D_{11} \\ 0 \end{bmatrix}.$$

Then, (C_1, D_1) is a pair with the same controllability indices and the same nontrivial invariant factors as $[L(\lambda), B]$. Furthermore, if $H_1 = [H_{11}\, 0] \in \mathbf{F}^{n\times np}$, then $H_1(\lambda I_{np} - C_1)^{-1}D_1$ is a state-space realization of $L(\lambda)^{-1}B$. It should be observed that this realization is such that (C_1, H_1) is not c.o. Indeed the assumption of complete observability in the previous statements was made to ensure that the controllable part of the system is also observable, but everything would be still right by removing the condition of observability from the uncontrollable part of the system.

We conclude with three final remarks:

(a) There are pairs (C_1, D_1) with the same nontrivial invariant factors and controllability indices as $[L(\lambda), B]$ such that C_1 is not a linearization of $L(\lambda)$.

(b) Conditions (i) and (ii) in Theorem 4 are equivalent to the existence of matrices $P_1 \in \mathbf{F}^{s\times s}$, $P_2 \in \mathbf{F}^{q\times q}$, $Q \in \mathbf{F}^{m\times m}$, and $R \in \mathbf{F}^{m\times q}$, P_1, P_2, and Q nonsingular, such that

$$\begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}\begin{bmatrix} C_{11} & C_{12} & D_{11} \\ 0 & C_{13} & 0 \end{bmatrix}\begin{bmatrix} P_1^{-1} & 0 & 0 \\ 0 & P_2^{-1} & 0 \\ 0 & R & Q \end{bmatrix} = \begin{bmatrix} E_1 & C_2 & D_{12} \\ 0 & C_3 & 0 \end{bmatrix}.$$

(c) If C_1 is an $np \times np$ matrix such that there is $D_1 \in \mathbf{F}^{np\times m}$ with (C_1, D_1) having the nontrivial invariant factors and controllability indices of $[L(\lambda), B]$, then C_1 is similar to a two block-triangular matrix (its Kalman form) with the diagonal blocks prescribed up to similarity. The characterization of the invariant factors of this type of matrices is a very interesting and important problem in matrix theory, and the connections of this problem with others in the different branches of mathematics is amazing. For an interesting exposition of such connections the reader is referred to [7].

REFERENCES

[1] W. A. Coppel, *Linear systems*; *some algebraic aspects*, Linear Algebra Appl., 40 (1981), pp. 257–273.

[2] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.

[3] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Invariant Subspaces with Applications*, John Wiley, New York, 1986.

[4] P. LANCASTER AND J. MAROULAS, *Selective perturbation of spectral properties of vibrating systems using feedback*, Linear Algebra Appl., 98 (1988), pp. 309–330.

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

[6] H. H. ROSENBROCK AND G. E. HAYTON, *The general problem of pole assignment*, Internat. J. Control, 7 (1978), pp. 837–852.

[7] R. C. THOMPSON, *Invariant Factors of Algebraic Combinations of Matrices*, in Frequency Domain and State Space Methods for Linear Systems, Elsevier Science Publishing Co., Inc., New York, 1986, pp. 73–87.

[8] I. ZABALLA, *Matrices with prescribed rows and invariant factors*, Linear Algebra Appl., 87 (1987), pp. 113–146.

[9] ———, *Interlacing inequalities and control theory*, Linear Algebra Appl., 101 (1988), pp. 9-31.

[10] ———, *Interlacing and majorization in invariant factor assignment problems*, Linear Algebra Appl., to appear.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 2, pp. 155-164, April 1989

003

LARGE GROWTH FACTORS IN GAUSSIAN ELIMINATION WITH PIVOTING*

NICHOLAS J. HIGHAM† AND DESMOND J. HIGHAM‡

Abstract. The growth factor plays an important role in the error analysis of Gaussian elimination. It is well known that when partial pivoting or complete pivoting is used the growth factor is usually small, but it can be large. The examples of large growth usually quoted involve contrived matrices that are unlikely to occur in practice. We present real and complex $n \times n$ matrices arising from practical applications that, for any pivoting strategy, yield growth factors bounded below by $n/2$ and n, respectively. These matrices enable us to improve the known lower bounds on the largest possible growth factor in the case of complete pivoting. For partial pivoting, we classify the set of real matrices for which the growth factor is 2^{n-1}. Finally, we show that large element growth does not necessarily lead to a large backward error in the solution of a particular linear system, and we comment on the practical implications of this result.

Key words. Gaussian elimination, growth factor, partial pivoting, complete pivoting, backward error analysis, stability

AMS(MOS) subject classifications. primary 65F05, 65G05

1. Introduction. In his famous backward error analysis, Wilkinson proved that if the linear system $Ax = b$, where A is $n \times n$, is solved in floating point arithmetic by Gaussian elimination with partial pivoting or complete pivoting, then the computed solution $\hat{x}$ satisfies (see, for example, [27, p. 108])

$$(A+E)\hat{x}=b, \tag{1.1a}$$

where

$$\|E\|_\infty \leqq \rho_n p(n) u \|A\|_\infty. \tag{1.1b}$$

Here, $p(n)$ is a cubic polynomial in n, u is the unit roundoff, and ρ_n is the *growth factor*, defined in terms of the quantities $a_{ij}^{(k)}$ occurring during the elimination by

$$\rho_n = \rho_n(A) = \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|}.$$

As Wilkinson notes, the term $p(n)$ arises from bounds in the analysis that are rarely attained, and for practical purposes we can replace $p(n)$ by n in (1.1b). Hence whether or not the bound in (1.1b) compares favourably with the "ideal" bound $\|E\|_\infty \leqq u\|A\|_\infty$ depends on the size of the growth factor.

Although the growth factor is one of the most well-known quantities in numerical analysis, its behaviour when pivoting is used is not completely understood. Current

* Received by the editors May 16, 1988; accepted for publication (in revised form) September 9, 1988. The work of the second author was supported by a Science and Engineering Research Council Research Studentship.

† Department of Mathematics, University of Manchester, Manchester M13 9PL, United Kingdom. Present address, Department of Computer Science, Cornell University, Ithaca, New York 14853 (na.nhigham@na-net.stanford.edu).

‡ Department of Mathematics, University of Manchester, Manchester M13 9PL, United Kingdom. Present address, Department of Computer Science, University of Toronto, Toronto, Canada M5S 1A4 (na.dhigham@na-net.stanford.edu).

knowledge, in the context of general, dense matrices, can be summarised as follows. For clarity we will denote the growth factors for partial and complete pivoting by ρ_n^p and ρ_n^c, respectively.

Partial pivoting. (The pivot element is selected as the element of largest absolute value in the active part of the pivot column.) The bound $\rho_n^p \leqq 2^{n-1}$ holds and is attained for matrices $A_n \in \mathbf{R}^{n \times n}$ of the following form [28, p. 212]:

$$(1.2) \qquad A_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix},$$

and also for $\tilde{A}_n = DA_nD$, $D = \operatorname{diag}(1, -1, 1, -1, \cdots, (-1)^{n+1})$ [26, p. 289]. Concerning the size of ρ_n^p in practice, Wilkinson [28, pp. 213–214] says: "It is our experience that any substantial increase in size of elements of successive A_r is extremely uncommon even with partial pivoting . . . No example which has arisen naturally has in my experience given an increase by a factor as large as 16." We are aware of no reports in the literature of experiences contrary to these related by Wilkinson over two decades ago. The largest growth factor that we have seen reported for a matrix not of the type (1.2) is $\rho_{100}^p = 35.1$, occurring for a symmetric matrix with elements from the uniform distribution on $[-1, 1]$ [18]; an earlier "record" value is $\rho_{40}^p = 23$, occurring for a random matrix of 1*s*, 0*s* and -1*s* [10, p. 1.21].

Complete pivoting. (The pivot element is selected as the element of largest absolute value in the whole of the remaining square submatrix.) Wilkinson [26, pp. 282–285] has shown that with complete pivoting

$$\rho_n^c \leqq n^{1/2}(2^1 3^{1/2} 4^{1/3} \cdots n^{1/(n-1)})^{1/2} \sim C n^{1/2} n^{1/4 \log n},$$

and that this bound is not attainable. He states in [26, p. 285] that "no matrix has been encountered in practice for which p_1/p_n was as large as 8," and in [28, p. 213] that "no matrix has yet been discovered for which $f(r) > r$." ($p_i = (n - i + 1)$st pivot, $f(r) \equiv \rho_r^c$.)

Cryer [7] defines

$$(1.3) \qquad g(n) = \sup_{A \in \mathbf{R}^{n \times n}} \rho_n^c(A).$$

The following results are known:

- $g(2) = 2$ (trivial).
- $g(3) = 2\frac{1}{4}$; Tornheim (see [7]) and Cohen [6].
- $g(4) = 4$; Cryer [7].
- $g(5) < 5.005$; Cohen [6].

Tornheim (see [7]) has shown that $\rho_n^c(H_n) \geqq n$ for any $n \times n$ Hadamard matrix H_n. H_n is a Hadamard matrix if each $h_{ij} \in \{-1, 1\}$ and the rows of H_n are mutually orthogonal. Hadamard matrices exist only for certain n; a necessary condition for their existence if $n > 2$ is that n is a multiple of four. For more about Hadamard matrices see [14, Chap. 14] and [25].

Cryer [7] conjectured that for real matrices $\rho_n^c(A) \leqq n$, with equality if and only if A is a Hadamard matrix. This conjecture is known to be false for complex matrices because Tornheim has constructed a 3×3 complex matrix A for which $\rho_3^c(A) > 3$ (see [7]).

As the summary above indicates, most of what is known about the growth factor had been discovered by the early 1970s. Recently, Trefethen [23] has drawn attention to the shortcomings of our knowledge about the growth factor and asked, as one of his three mysteries, "Why is the growth of elements during elimination [with partial pivoting] negligible in practice?" Trefethen and Schreiber [24] have proposed a statistical analysis to explain why the growth factor is usually small for partial pivoting.

In this work we take a different approach from that of Trefethen and Schreiber. Instead of trying to explain small growth we pursue examples of large growth, and we investigate the implications of a large growth factor for numerical stability.

In § 2 we present several families of real matrices for which ρ_n^c is bounded below by approximately $n/2$, and one family of complex matrices for which $\rho_n^c \geqq n$. Thus we obtain new lower bounds for $g(n)$ valid for all n. We also classify the real matrices for which $\rho_n^p = 2^{n-1}$, finding this to be a much richer class than might at first be thought.

In § 3 we reappraise the role of the growth factor in the backward error analysis of Gaussian elimination. We demonstrate that when solving linear systems by Gaussian elimination with partial pivoting large growth does not always induce a large backward error—there are certain, special right-hand sides for which the growth has no detrimental effect on the solution. We discuss the practical implications of this property for linear equation solvers.

2. Matrices with a large growth factor.

We begin with a result that shows how to obtain a lower bound for the growth factor in Gaussian elimination. The bound applies whatever strategy is used for interchanging rows and columns, but we will be concerned only with partial and complete pivoting.

THEOREM 2.1. *Let $A \in \mathbf{C}^{n \times n}$ be nonsingular, and set $\alpha = \max_{i,j} |a_{ij}|$, $\beta = \max_{i,j} |(A^{-1})_{ij}|$, and $\theta = (\alpha\beta)^{-1}$. Then $\theta \leqq n$, and for any permutation matrices P and Q such that PAQ has an* LU *factorisation, the growth factor ρ_n for Gaussian elimination without pivoting on PAQ satisfies $\rho_n \geqq \theta$.*

Proof. The inequality $\theta \leqq n$ follows from $\sum_{j=1}^{n} a_{ij}(A^{-1})_{ji} = 1$. Consider an LU factorisation $PAQ = LU$ computed by Gaussian elimination. We have

$$\begin{aligned} |u_{nn}^{-1}| &= |e_n^T U^{-1} e_n| = |e_n^T U^{-1} L^{-1} e_n| = |e_n^T Q^T A^{-1} P^T e_n| \\ &= |(A^{-1})_{ij}| \quad \text{for some } i,j \\ &\leqq \beta. \end{aligned}$$

Hence $\max_{i,j,k} |a_{ij}^{(k)}| \geqq |u_{nn}| \geqq \beta^{-1}$, and the result follows. □

Remarks. (1) $\theta^{-1} = \alpha\beta$ satisfies $\kappa_\infty(A)/n^2 \leqq \theta^{-1} \leqq \kappa_\infty(A)$, where the condition number $\kappa_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty$. Clearly, A has to be very well-conditioned for the theorem to provide a lower bound θ near the maximum of n.

(2) In the case of partial pivoting $Q = I$, and the proof of Theorem 2.1 shows that we can take $\beta = \max_j |(A^{-1})_{nj}|$, which leads to a lower bound θ potentially larger than the one in the theorem.

(3) The relation $u_{nn}^{-1} = (A^{-1})_{ij}$ is used also in [4], with the aim of investigating cases where u_{nn} is *small*.

To illustrate the theorem, consider a Hadamard matrix H_n. We have $H_n H_n^T = nI$, and so $H_n^{-1} = n^{-1} H_n^T$. Since $|h_{ij}| \equiv 1$, the theorem gives $\rho_n \geqq n$. As a special case we obtain $\rho_n^c(H_n) \geqq n$, as in [7] (this derivation is essentially the same as the one in [7]).

We present six further matrices to which the theorem can profitably be applied:

$$C_1 = \left(\cos\left(\frac{(i-1)(j-1)\pi}{n-1}\right)\right)_{i,j=1}^{n}, \qquad \rho_n \geqq \frac{n-1}{2}, \tag{2.1}$$

$$C_2 = \left(\cos\left(\frac{(i-1)(j-\frac{1}{2})\pi}{n}\right)\right)_{i,j=1}^{n}, \qquad \rho_n \geqq \frac{n}{2}, \tag{2.2}$$

$$S = \sqrt{\frac{2}{n+1}}\left(\sin\left(\frac{ij\pi}{n+1}\right)\right)_{i,j=1}^{n}, \qquad \rho_n \geqq \frac{n+1}{2}, \tag{2.3}$$

$$Q = \frac{2}{\sqrt{2n+1}}\left(\sin\left(\frac{2ij\pi}{2n+1}\right)\right)_{i,j=1}^{n}, \qquad \rho_n \geqq \frac{2n+1}{4}, \tag{2.4}$$

$$F = (f_{ij})_{i,j=1}^{n=2m},$$
$$f_{ij} = \begin{cases} \cos\left(\dfrac{(i-1)(j-1)\pi}{m}\right), & 1 \leqq i \leqq m+1 \\[2ex] \sin\left(\dfrac{(i-m-1)(j-1)\pi}{m}\right), & m+2 \leqq i \leqq n \end{cases}, \qquad \rho_n \geqq \frac{n}{2}, \tag{2.5}$$

$$V = \left(\exp\left(2\pi i(r-1)\frac{(s-1)}{n}\right)\right)_{r,s=1}^{n}, \qquad \rho_n \geqq n. \tag{2.6}$$

C_1 and C_2 are examples of Vandermonde-like matrices $C(\alpha_1, \alpha_2, \cdots, \alpha_n) = (T_{i-1}(\alpha_j))$ based on the Chebyshev polynomials T_k. (For further details of Vandermonde-like matrices and their applications see [16].) For C_1 the points $\alpha_j = \cos((j-1)\pi/(n-1))$ are the extrema of T_{n-1}, and for C_2 the points $\alpha_j = \cos((j-\frac{1}{2})\pi/n)$ are the zeros of T_n. The Chebyshev polynomials satisfy orthogonality conditions over both these sets of points [15, pp. 472–473]. Using these orthogonality properties, we can show that

$$C_1 D C_1 = \frac{(n-1)}{2} D^{-1}, \qquad D = \operatorname{diag}(\tfrac{1}{2}, 1, 1, \cdots, 1, \tfrac{1}{2}),$$

and

$$C_2 C_2^T = n \operatorname{diag}(1, \tfrac{1}{2}, \tfrac{1}{2}, \cdots, \tfrac{1}{2}).$$

Hence $C_1^{-1} = (2/(n-1)) D C_1 D$, and Theorem 2.1 yields $\rho_n(C_1) \geqq (n-1)/2$. It is not hard to show that for partial pivoting $u_{nn} = n-1$, and so $\rho_n^p(C_1) \geqq n-1$. Similarly, $C_2^{-1} = n^{-1} C_2^T \operatorname{diag}(1, 2, 2, \cdots, 2)$, and Theorem 2.1 gives $\rho_n(C_2) \geqq n/2$.

S is the symmetric, orthogonal eigenvector matrix for the second difference matrix (the tridiagonal matrix with typical row $(-1, 2, -1)$) [22, p. 457]. Theorem 2.1 gives $\rho_n(S) \geqq (n+1)/2$. Another application in which S and C_2 appear is the analysis of time series [1, § 6.5].

Q is symmetric and orthogonal [19] and Theorem 2.1 yields $\rho_n(Q) \geqq (2n+1)/4$.

The matrix F, of even order $n = 2m$, arises in the derivation of approximations to linear operators for periodic functions. Hamming [15, pp. 522–524] shows that

$$F^{-1} = \frac{2}{n} F^T \operatorname{diag}(d_i), \qquad d_i = \begin{cases} \frac{1}{2}, & i = 1, m, \\ 1 & \text{otherwise.} \end{cases}$$

Hence Theorem 2.1 yields $\rho_n(F) \geqq n/2$.

Finally, V is a complex Vandermonde matrix based on the roots of unity. It occurs in Fast Fourier Transform theory [22, pp. 292, 448]. $V^HV = nI$, so $V^{-1} = n^{-1}V^H$ and Theorem 2.1 gives $\rho_n(V) \geqq n$.

These matrices are not isolated examples: for each, the lower bound θ for ρ_n is insensitive to small perturbations of the matrix. To see this, note that for $\|A^{-1}E\|_\infty < 1$ (say), in the notation of Theorem 2.1,

$$\begin{aligned}\theta(A+E)^{-1} &= \alpha(A+E)\beta(A+E) \leqq (\alpha(A)+\alpha(E))(\beta(A)+O(\|E\|_\infty)) \\ &= \theta(A)^{-1}(1+O(\|E\|_\infty)).\end{aligned}$$

Regarding the perturbation E as the backward error in a computed LU factorisation, it follows also that, as long as E is not too large, the computed growth factors will satisfy the theoretical lower bounds to within roundoff.

It is natural to ask what are the actual growth factors ρ_n^p and ρ_n^c for the matrices above. In numerical tests we found ρ_n^p and ρ_n^c generally to be bigger than the lower bounds, but appreciably less than n, except in the case of V in (2.6) for which numerical evidence suggests that $\rho_n^p(V) = \rho_n^c(V) = n$.

All the above matrices are natural, noncontrived ones that arise in practical applications. For $n = 50$ (say), for both partial and complete pivoting, each of the matrices produces growth factors which exceed the generally accepted "maximum values in practice", such as the value 16 mentioned by Wilkinson in [28]. It is rather surprising that the growth factor properties of these examples have not previously been recognised. One possible explanation is that since each of the matrices is either an orthogonal or a diagonal scaling of an orthogonal matrix, Gaussian elimination may rarely have been applied to these matrices. (The growth factor properties of C_1 and C_2 were discovered incidentally when making a numerical comparison between Gaussian elimination with partial pivoting and a fast $O(n^2)$ algorithm [16].)

These examples provide new lower bounds for the maximum growth factor with complete pivoting. Specifically, we have, for $g(n)$ in (1.3),

$$g(n) \geqq \rho_n^c(S) \geqq \frac{n+1}{2} \quad \text{for all } n.$$

N. I. M. Gould (private communication) has suggested a way to obtain slightly sharper bounds: it is easy to show that

$$\theta(B) = \theta\left(\begin{bmatrix} A & A \\ A & -A \end{bmatrix}\right) = 2\theta(A),$$

and so, taking $A = S$, $g(2n) \geqq \rho_{2n}^c(B) \geqq \theta(B) = 2\theta(S) = n + 1$, which improves on the lower bound $(2n + 1)/2$. (Of course, for n such that a Hadamard matrix H_n exists, $g(n) \geqq n$ is a better bound; and for $n \leqq 5$ see the results quoted in § 1.) Furthermore, defining

$$\bar{g}(n) = \sup_{A \in \mathbf{C}^{n \times n}} \rho_n^c(A),$$

we have

$$\bar{g}(n) \geqq \rho_n^c(V) \geqq n.$$

The growth factors discussed above are relatively mild in the context of partial pivoting, since $O(n)$ growth falls significantly short of the potential $O(2^n)$. To investigate larger growth factors we have to make specific use of the properties of partial pivoting.

The following result shows that Wilkinson's example in which $\rho_n^p = 2^{n-1}$ is attained is just one from a nontrivial class of matrices with this property.

THEOREM 2.2. *All real $n \times n$ matrices A for which $\rho_n^p(A) = 2^{n-1}$ are of the form*

$$A = DM\begin{bmatrix} T & \vdots & \theta d \\ 0 & \vdots & \end{bmatrix},$$

where $D = \operatorname{diag}(\pm 1)$, M is unit lower triangular with $m_{ij} = -1$ for $i > j$, T is a nonsingular upper triangular matrix of order $n-1$, $d = (1, 2, 4, \cdots, 2^{n-1})^T$, and θ is a scalar such that $\theta = |a_{1n}| = \max_{i,j} |a_{ij}|$.

Proof. Gaussian elimination with partial pivoting applied to a matrix A gives a factorisation $B := PA = LU$, where P is a permutation matrix. It is easy to show that $|u_{ij}| \leqq 2^{i-1} \max_{r \leq i} |b_{rj}|$, with equality for $i = s$ only if there is equality for $i = 1, 2, \cdots, s-1$. Thus $\rho_n = 2^{n-1}$ implies that the last column of U has the form θDd, and also that $|b_{1n}| = \max_{i,j} |b_{ij}|$. By considering the final column of B, and imposing the requirement that $|l_{ij}| \leqq 1$, it is easy to show that the unit lower triangular matrix L must have the form $L = DMD$. It follows that at each stage of the reduction every multiplier is ± 1; hence no interchanges are performed, that is, $P = I$. The only requirement on T is that it be nonsingular, for if $t_{ii} = 0$ then the ith elimination stage would be skipped because of a zero pivot column, and no growth would be produced on that stage. □

In the case $n = 5$, the general form of A is

$$A = D\begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} & \theta \\ -t_{11} & -t_{12}+t_{22} & -t_{13}+t_{23} & -t_{14}+t_{24} & \theta \\ -t_{11} & -t_{12}-t_{22} & -t_{13}-t_{23}+t_{33} & -t_{14}-t_{24}+t_{34} & \theta \\ -t_{11} & -t_{12}-t_{22} & -t_{13}-t_{23}-t_{33} & -t_{14}-t_{24}-t_{34}+t_{44} & \theta \\ -t_{11} & -t_{12}-t_{22} & -t_{13}-t_{23}-t_{33} & -t_{14}-t_{24}-t_{34}-t_{44} & \theta \end{bmatrix}.$$

We mention that it is straightforward to extend Theorem 2.2 to complex matrices.

As well as being of theoretical interest, the matrices given in this section are useful test matrices for linear equation solvers. Note that $\kappa_\infty(A)$ can be bounded above and below by multiples of $\kappa_\infty(T)$, so T can be used to vary the condition of A. By varying the elements m_{ij} $(i > j)$ and the vector d in Theorem 2.2 we can construct matrices for which ρ_n^p achieves any desired value between 1 and 2^{n-1}. Indeed in practice it is expedient to modify M in Theorem 2.2 so that $|m_{ij}| < 1$ for $i > j$, to ensure that rounding errors do not affect the pivot sequence (and hence the computed growth factor).

3. Implications of a large growth factor. If the growth factor ρ_n is large then in the backward error result (1.1) the bound for $\|E\|_\infty$ is large. Whether or not $\|E\|_\infty$ itself is large when ρ_n is large depends on the sharpness of the bound. Since the bound is independent of b, and E clearly is not, we might suspect that the bound can be weak; in this section we will show that this is indeed the case.

We need to make use of an elementwise form of backward error analysis. Let $A \in \mathrm{R}^{n \times n}$. From [8] the computed solution $\hat{x}$ from Gaussian elimination (assuming, without loss of generality, no interchanges) satisfies

$$(A+F)\hat{x} = b, \tag{3.1a}$$

where

$$|F| \leqq \gamma(2+\gamma)|\hat{L}|\,|\hat{U}|, \qquad \gamma = nu/(1-nu), \tag{3.1b}$$

and where $A \approx \hat{L}\hat{U}$ is the computed LU factorisation and $|F| = (|f_{ij}|)$.

As our use of the notation E in (1.1) and F in (3.1) suggests, the backward error for solution of $Ax = b$ is not uniquely defined: G satisfying $(A + G)\hat{x} = b$ can be replaced by $G + H$ for any H whose rows are orthogonal to $\hat{x}$. However, it is well known that of the infinitely many backward error matrices there is a unique one of minimal Frobenius norm,

$$(3.2) \qquad G = \frac{r\hat{x}^T}{\hat{x}^T\hat{x}}, \qquad \|G\|_F = \frac{\|r\|_2}{\|\hat{x}\|_2},$$

where the residual $r = b - A\hat{x}$, and $\|G\|_F = (\sum_{i,j} g_{ij}^2)^{1/2}$ (see [9, p. 171] for a proof and discussion, albeit in a different context). Of course, for the minimal Frobenius norm backward error matrix G to be an appropriate one to consider, A should be reasonably well-scaled.

Our aim is to obtain an informative bound for the minimal backward error $|G|$. To do this we write $r = b - A\hat{x} = F\hat{x}$, from (3.1a), and invoke the bound (3.1b), obtaining

$$|r| \leqq |F|\,|\hat{x}| \leqq \gamma(2+\gamma)|\hat{L}|\,|\hat{U}|\,|\hat{x}|.$$

Hence

$$(3.3) \qquad |G| = \frac{|r|\,|\hat{x}|^T}{\|\hat{x}\|_2^2} \leqq \frac{\gamma(2+\gamma)}{\|\hat{x}\|_2^2}|\hat{L}|\,|\hat{U}|\,|\hat{x}|\,|\hat{x}|^T.$$

Our observation is that any large growth, which necessarily takes the form of large elements of $\hat{U}$ when partial pivoting is used, will not fully affect the backward error for a particular $\hat{x}$ if

$$\|\,|\hat{U}|\,|\hat{x}|\,\|_\infty \ll \|\hat{U}\|_\infty \|\hat{x}\|_\infty.$$

Since $|\hat{u}_{ij}| \leqq 2^{i-1}\max_{r \leqq i}|a_{rj}|$, large growth can occur only toward the (n, n) position of $\hat{U}$; consequently any $\hat{x}$ bounded by (say)

$$|\hat{x}| \leqq \|\hat{x}\|_\infty(1, 2^{-1}, 2^{-2}, \cdots, 2^{1-n})^T$$

can be shown to satisfy $\|\,|\hat{U}|\,|\hat{x}|\,\|_\infty \leqq 2\|A\|_\infty\|\hat{x}\|_\infty$, no matter how large $\|\hat{U}\|_\infty$.

For example, for any A, consider the use of partial pivoting for the particular system $Ax = b$ with $x = e_1$. Assume $\delta x = x - \hat{x}$ satisfies $\|\delta x\|_\infty \leqq 2^{1-n}/n$; this will certainly be the case if, making use of (1.1), $\kappa_\infty(A)4^{n-1}p(n)nu < 1$. Then

$$|\hat{L}|\,|\hat{U}|\,|\hat{x}| = |\hat{L}|\,|\hat{U}|\,|e_1 - \delta x| \leqq \max_r |a_{r1}|e + |\hat{L}|\,|\hat{U}|\,|\delta x|,$$

where $e = (1, 1, \cdots, 1)^T$, and thus

$$\|\,|\hat{L}|\,|\hat{U}|\,|\hat{x}|\,\|_\infty \leqq \|A\|_\infty + n2^{n-1}\|A\|_\infty\|\delta x\|_\infty \leqq 2\|A\|_\infty.$$

Hence, using (3.3), we have

$$\|G\|_\infty \leqq 2\gamma(2+\gamma)\|A\|_\infty(1 + O(2^{-n})),$$

which is an ideal backward error result, containing no growth factor term.

To illustrate the analysis we describe some numerical experiments performed using Gaussian elimination with partial pivoting and the perturbation $B_n = A_n + 0.1\,e_n e_n^T$ of Wilkinson's extreme growth matrix A_n in (1.2). This perturbation of the (n, n) element has the effect of causing rounding errors to be committed in the computation of the LU factorisation. Note that element growth occurs only in the *last* column of B_n during Gaussian elimination with partial pivoting. For several n we solved five different linear systems $B_n x = b$, and computed the backward error for the LU factorisation, $\|B_n - \hat{L}\hat{U}\|_F / \|B_n\|_F$ (note that this is unique for a given norm), and the minimal backward error in the Frobenius norm for each system solved, $\|r\|_2/(\|B_n\|_F\|\hat{x}\|_2)$. For four of the linear systems, we selected x or b as vectors suggested by the analysis; for the final system we used a random b with elements from the uniform distribution on $[0, 1]$.

The computations were performed using the WATFOR-77 Fortran 77 compiler on a PC-AT compatible machine. Solutions were computed in single precision (IEEE standard, $u \approx 1.19 \times 10^{-7}$), using LINPACK's SGEFA/SGESL. The residuals r and $B_n - \hat{L}\hat{U}$ were computed in double precision. The results are displayed in Table 3.1.

The backward errors for the LU factorisation are seen to be somewhat smaller than the large growth factor might lead us to expect, though still "alarmingly" large, except for $n = 10$. For $x = e_1$ the backward errors are all identically zero and $\hat{x} = x$; in this example the errors in the LU factorisation are nullified in the substitutions. The backward errors are also perfectly acceptable for $b = e_n$. Here the explanation is that $x_n = (B_n^{-1})_{nn} = u_{nn}^{-1}$, so that $u_{nn}x_n = 1$; thus the large elements in the last column of $\hat{U}$ vanish in the product $|\hat{U}|\,|\hat{x}|$ in (3.3). The backward errors for $b = e_1$, $b = e$, and the random b, all reflect the large backward error in the LU factorisation, as we would expect: the nonnegligible x_n components pick out the large last column of $\hat{U}$ in the product $|\hat{U}|\,|\hat{x}|$.

To summarise, we have shown the following: When a linear system $Ax = b$ is solved by Gaussian elimination with partial pivoting, the backward error for the computed solution $\hat{x}$, $\|b - A\hat{x}\|_2/(\|A\|_F\|\hat{x}\|_2)$, can, in certain special cases, be substantially smaller than the backward error for the LU factorisation, $\|A - \hat{L}\hat{U}\|_F/\|A\|_F$, if the latter is large. Thus, strictly, the growth factor, or any other quantity appearing in a measure or bound of $A - \hat{L}\hat{U}$, is an unreliable indicator of the stability of a particular solution $\hat{x}$. We do not claim that this result is new, nor do we think that it will surprise anyone who has worked in backward error analysis. Examples of references that allude to the result in some way are [11, p. 73] and [20]. However we are not aware of a published analysis like the one above, and we feel that the result deserves to be better known.

It is important to stress that large growth is indeed very uncommon with partial pivoting (see the quotation from [28] in § 1), and that when it does occur there is a high probability that it will adversely affect the stability of the computed solution $\hat{x}$. Nevertheless, the result above has implications for how one uses a linear equation solver.

For example, consider the use of threshold versions of partial pivoting (including no pivoting at all); here large growth factors are much more common, and it is standard practice to monitor stability by estimating the error in the factorisation, $A - \hat{L}\hat{U}$ [5], [11]–[13]. If the estimate is large then a popular course of action is to carry out a

TABLE 3.1
Results.
($u \approx 1.19 \times 10^{-7}$)

n	ρ_n^p	$\dfrac{\|B_n - \hat{L}\hat{U}\|_F}{\|B_n\|_F}$	$\|r\|_2/(\|B_n\|_F\|\hat{x}\|_2)$				
			$z = e_1$	$b = e_n$	$b = e_1$	$b = e$	b random
10	4.7E2	3.0E-6	0.0	1.5E-8	3.1E-6	4.4E-6	2.3E-6
20	4.8E5	1.7E-3	0.0	5.5E-9	1.2E-3	1.6E-3	2.4E-4
30	4.9E8	4.5E-3	0.0	1.5E-11	3.2E-3	4.5E-3	1.1E-1
40	5.0E11	3.4E-3	0.0	1.1E-14	2.4E-3	3.4E-3	3.7E-2
50	5.1E14	2.7E-3	0.0	1.1E-17	1.9E-3	2.7E-3	4.4E-2
60	5.2E17	5.7E-2	0.0	1.5E-19	1.6E-3	2.3E-3	8.9E-2

refactorisation with a different pivot sequence. Our view is that if just a *single* system involving A must be solved, it is worthwhile to proceed with the substitutions and to base refactorisation decisions on the easily computed *actual* backward error (3.2) rather than on (estimates of) $A - \hat{L}\hat{U}$, which may be misleading, as we have shown. For example, having computed $\hat{x}$ we might form $r = b - A\hat{x}$ (in single precision), evaluate the backward error $\|G\|_F = \|r\|_2/\|\hat{x}\|_2$, and test whether $\|G\|_F \leqq \delta\|A\|_F$, where δ is an appropriate tolerance (depending on the unit roundoff, at least). Even if $\hat{x}$ is unacceptable, the substitutions need not have been wasted, for we may be able to achieve stability through the use of a few steps of iterative refinement [2], [3], [17], [21].

A more general way to express these views is that it is better to use *a posteriori* estimates that reflect the actual rounding errors encountered, rather than error estimates based on *a priori* analysis, such as (1.1). For a discussion of this philosophy we can do no better than refer the reader to Wilkinson's eloquent exposition in [29].

REFERENCES

[1] T. W. ANDERSON, *The Statistical Analysis of Time Series*, John Wiley, New York, 1971.

[2] M. ARIOLI, J. W. DEMMEL, AND I. S. DUFF, *Solving sparse linear systems with sparse backward error*, Report CSS 214, Computer Science and Systems Division, Harwell Laboratory, AERE Harwell, Didcot, UK, 1988.

[3] Å. BJÖRCK, *Iterative refinement and reliable computing*, in Reliable Numerical Computation, M. G. Cox and S. J. Hammarling, eds., Oxford University Press, London, 1989.

[4] T. F. CHAN, *On the existence and computation of* LU-*factorizations with small pivots*, Math. Comp., 42 (1985), pp. 535–547.

[5] E. CHU AND A. GEORGE, *A note on estimating the error in Gaussian elimination without pivoting*, ACM SIGNUM Newsletter, 20 (1985), pp. 2–7.

[6] A. M. COHEN, *A note on pivot size in Gaussian elimination*, Linear Algebra Appl., 8 (1974), pp. 361–368.

[7] C. W. CRYER, *Pivot size in Gaussian elimination*, Numer. Math., 12 (1968), pp. 335–345.

[8] C. DE BOOR AND A. PINKUS, *Backward error analysis for totally positive linear systems*, Numer. Math., 27 (1977), pp. 485–490.

[9] J. E. DENNIS, JR. AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[10] 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.

[11] I. S. DUFF, A. M. ERISMAN, AND J. K. REID, *Direct Methods for Sparse Matrices*, Oxford University Press, London, 1986.

[12] A. M. ERISMAN, R. G. GRIMES, J. G. LEWIS, W. G. POOLE, AND H. D. SIMON, *Evaluation of orderings for unsymmetric sparse matrices*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 600–624.

[13] A. M. ERISMAN AND J. K. REID, *Monitoring the stability of the triangular factorization of a sparse matrix*, Numer. Math., 22 (1974), pp. 183–186.

[14] M. HALL, JR., *Combinatorial Theory*, Blaisdell, Waltham, MA, 1967.

[15] R. W. HAMMING, *Numerical Methods for Scientists and Engineers*, 2nd ed., McGraw-Hill, New York, 1973.

[16] N. J. HIGHAM, *Stability analysis of algorithms for solving confluent Vandermonde-like systems*, Numerical Analysis Report 148, University of Manchester, Manchester, UK, 1987.

[17] M. JANKOWSKI AND H. WOŹNIAKOWSKI, *Iterative refinement implies numerical stability*, BIT, 17 (1977), pp. 303–311.

[18] A. J. MACLEOD, *The distribution of the growth factor in Gaussian elimination with partial pivoting*, Technical Report, Department of Mathematics and Statistics, Paisley College of Technology, Paisley, Scotland, 1988.

[19] R. B. POTTS, *Symmetric square roots of the finite identity matrix*, Utilitas Math., 9 (1976), pp. 73–86.

[20] R. D. SKEEL, *Scaling for numerical stability in Gaussian elimination*, J. Assoc. Comput. Mach., 26 (1979), pp. 494–526.

[21] ———, *Iterative refinement implies numerical stability for Gaussian elimination*, Math. Comp., 35 (1980), pp. 817–832.

[22] G. STRANG, *Introduction to Applied Mathematics*, Wellesley-Cambridge Press, Wellesley, MA, 1986.

[23] L. N. TREFETHEN, *Three mysteries of Gaussian elimination*, ACM SIGNUM Newsletter, 20 (1985), pp. 2–5.

[24] L. N. TREFETHEN AND R. S. SCHREIBER, *Average-case stability of Gaussian elimination*, Numerical Analysis Report 88-3, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, 1987; SIAM J. Matrix Anal. Appl., to appear.

[25] W. D. WALLIS, A. P. STREET, AND J. S. WALLIS, *Combinatorics: Room Squares, Sum-Free Sets, Hadamard Matrices*, Lecture Notes in Mathematics 292, Springer-Verlag, Berlin, 1972.

[26] J. H. WILKINSON, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8 (1961), pp. 281–330.

[27] ———, *Rounding Errors in Algebraic Processes*, Notes on Applied Science No. 32, Her Majesty's Stationery Office, London, 1963.

[28] ———, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

[29] ———, *Error analysis revisited*, IMA Bulletin, 22 (1987), pp. 192–200.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 2, pp. 165–190, April 1989

004

SOLVING SPARSE LINEAR SYSTEMS WITH SPARSE BACKWARD ERROR*

M. ARIOLI†, J. W. DEMMEL‡, AND I. S. DUFF§

Abstract. When solving sparse linear systems, it is desirable to produce the solution of a nearby sparse problem with the same sparsity structure. This kind of backward stability helps guarantee, for example, that a problem with the same physical connectivity as the original has been solved. Theorems of Oettli, Prager [*Numer Math.*, 6 (1964), pp. 405–409] and Skeel [*Math. Comput.*, 35 (1980), pp. 817–832] show that one step of iterative refinement, even with single precision accumulation of residuals, guarantees such a small backward error if the final matrix is not too ill-conditioned and the solution components do not vary too much in magnitude. These results are incorporated into the stopping criterion of the iterative refinement step of a direct sparse matrix solver, and numerical experiments verify that the algorithm frequently stops after one step of iterative refinement with a componentwise relative backward error at the level of the machine precision. Furthermore, calculating this stopping criterion is very inexpensive. A condition estimator corresponding to this new backward error is discussed that provides an error estimate for the computed solution. This error estimate is generally tighter than estimates provided by standard condition estimators. We also consider the effects of using a drop tolerance during the LU decomposition.

Key words. sparse matrix, backward error, iterative refinement, componentwise error, error estimate, condition number

AMS(MOS) subject classifications. 65F05, 65G05, 65F35

1. Introduction. When solving systems of n linear equations $\mathbf{Ax} = \mathbf{b}$ by means of Gaussian elimination with pivoting, a classical analysis (Wilkinson (1961)) shows that we should expect to get the exact solution $\hat{\mathbf{x}}$ of a slightly different linear system $(\mathbf{A} + \delta\mathbf{A})\hat{\mathbf{x}} = \mathbf{b} + \delta\mathbf{b}$ where $\delta\mathbf{A}$ and $\delta\mathbf{b}$ are both small with respect to $\mathbf{A}$ and $\mathbf{b}$. By small we mean small in norm, i.e., $\|\delta\mathbf{A}\| \leqq k\varepsilon\|\mathbf{A}\|$ and $\|\delta\mathbf{b}\| \leqq k\varepsilon\|\mathbf{b}\|$ where $\|\cdot\|$ is a matrix norm, ε is the machine precision (that is, the greatest positive number such that $fl(1+\varepsilon)$, the floating-point representation of $(1+\varepsilon)$, equals one), and k is the product of the pivot growth factor and a modestly growing function of the dimension n. This classical view permits any entry of $\delta\mathbf{A}$ or $\delta\mathbf{b}$ to be equally large, and in particular $\mathbf{A} + \delta\mathbf{A}$ may be dense even if $\mathbf{A}$ is quite sparse. This is unsatisfactory because zero entries of $\mathbf{A}$ may represent nonexistent physical connections in a system being modeled, and so may be known exactly.

A more satisfying approach to backward error than merely bounding $\|\delta\mathbf{A}\|$ and $\|\delta\mathbf{b}\|$ would permit the user to specify scaling factors $e_{ij} \geqq 0$ and $f_i \geqq 0$ for each entry of $\delta\mathbf{A}$ and $\delta\mathbf{b}$, and would compute the smallest $\omega \geqq 0$ such that

$$(1) \qquad |\delta a_{ij}| \leqq \omega e_{ij}, \qquad |\delta b_i| \leqq \omega f_i.$$

By setting some e_{ij} to zero, we can insist that, if $\omega < \infty$, the corresponding a_{ij} are known exactly. For example, if $e_{ij} = |a_{ij}|$ and $f_i = |b_i|$, ω bounds the relative perturbation in each component of $\mathbf{A}$ and $\mathbf{b}$ needed to make $\hat{\mathbf{x}}$ an exact solution, and, in particular, $\delta\mathbf{A}$

* Received by the editors March 23, 1988; accepted for publication (in revised form) September 9, 1988.

† Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique (CERFACS), 42 Ave. G. Coriolis, 31057 Toulouse, France (na.arioli@na-net.stanford.edu). This author was a visiting researcher at Harwell Laboratory and was funded by a grant from the Italian National Council of Research (CNR), Pisa, Italy.

‡ Computer Science Department, Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, New York 10012 (na.demmel@na-net.stanford.edu).

§ Computer Science and Systems Division, Harwell Laboratory, United Kingdom Atomic Energy Authority, Oxfordshire OX11 0RA, United Kingdom (duff@mcs.anl.gov and na.duff.na-net.stanford.edu).

and $\delta\mathbf{b}$ have the same sparsity structures as $\mathbf{A}$ and $\mathbf{b}$. We will call this ω the *componentwise relative backward error*. It is important to use this different error estimate when considering these restricted perturbations, since Gear (1975) has shown that the conventional error bounds are not appropriate in this case. It turns out that the backward error ω is quite easy to compute, and in fact costs as little as two matrix-vector multiplications.

In the following, if $\mathbf{u}$ and $\mathbf{v}$ are vectors of entries u_i and v_i and $\mathbf{Q}$ and $\mathbf{P}$ are matrices of entries q_{ij} and p_{ij}, $|\mathbf{u}|$ is the vector of entries $|u_i|$, $|\mathbf{Q}|$ is the matrix of entries $|q_{ij}|$. $\mathbf{u} \leqq \mathbf{v}$ means $u_i \leqq v_i$ for all i, and $\mathbf{Q} \leqq \mathbf{P}$ means $q_{ij} \leqq p_{ij}$ for all i and j.

THEOREM 1 (Oettli and Prager (1964)). *The smallest ω satisfying* (1) *is given by*

$$\omega = \max_i \frac{|\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}|_i}{(\mathbf{E}|\hat{\mathbf{x}}| + \mathbf{f})_i}. \tag{2}$$

In this expression, $0/0$ *should be interpreted as* 0 *and* $\zeta/0$ ($\zeta \neq 0$) *as infinity*. $\omega = \infty$ *implies that no ω satisfying* (1) *exists*. *In particular, the smallest componentwise relative perturbation of* $\mathbf{A}$ *and* $\mathbf{b}$ *that makes* $\hat{\mathbf{x}}$ *an exact solution is*

$$\omega = \max_i \frac{|\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}|_i}{(|\mathbf{A}|\,|\hat{\mathbf{x}}| + |\mathbf{b}|)_i}. \tag{3}$$

Thus, this theorem gives an a posteriori measure of the backward error that is cheap to compute.

Gaussian elimination with pivoting does not guarantee that the backward error ω will be small for all possible $\mathbf{E}$ and $\mathbf{f}$. However, a theorem of Skeel (1980) shows that as long as $\mathbf{A}$ is not too ill-conditioned, and as long as the quantities $(|\mathbf{A}|\,|\hat{\mathbf{x}}|)_i$ in the denominator of (3) do not vary too much in magnitude, then one step of iterative refinement is enough to guarantee that ω will be small for the componentwise relative backward error in (3). This is true even if the residual $\mathbf{r} = \mathbf{A}\hat{\mathbf{x}} - \mathbf{b}$ is computed in the same arithmetic precision as used for the Gaussian elimination. The actual conditions under which the following theorem is true are quite complicated, and we refer for details to Skeel (1980, Thm. 5.1).

THEOREM 2 (Skeel (1980)). *Let ε be the machine precision, and let the arithmetic be such that the floating-point result* fl $(a \Diamond b)$ *of the operation* $a \Diamond b$, ($\Diamond \in \{+, -, \times, /\}$) *satisfies* fl $(a \Diamond b) = (a \Diamond b)(1 + e)$, *with* $|e| \leqq \varepsilon$. *There is a function* $f(\mathbf{A}, \mathbf{b})$, *typically behaving as* $O(n)$, *such that when the product of* $\hat{\kappa}(\mathbf{A}) \equiv \|\,|\mathbf{A}|\,|\mathbf{A}^{-1}|\,\|$ *and* $\sigma(\mathbf{A}, \mathbf{x}) \equiv \max_i (|\mathbf{A}|\,|\mathbf{x}|)_i / \min_i (|\mathbf{A}|\,|\mathbf{x}|)_i$ *is less than* $(f(\mathbf{A}, \mathbf{b})\varepsilon)^{-1}$, *and there is no overflow or underflow, the following iterative refinement algorithm will converge after one update of* $\hat{\mathbf{x}}$:

```
Solve $\mathbf{Ax} = \mathbf{b}$ using Gaussian elimination, obtaining solution $\hat{\mathbf{x}}$ and saving the LU factors;
Compute the residual $\mathbf{r} = \mathbf{A}\hat{\mathbf{x}} - \mathbf{b}$ (using arithmetic of machine precision $\varepsilon$);
while  $\omega = \max_i |r_i| / (|\mathbf{A}|\,|\hat{\mathbf{x}}| + |\mathbf{b}|)_i > (n + 1)\varepsilon$  do
begin
    Solve $\mathbf{Ad} = \mathbf{r}$ for $\mathbf{d}$ using the saved LU factors of $\mathbf{A}$;
    Update $\hat{\mathbf{x}} = \hat{\mathbf{x}} - \mathbf{d}$;
    Compute the residual $\mathbf{r} = \mathbf{A}\hat{\mathbf{x}} - \mathbf{b}$ (using arithmetic of machine precision $\varepsilon$);
end;
```

This theorem may also be extended to take into account underflow and the possibility that, for lack of a guard digit in the hardware, we can only assert that

$$\mathrm{fl}\,(a \pm b) = a(1 + e_1) \pm b(1 + e_2),$$

where $|e_i| \leqq \varepsilon$, (Demmel (1984)).

For sparse systems, it is also possible to improve the stopping criterion of Theorem 2 by changing n to γ, the maximum number of nonzero entries in one row of $\mathbf{A}$.

Note that this theorem contradicts the usual advice that iterative refinement is not worth doing unless the residual $\mathbf{r} = \mathbf{A}\hat{\mathbf{x}} - \mathbf{b}$ is computed using arithmetic of machine precision ε^2. Note also that the theorem does not say that the refined solution will be more accurate, just that it reflects the structure of the original problem more closely than the unrefined solution. If each of the nonzero entries of the original $\mathbf{A}$ is uncertain in its least significant bit and if $\omega \approx \varepsilon$, then we could say that we have computed the solution as accurately as the data warrants, since the answer is exact for a problem indistinguishable from the problem we really wanted to solve.

To use Theorem 2 as the basis of a practical scheme for solving sparse linear systems, some modifications are necessary. In particular, when solving sparse linear systems where both $\mathbf{A}$ and $\mathbf{b}$ are sparse (or $\mathbf{b}$ has components of widely varying magnitude), it often happens that the quantity $\sigma(\mathbf{A}, \mathbf{x})$ in Theorem 2 is huge, and convergence does not occur. Therefore, since our main goal is to guarantee sparsity in $\mathbf{A}$, we must make another choice for $\mathbf{f}$, taking less account of the smaller components b_i. This can be done quite easily using a modification of Theorem 1, and is discussed in § 2.2.

There is a new condition number corresponding to the new definition of backward error in (1). In the case of $\mathbf{E} = |\mathbf{A}|$ and $\mathbf{f} = |\mathbf{b}|$, this condition number is just $\| |\mathbf{A}^{-1}| |\mathbf{A}| \|$. This new condition number is no larger than the traditional condition number $\|\mathbf{A}^{-1}\| \|\mathbf{A}\|$. In fact, it may be much smaller than $\|\mathbf{A}^{-1}\| \|\mathbf{A}\|$ if the rows of $\mathbf{A}$ are badly scaled. Thus, combining the componentwise relative backward error with the new condition number, we obtain bounds for the real error that are independent of row scaling. We discuss this further in § 2.1.

It has become common to use inexpensive estimators for the usual condition number $\|\mathbf{A}^{-1}\| \|\mathbf{A}\|$ to estimate a bound for the error in the computed solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ (Cline et al. (1979), Higham (1987a), Dongarra et al. (1979)). In § 4, we present an inexpensive and accurate condition estimator for the new condition number $\| |\mathbf{A}^{-1}| |\mathbf{A}| \|$ (and its variations). The new condition estimator is based on recent work by Hager (1984) and Higham (1987a), (1987b).

Finally, we tested our algorithm and associated condition estimator in a modified version of the sparse linear system solver MA28 (Duff 1977) from the Harwell Subroutine Library, which uses the pivotal strategy of Markowitz (1957) and a relative pivot test

$$|a_{kk}^{(k)}| \geqq u \max_{j>k} |a_{kj}^{(k)}|$$

on the elements $a_{kj}^{(k)}$ of the kth pivot row. Here u (the threshold parameter) is a preassigned factor, usually set to 0.1. MA28 can also drop entries of $\mathbf{L}$ and $\mathbf{U}$ that fall below a "*drop tolerance*" to attempt to further decrease the fill-in. The $\mathbf{L}$ and $\mathbf{U}$ factors are used to solve $\mathbf{A}\mathbf{x} = \mathbf{b}$ for $\mathbf{x}$ by forward and back substitution in the usual way, followed by some steps of iterative refinement. We report on the details of the experiments in § 5. Our conclusion is that a stopping criterion such as the one in Theorem 2 (but suitably modified as discussed in § 5) is a reliable and inexpensive stopping criterion for iterative refinement, often stopping after one or no update of $\mathbf{x}$. When drop tolerances are used and we have convergence, the rate of convergence degrades slightly but is still quite good. The new condition estimator of § 4 also proves to be inexpensive to calculate and is an accurate estimate on our test matrices, usually providing good accuracy for the cost of a few forward and back substitutions with the $\mathbf{LU}$ factors of $\mathbf{A}$.

The rest of this paper is organized as follows. Section 2 discusses the componentwise backward error further and also the conditioning of $\mathbf{A}\mathbf{x} = \mathbf{b}$ with respect to this backward error measure. We extend the analysis of Skeel (1980) to allow sparsity in the right-hand sides and introduce an automatic scheme for permitting larger perturbations in the right-hand sides where this is needed to maintain the sparsity of the matrix in the error bounds. Section 3 examines how the statement of Theorem 2 must change when either the floating-point arithmetic has no guard digit (such as on the CRAY) or underflow occurs. Section 4 presents a condition estimator corresponding to componentwise relative backward error. Section 5 discusses the numerical experiments. Section 6 has conclusions.

2. Backward error and conditioning.

2.1. Condition number. The condition number of a problem is the least upper bound of the ratio of the norm of perturbation in the solution to the norm of the perturbation in the input data, in the limit as the perturbation in the input data goes to zero. To compute it, we need a norm for the perturbation $\Delta\mathbf{x}$ in the solution as well as a norm for the perturbations $\Delta\mathbf{A}$ and $\Delta\mathbf{b}$ in the input data. We adopt the notation Δ rather than δ at this point because $\Delta\mathbf{A}$ and $\Delta\mathbf{b}$ are allowed any values whereas we use the δ notation to indicate particular values associated with the algorithmic error. The norm for the input data will depend on $\mathbf{E}$ and $\mathbf{f}$ as described above: $\|(\Delta\mathbf{A},\Delta\mathbf{b})\|_{\mathbf{E},\mathbf{f}}$ is defined as the smallest ω such that $|\Delta\mathbf{A}| \leqq \omega\mathbf{E}$ and $|\Delta\mathbf{b}| \leqq \omega\mathbf{f}$. For the norm of the output, we choose the usual sup norm $\|\mathbf{x}\|_\infty \equiv \max_i |x_i|$, to cater for zero components in $\mathbf{x}$. With this notation we can write

$$\kappa_{\mathbf{E},\mathbf{f}}(\mathbf{A},\mathbf{b}) \equiv \limsup_{\substack{\Delta\mathbf{A}\to\mathbf{0}\\ \Delta\mathbf{b}\to\mathbf{0}}} \frac{\|\Delta\mathbf{x}\|_\infty/\|\mathbf{x}\|_\infty}{\|(\Delta\mathbf{A},\Delta\mathbf{b})\|_{\mathbf{E},\mathbf{f}}} \tag{4}$$

where $\mathbf{x}+\Delta\mathbf{x} = (\mathbf{A}+\Delta\mathbf{A})^{-1}(\mathbf{b}+\Delta\mathbf{b})$. Following Skeel (1979), this may be easily evaluated as follows:

$$\kappa_{\mathbf{E},\mathbf{f}}(\mathbf{A},\mathbf{b}) \equiv \frac{\|\,|\mathbf{A}^{-1}|\mathbf{E}|\mathbf{x}| + |\mathbf{A}^{-1}|\,\mathbf{f}\,\|_\infty}{\|\mathbf{x}\|_\infty}. \tag{5}$$

For example, if we choose $\mathbf{E} = |\mathbf{A}|$ and $\mathbf{f} = |\mathbf{b}|$ for the componentwise relative error,

$$\kappa_{|\mathbf{A}|,|\mathbf{b}|}(\mathbf{A},\mathbf{b}) = \frac{\|\,|\mathbf{A}^{-1}|\,|\mathbf{A}|\,|\mathbf{x}| + |\mathbf{A}^{-1}|\,|\mathbf{b}|\,\|_\infty}{\|\mathbf{x}\|_\infty}. \tag{6}$$

Sometimes it is convenient to have a condition number which is independent of the right-hand side $\mathbf{b}$. Since

$$\frac{\|\,|\mathbf{A}^{-1}|\,|\mathbf{A}|\,|\mathbf{x}|\,\|_\infty}{\|\mathbf{x}\|_\infty} \leqq \kappa_{|\mathbf{A}|,|\mathbf{b}|}(\mathbf{A},\mathbf{b}) \leqq 2\frac{\|\,|\mathbf{A}^{-1}|\,|\mathbf{A}|\,|\mathbf{x}|\,\|_\infty}{\|\mathbf{x}\|_\infty}, \tag{7}$$

and $\|\,|\mathbf{A}^{-1}|\,|\mathbf{A}|\,|\mathbf{x}|\,\|_\infty/\|\mathbf{x}\|_\infty \leqq \|\,|\mathbf{A}^{-1}|\,|\mathbf{A}|\,\|_\infty$, we get the simpler condition number

$$\kappa_{|\mathbf{A}|}(\mathbf{A}) \equiv \|\,|\mathbf{A}^{-1}|\,|\mathbf{A}|\,\|_\infty \geqq 0.5\,\kappa_{|\mathbf{A}|,|\mathbf{b}|}(\mathbf{A},\mathbf{b}). \tag{8}$$

The purpose of the condition number is, of course, to provide error bounds: if $\mathbf{A}$ is perturbed by $|\delta\mathbf{A}| \leqq \omega|\mathbf{A}|$ and $\mathbf{b}$ by $|\delta\mathbf{b}| \leqq \omega|\mathbf{b}|$, and if ω is small enough, then $\mathbf{x}$ will be perturbed by no more than about $\omega\kappa_{|\mathbf{A}|,|\mathbf{b}|}(\mathbf{A},\mathbf{b})$. More rigorously, Skeel (1979) shows that, for ω defined as in (3),

$$\frac{\|\delta\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \leqq \frac{\omega\kappa_{|\mathbf{A}|,|\mathbf{b}|}(\mathbf{A},\mathbf{b})}{1-\omega\kappa_{|\mathbf{A}|}(\mathbf{A})}. \tag{9}$$

Similarly, if we define

$$\kappa_{\mathbf{E}}(\mathbf{A}) \equiv \|\,|\mathbf{A}^{-1}|\mathbf{E}\|_\infty, \tag{10}$$

we have, for ω defined as in (2),

$$\frac{\|\delta\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \leqq \frac{\omega\kappa_{\mathbf{E},\mathbf{f}}(\mathbf{A},\mathbf{b})}{1-\omega\kappa_{\mathbf{E}}(\mathbf{A})}. \tag{11}$$

It is easy to see that the problem is no more badly conditioned with respect to the componentwise relative backward error measure than with respect to the usual normed backward error measure. This is because

$$\kappa(\mathbf{A}) \equiv \|\mathbf{A}^{-1}\|_\infty \|\mathbf{A}\|_\infty \geqq \| \, |\mathbf{A}^{-1}| \, |\mathbf{A}| \, \|_\infty = \kappa_{|\mathbf{A}|}(\mathbf{A}). \tag{12}$$

It is possible for $\kappa_{|\mathbf{A}|}(\mathbf{A})$ to be much smaller than $\kappa(\mathbf{A})$. For example, we can make $\kappa(\mathbf{A})$ arbitrarily large by multiplying one of the rows of $\mathbf{A}$ by a large enough constant. However, $\kappa_{|\mathbf{A}|}(\mathbf{A})$ is independent of the row scaling of $\mathbf{A}$.

2.2. Backward error. As stated in the Introduction, in practice it is necessary to modify the choice $\mathbf{f} = |\mathbf{b}|$ of the componentwise relative backward error. This need arises because of the factor $\sigma(\mathbf{A}, \mathbf{x})$ in Theorem 2; when $\sigma(\mathbf{A}, \mathbf{x})$ is large, convergence of the backward error ω in (3) to the roundoff level is not guaranteed. Take, for example, $\mathbf{A}$ sparse and irreducible, and $\mathbf{x}$ sparse such that some $b_i = \sum_j a_{ij}x_j$ are zero because each $a_{ij}x_j = 0$. Since $\mathbf{A}^{-1}$ is structurally full (Duff et al. (1985)), $\mathbf{x}$ will be structurally full as well, so that a computed component $\hat{x}_k$ can be zero only through exact cancellation. In practice, this means that all components of the computed solution $\hat{\mathbf{x}}$ will be nonzero, with the entries that should be zero containing roundoff error of unpredictable sign. Therefore both $r_i = (\mathbf{A}\hat{\mathbf{x}} - \mathbf{b})_i$ and $(|\mathbf{A}|\,|\hat{\mathbf{x}}| + |\mathbf{b}|)_i$ may be small but of similar orders of magnitude, so that ω stays large even after some steps of iterative refinement.

Ideally, we would like to choose $\mathbf{f}$ to satisfy the following four criteria:

(i) The backward error ω (in (2)) usually converges to machine precision after one step of iterative refinement;

(ii) $\omega\mathbf{f}$ is "small" compared to $\mathbf{b}$;

(iii) the resulting error bound in (11) is as small as possible; and

(iv) ω is row-scaling independent.

We have experimented with two choices for $\mathbf{f}$ that come close to meeting these four criteria; this will be borne out by the numerical experiments in § 5. It turns out we must sacrifice the sparsity structure of $\mathbf{b}$ to guarantee a small backward error bound ω (criterion (i)). A trivial way to do this is to set $\mathbf{E} = \mathbf{0}$ and $\mathbf{f} = |\mathbf{r}|/\varepsilon = |\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}|/\varepsilon$, whence $\delta\mathbf{A} = \mathbf{0}$, $\delta\mathbf{b} = \mathbf{r}$ and $\omega = \varepsilon$. Of course this is unsatisfactory because $\delta\mathbf{b} = \mathbf{r}$ may be much larger in norm than $\mathbf{b}$ if the system is ill-conditioned, violating criterion (ii). Our approach is to keep $\mathbf{E} = |\mathbf{A}|$ and choose f_i larger than $|b_i|$ only if it is necessary to keep ω small.

We will choose $\mathbf{f}$ in an a posteriori way, letting it depend on the computation as follows. Let $\mathbf{w} = |\mathbf{A}|\,|\hat{\mathbf{x}}| + |\mathbf{b}|$ be the vector of denominators in (3). We then choose a threshold τ_i for each w_i, so that when $w_i > \tau_i$ we can use the usual scaling factor $f_i = |b_i|$. Otherwise, when $w_i \leqq \tau_i$, we choose a larger f_i. Correspondingly, we divide the equations of $\mathbf{Ax} = \mathbf{b}$ into two categories, those where $w_i > \tau_i$, and those where $w_i \leqq \tau_i$. We may assume without loss of generality that the leading m equations of $\mathbf{Ax} = \mathbf{b}$, which we denote by $\mathbf{A}^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$, belong to the first category, and the remaining $n - m$ equations $\mathbf{A}^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$ belong to the second. As stated above, we will let $\mathbf{f}^{(1)} = |\mathbf{b}^{(1)}|$ in the first category. There are several possibilities for τ_i, but in practice the following one has worked well: $\tau_i = 1000n\varepsilon(\|\mathbf{A}_{i.}\|_\infty \|\hat{\mathbf{x}}\|_\infty + |b_i|)$, where $\mathbf{A}_{i.}$ is the ith row of $\mathbf{A}$. Note that τ_i is about 1000 times larger than the maximum possible roundoff error committed in computing w_i, and w_i can only be less than τ_i if each product $a_{ij}\hat{x}_j$ is tiny. We performed other runs to check the sensitivity of this choice and found that a change of say a factor of 10 (to 100) could occasionally change the number of iterations and the error estimate but usually not by much. We note, however, that this can be viewed as a local choice and could be varied while performing iterative refinement, possibly increasing it in order to decrease ω.

Given the vector τ of the thresholds τ_i, we can choose $\mathbf{f}^{(2)}$ in at least two ways. First, we let $\mathbf{f}^{(2)} = |\mathbf{A}^{(2)}|\mathbf{e}\|\hat{\mathbf{x}}\|_\infty$, where $\mathbf{e}$ is the column vector of all ones. This corresponds to the usual normwise backward error, and so the components r_i of the residual are almost guaranteed to be small compared to these $f_i^{(2)}$, insofar as Gaussian elimination alone guarantees a small residual in the norm sense. Since we have not modified the definition of $\mathbf{E}$, we are further guaranteed a solution $\hat{\mathbf{x}}$ that preserves the sparsity structure of $\mathbf{A}$.

There is a difficulty with this choice of $\mathbf{f}$, however: we are no longer guaranteed that $\|\delta\mathbf{b}\|_\infty$ is small compared to $\|\mathbf{b}\|_\infty$. This can only happen when $\mathbf{A}$ is very ill-conditioned, since $\|\mathbf{A}^{(2)}\|_\infty\|\hat{\mathbf{x}}\|_\infty/\|\mathbf{b}\|_\infty$ is a lower bound on the condition number $\|\mathbf{A}^{-1}\|_\infty\|\mathbf{A}\|_\infty$ of $\mathbf{A}$. We have constructed artificial examples where this happens, but not observed it in practice. There is also the possibility that large components in $\mathbf{f}$ will make the condition number $\kappa_{|\mathbf{A}|,\mathbf{f}}(\mathbf{A}, \mathbf{b})$ too large and so make the error estimate $\omega\kappa_{|\mathbf{A}|,\mathbf{f}}(\mathbf{A}, \mathbf{b})$ too pessimistic, but note that this condition number is still bounded by $2\kappa_{|\mathbf{A}|}(\mathbf{A})$. We may avoid this possibility as follows. Given the two backward errors

$$\omega_i \equiv \max_j \frac{|\mathbf{A}^{(i)}\hat{\mathbf{x}} - \mathbf{b}^{(i)}|_j}{(|\mathbf{A}^{(i)}||\hat{\mathbf{x}}| + |\mathbf{f}^{(i)}|)_j}, \qquad i = 1,2, \tag{13}$$

the residual satisfies

$$|\mathbf{r}| = \begin{pmatrix} |\mathbf{A}^{(1)}\hat{\mathbf{x}} - \mathbf{b}^{(1)}| \\ |\mathbf{A}^{(2)}\hat{\mathbf{x}} - \mathbf{b}^{(2)}| \end{pmatrix} \leqq \begin{pmatrix} \omega_1(|\mathbf{A}^{(1)}||\hat{\mathbf{x}}| + |\mathbf{b}^{(1)}|) \\ \omega_2(|\mathbf{A}^{(2)}||\hat{\mathbf{x}}| + |\mathbf{A}^{(2)}|\mathbf{e}\|\hat{\mathbf{x}}\|_\infty) \end{pmatrix} \tag{14}$$

and, to first order, the error is bounded by

(15)

$$\begin{aligned} \frac{\|\delta\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} &= \frac{\|\mathbf{A}^{-1}\mathbf{r}\|_\infty}{\|\mathbf{x}\|_\infty} \leqq \frac{\||\mathbf{A}^{-1}||\mathbf{r}|\|_\infty}{\|\mathbf{x}\|_\infty} \\ &\leqq \omega_1 \frac{\left\| |\mathbf{A}^{-1}| \begin{pmatrix} |\mathbf{A}^{(1)}||\hat{\mathbf{x}}| + |\mathbf{b}^{(1)}| \\ \mathbf{0} \end{pmatrix} \right\|_\infty}{\|\hat{\mathbf{x}}\|_\infty} + \omega_2 \frac{\left\| |\mathbf{A}^{-1}| \begin{pmatrix} \mathbf{0} \\ |\mathbf{A}^{(2)}||\hat{\mathbf{x}}| + \mathbf{f}^{(2)} \end{pmatrix} \right\|_\infty}{\|\hat{\mathbf{x}}\|_\infty} \\ &\equiv \omega_1\kappa_{\omega_1} + \omega_2\kappa_{\omega_2}. \end{aligned}$$

The advantage of this formulation is that components of $\mathbf{f}^{(2)}$ may be very large compared to the components of $\mathbf{b}^{(2)}$, causing ω_2 to be very small and κ_{ω_2} to be correspondingly large but without affecting ω_1 or κ_{ω_1}. This formulation is tested in the numerical experiments in § 5.

A second possible choice for $\mathbf{f}^{(2)}$ is to use $\mathbf{f}^{(2)} = \|\mathbf{b}\|_\infty\mathbf{e}$. This choice of $\mathbf{f}^{(2)}$ assures us that a small backward error indeed means $\|\delta\mathbf{b}\|_\infty/\|\mathbf{b}\|_\infty$ will be small, but gives us less assurance that the backward error will converge to machine precision. We have not seen it fail in practice. As with the other choice of $\mathbf{f}$, we can bound the error using two backward errors defined as in (13) and the sum of their products with two condition numbers as in (15). Section 5 also reports on numerical experience with this backward error measure.

Both the previous choices for $\mathbf{f}^{(2)}$ can violate one of the criteria (ii) or (iv). The choice $\mathbf{f}^{(2)} = |\mathbf{A}^{(2)}|\mathbf{e}\|\hat{\mathbf{x}}\|_\infty$ guarantees that ω_i, $i = 1, 2$, are row-scaling independent (criterion (iv)), while it can violate criterion (ii). The choice $\mathbf{f}^{(2)} = \|\mathbf{b}\|_\infty\mathbf{e}$ satisfies criterion (ii), but the corresponding ω_2 is row-scaling dependent. Both, as we shall see, satisfy criteria (i) and (iii).

We also see that the bound depends on the accuracy with which we can compute the residual $\mathbf{r}$ and the backwards error ω in (2). How much can roundoff contaminate the computed ω, especially when $\mathbf{r} = \mathbf{A}\hat{\mathbf{x}} - \mathbf{b}$ is computed by an arithmetic with machine precision ε? A standard error analysis shows that the error in the computed $\mathbf{r}$, $\delta\mathbf{r}$, is bounded by $(\gamma + 1)\varepsilon(|\mathbf{A}|\,|\hat{\mathbf{x}}| + |\mathbf{b}|)$, where γ is the maximum number of nonzero entries in a row of $\mathbf{A}$. When $\mathbf{E} = |\mathbf{A}|$ and $\mathbf{f} = |\mathbf{b}|$, this means that the computed ω cannot differ from the true ω by more than about $\pm(\gamma + 1)\varepsilon$ which will be within the tolerance of our sparse modification of Skeel's stopping criterion in Theorem 2. Since the computed ω is almost certainly at least about $\gamma\varepsilon$, the final error bound $\omega\kappa_{|\mathbf{A}|,|\mathbf{b}|}(\mathbf{A}, \mathbf{b})$, can be low by no more than a factor of two. The same is true for ω_i, $i = 1, 2$.

At this point, we might ask what choice of $\mathbf{E}$ and $\mathbf{f}$ minimizes the resulting error estimate (11). It is easy to see that any choice of $\mathbf{E}$ and $\mathbf{f}$ such that $\mathbf{E}|\mathbf{x}| + \mathbf{f}$ is a multiple of $|\mathbf{r}|$, say $\mathbf{E} = \mathbf{0}$ and $\mathbf{f} = |\mathbf{r}|$, yields the following minimum product:

$$\omega\kappa_{\mathbf{E},\mathbf{f}}(\mathbf{A},\mathbf{b}) = \|\,|\mathbf{A}^{-1}|\,|\mathbf{r}|\,\|_\infty / \|\mathbf{x}\|_\infty .$$

Since the true error is $\|\delta\mathbf{x}\|_\infty / \|\mathbf{x}\|_\infty = \|\mathbf{A}^{-1}\mathbf{r}\|_\infty / \|\mathbf{x}\|_\infty$, we see that the bound is as tight as ignoring signs in $\mathbf{r}$ allows. For this special choice of $\mathbf{E}$ and $\mathbf{f}$, we should also add $(\gamma + 1)\varepsilon(|\mathbf{A}|\,|\hat{\mathbf{x}}| + |\mathbf{b}|)$ to $|\mathbf{r}|$ since roundoff may lower the computed value of $|\mathbf{r}|$ by the same amount. The choice $\mathbf{E} = \mathbf{0}$ and $\mathbf{f} = |\mathbf{r}| + (\gamma + 1)\varepsilon(|\mathbf{A}|\,|\hat{\mathbf{x}}| + |\mathbf{b}|)$ yields a new error bound of

$$(16) \qquad \frac{\|\delta\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \leqq \frac{\|\,|\mathbf{A}^{-1}|\,|\mathbf{r}|\,\|_\infty}{\|\mathbf{x}\|_\infty} + (\gamma + 1)\varepsilon\kappa_{|\mathbf{A}|,|\mathbf{b}|}(\mathbf{A},\mathbf{b}).$$

Thus we see that the condition number $\kappa_{|\mathbf{A}|,|\mathbf{b}|}(\mathbf{A}, \mathbf{b})$ plays a central role independent of the notion of backward error, just because it reflects the possible roundoff errors in the computed residual. Furthermore, after only a few steps of iterative refinement Theorem 2 guarantees that, to first order, the bound (9) will be about the same as the bound (16). In our experiments we have seen that, usually, the estimates of the real error given by (9) and (15) have the same order of accuracy as the estimates obtained by the bound (16).

Note that if we set $e_{ij} = \|\mathbf{A}\|_\infty$ and $f_i = \|\mathbf{b}\|_\infty$, the backward error of $\hat{\mathbf{x}}$ with respect to $\mathbf{E}$ and $\mathbf{f}$ is given by $\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}\|_\infty / (\|\mathbf{A}\|_\infty\|\hat{\mathbf{x}}\|_1 + \|\mathbf{b}\|_\infty)$. It is also easy to see that

$$(17) \qquad \kappa_{\mathbf{E},\mathbf{f}}(\mathbf{A},\mathbf{b}) = \frac{\|\mathbf{A}^{-1}\|_\infty\|\mathbf{A}\|_\infty\|\mathbf{x}\|_1 + \|\mathbf{A}^{-1}\|_\infty\|\mathbf{b}\|_\infty}{\|\mathbf{x}\|_\infty},$$

which is within a factor of $2n$ of $\|\mathbf{A}^{-1}\|_\infty\|\mathbf{A}\|_\infty$. Thus, this choice of $\mathbf{E}$ and $\mathbf{f}$, which permits equally large perturbations in all entries of $\mathbf{A}$ and $\mathbf{b}$, gives essentially the same backward error and condition number as the usual normed backward error.

We note, in conclusion, that Skeel's original motivation (Skeel 1979) was to analyze the effects of row and column scaling of $\mathbf{A}$ on the accuracy and the stability of the LU factorization. He concluded that the optimal way to scale depended on the solution: the columns should be scaled (thus scaling the solution components) so that the components of the scaled solution are all equal in magnitude, and the rows should be scaled so each component of $|\mathbf{A}|\,|\mathbf{x}|$ ($\mathbf{x}$ is the solution) is equal in magnitude. This is unfortunately hard to use in practice since it requires much information about the solution. Fortunately, one step of iterative refinement tends to overcome the effects of bad row scaling, as we have seen.

3. Different models of floating-point arithmetic. Theorem 2 assumes that arithmetic is implemented rather cleanly, i.e., that the floating-point result fl $(a \diamond b)$ of the operation $a \diamond b$, $(\diamond \in \{+, -, \times, /\})$ satisfies

$$\text{fl}\,(a \diamond b) = (a \diamond b)(1+e) \tag{18}$$

with $|e| \leqq \varepsilon$, where ε is the machine precision. This model eliminates both the possibility of underflow as well as machines like the CRAYs, where for lack of a guard digit in the hardware we can only assert that

$$\text{fl}\,(a \pm b) = a(1+e_1) \pm b(1+e_2) \tag{19}$$

where $|e_i| \leqq \varepsilon$. Thus, when a and b are very close and we are subtracting, this model permits a large relative error in the computed difference. For example, on any CRAY or many CDC machines, the computed difference of any power of two and the next smaller floating-point number is wrong by a factor of two (see, Kahan (1981)).

Despite this difficulty, it is possible to carry through the proof of Theorem 2 using the weaker model (19) instead of (18) and arrive at essentially the same conclusion: one step of iterative refinement, even without computing the residual using arithmetic of machine precision ε^2, is enough to guarantee a small componentwise relative backward error as long as the matrix is not too ill-conditioned and $\sigma(\mathbf{A}, \mathbf{x})$ is not too large. We might expect problems in bounding the error in the computed residual fl $(\mathbf{A}\hat{\mathbf{x}} - \mathbf{b})$, since the result might be off by a factor of two, but in the analysis this potential error is dominated by the error in computing $\mathbf{A}\hat{\mathbf{x}}$, so the proof goes through. Similarly, the error in updating $\hat{\mathbf{x}} - \mathbf{d}$ is swamped by larger errors.

The other exception to the model in (18) is underflow. The extension of error analysis to include underflow is discussed in some detail by Demmel (1984), and we just summarize the results here. In place of (18) we use the model

$$\text{fl}\,(a \diamond b) = (a \diamond b)(1+e) + \nu \tag{20}$$

where $|e| \leqq \varepsilon$ as before, and ν represents the underflow error. Let λ be the underflow threshold, that is the smallest positive, normalized floating-point number. Then, on machines where computed quantities which would be smaller than λ are replaced by zero, $|\nu|$ is bounded by λ. On machines with IEEE standard floating-point arithmetic (see (IEEE 1985), (IEEE 1987)), gradual underflow lowers the bound on $|\nu|$ to $\varepsilon\lambda$.

The statement of Theorem 2 must be modified as follows to account for underflow. For gradual underflow, we can say the following. If the inputs $\mathbf{A}$ and $\mathbf{b}$ and the output $\hat{\mathbf{x}}$ are normalized (that is, exceed λ in magnitude), and if the residuals are computed by an arithmetic of machine precision either ε or ε^2, then gradual underflow can only degrade performance to the level of the residual computation using the arithmetic of machine precision ε. For conventional underflow, the norms of $\mathbf{A}$, $\mathbf{b}$, and $\hat{\mathbf{x}}$ must exceed λ/ε for this statement to be true.

The use of extended range and precision in intermediate computations does not change these conclusions. Assuming $\mathbf{r}$ and $\mathbf{d}$ are stored in the same format as $\mathbf{A}$, $\mathbf{b}$ and $\hat{\mathbf{x}}$, underflows in $\mathbf{r}$ and $\mathbf{d}$ have the same potential effects on performance as they did when they were not computed in extended format.

We have not yet considered the effect of underflow on the rate of convergence of the iteration. There are matrices for which the iteration converges only if underflows do not occur, but the matrices are so ill-conditioned as to make the computed solution untrustworthy anyway. As long as some entry of **A** is large enough (λ for gradual underflow and λ/ε for conventional underflow) then underflows will have an effect on the convergence rate comparable to roundoff.

4. An estimator for $\kappa_{|\mathbf{A}|,|\mathbf{b}|}(\mathbf{A}, \mathbf{b})$.

To estimate the accuracy of a computed solution of $\mathbf{Ax} = \mathbf{b}$, two ingredients are needed: a bound on the backward error (however it is measured) and a condition number with respect to the choice of backward error. As discussed in § 2.2, the product of the two previous quantities provides an approximate upper bound on the relative error in the computed solution.

In the case of the conventional normwise backward error, the condition number is essentially given by $\kappa(\mathbf{A}) = \|\mathbf{A}^{-1}\|_\infty \|\mathbf{A}\|_\infty$. There has been much work on such estimators for $\kappa(\mathbf{A})$ in recent years (for example, Cline et al. 1979; see Higham (1987a) for a complete list of references), and cheap, reliable estimators are available in standard software packages such as LINPACK (Dongarra et al. (1979)). It is natural to seek an analogous estimator for $\kappa_{|\mathbf{A}|,|\mathbf{b}|}(\mathbf{A}, \mathbf{b})$.

From (5) we see that the quantity we need to estimate is

$$\| |\mathbf{A}^{-1}| \mathbf{E} |\mathbf{x}| + |\mathbf{A}^{-1}| \mathbf{f} \|_\infty = \| |\mathbf{A}^{-1}| (\mathbf{E} |\mathbf{x}| + \mathbf{f}) \|_\infty . \tag{21}$$

In place of the true solution $\mathbf{x}$, we may use its computed approximation $\hat{\mathbf{x}}$. In the case of componentwise relative backward error, we may also just use the simpler condition number $\kappa_{|\mathbf{A}|}(\mathbf{A})$ that requires us to estimate

$$\| |\mathbf{A}^{-1}| |\mathbf{A}| \|_\infty = \| |\mathbf{A}^{-1}| |\mathbf{A}| \mathbf{e} \|_\infty \tag{22}$$

where $\mathbf{e}$ is the vector of all ones. Either way, we need to be able to estimate

$$\| |\mathbf{A}^{-1}| \mathbf{g} \|_\infty \tag{23}$$

where $\mathbf{g}$ is a nonnegative vector that is easy to compute (in the above examples it costs just one matrix-vector multiply).

Let $\mathbf{G} = \operatorname{diag}(g_1, \cdots, g_n)$. Then $\mathbf{g} = \mathbf{Ge}$ and

$$\| |\mathbf{A}^{-1}| \mathbf{g} \|_\infty = \| |\mathbf{A}^{-1}| \mathbf{Ge} \|_\infty = \| |\mathbf{A}^{-1}\mathbf{G}| \mathbf{e} \|_\infty = \| |\mathbf{A}^{-1}\mathbf{G}| \|_\infty = \|\mathbf{A}^{-1}\mathbf{G}\|_\infty . \tag{24}$$

$\|\mathbf{A}^{-1}\mathbf{G}\|_\infty$ can be estimated by the algorithm of Hager (1984) and Higham (1987a), (1987b) that estimates the one-norm (or infinity-norm) of a $n \times n$ matrix given the ability to multiply a vector by both the matrix and its transpose. We can multiply any vector $\mathbf{z}$ by the operator $\mathbf{A}^{-1}\mathbf{G}$ by multiplying $\mathbf{z}$ by the diagonal matrix $\mathbf{G}$, and then solving $\mathbf{Ay} = \mathbf{Gz}$ using the LU factorization of $\mathbf{A}$. Multiplying by $(\mathbf{A}^{-1}\mathbf{G})^T$ is equally easy.

Our estimate of condition numbers $\kappa_{|\mathbf{A}|,|\mathbf{b}|}(\mathbf{A}, \mathbf{b})$ includes a dependence on the calculated solution. We also performed runs for different solutions (for example, $x_i = i^2$, $i = 1, \cdots, n$) and found little sensitivity. Note that the experiments in Set 1 in § 5 give us results close to the upper bound of twice $\kappa_{|\mathbf{A}|}$.

5. Numerical experiments.

In this section, we discuss numerical experiments supporting our earlier analysis and discussion. The issues involved are the effectiveness of using arithmetic of machine precision ε in the computation of the residual, the choice of values for $\mathbf{f}$ (§ 2.2), the effectiveness of our new condition numbers, and the use of drop tolerances.

To do this investigation, we group our experiments into four sets. In Set 1, we do not have any sparsity in the right-hand side: we show that the computation of the residual in arithmetic of machine precision ε is satisfactory and illustrate our analysis in the standard case assumed by Skeel (1980). In Sets 2 and 3, we show the extension to sparse right-hand sides and test different choices for **f**. Finally, we examine the use of drop tolerances in the runs in Set 4.

We perform the experiments by modifying the sparse linear system solver MA28 in the Harwell Subroutine Library (Duff (1977)). The Set 4 runs are easy to perform because MA28 can drop entries of **L** and **U** that fall below a tolerance, called *drop tol* in our tables (*drop tol* = 0 corresponds to standard Gaussian elimination). The resulting **L** and **U** factors are then used to solve $\mathbf{Ax} = \mathbf{b}$ for **x** by forward and back substitution in the usual way, followed by some steps of iterative refinement.

All our runs are on a common set of test matrices from the Harwell–Boeing test set (Duff, Grimes, and Lewis (1987)). Their names, number of nonzero entries and condition numbers $\kappa(\mathbf{A})$ and $\kappa_{|\mathbf{A}|}(\mathbf{A})$ are given in Table 1. The name of each matrix includes its dimensions, for example, GRE115 is 115 by 115. The two matrices of order 216 have the same structure, but they have quite different numerical values. We also ran our tests on some other matrices from the set and obtained results broadly comparable with these displayed.

For each run, we chose the value of the solution **x** and then we computed the right-hand side **b** by multiplying the solution by the test matrix. All matrices have also been scaled before computing the right-hand side, thus obtaining two test problems for each matrix. The scaling is computed using the Harwell routine MC19, which makes the nonzeros of the scaled matrix near to unity by minimizing the sum of the squares of logarithms of the moduli of the nonzeros (Curtis and Reid (1972)). This scaling does not guarantee that $\kappa(\mathbf{A})$ and $\kappa_{|\mathbf{A}|}(\mathbf{A})$ must decrease (see Table 1) although on many matrices the effect is very beneficial, particularly for the classical condition number. This

TABLE 1
Condition numbers before and after scaling.

	Nonzeros	Before scaling		After scaling	
		$\kappa(\mathbf{A})$	$\kappa_{\lvert\mathbf{A}\rvert}(\mathbf{A})$	$\kappa(\mathbf{A})$	$\kappa_{\lvert\mathbf{A}\rvert}(\mathbf{A})$
GRE115	421	0.93D+02	0.86D+02	0.69D+04	0.13D+03
GRE185	975	0.38D+06	0.15D+06	0.39D+06	0.14D+06
GRE216A	812	0.28D+03	0.22D+03	0.20D+03	0.18D+03
GRE216B	812	0.83D+15	0.29D+14	0.56D+08	0.85D+07
GRE343	1310	0.47D+03	0.37D+03	0.30D+03	0.26D+03
GRE512	1976	0.46D+03	0.37D+03	0.40D+03	0.36D+03
GRE1107	5664	0.18D+09	0.98D+08	0.77D+10	0.24D+09
WEST67	294	0.91D+03	0.31D+03	0.30D+03	0.13D+03
WEST132	413	0.11D+13	0.80D+07	0.94D+04	0.21D+04
WEST156	362	0.12D+32	0.38D+09	0.91D+12	0.15D+09
WEST167	506	0.69D+11	0.52D+06	0.46D+04	0.12D+04
WEST381	2134	0.53D+07	0.38D+05	0.38D+06	0.53D+04
WEST479	1888	0.49D+12	0.37D+07	0.27D+06	0.20D+05
WEST497	1721	0.38D+12	0.13D+07	0.42D+07	0.63D+04
WEST655	2808	0.49D+12	0.37D+07	0.42D+06	0.36D+05
WEST989	3518	0.13D+13	0.10D+08	0.58D+06	0.52D+05
WEST1505	5414	0.14D+13	0.10D+08	0.67D+08	0.21D+07
WEST2021	7310	0.28D+13	0.21D+08	0.86D+06	0.10D+06

is particularly so for the GRE216B example, where, before the scaling, the matrix was essentially singular. Note in general that many of the matrices are poorly conditioned, particularly before scaling.

In all the runs, the standard normwise backward error

$$\eta = \frac{\|\mathbf{r}\|_\infty}{\|\mathbf{A}\|_\infty \|\hat{\mathbf{x}}\|_\infty + \|\mathbf{b}\|_\infty}, \tag{25}$$

the condition number $\kappa(\mathbf{A})$ and the error bound $\eta\kappa(\mathbf{A})$ were computed and compared to the other backward errors, condition numbers and error bounds.

All tests were done on an IBM 3084. In single precision the machine precision ε is $16^{-5} \approx 10^{-6}$. In double precision it is $16^{-13} \approx 2 \times 10^{-16}$. The main data for our numerical experiments are presented in Tables A1–A15 in the Appendix. In this section, we display summaries of these results.

In all cases, the stopping criterion was

Stop if $\omega \leqq \varepsilon$ or ω does not decrease by at least a factor of two.

All the runs used IBM double precision, except for the experiments in single and mixed precision in Set 1. This stopping criterion differs from that used in Theorem 2 ($\omega \leqq (n + 1)\varepsilon$). The value in Theorem 2 can be too large, especially for very large and sparse matrices, and the iterative refinement could stop too early. Generally, our stopping criterion terminates the iterative refinement with a value of ω less than ε. If the convergence is slow (for example, using double precision, the GRE216B matrix in Table A7 stops after four iterations with $\omega = 0.4 \times 10^{-15} \approx 2\varepsilon$), our stopping criterion recognizes this early. However, the final value of ω is still of order ε. Somewhat surprisingly we find there is no advantage in including a factor $(\gamma + 1)$ in our stopping criterion. Indeed, its inclusion would often result in no iterations, and there are only few occasions in Sets 1–3 where the $\omega \leqq \varepsilon$ criterion is not met. Note that, in the runs in Sets 2–4, ω is replaced by $\omega_1 + \omega_2$ (as in (13)–(15)). If we used a similar condition on η, in most of the examples we did not perform any steps of iterative refinement because the first solution satisfied the stopping criterion, but, before scaling, the estimation of the error $\|\delta\mathbf{x}\|_\infty / \|\mathbf{x}\|_\infty$ given by $\eta\kappa(\mathbf{A})$ was very poor because of the very large value of $\kappa(\mathbf{A})$.

We discuss the experiments for each of our four sets of values in turn. In all the following tables, the row corresponding to "Num. iter." gives the number of steps performed by the iterative refinement algorithm and the row corresponding to "Num. cases" gives the number of examples for which the iterative refinement performed that number of iterations. By "Error" we denote the max-norm of the difference between the computed solution and the actual solution used to generate the right-hand side, divided by the max-norm of the actual solution.

In the following, we denote by $\omega_i^{(j)}$ and by $\kappa_{\omega_i}^{(j)}$, $i = 1, 2, j = 1, 2, 3, 4$, the componentwise backward errors defined by (13) and the corresponding condition numbers defined by (15). The superscript identifies the set of tests.

Set 1.

For these tests, the right-hand sides $\mathbf{b}$ were chosen so that the true solution $\mathbf{x}$ had all components equal to one, so that all equations belonged to category 1. Thus the backwards error was given by $\omega_1^{(1)}$ as defined in (13), the condition number $\kappa_{\omega_1}^{(1)}$ and the error bound by $\omega_1^{(1)}\kappa_{\omega_1}^{(1)}$ as defined in (15). Because all the equations belong to category 1, $\kappa_{\omega_1}^{(1)} = \kappa_{|\mathbf{A}|,|\mathbf{b}|}(\mathbf{A}, \mathbf{b})$, and $\omega_2^{(1)} = 0$. The drop tolerance was zero. These tests were run in single precision, double precision, and mixed precision (all single precision, except for double precision computation of residuals). The Tables A1–A5 in the Appendix are

TABLE 2
Summary of results for the condition numbers for Set 1.

	min	avr	max
$Log_{10}\frac{\kappa(A)\,(Before\ scaling)}{\kappa(A)\,(After\ scaling)}$	-1.9	4.1	19
$Log_{10}\frac{\kappa^{(1)}_{\omega_1}\,(Before\ scaling)}{\kappa^{(1)}_{\omega_1}\,(After\ scaling)}$	-0.38	1.4	6.5

	Before scaling			After scaling		
	min	avr	max	min	avr	max
$Log_{10}(\kappa(A)/\kappa^{(1)}_{\omega_1})$	-0.26	3.6	22	-0.26	0.91	3.5

relative to Set 1. In Table 2, summarizing the results in Table A1, we observe that the condition number $\kappa^{(1)}_{\omega_1}$ is always less, for both scaled and unscaled matrices, than twice the classical condition number $\kappa(A)$, as must be the case from the theory. In some examples, $\kappa^{(1)}_{\omega_1}$ is much better than $\kappa(A)$ (for example, in the WEST156 example before scaling $\kappa^{(1)}_{\omega_1} < 3.2 \times 10^{-23}\kappa(A)$). Moreover, Table 2 shows that the classical condition number $\kappa(A)$, without any form of scaling, is rather unreliable as a measure of the ill-conditioning of the system. Table 3 (summarizing the results in Tables A2 and A3) reflects the previous considerations, so that the estimation $\omega_1^{(1)}\kappa^{(1)}_{\omega_1}$ of the error is generally quite tight, while $\eta\kappa(A)$ can be too pessimistic before scaling. Note that it is possible for our bound to be less tight than that from the classical theory but, when this happens in the experiments, our bound is only three times greater than the classical one in the worst case.

Throughout, our estimate of condition numbers $\kappa_{|A|,|b|}(A, \mathbf{b})$ includes a dependence on the calculated solution. We also performed runs for different solutions (for example,

TABLE 3
Summary of results for Set 1.

	Before scaling			After scaling		
Num. iter.	0	1	≥ 2	0	1	≥ 2
Num. cases	0	17	1	1	16	1
	min	avr	max	min	avr	max
$Log_{10}(\eta)$	-18	-16	-16	-17	-16	-16
$Log_{10}(\omega_1^{(1)})$	-16	-16	-16	-16	-16	-16
$Log_{10}\frac{\eta\,\kappa(A)}{Error}$	0.78	4.7	22	0.93	2.0	4.1
$Log_{10}\frac{\omega_1^{(1)}\,\kappa^{(1)}_{\omega_1}}{Error}$	0.48	1.5	2.5	0.43	1.4	3.3
$Log_{10}\frac{\eta\,\kappa(A)}{\omega_1^{(1)}\,\kappa^{(1)}_{\omega_1}}$	-0.32	3.2	20	-0.41	0.53	3.0

$x_i = i^2, i = 1, \cdots, n$) and found little sensitivity. Note that our choice of $\mathbf{x}$ in Set 1 gives us results close to the upper bound of twice $\kappa_{|A|}$. In Tables A4 and A5, we report the results of the algorithm using single and mixed precision. Unfortunately, the test matrices are in many cases so ill-conditioned that the iterative refinement diverged, that is $\omega_1^{(1)}$ increased after some steps as in, for example, GRE1107 and the GRE216B example in Table A4. In practice, IBM single precision is too poor to produce good results, and the use of mixed precision does not help. Note, however, that our algorithm still terminates after only a few steps. In every case, we tried running the iterative refinement for twenty steps and in no cases did we get much improvement over the results shown. Our algorithm for computing the condition numbers encounters numerical difficulties partly because of the ill-conditioning of these matrices and partly because we use threshold pivoting in the LU factorization. We would have used iterative refinement in this computation, but this would be at variance with our desire for a cheap estimator. Our feeling is that single precision calculations are inappropriate here.

Set 2.

For these tests we chose $\tau_i = 1000n\varepsilon(\|\mathbf{A}_{i.}\|_\infty\|\hat{\mathbf{x}}\|_\infty + |b_i|)$ and $\mathbf{f}^{(2)} = |\mathbf{A}^{(2)}|\mathbf{e}\|\hat{\mathbf{x}}\|_\infty$, where $\mathbf{e}$ is the column vector of all ones. This leads to the backward errors $\omega_1^{(2)}$ and $\omega_2^{(2)}$ defined in (13) and the condition numbers $\kappa_{\omega_1}^{(2)}$ and $\kappa_{\omega_2}^{(2)}$ and error bound $\omega_1^{(2)}\kappa_{\omega_1}^{(2)} + \omega_2^{(2)}\kappa_{\omega_2}^{(2)}$ defined in (15). The right-hand sides were chosen so that the true solution $\mathbf{x}$ had every fifth entry equal to one ($x_1 = x_6 = x_{11} = \cdots = 1$) and the rest zero. The drop tolerance was zero. These tests were done in double precision only. Tables A6–A8 show the results of runs on Set 2. We present a summary of these results in Tables 4 and 5. We also ran all the test examples of Set 2 replacing zero with 10^{-16} in $\mathbf{x}$ and obtained similar results. It is necessary to emphasize that, in most of the examples of Set 2, the standard ω computed by (3) was very large (sometimes of order one), so that we would get no useful information if we use a very large value for τ_i. Note that, in all our runs, $\omega_2^{(2)}$ is very small compared with $\omega_1^{(2)}$, in agreement with our comments after (15).

It may appear that our error estimate is sometimes poor, but the relatively good solution obtained is really fortuitous as can be seen by the results in the Appendix using

TABLE 4
Summary of results for the condition numbers for Set 2.

	min	avr	max
$Log_{10}\dfrac{\kappa(A)\ (\textit{Before scaling})}{\kappa(A)\ (\textit{After scaling})}$	–1.9	4.1	19
$Log_{10}\dfrac{\kappa_{\omega_1}^{(2)}\ (\textit{Before scaling})}{\kappa_{\omega_1}^{(2)}\ (\textit{After scaling})}$	–0.37	1.3	7.0
$Log_{10}\dfrac{\kappa_{\omega_2}^{(2)}\ (\textit{Before scaling})}{\kappa_{\omega_2}^{(2)}\ (\textit{After scaling})}$	–0.43	1.6	6.1

	Before scaling			After scaling		
	min	avr	max	min	avr	max
$Log_{10}(\kappa(A)/\kappa_{\omega_1}^{(2)})$	0.45	4.3	23	0.26	1.5	3.8
$Log_{10}(\kappa(A)/\kappa_{\omega_2}^{(2)})$	0.52	4.3	23	0.30	1.8	5.2

TABLE 5
Summary of results for Set 2.

	Before scaling			After scaling		
Num. iter.	0	1	≥ 2	0	1	≥ 2
Num. cases	1	13	4	2	12	4
	min	avr	max	min	avr	max
$Log_{10}(\eta)$	–23	–17	–16	–17	–17	–16
$Log_{10}(\omega_1^{(2)})$	–16	–16	–15	–16	–16	–15
$Log_{10}(\omega_2^{(2)})$	–32	–27	–19	–31	–28	–19
$Log_{10}\dfrac{\eta\kappa(A)}{Error}$	0.65	4.5	19	0.97	2.2	4.0
$Log_{10}\dfrac{\omega_1^{(2)}\kappa_{\omega_1}^{(2)}+\omega_2^{(2)}\kappa_{\omega_2}^{(2)}}{Error}$	0.58	1.7	4.3	0.50	1.6	2.7
$Log_{10}\dfrac{\eta\kappa(A)}{\omega_1^{(2)}\kappa_{\omega_1}^{(2)}+\omega_2^{(2)}\kappa_{\omega_2}^{(2)}}$	–0.17	2.8	16	–0.23	0.63	2.4

the same matrix but with a different right-hand side (the examples shown by the GRE1107 results in Tables A3 and A8 and by the GRE216B results in Tables A2 and A7).

Set 3.

For these tests we chose $\tau_i = 1000n\epsilon(\|\mathbf{A}_{i.}\|_\infty\|\hat{\mathbf{x}}\|_\infty + |b_i|)$ just as in Set 2, and $\mathbf{f}^{(2)} = \|\mathbf{b}\|_\infty\mathbf{e}$, where $\mathbf{e}$ is the column vector of all ones. This leads to backward errors $\omega_1^{(3)}$ and $\omega_2^{(3)}$ defined in (13) and the condition numbers $\kappa_{\omega_1}^{(3)}$ and $\kappa_{\omega_2}^{(3)}$ and error bound $\omega_1^{(3)}\kappa_{\omega_1}^{(3)} + \omega_2^{(3)}\kappa_{\omega_2}^{(3)}$ defined in (15). The right-hand sides were chosen so that the true solution $\mathbf{x}$ had every fifth entry equal to one and the rest zero. The drop tolerance was

TABLE 6
Summary of results for the condition numbers for Set 3.

	min	avr	max
$Log_{10}\dfrac{\kappa(A)\,(\textit{Before scaling})}{\kappa(A)\,(\textit{After scaling})}$	–1.9	4.1	19
$Log_{10}\dfrac{\kappa_{\omega_1}^{(3)}\,(\textit{Before scaling})}{\kappa_{\omega_1}^{(3)}\,(\textit{After scaling})}$	–0.37	1.3	7.0
$Log_{10}\dfrac{\kappa_{\omega_2}^{(3)}\,(\textit{Before scaling})}{\kappa_{\omega_2}^{(3)}\,(\textit{After scaling})}$	–1.9	4.0	14

	Before scaling			After scaling		
	min	avr	max	min	avr	max
$Log_{10}(\kappa(A)/\kappa_{\omega_1}^{(3)})$	0.45	4.3	23	0.26	1.5	3.8
$Log_{10}(\kappa(A)/\kappa_{\omega_2}^{(3)})$	0.10	0.97	6.4	0.38	0.86	2.6

zero. These tests were done in double precision only. The Tables A9–A11 are relative to Set 3 of parameters, and we summarize these in Tables 6 and 7. Comparing Tables 4 and 6 we observe that, while $\kappa_{\omega_1}^{(2)}$ and $\kappa_{\omega_2}^{(2)}$ are usually quite close, $\kappa_{\omega_2}^{(3)}$ can be much larger than $\kappa_{\omega_1}^{(3)}$ (for example, see the WEST156 example before scaling, where $\kappa_{\omega_2}^{(3)}$ is 10^{16} times larger than $\kappa_{\omega_1}^{(3)}$) and the error estimation can be pessimistic. Also note that, comparing line 7 of Tables 5 and 7, this choice of $\mathbf{f}$ does not give as good a bound as our choice for $\mathbf{f}$ in Set 2, although the difference is minimal after scaling.

Set 4.

For these tests we used nonzero drop tolerances (*drop tol* = 10^{-5}, *drop tol* = 10^{-3}). We changed τ_i from its earlier value to $\tau_i = 1000n(\varepsilon + \textit{drop tol})(\|\mathbf{A}_{i.}\|_\infty \|\hat{\mathbf{x}}\|_\infty + |b_i|)$ and used $\mathbf{f}^{(2)} = |\mathbf{A}^{(2)}|\mathbf{e}\|\hat{\mathbf{x}}\|_\infty$, where $\mathbf{e}$ is the column vector of all ones. The entries of $\mathbf{b}$ and $\mathbf{x}$ were chosen as in Set 3. Double precision was used. Tables A12–A15 are the results of runs using this set of parameters, and the results are summarized in Table 8. Note that, in this set, we nearly always have $\omega_1^{(4)} = 0$. This corresponds to putting all of the error into $\mathbf{b}$, that is, $\delta\mathbf{A} = \mathbf{0}$ and $\delta\mathbf{b} = \mathbf{A}\hat{\mathbf{x}} - \mathbf{b}$, obtaining the situation discussed at the beginning of § 2.2. In this case, $\mathbf{f}$ does not depend on $\mathbf{b}$ explicitly, but our bounds are still good. Note again that our stopping criterion terminates after only a few iterations if the iteration diverges. We checked this divergence by forcing more iterations and observed either oscillation or divergence.

We observed, contrary to Zlatev, Wasniewski, and Schaumburg (1986), that little gain in sparsity was obtained (see, for example, Table A15), while even moderate values of drop tolerance caused divergence of the iterative refinement. A drop tolerance strategy appears to work well only on very structured sparse matrices such as those resulting from discretizations of partial differential equations. We confirmed this with a few test runs. See, for example, the results in Table 9.

Finally, Duff, Erisman, and Reid (1986, p. 276) have described an example of Gear (1975) where the error matrix for minimizing the Frobenius norm of the error becomes

TABLE 7
Summary of results for Set 3.

	Before scaling			After scaling		
Num. iter.	0	1	≥ 2	0	1	≥ 2
Num. cases	1	13	4	2	12	4
	min	avr	max	min	avr	max
$Log_{10}(\eta)$	–23	–17	–16	–17	–17	–16
$Log_{10}(\omega_1^{(3)})$	–16	–16	–15	–16	–16	–15
$Log_{10}(\omega_2^{(3)})$	–30	–27	–17	–31	–28	–19
$Log_{10}\dfrac{\eta\,\kappa(A)}{Error}$	0.65	4.5	19	0.97	2.2	4.0
$Log_{10}\dfrac{\omega_1^{(3)}\,\kappa_{\omega_1}^{(3)} + \omega_2^{(3)}\,\kappa_{\omega_2}^{(3)}}{Error}$	0.58	2.2	7.9	0.50	1.6	2.7
$Log_{10}\dfrac{\eta\,\kappa(A)}{\omega_1^{(3)}\,\kappa_{\omega_1}^{(3)} + \omega_2^{(3)}\,\kappa_{\omega_2}^{(3)}}$	–0.17	2.3	11	–0.23	0.63	2.4

TABLE 8

Summary of results for Set 4. The $-\infty$ entries correspond to values of $\omega_1^{(4)} = 0$.

	drop tol. = 10^{-5}			*drop tol.* = 10^{-3}		
Num. iter.	0	1	≥ 2	0	1	≥ 2
Num. cases	2	6	10	2	1	15
	min	avr	max	min	avr	max
$Log_{10}(\eta)$	–18	–16	–16	–18	–15	–4.6
$Log_{10}(\omega_1^{(4)})$	$-\infty$	$-\infty$	–17	$-\infty$	$-\infty$	$-\infty$
$Log_{10}(\omega_2^{(4)})$	–16	–16	–15	–16	–15	–2.8
$Log_{10}\dfrac{\eta\,\kappa(A)}{Error}$	0.66	2.3	3.7	0.90	2.1	4.6
$Log_{10}\dfrac{\omega_1^{(4)}\,\kappa_{\omega_1}^{(4)}+\omega_2^{(4)}\,\kappa_{\omega_2}^{(4)}}{Error}$	0.66	1.6	2.8	0.64	1.6	3.8
$Log_{10}\dfrac{\eta\,\kappa(A)}{\omega_1^{(4)}\,\kappa_{\omega_1}^{(4)}+\omega_2^{(4)}\,\kappa_{\omega_2}^{(4)}}$	–0.14	0.64	2.5	–0.95	0.45	2.9

arbitrarily large if the perturbations are constrained to the original pattern. On this example, after one step of iterative refinement, using as a starting point the solution

$$\hat{\mathbf{x}} = \begin{pmatrix} (\delta-\sigma)/\delta \\ 1/\delta \\ 1/\delta \\ (\delta-\sigma)/\delta \end{pmatrix}, \qquad \sigma = 10^{-15},$$

we can guarantee that the error matrix $\mathbf{E}$ has the same pattern as the original matrix. That is,

$$\mathbf{E} \leqq \omega \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & |\delta| & 0 & 0 \\ 0 & 0 & |\delta| & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} = \omega\,|\mathbf{A}|,$$

with $\omega \leqq 10^{-16}$, $\delta = 10^{-8}$. It is interesting to note that $\kappa(\mathbf{A}) = 1 + 1/\delta$ and $\kappa_{|\mathbf{A}|}(\mathbf{A}) = 4$.

TABLE 9

Fill-in, numbers of iterations and error for the five point operator on a 30×30 *grid, using* $x_i = 1$, $i = 1, \cdots, n$ *and different values of drop tol.*

drop tol	0	10^{-2}	10^{-1}
Fill-in	23619	16085	4697
Num. iter.	2	14	16
Error	0.32D-14	0.25D-14	0.29D-01

6. Conclusions. We have shown that, when the iterative refinement is converging, it is possible and inexpensive to guarantee solutions of sparse linear systems that are exact solutions of a nearby system whose matrix has the same sparsity structure. Thus we have answered the open problem posed by Duff, Erisman, and Reid (1986, p. 276) concerning obtaining bounded perturbations while maintaining sparsity. If the equations arise from the discretization of a partial differential equation, then a componentwise tiny error should indicate that the solution obtained is that of a neighbouring partial differential equation, a conclusion that would not be available if classical error bounds were being used.

We have extended the work of Skeel (1980) and Demmel (1984) to include the possibility of having sparse right-hand sides and solutions vectors and have shown that, although we cannot always guarantee the solution to a nearby problem whose right-hand side sparsity is the same, we can develop suitable bounds for perturbations in the right-hand side.

We discuss methods of inexpensively and accurately calculating a condition number appropriate to this tighter backward error. This condition number is not bigger than that of Wilkinson and can indeed be much smaller, particularly if the matrix is badly row-scaled. For example, in Set 1, the average of the logarithms of the ratio of the classical condition number before and after scaling is 4.1, while for the Skeel condition number the corresponding value is 1.4.

We have incorporated our backward error estimator in the iterative refinement step of a direct sparse matrix solver and have found that we often require zero or one step of iterative refinement to guarantee that the computed solution is the solution of a nearby system with the same sparsity structure as the original matrix. We also have observed that we do not require any extra precision in calculating residuals, thus confirming remarks made by Skeel (1980). Additionally, when combined with our condition number estimator, a good estimate of the actual error is obtained. Furthermore, when iterative refinement diverges, our stopping criterion recognizes this early.

We observed, contrary to Zlatev, Wasniewski, and Schaumburg (1986), that little gain in sparsity was obtained while even moderate values of drop tolerance caused divergence of the iterative refinement. A drop tolerance strategy appears to work well only on very structured sparse matrices such as those resulting from discretizations of partial differential equations.

In this paper, we have been using iterative refinement to improve the solution obtained using an LU factorization. We have also considered the case when our LU factorization can be quite inaccurate (Set 4). In this case, we could use other techniques including SOR and CG and it is a open question as to how far our analysis could be continued to cover these cases.

Appendix. Tables of results of numerical experiments. In Tables A1–A15, the column corresponding to "Num.iter." gives the number of steps performed by the iterative refinement algorithm. By "Error" we denote the max-norm of the difference between the computed solution and the actual solution used to generate the right-hand side, divided by the max-norm of the actual solution.

TABLE A1

Set 1. Condition numbers before and after scaling.

	Before scaling		After scaling	
	$\kappa(A)$	$\kappa_{\omega_1}^{(1)}$	$\kappa(A)$	$\kappa_{\omega_1}^{(1)}$
GRE115	0.93D+02	0.17D+03	0.69D+04	0.26D+03
GRE185	0.38D+06	0.30D+06	0.39D+06	0.29D+06
GRE216A	0.28D+03	0.44D+03	0.20D+03	0.35D+03
GRE216B	0.83D+15	0.58D+14	0.56D+08	0.17D+08
GRE343	0.47D+03	0.74D+03	0.30D+03	0.51D+03
GRE512	0.46D+03	0.73D+03	0.40D+03	0.72D+03
GRE1107	0.18D+09	0.20D+09	0.77D+10	0.48D+09
WEST67	0.91D+03	0.15D+03	0.30D+03	0.16D+03
WEST132	0.11D+13	0.12D+08	0.94D+04	0.33D+04
WEST156	0.12D+32	0.38D+09	0.91D+12	0.30D+09
WEST167	0.69D+11	0.80D+06	0.46D+04	0.18D+04
WEST381	0.53D+07	0.75D+05	0.38D+06	0.85D+04
WEST479	0.49D+12	0.57D+07	0.27D+06	0.25D+05
WEST497	0.38D+12	0.20D+07	0.42D+07	0.12D+05
WEST655	0.49D+12	0.57D+07	0.42D+06	0.41D+05
WEST989	0.13D+13	0.16D+08	0.58D+06	0.70D+05
WEST1505	0.14D+13	0.16D+08	0.67D+08	0.35D+07
WEST2021	0.28D+13	0.32D+08	0.86D+06	0.12D+06

TABLE A2

Set 1. $x_i = 1, i = 1, \cdots, n$, *double precision before scaling.*

	Num. iter.	η	$\eta\kappa(A)$	$\omega_1^{(1)}$	$\omega_1^{(1)}\kappa_{\omega_1}^{(1)}$	Error
GRE115	1	0.52D-16	0.48D-14	0.59D-16	0.10D-13	0.79D-15
GRE185	1	0.12D-15	0.47D-10	0.16D-15	0.48D-10	0.16D-12
GRE216A	1	0.67D-16	0.19D-13	0.67D-16	0.29D-13	0.26D-15
GRE216B	1	0.73D-16	0.61D-01	0.11D-15	0.64D-02	0.21D-02
GRE343	1	0.10D-15	0.47D-13	0.10D-15	0.74D-13	0.50D-15
GRE512	1	0.83D-16	0.38D-13	0.83D-16	0.61D-13	0.26D-15
GRE1107	1	0.93D-16	0.17D-07	0.11D-15	0.22D-07	0.74D-10
WEST67	1	0.49D-16	0.45D-13	0.89D-16	0.13D-13	0.24D-14
WEST132	1	0.93D-17	0.98D-05	0.15D-15	0.18D-08	0.18D-09
WEST156	1	0.77D-18	0.90D+13	0.11D-15	0.42D-07	0.38D-09
WEST167	1	0.80D-16	0.55D-05	0.12D-15	0.95D-10	0.48D-11
WEST381	2	0.45D-16	0.24D-09	0.16D-15	0.12D-10	0.23D-11
WEST479	1	0.19D-16	0.94D-05	0.17D-15	0.96D-09	0.42D-10
WEST497	1	0.77D-16	0.29D-04	0.11D-15	0.22D-09	0.23D-10
WEST655	1	0.19D-16	0.94D-05	0.21D-15	0.12D-08	0.54D-10
WEST989	1	0.95D-16	0.13D-03	0.13D-15	0.21D-08	0.17D-09
WEST1505	1	0.93D-16	0.13D-03	0.16D-15	0.26D-08	0.17D-09
WEST2021	1	0.98D-16	0.27D-03	0.16D-15	0.52D-08	0.88D-10

TABLE A3

Set 1. $x_i = 1, i = 1, \cdots, n$, *double precision after scaling.*

	Num. iter.	η	$\eta\kappa(A)$	$\omega_1^{(1)}$	$\omega_1^{(1)}\kappa_{\omega_1}^{(1)}$	Error
GRE115	1	0.64E-16	0.44E-12	0.83E-16	0.22E-13	0.42E-14
GRE185	1	0.62E-16	0.24E-10	0.64E-16	0.18E-10	0.54E-13
GRE216A	1	0.54E-16	0.11E-13	0.79E-16	0.28E-13	0.13E-14
GRE216B	1	0.89E-16	0.50E-08	0.93E-16	0.16E-08	0.17E-09
GRE343	1	0.76E-16	0.23E-13	0.88E-16	0.45E-13	0.10E-14
GRE512	1	0.76E-16	0.31E-13	0.93E-16	0.66E-13	0.27E-14
GRE1107	1	0.39E-16	0.30E-06	0.10E-15	0.48E-07	0.25E-10
WEST67	1	0.35E-16	0.11E-13	0.14E-15	0.21E-13	0.89E-15
WEST132	1	0.28E-16	0.26E-12	0.98E-16	0.33E-12	0.73E-14
WEST156	0	0.57E-16	0.52E-04	0.16E-15	0.48E-07	0.98E-08
WEST167	1	0.29E-16	0.13E-12	0.11E-15	0.20E-12	0.44E-14
WEST381	1	0.15E-15	0.58E-10	0.17E-15	0.15E-11	0.56E-12
WEST479	1	0.35E-16	0.94E-11	0.22E-15	0.56E-11	0.12E-12
WEST497	1	0.17E-16	0.70E-10	0.11E-15	0.13E-11	0.26E-12
WEST655	1	0.52E-16	0.22E-10	0.19E-15	0.80E-11	0.19E-12
WEST989	1	0.25E-16	0.15E-10	0.12E-15	0.80E-11	0.33E-12
WEST1505	1	0.50E-16	0.34E-08	0.17E-15	0.60E-09	0.82E-10
WEST2021	1	0.50E-16	0.43E-10	0.18E-15	0.22E-10	0.19E-12

TABLE A4

Set 1. $x_i = 1, i = 1, \cdots, n$, *single precision after scaling.*

	Num. iter.	η	$\eta\kappa(A)$	$\omega_1^{(1)}$	$\omega_1^{(1)}\kappa_{\omega_1}^{(1)}$	Error
GRE115	1	0.15E-06	0.10E-02	0.29E-06	0.77E-04	0.18E-04
GRE185	2	0.33E-06	0.13E+00	0.33E-06	0.95E-01	0.40E-02
GRE216A	1	0.36E-06	0.73E-04	0.39E-06	0.14E-03	0.43E-05
GRE216B	2	0.59E-06	0.33E+02	0.83E-06	0.11E+02	0.43E-01
GRE343	1	0.39E-06	0.11E-03	0.42E-06	0.17E-03	0.29E-05
GRE512	1	0.74E-06	0.30E-03	0.74E-06	0.42E-03	0.15E-04
GRE1107	4	0.18E-05	0.13E+04	0.11E-03	0.13E+03	0.86E+00
WEST67	1	0.15E-06	0.45E-04	0.46E-06	0.19E-05	0.97E-05
WEST132	1	0.18E-06	0.17E-02	0.47E-06	0.41E-04	0.82E-04
WEST156	0	0.22E-07	0.20E+05	0.54E-06	0.42E+01	0.95E+00
WEST167	1	0.84E-07	0.38E-03	0.41E-06	0.19E-04	0.40E-04
WEST381	1	0.48E-07	0.19E-01	0.51E-06	0.11E-03	0.23E-02
WEST479	1	0.22E-06	0.61E-01	0.95E-06	0.62E-03	0.83E-03
WEST497	1	0.12E-06	0.49E+00	0.50E-06	0.15E-03	0.17E-02
WEST655	1	0.74E-07	0.31E-01	0.73E-06	0.78E-03	0.77E-03
WEST989	1	0.11E-06	0.63E-01	0.49E-06	0.89E-03	0.72E-03
WEST1505	1	0.11E-06	0.73E+01	0.70E-06	0.63E-01	0.10E+00
WEST2021	1	0.11E-06	0.93E-01	0.72E-06	0.22E-02	0.56E-03

TABLE A5

Set 1. $x_i = 1, i = 1, \cdots, n$, *mixed precision after scaling.*

	Num. iter.	η	$\eta\,\kappa(A)$	$\omega_1^{(1)}$	$\omega_1^{(1)}\,\kappa_{\omega_1}^{(1)}$	Error
GRE115	1	0.20E-06	0.14E-02	0.40E-06	0.10E-03	0.57E-05
GRE185	2	0.26E-06	0.10E+00	0.58E-06	0.17E+00	0.16E-02
GRE216A	1	0.33E-06	0.66E-04	0.72E-06	0.25E-03	0.31E-05
GRE216B	4	0.16E-06	0.89E+01	0.11E-05	0.14E+02	0.62E-01
GRE343	1	0.33E-06	0.97E-04	0.72E-06	0.27E-03	0.26E-05
GRE512	2	0.25E-06	0.10E-03	0.60E-06	0.31E-03	0.72E-05
GRE1107	4	0.17E-05	0.12E+04	0.20E-03	0.24E+03	0.84E+00
WEST67	1	0.20E-06	0.60E-04	0.51E-06	0.21E-05	0.86E-05
WEST132	1	0.15E-06	0.14E-02	0.75E-06	0.66E-04	0.13E-03
WEST156	1	0.11E-07	0.98E+04	0.59E-06	0.46E+01	0.18E+01
WEST167	1	0.12E-06	0.53E-03	0.58E-06	0.28E-04	0.16E-04
WEST381	1	0.17E-06	0.67E-01	0.73E-06	0.16E-03	0.31E-03
WEST479	1	0.77E-07	0.21E-01	0.63E-06	0.41E-03	0.24E-03
WEST497	1	0.12E-06	0.51E+00	0.67E-06	0.20E-03	0.21E-03
WEST655	1	0.74E-07	0.31E-01	0.82E-06	0.89E-03	0.69E-03
WEST989	1	0.94E-07	0.55E-01	0.88E-06	0.16E-02	0.65E-03
WEST1505	1	0.12E-06	0.80E+01	0.79E-06	0.71E-01	0.12E+00
WEST2021	1	0.99E-07	0.86E-01	0.80E-06	0.25E-02	0.15E-03

TABLE A6

Set 2. *Condition numbers before and after scaling.*

	Before scaling			After scaling		
	$\kappa(A)$	$\kappa_{\omega_1}^{(2)}$	$\kappa_{\omega_2}^{(2)}$	$\kappa(A)$	$\kappa_{\omega_1}^{(2)}$	$\kappa_{\omega_2}^{(2)}$
GRE115	0.93D+02	0.33D+02	0.23D+02	0.69D+04	0.58D+02	0.56D+02
GRE185	0.38D+06	0.50D+05	0.54D+05	0.39D+06	0.46D+05	0.52D+05
GRE216A	0.28D+03	0.90D+02	0.82D+02	0.20D+03	0.11D+03	0.10D+03
GRE216B	0.83D+15	0.37D+14	0.48D+13	0.56D+08	0.35D+07	0.37D+07
GRE343	0.47D+03	0.16D+03	0.13D+03	0.30D+03	0.10D+03	0.11D+03
GRE512	0.46D+03	0.14D+03	0.14D+03	0.40D+03	0.14D+03	0.14D+03
GRE1107	0.18D+09	0.40D+08	0.31D+08	0.77D+10	0.91D+08	0.83D+08
WEST67	0.91D+03	0.54D+02	0.78D+02	0.30D+03	0.51D+02	0.41D+02
WEST132	0.11D+13	0.26D+07	0.25D+07	0.94D+04	0.61D+03	0.83D+03
WEST156	0.12D+32	0.12D+09	0.13D+09	0.91D+12	0.28D+09	0.54D+07
WEST167	0.69D+11	0.45D+05	0.35D+06	0.46D+04	0.86D+03	0.40D+03
WEST381	0.53D+07	0.16D+05	0.63D+04	0.38D+06	0.23D+04	0.13D+04
WEST479	0.49D+12	0.12D+06	0.22D+07	0.27D+06	0.57D+04	0.34D+04
WEST497	0.38D+12	0.75D+06	0.33D+06	0.42D+07	0.73D+03	0.54D+04
WEST655	0.49D+12	0.66D+06	0.14D+07	0.42D+06	0.12D+05	0.32D+04
WEST989	0.13D+13	0.45D+07	0.47D+07	0.58D+06	0.21D+05	0.11D+05
WEST1505	0.14D+13	0.49D+07	0.53D+07	0.67D+08	0.27D+07	0.17D+05
WEST2021	0.28D+13	0.50D+07	0.89D+07	0.86D+06	0.42D+05	0.11D+05

TABLE A7

Set 2. $x_i = 1, i = 1, 6, \cdots$, *else* $x_i = 0$, *before scaling*.

	Num.	η	$\eta\,\kappa(\mathrm{A})$	$\omega_1^{(2)}$	$\omega_2^{(2)}$	$\omega_1^{(2)}\,\kappa_{\omega_1}^{(2)}+$	Error
	iter.					$\omega_2^{(2)}\,\kappa_{\omega_2}^{(2)}$	
GRE115	1	0.35D-16	0.32D-14	0.84D-16	0.89D-28	0.27D-14	0.71D-15
GRE185	1	0.94D-16	0.35D-10	0.19D-15	0.24D-25	0.94D-11	0.14D-12
GRE216A	1	0.12D-16	0.34D-14	0.56D-16	0.64D-27	0.50D-14	0.13D-15
GRE216B	4	0.51D-16	0.42D-01	0.41D-15	0.25D-26	0.15D-01	0.76D-06
GRE343	1	0.14D-16	0.65D-14	0.56D-16	0.74D-26	0.90D-14	0.11D-15
GRE512	1	0.25D-16	0.11D-13	0.83D-16	0.27D-25	0.12D-13	0.19D-15
GRE1107	2	0.42D-16	0.78D-08	0.20D-15	0.58D-24	0.82D-08	0.83D-10
WEST67	1	0.42D-16	0.38D-13	0.16D-15	0.27D-30	0.88D-14	0.12D-14
WEST132	1	0.24D-16	0.25D-04	0.13D-15	0.80D-28	0.34D-09	0.16D-10
WEST156	1	0.12D-22	0.14D+09	0.86D-16	0.15D-31	0.10D-07	0.10D-10
WEST167	0	0.28D-17	0.19D-06	0.20D-15	0.25D-18	0.92D-11	0.37D-12
WEST381	1	0.78D-17	0.41D-10	0.15D-15	0.40D-29	0.24D-11	0.29D-12
WEST479	3	0.33D-19	0.16D-07	0.33D-15	0.14D-28	0.39D-10	0.91D-12
WEST497	1	0.12D-17	0.44D-06	0.16D-15	0.28D-28	0.12D-09	0.30D-11
WEST655	3	0.88D-19	0.43D-07	0.26D-15	0.15D-25	0.17D-09	0.29D-11
WEST989	1	0.14D-16	0.19D-04	0.14D-15	0.29D-27	0.61D-09	0.26D-10
WEST1505	1	0.23D-16	0.31D-04	0.20D-15	0.67D-27	0.99D-09	0.46D-10
WEST2021	1	0.19D-16	0.52D-04	0.22D-15	0.32D-27	0.11D-08	0.24D-10

TABLE A8

Set 2. $x_i = 1, i = 1, 6, \cdots$, *else* $x_i = 0$, *after scaling*.

	Num.	η	$\eta\,\kappa(\mathrm{A})$	$\omega_1^{(2)}$	$\omega_2^{(2)}$	$\omega_1^{(2)}\,\kappa_{\omega_1}^{(2)}+$	Error
	iter.					$\omega_2^{(2)}\,\kappa_{\omega_2}^{(2)}$	
GRE115	1	0.32E-17	0.22E-13	0.96E-16	0.36E-27	0.56E-14	0.29E-15
GRE185	1	0.64E-16	0.25E-10	0.11E-15	0.41E-24	0.52E-11	0.57E-13
GRE216A	2	0.60E-16	0.12E-13	0.15E-15	0.10E-28	0.16E-13	0.81E-15
GRE216B	1	0.12E-15	0.68E-08	0.14E-15	0.94E-25	0.50E-09	0.77E-10
GRE343	1	0.60E-16	0.18E-13	0.22E-15	0.48E-26	0.23E-13	0.67E-15
GRE512	1	0.86E-16	0.35E-13	0.22E-15	0.25E-25	0.31E-13	0.67E-15
GRE1107	3	0.77E-16	0.59E-06	0.20E-14	0.18E-22	0.18E-06	0.10E-08
WEST67	1	0.40E-16	0.12E-13	0.16E-15	0.28E-30	0.79E-14	0.13E-14
WEST132	1	0.17E-16	0.16E-12	0.17E-15	0.78E-31	0.11E-12	0.54E-14
WEST156	0	0.61E-17	0.56E-05	0.10E-15	0.14E-29	0.30E-07	0.32E-08
WEST167	0	0.21E-16	0.94E-13	0.18E-15	0.50E-19	0.16E-12	0.24E-14
WEST381	1	0.35E-16	0.13E-10	0.12E-15	0.57E-29	0.27E-12	0.86E-13
WEST479	2	0.37E-17	0.10E-11	0.16E-15	0.33E-30	0.90E-12	0.28E-13
WEST497	1	0.52E-17	0.22E-10	0.11E-15	0.13E-30	0.81E-13	0.22E-14
WEST655	2	0.13E-16	0.55E-11	0.19E-15	0.60E-29	0.22E-11	0.61E-14
WEST989	1	0.32E-16	0.19E-10	0.20E-15	0.63E-29	0.43E-11	0.48E-13
WEST1505	1	0.32E-16	0.21E-08	0.20E-15	0.36E-28	0.54E-09	0.97E-11
WEST2021	1	0.32E-16	0.27E-10	0.20E-15	0.95E-29	0.85E-11	0.18E-13

TABLE A9

Set 3. *Condition numbers before and after scaling.*

	Before scaling			After scaling		
	$\kappa(A)$	$\kappa^{(3)}_{\omega_1}$	$\kappa^{(3)}_{\omega_2}$	$\kappa(A)$	$\kappa^{(3)}_{\omega_1}$	$\kappa^{(3)}_{\omega_2}$
GRE115	0.93D+02	0.33D+02	0.38D+02	0.69D+04	0.58D+02	0.29D+04
GRE185	0.38D+06	0.50D+05	0.93D+05	0.39D+06	0.46D+05	0.92D+05
GRE216A	0.28D+03	0.90D+02	0.84D+02	0.20D+03	0.11D+03	0.82D+02
GRE216B	0.83D+15	0.37D+14	0.18D+15	0.56D+08	0.35D+07	0.19D+08
GRE343	0.47D+03	0.16D+03	0.10D+03	0.30D+03	0.10D+03	0.85D+02
GRE512	0.46D+03	0.14D+03	0.14D+03	0.40D+03	0.14D+03	0.12D+03
GRE1107	0.18D+09	0.40D+08	0.42D+08	0.77D+10	0.91D+08	0.21D+10
WEST67	0.91D+03	0.54D+02	0.45D+02	0.30D+03	0.51D+02	0.24D+02
WEST132	0.11D+13	0.26D+07	0.39D+11	0.94D+04	0.61D+03	0.27D+04
WEST156	0.12D+32	0.12D+09	0.44D+25	0.91D+12	0.28D+09	0.23D+11
WEST167	0.69D+11	0.45D+05	0.68D+09	0.46D+04	0.86D+03	0.15D+04
WEST381	0.53D+07	0.16D+05	0.29D+07	0.38D+06	0.23D+04	0.30D+05
WEST479	0.49D+12	0.12D+06	0.28D+12	0.27D+06	0.57D+04	0.28D+05
WEST497	0.38D+12	0.75D+06	0.10D+12	0.42D+07	0.73D+03	0.85D+06
WEST655	0.49D+12	0.66D+06	0.18D+12	0.42D+06	0.12D+05	0.20D+05
WEST989	0.13D+13	0.45D+07	0.73D+12	0.58D+06	0.21D+05	0.11D+06
WEST1505	0.14D+13	0.49D+07	0.11D+13	0.67D+08	0.27D+07	0.17D+06
WEST2021	0.28D+13	0.50D+07	0.14D+13	0.86D+06	0.42D+05	0.12D+06

TABLE A10

Set 3. $x_i = 1, i = 1, 6, \cdots$, *else* $x_i = 0$, *before scaling.*

	Num. iter.	η	$\eta\,\kappa(A)$	$\omega_1^{(3)}$	$\omega_2^{(3)}$	$\omega_1^{(3)}\,\kappa^{(3)}_{\omega_1}+$ $\omega_2^{(3)}\,\kappa^{(3)}_{\omega_2}$	Error
GRE115	1	0.35D-16	0.32D-14	0.84D-16	0.89D-28	0.27D-14	0.71D-15
GRE185	1	0.94D-16	0.35D-10	0.19D-15	0.24D-25	0.94D-11	0.14D-12
GRE216A	1	0.12D-16	0.34D-14	0.56D-16	0.64D-27	0.50D-14	0.13D-15
GRE216B	4	0.51D-16	0.42D-01	0.41D-15	0.25D-26	0.15D-01	0.76D-06
GRE343	1	0.14D-16	0.65D-14	0.56D-16	0.12D-25	0.90D-14	0.11D-15
GRE512	1	0.25D-16	0.11D-13	0.83D-16	0.34D-25	0.12D-13	0.19D-15
GRE1107	2	0.42D-16	0.78D-08	0.20D-15	0.58D-24	0.82D-08	0.83D-10
WEST67	1	0.42D-16	0.38D-13	0.16D-15	0.50D-30	0.88D-14	0.12D-14
WEST132	1	0.24D-16	0.25D-04	0.13D-15	0.80D-28	0.34D-09	0.16D-10
WEST156	1	0.12D-22	0.14D+09	0.86D-16	0.17D-27	0.75D-03	0.10D-10
WEST167	0	0.28D-17	0.19D-06	0.20D-15	0.18D-16	0.12D-07	0.37D-12
WEST381	1	0.78D-17	0.41D-10	0.15D-15	0.40D-29	0.24D-11	0.29D-12
WEST479	3	0.33D-19	0.16D-07	0.33D-15	0.14D-28	0.39D-10	0.91D-12
WEST497	1	0.12D-17	0.44D-06	0.16D-15	0.28D-28	0.12D-09	0.30D-11
WEST655	3	0.88D-19	0.43D-07	0.26D-15	0.15D-25	0.17D-09	0.29D-11
WEST989	1	0.14D-16	0.19D-04	0.14D-15	0.29D-27	0.61D-09	0.26D-10
WEST1505	1	0.23D-16	0.31D-04	0.20D-15	0.67D-27	0.99D-09	0.46D-10
WEST2021	1	0.19D-16	0.52D-04	0.22D-15	0.32D-27	0.11D-08	0.24D-10

TABLE A11

Set 3. $x_i = 1, i = 1, 6, \cdots$, *else* $x_i = 0$, *after scaling.*

	Num. iter.	η	$\eta\kappa(A)$	$\omega_1^{(3)}$	$\omega_2^{(3)}$	$\omega_1^{(3)}\kappa_{\omega_1}^{(3)}+$ $\omega_2^{(3)}\kappa_{\omega_2}^{(3)}$	Error
GRE115	1	0.32E-17	0.22E-13	0.96E-16	0.36E-27	0.56E-14	0.29E-15
GRE185	1	0.64E-16	0.25E-10	0.11E-15	0.41E-24	0.52E-11	0.57E-13
GRE216A	2	0.60E-16	0.12E-13	0.15E-15	0.12E-28	0.16E-13	0.81E-15
GRE216B	1	0.12E-15	0.68E-08	0.14E-15	0.94E-25	0.50E-09	0.77E-10
GRE343	1	0.60E-16	0.18E-13	0.22E-15	0.71E-26	0.23E-13	0.67E-15
GRE512	1	0.86E-16	0.35E-13	0.22E-15	0.31E-25	0.31E-13	0.67E-15
GRE1107	3	0.77E-16	0.59E-06	0.20E-14	0.18E-22	0.18E-06	0.10E-08
WEST67	1	0.40E-16	0.12E-13	0.16E-15	0.57E-30	0.79E-14	0.13E-14
WEST132	1	0.17E-16	0.16E-12	0.17E-15	0.78E-31	0.11E-12	0.54E-14
WEST156	0	0.61E-17	0.56E-05	0.10E-15	0.14E-29	0.30E-07	0.32E-08
WEST167	0	0.21E-16	0.94E-13	0.18E-15	0.50E-19	0.16E-12	0.24E-14
WEST381	1	0.35E-16	0.13E-10	0.12E-15	0.57E-29	0.27E-12	0.86E-13
WEST479	2	0.37E-17	0.10E-11	0.16E-15	0.33E-30	0.90E-12	0.28E-13
WEST497	1	0.52E-17	0.22E-10	0.11E-15	0.13E-30	0.81E-13	0.22E-14
WEST655	2	0.13E-16	0.55E-11	0.19E-15	0.60E-29	0.22E-11	0.61E-14
WEST989	1	0.32E-16	0.19E-10	0.20E-15	0.63E-29	0.43E-11	0.48E-13
WEST1505	1	0.32E-16	0.21E-08	0.20E-15	0.36E-28	0.54E-09	0.97E-11
WEST2021	1	0.32E-16	0.27E-10	0.20E-15	0.95E-29	0.85E-11	0.18E-13

TABLE A12

Set 4. *Condition numbers after scaling for drop tol.* = 10^{-5} *and drop tol.* = 10^{-3}.

		drop tol $=10^{-5}$		*drop tol* $=10^{-3}$	
	$\kappa(A)$	$\kappa_{\omega_1}^{(4)}$	$\kappa_{\omega_2}^{(4)}$	$\kappa_{\omega_1}^{(4)}$	$\kappa_{\omega_2}^{(4)}$
GRE115	0.69E+04	0.00E+00	0.12E+03	0.00E+00	0.12E+03
GRE185	0.39E+06	0.00E+00	0.17E+06	0.00E+00	0.14E+06
GRE216A	0.20E+03	0.00E+00	0.21E+03	0.00E+00	0.21E+03
GRE216B	0.84E+08	0.00E+00	0.15E+08	0.00E+00	0.10E+07
GRE343	0.30E+03	0.00E+00	0.31E+03	0.00E+00	0.26E+03
GRE512	0.40E+03	0.00E+00	0.43E+03	0.00E+00	0.37E+03
GRE1107	0.63E+10	0.00E+00	0.23E+09	0.00E+00	0.55E+07
WEST67	0.30E+03	0.29E+01	0.16E+03	0.00E+00	0.14E+03
WEST132	0.94E+04	0.00E+00	0.24E+04	0.00E+00	0.22E+04
WEST156	0.91E+12	0.00E+00	0.29E+09	0.00E+00	0.16E+06
WEST167	0.46E+04	0.00E+00	0.16E+04	0.00E+00	0.13E+04
WEST381	0.38E+06	0.00E+00	0.65E+04	0.00E+00	0.54E+04
WEST479	0.27E+06	0.00E+00	0.23E+05	0.00E+00	0.20E+05
WEST497	0.42E+07	0.00E+00	0.65E+04	0.00E+00	0.63E+04
WEST655	0.42E+06	0.00E+00	0.43E+05	0.00E+00	0.37E+05
WEST989	0.58E+06	0.00E+00	0.63E+05	0.00E+00	0.53E+05
WEST1505	0.67E+08	0.00E+00	0.35E+07	0.00E+00	0.21E+07
WEST2021	0.86E+06	0.00E+00	0.12E+06	0.00E+00	0.10E+06

TABLE A13

Set 4. $x_i = 1, i = 1, 6, \cdots$, *else* $x_i = 0$, *after scaling and drop tol.* $= 10^{-5}$.

	Num. iter.	η	$\eta\,\kappa(A)$	$\omega_1^{(4)}$	$\omega_2^{(4)}$	$\omega_1^{(4)}\,\kappa_{\omega_1}^{(4)}+$ $\omega_2^{(4)}\,\kappa_{\omega_2}^{(4)}$	Error
GRE115	2	0.99E-18	0.68E-14	0.00E+00	0.69E-16	0.85E-14	0.15E-14
GRE185	3	0.55E-16	0.22E-10	0.00E+00	0.50E-16	0.83E-11	0.80E-13
GRE216A	1	0.90E-16	0.18E-13	0.00E+00	0.83E-16	0.18E-13	0.88E-15
GRE216B	29	0.10E-15	0.84E-08	0.00E+00	0.50E-15	0.77E-08	0.63E-10
GRE343	1	0.90E-16	0.27E-13	0.00E+00	0.83E-16	0.26E-13	0.81E-15
GRE512	1	0.86E-16	0.35E-13	0.00E+00	0.11E-15	0.48E-13	0.68E-15
GRE1107	15	0.62E-16	0.39E-06	0.00E+00	0.19E-15	0.44E-07	0.27E-09
WEST67	1	0.50E-16	0.15E-13	0.13E-16	0.61E-16	0.10E-13	0.85E-15
WEST132	2	0.36E-16	0.33E-12	0.00E+00	0.67E-16	0.16E-12	0.46E-14
WEST156	0	0.61E-17	0.56E-05	0.00E+00	0.54E-16	0.16E-07	0.32E-08
WEST167	0	0.21E-16	0.94E-13	0.00E+00	0.67E-16	0.11E-12	0.24E-14
WEST381	2	0.23E-16	0.89E-11	0.00E+00	0.54E-16	0.36E-12	0.78E-13
WEST479	3	0.26E-16	0.71E-11	0.00E+00	0.57E-16	0.13E-11	0.55E-13
WEST497	1	0.58E-17	0.25E-10	0.00E+00	0.55E-16	0.36E-12	0.47E-14
WEST655	2	0.55E-16	0.23E-10	0.00E+00	0.91E-16	0.39E-11	0.22E-13
WEST989	1	0.13E-15	0.75E-10	0.00E+00	0.19E-15	0.12E-10	0.18E-13
WEST1505	2	0.64E-16	0.43E-08	0.00E+00	0.10E-15	0.35E-09	0.10E-10
WEST2021	2	0.95E-16	0.82E-10	0.00E+00	0.13E-15	0.16E-10	0.59E-13

TABLE A14

Set 4. $x_i = 1, i = 1, 6, \cdots$, *else* $x_i = 0$, *after scaling and drop tol.* $= 10^{-3}$.

	Num. iter.	η	$\eta\,\kappa(A)$	$\omega_1^{(4)}$	$\omega_2^{(4)}$	$\omega_1^{(4)}\,\kappa_{\omega_1}^{(4)}+$ $\omega_2^{(4)}\,\kappa_{\omega_2}^{(4)}$	Error
GRE115	4	0.35E-17	0.24E-13	0.00E+00	0.48E-16	0.59E-14	0.80E-15
GRE185	15	0.46E-16	0.15E-10	0.00E+00	0.61E-16	0.87E-11	0.14E-12
GRE216A	1	0.65E-16	0.13E-13	0.00E+00	0.74E-16	0.16E-13	0.11E-14
GRE216B	3	0.26E-04	0.15E+03	0.00E+00	0.11E-02	0.12E+04	0.22E+01
GRE343	3	0.66E-16	0.20E-13	0.00E+00	0.87E-16	0.23E-13	0.72E-15
GRE512	4	0.63E-16	0.26E-13	0.00E+00	0.89E-16	0.32E-13	0.79E-15
GRE1107	3	0.64E-05	0.10E+04	0.00E+00	0.16E-02	0.90E+04	0.13E+01
WEST67	2	0.37E-16	0.11E-13	0.00E+00	0.45E-16	0.61E-14	0.14E-14
WEST132	3	0.25E-16	0.23E-12	0.00E+00	0.52E-16	0.11E-12	0.21E-14
WEST156	0	0.59E-18	0.73E-08	0.00E+00	0.54E-16	0.87E-11	0.18E-12
WEST167	0	0.21E-16	0.94E-13	0.00E+00	0.67E-16	0.84E-13	0.24E-14
WEST381	4	0.17E-16	0.67E-11	0.00E+00	0.53E-16	0.29E-12	0.33E-13
WEST479	7	0.34E-17	0.91E-12	0.00E+00	0.55E-16	0.11E-11	0.51E-13
WEST497	4	0.30E-17	0.13E-10	0.00E+00	0.58E-16	0.36E-12	0.36E-14
WEST655	5	0.28E-16	0.12E-10	0.00E+00	0.65E-16	0.24E-11	0.55E-13
WEST989	5	0.32E-16	0.19E-10	0.00E+00	0.64E-16	0.34E-11	0.11E-12
WEST1505	10	0.32E-16	0.20E-08	0.00E+00	0.90E-16	0.19E-09	0.23E-10
WEST2021	5	0.32E-16	0.27E-10	0.00E+00	0.94E-16	0.98E-11	0.72E-13

TABLE A15

Set 4. *Number of nonzero entries in the original matrices and fill-in for drop tol.* = 0.0, *drop tol.* = 10^{-5} *and drop tol.* = 10^{-3} *after scaling.*

	Nonzeros	Fill-in		
		drop tol = 0.0	*drop tol* = 10^{-5}	*drop tol* = 10^{-3}
GRE115	421	647	651	605
GRE185	975	3173	3028	2929
GRE216A	812	2544	2263	2262
GRE216B	812	2767	2580	2180
GRE343	1310	5334	4891	4890
GRE512	1976	11535	11020	11007
GRE1107	5664	47603	45255	41181
WEST67	294	267	202	204
WEST132	413	89	87	83
WEST156	362	27	20	15
WEST167	506	96	96	92
WEST381	2134	2057	1867	1711
WEST479	1888	1121	982	790
WEST497	1721	279	263	252
WEST655	2808	2092	1791	1709
WEST989	3518	1156	1139	1135
WEST1505	5414	2032	1934	1821
WEST2021	7310	2539	2466	2410

Acknowledgment. We would like to thank John Reid for reading a draft of this paper and making some helpful remarks.

REFERENCES

A. K. CLINE, C. B. MOLER, G. W. STEWART, AND J. H. WILKINSON (1979), *An estimate for the condition number of a matrix*, SIAM J. Numer. Anal., 16, pp. 368–375.

A. R. CURTIS AND J. K. REID (1972), *On the automatic scaling of matrices for Gaussian elimination*, J. Inst. Maths. Applics., 10, pp. 118–124.

J. W. DEMMEL (1984), *Underflow and the reliability of numerical software*, SIAM J. Sci. Statist. Comput., 5, pp. 887–919.

J. J. DONGARRA, J. R. BUNCH, C. B. MOLER, AND G. W. STEWART (1979), LINPACK *User's Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.

I. S. DUFF (1977), MA28—*a set of Fortran subroutines for sparse unsymmetric linear equations*, Report AERE R8730, Her Majesty's Stationery Office, London.

I. S. DUFF, A. M. ERISMAN, C. W. GEAR, AND J. K. REID (1985), *Some remarks on inverses of sparse matrices*, Report CSS 171, CSS Division, Harwell Laboratory, England; ACM SIGNUM newsletter, 23 (1988), pp. 2–8.

I. S. DUFF, A. M. ERISMAN, AND J. K. REID (1986), *Direct methods for sparse matrices*, Oxford University Press, London.

I. S. DUFF, R. G. GRIMES, AND J. G. LEWIS (1987), *Sparse matrix test problems*, Report CSS 191, CSS Division, Harwell Laboratory, England; ACM Trans. Math. Software, to appear.

C. W. GEAR (1975), *Numerical errors in sparse linear equations*, Report UIUCDCS-F-75-885, Department of Computer Science, University of Illinois, Urbana-Champaign, IL.

W. W. HAGER (1984), *Condition estimators*, SIAM J. Sci. Statist. Comput., 5, pp. 311–316.

N. J. HIGHAM (1987a), *A survey of condition number estimation for triangular matrices*, SIAM Rev., 29, pp. 575–596.

——— (1987b), *Fortran codes for estimating the one-norm of a real or complex matrix, with applications to condition estimation*, Numerical Analysis Report No. 135, University of Manchester, England.

IEEE (1985), *Standard for Binary Floating Point Arithmetic*, ANSI/IEEE Std 754-1985, IEEE, NY.

——— (1987), *Radix and Format Independent Standard for Floating Point Arithmetic*, ANSI/IEEE Std 854-1987, IEEE, NY.

W. KAHAN (1981), *Why do we need a floating point arithmetic standard*? IEEE Floating Point Subcommittee Working Document P754/81-2.8, IEEE, NY.

H. M. MARKOWITZ, (1957), *The elimination form of the inverse and its application to linear programming*. Management Sci., 3, pp. 255–269.

W. OETTLI AND W. PRAGER (1964), *Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides*, Numer. Math., 6, pp. 405–409.

R. D. SKEEL (1979), *Scaling for numerical stability in Gaussian elimination*, J. Assoc. Comput. Mach., 26, pp. 494–526.

——— (1980), *Iterative refinement implies numerical stability for Gaussian elimination*, Math. Comput., 35, pp. 817–832.

J. H. WILKINSON (1961), *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8, pp. 281–330.

Z. ZLATEV, J. WASNIEWSKI, AND K. SCHAUMBURG (1986), *Condition number estimators in a sparse matrix software*, SIAM J. Sci. Statist. Comput., 7, pp. 1175–1189.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 2, pp. 191–209, April 1989

005

CONDITION ESTIMATES FOR MATRIX FUNCTIONS*

CHARLES KENNEY† AND ALAN J. LAUB†

Abstract. A sensitivity theory based on Fréchet derivatives is presented that has both theoretical and computational advantages. Theoretical results such as a generalization of Van Loan's work on the matrix exponential are easily obtained: matrix functions are least sensitive at normal matrices. Computationally, the central problem is to estimate the norm of the Fréchet derivative, since this is equal to the function's condition number. Two norm-estimation procedures are given; the first is based on a finite-difference approximation of the Fréchet derivative and costs only two extra function evaluations. The second method was developed specifically for the exponential and logarithmic functions; it is based on a trapezoidal approximation scheme suggested by the chain rule for the identity $e^X = (e^{X/2^n})^{2^n}$. This results in an infinite sequence of coupled Sylvester equations that, when truncated, is uniquely suited to the "scaling and squaring" procedure for e^X or the "inverse scaling and squaring" procedure for log X.

Both the trapezoid approximation method and the more general finite-difference approach yield excellent condition estimates for a large class of problems taken from the literature. The problems in this set illustrate that condition estimates based on the Fréchet derivative have the virtue of reliability and general applicability.

Key words. condition estimation, matrix-valued function, exponential, logarithm, Fréchet derivative

AMS(MOS) subject classifications. 65F35, 65F30, 15A12

1. Introduction.

In this paper, we are concerned with the effects of perturbations on matrix functions

$$F(X) \equiv \sum_{n=0}^{\infty} a_n X^n \tag{1.1}$$

where $a_n \in \mathbb{R}$ and $X \in \mathbb{R}^{p \times p}$. We assume that the scalar power series

$$F(x) = \sum_{n=0}^{\infty} a_n x^n \tag{1.2}$$

is absolutely convergent for $|x| < r$ for some $r > 0$. We are interested in estimating the "worst case" perturbations as defined by the condition numbers [28]

$$\begin{aligned} K_\delta = K_\delta(F,X) &\equiv \max_{\|Z\| \le 1} \frac{\|F(X+\delta Z) - F(X)\|}{\delta}, \\ K = K(F,X) &\equiv \lim_{\delta \to 0^+} K_\delta(F,X) \end{aligned} \tag{1.3}$$

where we assume that $\delta > 0$ and $\|X\| + \delta < r$, so that $F(X + \delta Z)$ is well defined. We shall use the Frobenius matrix norm

$$\|M\|^2 \equiv \sum M_{ij}^2 \tag{1.4}$$

throughout the paper unless explicitly noted otherwise, since this norm has nice properties vis-à-vis the Kronecker matrix product.

* Received by the editors February 9, 1988; accepted for publication (in revised form) September 13, 1988. This research was supported by the Office of Naval Research under contract N00014-85-K-0553 and by the National Science Foundation (and the Air Force Office of Scientific Research) under grant ECS87-18897.

† Department of Electrical and Computer Engineering, University of California, Santa Barbara, California 93106 (laub%lanczos@hub.ucsb.edu).

The condition number $K(F, X)$ of F at X is determined by the Fréchet derivative of F at X: we say that a linear mapping $L : \mathbb{R}^{p\times p} \rightarrow \mathbb{R}^{p\times p}$ is the Fréchet derivative of F at X (see [2], [12]) if for all Z in $\mathbb{R}^{p\times p}$

$$\lim_{\delta\rightarrow 0} \left\| \frac{F(X+\delta Z)-F(X)}{\delta} - L(Z) \right\| = 0. \tag{1.5}$$

When it is convenient to explicitly indicate the dependence of L on X, we will write $L(Z, X)$ instead of $L(Z)$. For brevity, we will refer to L as the derivative of F.

Example 1. The squaring function $F(X) = X^2$ satisfies $(F(X + \delta Z) - F(X))/\delta = XZ + ZX + \delta Z^2$, so its derivative at X is given by $L(Z) = XZ + ZX$.

Example 2. The derivative at X of the exponential function $F(X) = e^X$ is given by [31]

$$L(Z) = \int_0^1 e^{X(1-s)} Z e^{Xs}\, ds. \tag{1.6}$$

Other examples are given in Appendix B.

From the definition of the Fréchet derivative (see [31, Thm. 5]), we have

$$K(F,X) = \|L(\cdot,X)\| \equiv \max_{Z\neq 0} \frac{\|L(Z,X)\|}{\|Z\|}. \tag{1.7}$$

Because of this, most of our effort is devoted to studying L and methods for estimating its norm.

In § 2, the eigenvalues of X are used to obtain a lower bound on $K(F, X)$; this lower bound is in fact equal to $K(F, X)$ when X is normal. Thus matrix functions exhibit minimal sensitivity when they are evaluated at normal matrices, an effect demonstrated by Van Loan [31] for the exponential function $F(X) = e^X$. Similar results are given for large scale perturbations.

In § 2, we also lay the groundwork for estimating the norm of L via the power method: given Z_0 of unit norm, let

$$W \equiv L(Z_0, X), \tag{1.8}$$

$$Z_1 \equiv L(W, X^T). \tag{1.9}$$

For suitably chosen Z_0, $\|Z_1\|^{1/2} \cong \|L(\cdot, X)\|$, and more accurate estimates can be obtained by repeating the cycle with $Z_0 = Z_1/\|Z_1\|$.

The main problem with this approach is that evaluating $L(Z)$ directly may be rather difficult. For example, in the case of the matrix exponential, it is not at all clear how we should go about evaluating the integral representation in (1.6). In § 3, we consider the problem of forming $L(Z)$ for both the exponential and logarithmic matrix functions. For the exponential problem, $L(Z)$ can be accurately approximated by using a compound trapezoid approximation in (1.7); this approach can be efficiently implemented during the squaring phase of the "scaling and squaring" method of evaluating e^X. This association with scaling and squaring is quite natural because the trapezoid approximation can be derived from the chain rule for the identity $e^X = (e^{X/2^n})^{2^n}$. For the logarithmic problem, a similar approximation can be done during the square root phase of the "inverse scaling and squaring" method of evaluating $\log X$.

While these sensitivity estimation procedures can be easily incorporated into standard packages, such as MATEXP by Ward [32], the numerical effort involved in using them can vary considerably depending on the amount of scaling to be done. For example, one power method cycle of evaluating L and L^T for the matrix exponential can range in cost from $\frac{1}{4}$ to as much as three times the effort needed to evaluate e^X.

By contrast, there is another way of evaluating L such that, independent of the function F, $\|L\|$ can be estimated at a cost of only two extra function evaluations. The idea behind this method is to use the relation

$$\frac{F(X+\delta Z)-F(X)}{\delta}=L(Z,X)+O(\delta) \tag{1.10}$$

as a means of approximating $L(Z,X)$. Thus the power method steps (1.8) and (1.9) can be approximated by

$$W=\frac{F(X+\delta Z_0)-F(X)}{\delta}, \tag{1.11}$$

$$Z_1=\frac{F(X^T+\delta W)-F(X^T)}{\delta}, \tag{1.12}$$

for δ sufficiently small.

To provide a practical assessment of the trapezoid and finite-difference condition estimators, a large set of problems from [3], [7], [16], [25], and [32] was tested numerically; a selected subset of the results is given in § 4. For almost all of the examples, our condition estimates, based on one power method cycle, were within 90 percent of the actual condition number and none of the estimates was less than 25 percent of the actual condition number (see Tables 1 and 2). Of particular interest is an example considered by Ward [32, Example 3] that has shown that the sensitivity estimation scheme employed in the subroutine MATEXP can give very conservative bounds. In this case, Ward's method predicts that not more than 12 digits of accuracy would be lost in the computation of the matrix exponential, whereas one cycle of the power method predicted that at most four digits would be lost; in fact, the numerically computed result had lost exactly four digits of accuracy. This illustrates that condition estimates based on the Fréchet derivative appear to be extremely reliable.

2. General perturbation results. For $\|Z\|=1$ and $\|X\|+\delta<r$, we may write by (1.2),

(2.1)

$$F(X+\delta Z)=F(X)+\delta\sum_{n=1}^{\infty}a_n\sum_{k=0}^{n-1}X^kZX^{n-1-k}+\cdots$$
$$+\delta^m\sum_{n=m}^{\infty}a_n\sum_{k_1+\cdots+k_m=0}^{n-m}X^{k_1}ZX^{k_2}Z\cdots X^{k_m}ZX^{n-m-k_1-\cdots-k_m}+\cdots$$

where the absolute convergence of the series justifies the rearrangement of the terms in (2.1). From (2.1), and (1.5),

$$L(Z,X)=\sum_{n=1}^{\infty}a_n\sum_{k=0}^{n-1}X^kZX^{n-1-k}. \tag{2.2}$$

The discussion in the previous section has shown that the condition number (1.3) satisfies

$$K(F,X)=\|L(\cdot,X)\|=\max_{Z\neq 0}\frac{\|L(Z,X)\|}{\|Z\|}.$$

We use the Frobenius norm (1.4) because of its natural connection to the spectral or two-norm of the Kronecker form of the Fréchet derivative. Let Vec A denote the vector formed by stacking the columns of a matrix A, and define the Kronecker product of two matrices A and B by (see [15]) $A \otimes B \equiv [a_{ij}B]$. Then the Frobenius norm of a matrix Z is equal to the two-norm of Vec Z:

$$\|Z\| = \|\mathrm{Vec}\, Z\|_2. \tag{2.3}$$

Also, Vec $(AZB) = (B^T \otimes A)$ Vec Z. Thus,

$$\mathrm{Vec}\, L(Z,X) = D(X)\, \mathrm{Vec}\, Z \tag{2.4}$$

where $D(X)$ is the Kronecker form of the Fréchet derivative

$$D(X) \equiv \sum_{n=1}^{\infty} a_n \sum_{k=0}^{n-1} (X^T)^{n-1-k} \otimes X^k. \tag{2.5}$$

By (2.3) and (2.4), we have

$$\max_{Z \neq 0} \frac{\|L(Z,X)\|}{\|Z\|} = \max_{Z \neq 0} \frac{\|\mathrm{Vec}\, L(Z,X)\|_2}{\|\mathrm{Vec}\, Z\|_2} = \max_{Z \neq 0} \frac{\|D(X)\, \mathrm{Vec}\, Z\|_2}{\|\mathrm{Vec}\, Z\|_2},$$

so that the Frobenius norm of the Fréchet derivative is equal to the two-norm of its Kronecker matrix form:

$$\|L(\cdot,X)\| = \|D(X)\|_2. \tag{2.6}$$

The importance of this identity lies in the fact that the two-norm of a real matrix A is the square root of the largest eigenvalue λ of A^TA, and hence can be estimated by using the power method. For a given vector v_0 with $\|v_0\|_2 = 1$, compute the vectors $u_k \equiv Av_k$, $\tilde{v}_{k+1} \equiv A^Tu_k$, $v_{k+1} \equiv \tilde{v}_{k+1}/\|\tilde{v}_{k+1}\|_2$ for $k = 0, 1, 2, \cdots$. If v_0 is not orthogonal to the eigenspace E_λ of A^TA corresponding to λ where $\lambda^{1/2} = \|A\|_2$, then $\|\tilde{v}_k\|^{1/2} \rightarrow \|A\|_2$; and unless v_0 is poorly chosen, $\|\tilde{v}_1\|^{1/2} \cong \|A\|_2$. That is, one cycle of the power method provides an approximation of $\|A\|_2$ that is usually sufficient for the purposes of condition estimation [8].

Using $(A \otimes B)^T = A^T \otimes B^T$ and (2.5), we have

$$\begin{aligned}(D(X))^T &= \sum_{n=1}^{\infty} a_n \sum_{k=0}^{n-1} X^{n-1-k} \otimes (X^T)^k \\ &= D(X^T).\end{aligned} \tag{2.7}$$

Define v_1 by $u_0 \equiv D(X)v_0$, $\tilde{v}_1 \equiv (D(X))^Tu_0$, $v_1 \equiv \tilde{v}_1/\|\tilde{v}_1\|_2$. From (2.4) and (2.7) this is equivalent to forming Z_1 by

$$W \equiv L(Z_0,X), \quad \tilde{Z}_1 \equiv L(W,X^T), \quad Z_1 \equiv \tilde{Z}_1/\|\tilde{Z}_1\|, \tag{2.8}$$

where $v_0 = \mathrm{Vec}\,(Z_0)$ and $v_1 = \mathrm{Vec}\,(Z_1)$. This is fortunate, because it means that we can avoid dealing with the $p^2 \times p^2$ Kronecker matrix $D(X)$ when estimating the condition of F at X by the power method. Instead, we may use the more compact formulation (2.8).

Now we establish a lower bound on $K(F, X)$ and show that this lower bound is in fact equal to $K(F, X)$ when X is normal.

LEMMA 2.1. *Let v and w be nonzero vectors such that $Xv = \lambda v$ and $X^T w = \mu w$. Then $w \otimes v$ is an eigenvector of $D(X)$ with associated eigenvalue ν where*

$$\nu = F'(\lambda) \qquad \textit{for}\ \lambda = \mu,$$

$$\nu = \frac{F(\lambda) - F(\mu)}{\lambda - \mu} \qquad \textit{for}\ \lambda \neq \mu. \tag{2.9}$$

Proof. Since $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ for any compatible matrices A, B, C, D (see [15]), we have

$$\begin{aligned} D(X)(w \otimes v) &= \sum_{n=1}^{\infty} a_n \sum_{k=0}^{n-1} ((X^T)^{n-1-k} \otimes X^k)(w \otimes v) \\ &= \sum_{n=1}^{\infty} a_n \sum_{k=0}^{n-1} ((X^T)^{n-1-k} w) \otimes (X^k v) \\ &= \sum_{n=1}^{\infty} a_n \left(\sum_{k=0}^{n-1} \mu^{n-1-k} \lambda^k \right) (w \otimes v). \end{aligned}$$

Now, if $\mu = \lambda$, then $\sum_{k=0}^{n-1} \mu^{n-1-k} \lambda^k = n\lambda^{n-1}$ and

$$D(X)(w \otimes v) = \sum_{n=1}^{\infty} n a_n \lambda^{n-1} (w \otimes v) = F'(\lambda)(w \otimes v).$$

Otherwise, if $\mu \neq \lambda$, then $\sum_{k=0}^{n-1} \mu^{n-1-k} \lambda^k = (\lambda^n - \mu^n)/(\lambda - \mu)$ and

$$D(X)(w \otimes v) = \sum_{n=1}^{\infty} a_n \frac{\lambda^n - \mu^n}{\lambda - \mu} (w \otimes v) = \frac{F(\lambda) - F(\mu)}{\lambda - \mu} (w \otimes v). \qquad \square$$

COROLLARY 2.2. *Let $\nu_{\max}$ be defined by*

$$\nu_{\max} = \max_{\lambda, \mu \in \Lambda(X)} \left| \frac{F(\lambda) - F(\mu)}{\lambda - \mu} \right| \tag{2.10}$$

where $\Lambda(X)$ denotes the set of eigenvalues of X and the ratio in (2.10) is taken to be $|F'(\lambda)|$ when $\lambda = \mu$. Then the condition number of F at X is bounded below by $\nu_{\max}$: $\nu_{\max} \leqq K(F, X)$.

Proof. By (1.3), (1.7), and (2.6) we have that $K(F, X) = \|L(\cdot, X)\| = \|D(X)\|_2$, but the two-norm of $D(X)$ is bounded below by the absolute value of any eigenvalue of $D(X)$. Hence by Lemma 2.1 we must have $\nu_{\max} \leqq \|D(X)\|_2$. □

LEMMA 2.3. *If $X \in \mathbb{R}^{p \times p}$ is normal, that is, $X^T X = X X^T$, then $D(X)$ is normal.*

Proof. Use $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ together with $(A \otimes B)^T = A^T \otimes B^T$ and (2.5) to show that $D(X)$ and $D^T(X)$ commute. □

COROLLARY 2.4. *If X is normal, then the condition number of F at X is equal to $\nu_{\max}$ in (2.10):*

$$K(F, X) = \max_{\lambda, \mu \in \Lambda(X)} \left| \frac{F(\lambda) - F(\mu)}{\lambda - \mu} \right|.$$

Proof. By Lemma 2.3, $D(X)$ is normal. Thus its two-norm is equal to its spectral radius that, by Lemma 2.1, is just $\nu_{\max}$ in (2.10). □

The fact that the lower bound in Corollary 2.2 is attained for normal matrices indicates that the condition number of F is as small as possible when X is normal. This effect has been demonstrated for the exponential function $F(X) = e^X$ in the sensitivity study of Van Loan [31], by using the explicit representation (1.6) together with the property that when X is normal, $\| e^X\|_2 = e^{\alpha(X)}$ where $\alpha(X) \equiv \max_{\lambda \in \Lambda(X)} \operatorname{Re}(\lambda)$.

The preceding dealt with linear perturbation theory of the matrix function $F(X) = \sum_{n=0}^{\infty} a_n X^n$, by considering the limiting behavior of the finite-difference operator

$$DF(Z,X,\delta) \equiv \frac{F(X+\delta Z)-F(X)}{\delta}$$

as $\delta \to 0^+$. We conclude this section with similar results on the behavior of F with respect to large (i.e., nondifferential) perturbations, and our goal will be to bound $K_\delta(F,X)$ in (1.3).

LEMMA 2.5. *Let $\|X\| + \delta < r$. Let v and w be normalized eigenvectors such that $Xv = \lambda_1 v$ and $w^H X = \lambda_2 w^H$. Define $Z = vw^H$. Then $\mu Z = (F(X+\delta Z) - F(X))/\delta$ where*

$$\mu = \frac{F(\lambda_1)-F(\lambda_2)}{\lambda_1-\lambda_2} \quad \text{if } \lambda_1 \neq \lambda_2, \tag{2.11}$$

$$\mu = \frac{F(\lambda_1 + w^H v\delta)-F(\lambda_1)}{w^H v\delta} \quad \text{if } \lambda_1 = \lambda_2. \tag{2.12}$$

The right-hand side of (2.12) *is taken to be $F'(\lambda_1)$ if $w^H v = 0$. As a consequence, we have the lower bound*

$$\max |\mu| \leqq K_\delta(F,X). \tag{2.13}$$

Proof. The proof is essentially the same as that used in Lemma 2.1. □

The next lemma gives a simple upper bound for $K_\delta(F,X)$ in terms of the function F_+ defined by the associated "positive" series $F_+(x) \equiv \sum_{n=0}^{\infty} |a_n| x^n$.

LEMMA 2.6. *Let $\|Z\| \leqq 1$ and $\|X\| + \delta < r$. Then*

$$K_\delta(F,X) \leqq \frac{F_+(\|X\|+\delta)-F_+(\|X\|)}{\delta} \leqq F'_+(\|X\|+\delta). \tag{2.14}$$

Proof. From (2.1) and $\|Z\| \leqq 1$,

$$\begin{aligned}\|F(X+\delta Z)-F(X)\| &\leqq \delta \sum_{n=1}^{\infty} n|a_n|\,\|X\|^{n-1} + \cdots + \delta^m \sum_{n=m}^{\infty} \binom{n}{m} |a_n|\,\|X\|^{n-m} + \cdots \\ &= \delta F'_+(\|X\|) + \cdots + \frac{\delta^m}{m!} F_+^{m}(\|X\|) + \cdots \\ &= F_+(\|X\|+\delta) - F_+(\|X\|).\end{aligned}$$

Thus

$$\frac{\|F(X+\delta Z)-F(X)\|}{\delta} \leqq \frac{F_+(\|X\|+\delta)-F_+(\|X\|)}{\delta} = F'_+(\|X\|+\rho) \leqq F'_+(\|X\|+\delta)$$

for some $0 \leqq \rho \leqq \delta$ by the mean value theorem and the fact that F'_+ is nondecreasing. □

For example, if $F(x) = e^x$, then Lemma 2.6 gives

$$\|e^{X+\delta Z} - e^X\|/\delta \leqq e^{\|X\|+\delta} - e^{\|X\|}/\delta \leqq e^{\|X\|+\delta}.$$

This upper bound can be very conservative in some cases. However, the next lemma shows that there are situations where the upper bound in Lemma 2.6 coincides with the lower bound in Lemma 2.5 to give an exact value for $K_\delta(F, X)$.

LEMMA 2.7. *Let $X = X^T \geqq 0$ and let the series* (1.2) *have nonnegative coefficients, $a_n \geqq 0$, so that $F = F_+$. Then $K_\delta(F, X) = (F(\|X\| + \delta) - F(\|X\|))/\delta$.*

Proof. Since X is nonnegative definite symmetric there exists a real eigenvector v such that $Xv = \lambda v$ with $v^Tv = 1$ where $\lambda = \|X\|$. By Lemma 2.5, $\mu Z = (F(X + \delta Z) - F(X))/\delta$ where $Z = vv^T$ and $\mu = (F(\lambda + \delta) - F(\lambda))/\delta$. Note that $\mu > 0$, since F is nondecreasing. Thus, $K_\delta(F, X) \geqq \|\mu Z\| = \mu$, since $\|Z\| = \|vv^T\| = 1$. On the other hand, $\mu = (F(\|X\| + \delta) - F(\|X\|))/\delta$, since $\|X\| = \lambda$. Thus, since $F = F_+$, we have by Lemma 2.6 that $K_\delta(F, X) \leqq \mu$. This shows that we must have $K_\delta(F, X) = \mu = (F(\|X\| + \delta) - F(\|X\|))/\delta$. □

The next lemma shows that for $F = F_+$ and large δ, Lemma 2.7 is approximately true, not just for symmetric nonnegative definite matrices, but for *any* matrix X.

LEMMA 2.8. *Let $F = F_+$ with radius of convergence $r = \infty$. Then for $\delta > 2\|X\|$ and any real matrix X, we have*

$$\frac{F(\delta - \|X\|) - F(\|X\|)}{\delta} \leqq K_\delta(F, X) \leqq \frac{F(\delta + \|X\|) - F(\|X\|)}{\delta}. \tag{2.15}$$

Proof. The right-hand side inequality of (2.15) is simply a restatement of (2.14) in Lemma 2.6. To prove the left-hand side inequality in (2.15), let $Z_1 \equiv e_1e_1^T$ where $e_1 \equiv (1, 0, \cdots, 0)^T$ and set $Z = (1 - \varepsilon)Z_1 - X/\delta$ where ε is chosen so that $\|Z\| = 1$. Then $1 = \|Z\| \leqq 1 - \varepsilon + \|X\|/\delta$ so $\varepsilon \leqq \|X\|/\delta$. Now $X + \delta Z = \delta(1 - \varepsilon)Z_1$, so $F(X + \delta Z) = F(\delta(1 - \varepsilon))Z_1$, since $F(\alpha Z_1) = F(\alpha)Z_1$ for any scalar α. Moreover, since $\delta\varepsilon \leqq \|X\|$, $\|F(X + \delta Z)\| = F(\delta(1 - \varepsilon)) \geqq F(\delta - \|X\|)$ because $F = F_+$ is nondecreasing. However, $\|F(X + \delta Z)\| - \|F(X)\| \leqq \|F(X + \delta Z) - F(X)\|$. Thus,

$$F(\delta - \|X\|) - F(\|X\|) \leqq \|F(X + \delta Z)\| - \|F(X)\| \leqq \|F(X + \delta Z) - F(X)\|$$

because $\|F(X)\| \leqq F(\|X\|)$. Dividing by δ in the above completes the proof. □

Example. Let $F(X) = e^X$ with $\|X\| = 1$; then by Lemma 2.8, we have that

$$\frac{e^{\delta - 1} - e}{\delta} \leqq K_\delta(F, X) \leqq \frac{e^{\delta + 1} - e}{\delta},$$

which determines K_δ to within a factor of e^2 for large δ.

3. Exponential and logarithmic linear perturbation theory. In this section, we treat the problem of approximating the Fréchet derivatives of the exponential and logarithmic matrix functions.

The earliest representation of the exponential derivative appears to be due to Hausdorff [17]:

$$L(Z, X) = e^X \sum_{n=0}^{\infty} \frac{1}{(n+1)!}\{Z, X^n\} = e^X\{Z, (e^X - I)X^{-1}\} \tag{3.1}$$

where the nested Lie product $\{\cdot, \cdot\}$ is defined by $\{Z, X^n\} \equiv [\cdots[[Z, X], X], \cdots, X]$ with n factors of X appearing; $[Z, X]$ denotes the Lie bracket $ZX - XZ$. In the rightmost side of (3.1), the expression $(e^X - I)X^{-1}$ should be interpreted as the series $\sum_{m=0}^{\infty} X^m/(m + 1)!$ when X is not invertible. This Lie product expansion for the exponential derivative arose in connection with the Baker–Campbell–Hausdorff formula (see [23, pp. 656–658], [4])

$$e^Xe^Y = \exp\left(X + Y + \frac{1}{2}[X, Y] + \frac{1}{12}[[X, Y], Y - X] + \cdots\right).$$

From (2.2), with $F(X) = e^X$, we obtain another series representation:

$$L(Z,X) = \sum_{n=1}^{\infty} \frac{1}{n!} \sum_{k=0}^{n-1} X^k Z X^{n-1-k},$$

but this series and (3.1) are too hard to work with numerically. A much more useful representation is given by Van Loan in [31]: $L(Z, X) = \int_0^1 e^{X(1-s)} Z e^{Xs}\, ds$. This may be approximated by using the trapezoid rule (see [10])

$$(3.2) \qquad L(Z,X) \cong L_n(Z,X) \equiv \frac{1}{2^{n+1}} \left[e^X Z + 2 \sum_{k=1}^{2^n - 1} e^{kX/2^n} Z e^{(2^n - k)X/2^n} + Z e^X \right]$$

for $n = 0, 1, 2, \cdots$.

We have selected this method of approximation because it is uniquely suited to one of the most successful methods of computing e^X, namely the scaling and squaring method [24], [32]. In this method, X is scaled by a power of two, say 2^n, so that $e^{X/2^n}$ is easily evaluated by using, for example, a Padé approximation. The result is then squared n times: $e^X = (e^{X/2^n})^{2^n}$. During the squaring phase, we have available to us sequentially the computed values of the matrices $e^{X/2^n}$, $e^{X/2^{n-1}}$, $\cdots$, $e^{X/2}$, and e^X. This raises the possibility of evaluating the trapezoid approximant, $L_n(Z, X)$, for a given matrix Z, during the computation of e^X. This would not be practical if we implemented (3.2) directly, but fortunately there is an equivalent formulation for $L_n(Z, X)$ that is much easier to evaluate and only requires the matrices $e^{X/2^{n-k}}$ as they become available. Let

$$(3.3) \qquad W_0 \equiv (e^{X/2^n} Z + Z e^{X/2^n})/2^{n+1},$$

and for $j = n, n-1, \cdots, 1$, define

$$(3.4) \qquad W_{n+1-j} \equiv e^{X/2^j} W_{n-j} + W_{n-j} e^{X/2^j}.$$

Then from (3.2)–(3.4), $L_n(Z, X) = W_n$.

We will show that if n is large, then $L_n(Z, X)$ is near $L(Z, X)$ and $\| L_n(\cdot, X) \|$ provides a good estimate of $\| L(\cdot, X) \|$. For example, if n is large enough so that $\| e^{X/2^n} - I \| < \frac{1}{4}$, then our results give

$$0.950 \| L_n(\cdot, X) \| \leqq \| L(\cdot, X) \| \leqq 1.055 \| L_n(\cdot, X) \|.$$

Somewhat surprisingly, it seems to be the case that the easiest way to determine how well $L_n(\cdot, X)$ approximates $L(\cdot, X)$ is to study the approximation of $L^{-1}(\cdot, X)$ by $L_n^{-1}(\cdot, X)$. Both $L^{-1}(\cdot, X)$ and $L_n^{-1}(\cdot, X)$ arise naturally in the study of the inverse exponential or logarithmic problem $e^X \rightarrow X$. Such problems occur, for example, in a control theory setting wherein discrete samples from a continuous system are used to identify system parameters. See [20], [27], and [29]. Since the logarithm is a multivalued function, we need some restriction on X to ensure the existence of a unique real solution X to the problem $e^X = A$ (cf. [9], [13], [18], [33]). To do so, we shall assume throughout this section that $A \in \mathbb{R}^{p \times p}$ has no eigenvalues on the negative real axis including zero. This is sufficient to ensure that there exists a unique real matrix $X \in \mathbb{R}^{p \times p}$ such that $e^X = A$ with the eigenvalues of X confined to the strip $-\pi < \operatorname{Im}(z) < \pi$. (See Appendix A.)

Under the above assumptions,

$$(3.5) \qquad \log A = 2^n \log A^{1/2^n}$$

where $A^{1/2^n}$ denotes the unique real nth square root of A (see [11]) whose eigenvalues, $\lambda = \lambda(A^{1/2^n})$ lie in the sector $-\pi/2^n < \arg(\lambda) < \pi/2^n$. (See Appendix A.) This forms the basis of the "inverse scaling and squaring" method for approximating $\log A$. Take n square roots of A, so that $A^{1/2^n}$ is near the identity. Then $\log A^{1/2^n}$ can be computed by using, for example, a Padé approximation in the variable $Y \equiv I - A^{1/2^n}$. Multiplying the result by 2^n, we obtain $\log A$ as in (3.5). (See § 4 for more details.)

By Lemma A1, in Appendix A, the derivative of the logarithmic function $e^X \to X$ is the inverse of $L(\cdot, X)$ in (1.6) provided $L(\cdot, X)$ is invertible. However, $L(\cdot, X)$ is invertible if and only if the associated Kronecker matrix $D(X)$ given by (2.5) is nonsingular. By Lemma 2.1, $D(X)$ is singular for the exponential function if and only if $e^\lambda = 0$ or $(e^\lambda - e^\mu)/(\lambda - \mu) = 0$ for $\lambda, \mu \in \Lambda(X)$, $\mu \neq \lambda$. However, e^λ is never zero and $e^\lambda = e^\mu$ with $\mu \neq \lambda$ means that $\lambda = \mu + 2\pi i k$ for some nonzero integer k, which would violate the condition that $-\pi < \operatorname{Im}(\lambda), \operatorname{Im}(\mu) < \pi$. Thus $D(X)$ is nonsingular and $L(\cdot, X)$ is invertible whenever $\Lambda(X)$ is confined to the strip $-\pi < \operatorname{Im}(z) < \pi$. This strip condition also implies that $L_n(\cdot, X)$ is invertible. To see this, note that $L_n^{-1}(W, X)$, for a given matrix W, can be found by inverting the procedure in (3.3)–(3.4). That is, for $A = e^X$, set $W_n = W$ and solve sequentially for $W_{n-1}, \cdots, W_0$ and Z in

$$(3.6) \qquad W_{n+1-j} = A^{1/2^j} W_{n-j} + W_{n-j} A^{1/2^j},$$

$$(3.7) \qquad 2^{n+1} W_0 = A^{1/2^n} Z + Z A^{1/2^n}.$$

Then $L_n^{-1}(W, X) = Z$ because $L_n(Z, X) = W$ by (3.2)–(3.4).

From this we see that $L_n^{-1}(\cdot, X)$ is invertible whenever the Sylvester equations (3.6) and (3.7) are uniquely solvable. However, the strip condition on X forces the eigenvalues of $A^{1/2^j}$, for $j \geqq 1$, to lie in the open right-half complex plane. Consequently $\mu + \lambda \neq 0$ for $\mu, \lambda \in \Lambda(A^{1/2^j})$ and (3.6), (3.7) have unique solutions [21].

The sequence $W_0, \cdots, W_n$ has a nice representation that forms the basis of our analysis of the relationship between L and L_n and which originally inspired our work in this area.

LEMMA 3.1. *Let W_j be defined by* (3.6), (3.7) *with $W = W_n \equiv L(\hat{Z}, X)$, i.e., $W_n = \int_0^1 e^{X(1-s)} \hat{Z} e^{Xs}\, ds$. Then*

$$(3.8) \qquad W_{n-j} = \int_0^{1/2^j} e^{X(1/2^j - s)} \hat{Z} e^{Xs}\, ds$$

for $j = 1, \cdots, n$.

Proof. We show that (3.8) is valid for $j = 1$; for $j > 1$ use similar arguments. By (3.6),

$$W_n = A^{1/2} W_{n-1} + W_{n-1} A^{1/2},$$

which we may rewrite, using $A = e^X$, as $W_n = e^{X/2} W_{n-1} + W_{n-1} e^{X/2}$. Under the assumption that $\Lambda(X)$ lies in the strip $-\pi < \operatorname{Im}(z) < \pi$, this equation has a unique solution. Thus it is sufficient to show that $W_n = e^{X/2} \tilde{W}_{n-1} + \tilde{W}_{n-1} e^{X/2}$ where $\tilde{W}_{n-1} \equiv \int_0^{1/2} e^{X(1/2-s)} \hat{Z} e^{Xs}\, ds$. But

$$\begin{aligned} e^{X/2} \tilde{W}_{n-1} + \tilde{W}_{n-1} e^{X/2} &= \int_0^{1/2} e^{X(1-s)} \hat{Z} e^{Xs}\, ds + \int_0^{1/2} e^{X(1/2-s)} \hat{Z} e^{X(s+1/2)}\, ds \\ &= \int_0^{1/2} e^{X(1-s)} \hat{Z} e^{Xs}\, ds + \int_{1/2}^{1} e^{X(1-s)} \hat{Z} e^{Xs}\, ds \\ &= \int_0^1 e^{X(1-s)} \hat{Z} e^{Xs}\, ds \equiv W. \qquad \square \end{aligned}$$

We need two technical lemmas to prove our main result (Theorem 3.4).

LEMMA 3.2. *Let $[C, B]$ denote the Lie product $[C, B] \equiv CB - BC$. Then we may write*

$$\int_0^1 e^{B(1-s)}Ce^{Bs}\,ds = \frac{1}{2}(e^BC + Ce^B) - \frac{1}{2}\int_0^1 [e^{B(1-s)}, [e^{Bs}, C]]\,ds.$$

Proof. Expand the nested Lie product on the right-hand side and use the identity

$$\int_0^1 e^{B(1-s)}Ce^{Bs}\,ds = \int_0^1 e^{Bs}Ce^{B(1-s)}\,ds. \qquad \square$$

LEMMA 3.3. *Let $\mu(B) \equiv \frac{1}{2}\lambda_{\max}(B + B^T)$. Then we have*

$$\left\| \int_0^1 e^{B(1-s)}Ce^{Bs}\,ds - \frac{1}{2}(e^BC + Ce^B) \right\| \leqq \frac{1}{3}\|C\|\,\|B\|^2 e^{\mu(B)}. \tag{3.9}$$

Proof. Use the methods of Lemma 3 of [24, Appendix 2]. □

Using the preceding three lemmas we can now prove our main result on the approximation of $L^{-1}(\cdot, X)$ by $L_n^{-1}(\cdot, X)$.

THEOREM 3.4. *Let n be large enough so that $\omega \equiv \|I - e^{X/2^n}\| < 1$. Then for any $W \in \mathbb{R}^{p\times p}$, we have*

$$\|L^{-1}(W,X) - L_n^{-1}(W,X)\| \leqq \frac{1}{3}\left(\frac{1}{1-\omega}\log\left(\frac{1}{1-\omega}\right)\right)^2 \|L^{-1}(W,X)\|. \tag{3.10}$$

Proof. Let $L_n(Z, X) = W$ and $L(\hat{Z}, X) = W$ so that $Z = L_n^{-1}(W, X)$ and $\hat{Z} = L^{-1}(W, X)$. Now define $W_0, W_1, \cdots, W_{n-1}$ by (3.6), (3.7), so that by the definition of L_n^{-1} in (3.6), (3.7)

$$W_0 = \frac{(e^{X/2^n}Z + Ze^{X/2^n})}{2^{n+1}}. \tag{3.11}$$

However, by Lemma 3.1,

$$W_0 = \int_0^{1/2^n} e^{X(1/2^n - s)}\hat{Z}e^{Xs}\,ds. \tag{3.12}$$

By the change of variables, $s \rightarrow 2^n s$,

$$\int_0^{1/2^n} e^{X(1/2^n - s)}\hat{Z}e^{Xs}\,ds = \frac{1}{2^n}\int_0^1 e^{(X/2^n)(1-s)}\hat{Z}e^{(X/2^n)s}\,ds. \tag{3.13}$$

Now by Lemma 3.2 with $B = X/2^n$ and $C = \hat{Z}$,

$$\begin{aligned} \frac{1}{2^n}\int_0^1 e^{(X/2^n)(1-s)}\hat{Z}e^{(X/2^n)s}\,ds &= \frac{1}{2^{n+1}}(e^{X/2^n}\hat{Z} + \hat{Z}e^{X/2^n}) \\ &\quad - \frac{1}{2^{n+1}}\int_0^1 [e^{(X/2^n)(1-s)}, [e^{(X/2^n)s}, \hat{Z}]]\,ds. \end{aligned} \tag{3.14}$$

Combining (3.11)–(3.14), we obtain

$$e^{X/2^n}Z + Ze^{X/2^n} = e^{X/2^n}\hat{Z} + \hat{Z}e^{X/2^n} - \int_0^1 [e^{(X/2^n)(1-s)}, [e^{(X/2^n)s}, \hat{Z}]]\,ds.$$

This may be written as $\Omega(\hat{Z} - Z) = \int_0^1 [e^{(X/2^n)(1-s)}, [e^{(X/2^n)s}, \hat{Z}]]\, ds$ where $\Omega(V) \equiv e^{X/2^n}V + Ve^{X/2^n}$. Thus $\hat{Z} - Z = \Omega^{-1}(\int_0^1 [e^{(X/2^n)(1-s)}, [e^{(X/2^n)s}, \hat{Z}]]\, ds)$, so

$$\text{(3.15)} \qquad \begin{aligned} \|\hat{Z}-Z\| &\leqq \|\Omega^{-1}\| \left\| \int_0^1 [e^{(X/2^n)(1-s)}, [e^{(X/2^n)s}, \hat{Z}]]\, ds \right\| \\ &\leqq \|\Omega^{-1}\| \frac{2}{3} \left\| \frac{X}{2^n} \right\|^2 e^{\mu(X/2^n)} \|\hat{Z}\| \end{aligned}$$

as in the proof of Lemma 3.3. We now show that for $\omega \equiv \|I - e^{X/2^n}\| < 1$,

$$\text{(3.16)} \qquad \|\Omega^{-1}\| \leqq \frac{1}{2}\frac{1}{1-\omega},$$

$$\text{(3.17)} \qquad \left\| \frac{X}{2^n} \right\| \leqq \log\left(\frac{1}{1-\omega}\right),$$

$$\text{(3.18)} \qquad e^{\mu(X/2^n)} \leqq \frac{1}{1-\omega}.$$

When combined with (3.15) and $\hat{Z} = L^{-1}(W, X)$, $Z = L_n^{-1}(W, X)$, we shall have (3.10), thus completing the proof.

To show (3.16), let $Q = \Omega(V)$ so that $V = \Omega^{-1}(Q)$. Now, $2V = Q + YV + VY$ where $Y \equiv I - e^{X/2^n}$ and $\|Y\| = \omega < 1$. Thus, $2\|V\| \leqq \|Q\| + 2\omega\|V\|$, so $\|V\| \leqq \|Q\|/2(1 - \omega)$. Inequality (3.16) follows immediately since $\|V\| = \|\Omega^{-1}(Q)\|$.

To get (3.17), use $X/2^n = \log e^{X/2^n} = \log(I - Y)$, and

$$\|\log(I - Y)\| \leqq \sum_{m=1}^{\infty} \frac{\|Y\|^m}{m} = |\log(1 - \|Y\|)| = \log\left(\frac{1}{1-\omega}\right).$$

This also gives (3.18) because $e^{\mu(X/2^n)} \leqq e^{\|X/2^n\|} \leqq \exp(\log(1/(1-\omega))) = 1/(1-\omega)$. □

From Theorem 3.4, we can easily obtain a bound on the logarithmic condition number, $\|L^{-1}(\cdot, X)\|$ in terms of the norm, $\|L_n^{-1}(\cdot, X)\|$ of the inverse trapezoid approximant.

COROLLARY 3.5. *Let* $\omega \equiv \|I - e^{X/2^n}\| < 1$ *and define*

$$\omega_1 \equiv \tfrac{1}{3}(1/(1-\omega)\log(1/(1-\omega)))^2.$$

If n is large enough so that $\omega_1 < 1$, *then*

$$\|L_n^{-1}(\cdot, X)\|/(1+\omega_1) \leqq \|L^{-1}(\cdot, X)\| \leqq \|L_n^{-1}(\cdot, X)\|/(1-\omega_1).$$

Proof. Use (3.10) for the proof. □

As an example, if $\|I - e^{X/2^n}\| \leqq \frac{1}{4}$, then $0.953\|L_n^{-1}(\cdot, X)\| \leqq \|L^{-1}(\cdot, X)\| \leqq 1.052\|L_n^{-1}(\cdot, X)\|$. To obtain bounds on the exponential condition number $\|L(\cdot, X)\|$, we now return to the problem of how well $L_n(\cdot, X)$ approximates $L(\cdot, X)$.

THEOREM 3.6. *Let* $\omega_2 \equiv \omega_1/(1 - \omega_1)$ *for* ω *and* ω_1 *as in Corollary* 3.5. *Assume that n is large enough so that* ω, ω_1, *and* ω_2 *are less than one. Then for any* $Z \in \mathbb{R}^{p \times p}$, *we have*

$$\|L(Z, X) - L_n(Z, X)\| \leqq \omega_2 \|L(\cdot, X)\| \, \|Z\|.$$

Proof. Let Z be given and let $W \equiv L_n(Z, X)$, so that $L_n^{-1}(W, X) = Z$. Let $\hat{Z} \equiv L^{-1}(W, X)$ so that $L(\hat{Z}, X) = W = L_n(Z, X)$. Then by Theorem 3.4, $\|\hat{Z} - Z\| \leqq$

$\omega_1\|\hat{Z}\|$. But by Corollary 3.5, $\|\hat{Z}\| \leqq \|Z\|/(1-\omega_1)$ so $\|\hat{Z}-Z\| \leqq \omega_1\|Z\|/(1-\omega_1) \equiv \omega_2\|Z\|$. Thus,

$$\begin{aligned}\|L(Z,X)-L_n(Z,X)\| &= \|L(Z,X)-L(\hat{Z},X)\| \\ &= \|L(Z-\hat{Z},X)\| \\ &\leqq \|L(\cdot,X)\|\ \|Z-\hat{Z}\| \\ &\leqq \|L(\cdot,X)\|\omega_2\|Z\|. \qquad \square\end{aligned}$$

COROLLARY 3.7. *Under the assumptions of Theorem* 3.6,

$$\|L_n(\cdot,X)\|/(1+\omega_2) \leqq \|L(\cdot,X)\| \leqq \|L_n(\cdot,X)\|/(1-\omega_2).$$

Proof. Use standard norm arguments and Theorem 3.6 for the proof. □

As an example, if $\|I-e^{X/2^n}\| \leqq \frac{1}{4}$, then $0.950\|L_n(\cdot,X)\| \leqq \|L(\cdot,X)\| \leqq 1.055\|L_n(\cdot,X)\|$.

To illustrate the trapezoid approximation method and Theorem 3.6, let $X = \left[\begin{smallmatrix}0 & 1\\ 0 & 0\end{smallmatrix}\right]$. Then

$$e^X = \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix} \quad \text{and} \quad e^{X/2^n} = \begin{bmatrix}1 & 1/2^n\\ 0 & 1\end{bmatrix}.$$

If we impose a scaling condition of $\|I-e^{X/2^n}\| < \frac{1}{4}$, then we may take $n = 3$, in which case $\|I-e^{X/8}\| = \frac{1}{8}$. Let $Z = e^X e^{X^T} + e^{X^T} e^X = \left[\begin{smallmatrix}3 & 2\\ 2 & 3\end{smallmatrix}\right]$. (For reasons explained in the next section, this choice of Z can be expected to have a large component in the matrix direction which maximally perturbs e^X.) Using (3.3) and (3.4), we find (to four significant figures), $L_3(Z,X) = \left[\begin{smallmatrix}4 & 5.328\\ 2 & 4\end{smallmatrix}\right]$. To compare this with $L(Z,X)$, note that X is nilpotent with $X^2 = 0$. Thus from (2.2),

$$L(Z,X) = \sum_{n=1}^{\infty} \frac{1}{n!} \sum_{k=0}^{n-1} X^k Z X^{n-1-k} = Z + \frac{XZ+ZX}{2} + \frac{XZX}{6} = \begin{bmatrix}4 & 5.333\\ 2 & 4\end{bmatrix}.$$

This gives, as in Theorem 3.6, $0.005 = \|L(Z,X) - L_3(Z,X)\| \leqq \|L(\cdot,X)\|\omega_2\|Z\| = 0.064$, where $\|L(\cdot,X)\| = 1.609$ was determined by finding the largest singular value of the associated Kronecker matrix D in (2.5). It is interesting to note that one power method cycle, using L_3 and this Z, gives an estimate of 1.592 for the norm of $L(\cdot,X)$.

4. Numerical results. In this section, we discuss some of the details of testing the trapezoid approximation and finite-difference condition estimation procedures for the exponential and logarithmic matrix functions.

From § 3, we have implemented the trapezoid approximation method (3.3), (3.4) for the matrix exponential in conjunction with the subroutine MATEXP of Ward [32]. To avoid analytical complications in the sensitivity estimate resulting from the use of the balancing transformation BALANX (which is a modified version of the EISPACK subroutine BALANC [30]), we have implemented (3.3), (3.4) after the back substitution BALINV in MATEXP. For one cycle of the power method, this results in a condition estimate, which costs $4n + 4$ matrix multiplications, where the scaling parameter n is chosen so that $\|X/2^n\|_1 \leqq \log(5/4) \cong 0.223$. This ensures that $\|I-e^{X/2^n}\| \leqq \frac{1}{4}$, as seen by the following lemma. (By Corollary 3.7, this scaling condition forces the norm of the trapezoid approximation, $\|L_n(\cdot,X)\|$ to be within six percent of the exponential condition number, $\|L(\cdot,X)\|$.)

LEMMA 4.1. *If* $\|Z\| \leqq \log(1+\omega)$, *then* $\|I-e^Z\| \leqq \omega$.

Proof. $\|I-e^Z\| \leqq e^{\|Z\|} - 1 \leqq e^{\log(1+\omega)} - 1 = \omega$. □

The subroutine MATEXP needs about $8 + n$ matrix multiplications to evaluate e^X (to a relative precision of about 10^{-16}), so the sensitivity estimate via (3.3), (3.4) is about 1.9 times as expensive as evaluating e^X when $n = 6$, which was the average value of n for the examples we considered.

We also implemented the inverse trapezoid approximation (3.6), (3.7) to estimate the condition of the logarithm, subject to the scaling condition $\|I - A^{1/2^n}\| \leqq \frac{1}{4}$. The square root of a matrix can be obtained in a stable manner by using the Schur algorithm described in [5]. This involves finding the real Schur form of A: $A = QTQ^T$ where Q is orthogonal and T is quasi-upper-triangular. Once this is done, $A^{1/2} = QT^{1/2}Q^T$, where $T^{1/2}$ is found by a simple linear recursion involving the entries of T and the square roots of the main diagonal entries of T (including the 2×2 blocks corresponding to the complex conjugate eigenvalues of T; see [22]). Moreover, the jth square root satisfies $A^{1/2^j} = QT^{1/2^j}Q^T$, which means that the Schur decomposition need only be done once in the process of generating $A^{1/2^n}$. This is important because the Schur decomposition of a matrix of order p requires about $8p^3$ floating-point operations (flops), whereas the square root of a quasi-upper-triangular matrix of order p requires only about $p^3/6$ flops. The logarithm of $A^{1/2^n}$ can be approximated by truncating the slowly convergent Taylor series, $\log(I - Y) = -\sum_{m=1}^{\infty} Y^m/m$, but rational Padé approximants are generally superior. For example, it is shown in [20] that if $\|Y\| = \|I - A^{1/2^n}\| \leqq \frac{1}{4}$, then the eighth-order main diagonal Padé approximant $R_{88}(Y) \equiv P_{88}(Y)Q_{88}^{-1}(Y)$ differs from $\log(I - Y)$ by less than 10^{-18}, whereas the sixteenth-order Taylor approximant, which requires about the same amount of work, can be in error by as much as 5×10^{-12}. In the above,

$$P_{88}(Y) \equiv -Y + \frac{7}{2}Y^2 - \frac{73}{15}Y^3 + \frac{41}{12}Y^4 - \frac{743}{585}Y^5 + \frac{31}{130}Y^6 - \frac{111}{5775}Y^7 + \frac{761}{1801800}Y^8,$$

$$Q_{88}(Y) \equiv 1 - 4Y + \frac{98}{15}Y^2 - \frac{28}{5}Y^3 + \frac{35}{13}Y^4 - \frac{28}{39}Y^5 + \frac{14}{143}Y^6 - \frac{4}{715}Y^7 + \frac{1}{12870}Y^8.$$

Moreover, when $\|I - A^{1/2^n}\| \leqq \frac{1}{4}$, the Padé denominator matrix $Q_{88}(Y)$ is very well-conditioned with $K(Q_{88}(Y)) \equiv \|Q_{88}\| \, \|Q_{88}^{-1}\| \leqq 7.59$ (see [20]).

The inverse scaling and squaring procedure for evaluating the logarithm of a matrix takes about $11 + n/6$ matrix multiplications, whereas the first cycle of the power method of estimating the condition number $\|L^{-1}(\cdot, X)\|$ takes about $2 + 13/6n$ matrix multiplications. Thus the condition estimate takes about 1.2 times the effort needed to evaluate the logarithm when $n = 6$.

We have also implemented the "finite-difference" power method for the exponential and logarithmic functions. Given Z_0 define $\tilde{W}_0 \equiv (F(X + \delta Z_0) - F(X))/\delta$, $W_0 \equiv \tilde{W}_0 / \|\tilde{W}_0\|$, and $Z_1 \equiv (F(X^T + \delta W_0) - F(X^T))/\delta$. Then $\|Z_1\|$ provides a condition estimate of F at X, at a cost of two function evaluations beyond $F(X)$, when we use the fact that $F(X^T) = (F(X))^T$.

A common problem, for both the trapezoid and finite-difference approximation methods, is the choice of the initial matrix Z_0. The complex nature of both methods makes it difficult to use "look-ahead" procedures such as those described in [8] and [6]. Instead, we have tried two different methods of choosing Z_0. The first consists of letting Z_0 have random entries in the interval $[-1, 1]$. This practically guarantees that Z_0 has a nontrivial component in the matrix direction that maximizes $\|L(Z, X)\|$. Consequently, one power method cycle usually provides an estimate of $\|L(\cdot, X)\|$ that is sufficient for the purposes of condition estimation [8]. We found that this was the case for the problems that we tested and that for most of the examples, one cycle of the finite-difference power method with a random Z_0 produced a condition estimate that was within 90 percent of the true condition number while none of the one cycle estimates was less than 25 percent of the true value.

The second method of choosing Z_0, for the exponential function, consists of setting $Z_0 \equiv (e^{X^T}e^X + e^Xe^{X^T})/2$. The rationale behind this choice is that since

$$L(Z,X)=\int_0^1 e^{X(1-s)}Ze^{Xs}\,ds,$$

if we set $Z = I$, then $L(I, X) = e^X$. The adjoint step in the power method then gives

$$L(e^X,X^T)=\int_0^1 e^{X^T(1-s)}e^Xe^{X^Ts}\,ds \cong \frac{e^{X^T}e^X+e^Xe^{X^T}}{2}=Z_0.$$

Thus one cycle of the power method with Z_0 as above has approximately the effect of two cycles and the resulting condition estimate should be much nearer the true condition number. We found that this was indeed the case and the resulting condition estimates were always better than those obtained with random matrices. A similar procedure was used for the logarithmic problem.

To determine the true condition numbers for our problem set, the trapezoid power method was iterated until the estimates from one iteration to the next had a relative difference of less than 10^{-8}. (The resulting values were cross-checked by iterating the finite-difference method.)

In Tables 1 and 2, we give the following relative condition numbers:

$$K_{\text{TRAP}} \equiv \frac{\|X\|}{\|F(X)\|}\|L(\cdot,X)\|_{\text{TRAP}}, \tag{4.1}$$

$$K_{\text{FD}} \equiv \frac{\|X\|}{\|F(X)\|}\|L(\cdot,X)\|_{\text{FD}}, \tag{4.2}$$

$$K_{\text{EXACT}} \equiv \frac{\|X\|}{\|F(X)\|}\|L(\cdot,X)\|, \tag{4.3}$$

for the exponential and logarithmic functions where $\|L(\cdot, X)\|_{\text{TRAP}}$ and $\|L(\cdot, X)\|_{\text{FD}}$ refer to the one-cycle power method estimates of $\|L(\cdot, X)\|$ obtained by using the trapezoid and finite-difference approximation methods, respectively.

The problems tested included eight examples from the standard collection of matrices [16], four examples of Ward [32]; 10 examples arising from state space models [1] in control theory [3], [7], [25], [26]; and 1,000 randomly generated matrices of orders between two and 16. For brevity, we discuss only a representative subsample consisting of six problems.

TABLE 1
Condition estimates for $F(X) = e^X$.

Problem number	K_{TRAP} (from 4.1)	K_{FD} (from 4.2)	K_{EXACT} (from 4.3)
1	7.49	7.50	7.50
2	53.9	53.9	53.9
3	2×10^4	2×10^4	2×10^4
4	1.59	1.59	1.68
5	2×10^{11}	2×10^{11}	2×10^{11}
6	3×10^3	3×10^3	3×10^3

TABLE 2
Condition estimates for $F(X) = \log X$.

Problem number	K_{TRAP} (from 4.1)	K_{FD} (from 4.2)	K_{EXACT} (from 4.3)
1	5.15	5.17	5.25
2	9×10^6	9×10^6	9×10^6
3	6×10^9	6×10^9	6×10^9
4	3.76	3.76	4.03
5	3×10^{11}	3×10^{11}	3×10^{11}
6	6×10^6	6×10^6	6×10^6

The first four problems of Tables 1 and 2 were taken from [32]. Of these, Examples 3 and 4 are interesting because they show, as noted by Ward [32], that the condition estimation scheme used in the subroutine MATEXP can give very conservative bounds. For problem 3, MATEXP predicted that not more than 12 digits of accuracy would be lost in the computation of e^X, whereas one cycle of the power method for the Fréchet derivative (see K_{TRAP} and K_{FD}, Table 1, problem 3) predicted four digits would be lost. In fact, the computed result had lost exactly four digits of accuracy. Similarly, for problem 4, MATEXP predicted a loss of at most nine digits, the power method predicted a loss of one digit, and the computed result had lost one digit of accuracy. This illustrates that condition estimates based on the norm of the Fréchet derivative have the virtue of reliability. For the fifth problem, $X = \left[\begin{smallmatrix} 0 & \xi \\ 0 & \delta \end{smallmatrix}\right]$ with $\xi = 10^6$. This value of ξ was chosen because the exponential condition number is then very large. The excellent agreement between K_{TRAP}, K_{FD}, and K_{EXACT} in Tables 1 and 2 is reminiscent of the fact that inverse power method estimates of $\|A^{-1}\|$ become more accurate as A becomes more singular.

An interesting feature of this problem is the strong dependence of K_{FD} on δ, as illustrated in Table 3. For example, $K_{\text{FD}} = 7 \times 10^{36}$ when $\delta/\|X\| = 5 \times 10^{-9}$. This seems rather conservative since $K_{\text{EXACT}} \equiv K(F, X) = 2 \times 10^{11}$. However, the given values of K_{FD} are correct and appropriate, as the following two points will make clear. First, for a given value of δ, K_{FD} is a lower bound on K_δ in (1.3):

$$K_{\text{FD}} = (\|e^{X+\delta Z} - e^X\|/\delta)(\|X\|/\|e^X\|) \leqq K_\delta .$$

Thus K_{FD} estimates K_δ rather than $K(F, X) \equiv \lim_{\delta \to 0} K_\delta$. Normally, when $\delta = 5 \times 10^{-9}\|X\|$, the difference between K_δ and $K(F, X)$ is small. However, and this is the second point, for this example, K_δ grows dramatically with δ. To see this, let $Z = \left[\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right]$, then (after some algebra),

$$\frac{\|e^{X+\delta Z} - e^X\|}{\delta} \frac{\|X\|}{\|e^X\|} \cong \frac{\sqrt{\delta\xi}\, e^{\sqrt{\delta\xi}}}{2\delta^2} \leqq K_\delta .$$

TABLE 3
Perturbation estimates for Problem 5 with $\|X\| = 10^6$.

$\delta/\|X\|$	K_{FD}
5×10^{-9}	7×10^{36}
1×10^{-9}	9×10^{20}
5×10^{-10}	3×10^{17}
1×10^{-10}	1×10^{13}
5×10^{-11}	2×10^{12}
1×10^{-11}	3×10^{11}
5×10^{-12}	2×10^{11}

For example, if $\xi = 10^6$ and $\delta = 5 \times 10^{-9}\xi = 5 \times 10^{-3}$, then $\sqrt{\delta\xi}e^{\sqrt{\delta\xi}}/2\delta^2 = 7.24 \times 10^{36}$. In fact, for this problem, K_{FD} provides a reasonably good estimate of nonlinear, "large-scale" perturbation effects.

This points the way to choosing the right value of δ to use with K_{FD}: $\delta/\|X\|$ should be on the order of the uncertainty in the data, X, or if X is known exactly, $\delta/\|X\|$ should be near the machine epsilon, since this is the size of the error induced by machine representation. After extensive numerical testing, we found that good results were consistently obtained by taking $\delta = \varepsilon 10^3\|X\|$ where ε is the machine epsilon ($\cong 2.8 \times 10^{17}$ for double precision on a VAX 11/780). This value of δ is small enough so that $(F(X+\delta Z) - F(X))/\delta$ provides a good approximation to $L(Z, X)$, but not so small as to generate the truncation effects which occur when $\delta/\|X\|$ is at or below the machine epsilon. For extremely ill-conditioned problems (for example, problem 5 with $\xi \geqq 10^8$) even $\delta = \varepsilon 10^3\|X\|$ is too large to give a good estimate for $\|L(\cdot, X)\|$. In cases of this type, the trapezoid method provides a reliable means of estimating $\|L(\cdot, X)\|$ (see Corollary 3.5) since it does not depend on δ.

The last problem (#6) in Tables 1 and 2 is taken from [26] and illustrates the fact that condition estimates based on upper triangular canonical forms can be extremely conservative. For this problem,

$$X = \begin{bmatrix} 48 & -49 & 50 & 49 \\ 0 & -2 & 100 & 0 \\ 0 & -1 & -2 & 1 \\ -50 & 50 & 50 & -52 \end{bmatrix}.$$

Let $X = SJS^{-1}$ where J is the Jordan form of X. Petkov, Christov, and Konstantinov [26] show that the Jordan decomposition bound,

$$\frac{\|e^{X+\delta Z} - e^X\|_2}{\delta}\frac{\|X\|_2}{\|e^X\|_2} \leqq 16\delta\|S\|_2^2\|S^{-1}\|_2^2 e^{4\delta\|S\|_2\|S^{-1}\|_2}\ \frac{\|X\|_2}{\|e^X\|_2},$$

gives

$$K_\delta(F,X) \leqq 4 \times 10^{104}$$

for $\delta = 4 \times 10^{-3}$. However, Lyapunov arguments can be given to show that $K_\delta(F,X) \leqq 2 \times 10^6$; a lower bound for $K_\delta(F,X)$ is given by $K_{\mathrm{FD}} = 3.4 \times 10^3$ for $\delta = 4 \times 10^{-3}$.

5. Conclusion. The natural connection between the Fréchet derivatives of matrix functions and sensitivity allows us to develop a very general condition estimation procedure based on finite-difference approximations. This procedure is computationally reasonable since it only requires two extra function evaluations. As seen in the section on numerical tests, the ability to manipulate the "stepsize" δ in the finite-difference method can lead to sensitivity estimates even when the size of the perturbation is relatively large (see Table 3, § 4). This area needs further research, as does the related problem of condition estimation for perturbations that are restricted in some way, as in the theory of structured singular values.

We have also presented an alternative sensitivity estimation procedure for the matrix exponential and logarithmic functions. This method is based on a trapezoid approximation of the integral representation of the Fréchet derivative of the exponential function.

Because of its form, this method dovetails nicely with the "scaling and squaring" method of evaluating the matrix exponential and the "inverse scaling and squaring" method of evaluating the logarithm of a matrix. Both the finite-difference and trapezoid approaches require almost the same effort computationally. However, the trapezoid method has an advantage in that it does not depend on the stepsize δ, and consequently is a more reliable method for estimating the norm of the Fréchet derivative when the matrix function is very ill-conditioned, as in Example 5.

Appendix A. The square root and logarithm of a matrix. In this Appendix, we show that any real matrix $A \in \mathbb{R}^{p\times p}$, with no eigenvalues on the negative real axis including zero, has a unique real square root and a unique real logarithm. We also justify the inverse scaling and squaring formula $\log A = 2^n \log A^{1/2^n}$.

LEMMA A1. *Let $A \in \mathbb{R}^{p\times p}$ with no eigenvalues on the negative real axis, including zero. Then there exists a unique real matrix X, such that we have the following*:

(A1) (1) $X^2 = A$,

(A2) (2) *The eigenvalues of X are restricted to the sector* $-\pi/2 < \arg(z) < \pi/2$.

Proof. The existence of such a matrix X follows from the Cauchy integral formula for operators. This method was used by DePrima and Johnson in [11], in which this lemma was proved under the added condition that X satisfies: (3) $XS = SX$ whenever $AS = SA$. However, this condition can be shown to be a consequence of (A1) and (A2). □

LEMMA A2. *Let $A \in \mathbb{R}^{p\times p}$ with no eigenvalues on the negative real axis, including zero. Then there exists a unique real matrix X, such that we have the following*:

(A3) (1) $e^X = A$,

(A4) (2) *The eigenvalues of X lie in the strip* $-\pi < \operatorname{Im}(z) < \pi$.

Proof. The proof is similar to that of Lemma 6.1. □

LEMMA A3. *Let $A \in \mathbb{R}^{p\times p}$ with no eigenvalues on the negative real axis. Let $A^{1/2}$ and* $\log A$ *denote the unique real square root and logarithm of A as in Lemmas* A1 *and* A2, *respectively. Then*

$$A^{1/2} = e^{1/2 \log A} \tag{A5}$$

and

$$\log A = 2 \log A^{1/2}. \tag{A6}$$

Proof. Let $X = \log A$ satisfy (6.3) and (6.4). Then $e^{X/2}$ satisfies $e^{X/2}e^{X/2} = e^X = A$ and the eigenvalues of $e^{X/2}$ lie in the sector $-\pi/2 < \arg(z) < \pi/2$. This means that the real matrix $e^{X/2}$ satisfies (A1) and (A2) and so by Lemma A1, $A^{1/2} = e^{X/2} = e^{1/2 \log A}$, which proves (A5).

Now suppose that $\hat{X} = A^{1/2}$ satisfies (A1) and (A2). Using the Cauchy integral operator representation of the logarithm of $A^{1/2}$, we see that the eigenvalue condition (A2) implies that the eigenvalues of $\log A^{1/2}$ lie in the strip $-\pi/2 < \operatorname{Im}(z) < \pi/2$. Thus the matrix $X \equiv 2 \log A^{1/2}$ satisfies (A3) and (A4) and must be equal to $\log A$ by Lemma A2. This proves (A6). □

COROLLARY A4. *For A as in Lemma* A3, *the "inverse scaling and squaring" formula* $\log A = 2^n \log A^{1/2^n}$ *is valid*.

Proof. By Lemma A3, $\log A = 2 \log A^{1/2} = 4 \log A^{1/4} = \cdots = 2^n \log A^{1/2^n}$. □

Appendix B. Examples of Fréchet derivatives. The following lemma enables us to find the derivatives of the square root and logarithmic functions.

LEMMA B1. *Let F be diffeomorphic at X, that is, let F be invertible in a neighborhood of $Y \equiv F(X)$ and let the derivative, $L_F(\cdot, X)$, of F at X be nonsingular. Then the derivative, $L_{F^{-1}}(\cdot, Y)$ of F^{-1} at Y exists and is given by the inverse of the derivative of F at X*:

$$L_{F^{-1}}(\cdot, F(X)) = L_F^{-1}(\cdot, X).$$

Proof. Although the proof of this lemma is not hard, we omit it for the sake of brevity. □

Using this lemma, we may find the derivatives of the inverse of the functions considered in Examples 1 and 2 of the introductory section.

Example 3. Let $A \in \mathbb{R}^{p \times p}$ with no eigenvalues on the negative real axis including zero, and let $A^{1/2}$ denote the square root of A as in Lemma A1. Condition (A2) on the eigenvalues of $A^{1/2}$ ensures that the Sylvester operator $L(Z) = A^{1/2}Z + ZA^{1/2}$ is invertible [21]. Hence, the derivative, $L_{1/2}$ of the square root function $A \rightarrow A^{1/2}$ is the inverse of L in Example 1 of the Introduction: $L_{1/2}(W, A) = Z$, where $L(Z, A^{1/2}) = W$.

Example 4. Let $A \in \mathbb{R}^{p \times p}$ with no eigenvalues on the negative real axis including zero, and let $\log A$ denote the logarithm of A as in Lemma A2. Condition (A4) ensures that the exponential derivative operator, L, defined by (1.6) is invertible. Hence, the derivative, $L_{\log}$, of the logarithmic function $A \rightarrow \log A$ is the inverse of L: $L_{\log}(W, A) = Z$ where $L(Z, \log A) = W$. (See § 3 for more details.)

Example 5. Let $X \in \mathbb{R}^{p \times p}$ be invertible. Then

$$(X+\delta Z)^{-1} = X^{-1} - \delta X^{-1}ZX^{-1} + O(\delta^2),$$

so the derivative of the inverse function is given by $L(Z, X) = -X^{-1}ZX^{-1}$. It is interesting to note that the inverse function is invariant under the inversion operation and

$$L^{-1}(\cdot, X) = L(\cdot, X^{-1}).$$

Since the squaring and exponential functions are related via the identity $e^X = (e^{X/2})^2$, it is not surprising that there exists a chain rule relationship between their derivatives:

$$L_{\exp}(Z, X) = \tfrac{1}{2} L_s(L_{\exp}(Z, X/2), e^{X/2})$$

where L_s and $L_{\exp}$ denote the derivatives of the squaring and exponential functions, respectively. This relationship is a consequence of the following lemma.

LEMMA B2. *Let $F(X) \equiv g(f(X))$ where we assume that the derivatives of f and g exist at X and $Y = f(X)$, respectively. Then the derivative of F at X exists and is given by*

$$L_F(Z, X) = L_g(L_f(Z, X), Y)$$

where L_f, L_g, and L_F denote the derivatives of f, g, and F, respectively.

Proof. The proof follows rather easily from (1.5). □

Example 6. Let $A \in \mathbb{R}^{p \times p}$ with no eigenvalues on the negative real axis including zero. Then we may define a real qth power of A, say $X = A^q$ by setting $X = e^{q \log A}$. We may write $A^q = h(g(f(A)))$ where $f(A) = \log A$, $g(B) = qB$ and $h(C) = e^C$. Then the derivative, L_q, of the map $A \rightarrow A^q$ is given by $L_q(Z, A) = L_{\exp}(qL_{\log}(Z, A), q \log A) = qL_{\exp}(L_{\log}(Z, A), q \log A)$.

Note added in proof. We wish to thank N. J. Higham for pointing out to us that our method for computing a matrix square root based on [5], while arrived at independently, is essentially identical to that given in [19]. The latter's much more thorough analysis should be consulted for details.

REFERENCES

[1] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[2] P. ANSELONE, *Collectively Compact Operator Approximation Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[3] W. F. ARNOLD, *Numerical solution of algebraic Riccati equations*, Ph.D. thesis, University of Southern California, Los Angeles, CA, December 1983.

[4] J. BELINFANTE AND B. KOLMAN, *A Survey of Lie Groups and Lie Algebras with Applications and Computational Methods*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1972.

[5] A. BJÖRCK AND S. HAMMARLING, *A Schur method for the square root of a matrix*, Linear Algebra Appl., 52/53 (1983), pp. 127-140.

[6] R. BYERS, A LINPACK-style condition estimator for the equation $AX - XB^T = C$, IEEE Trans. Automat. Control, 29 (1984), pp. 926-928.

[7] J. C. CHUNG AND E. Y. SHAPIRO, *Constrained eigenvalue/eigenvector assignment—application to flight control systems*, in Proc. Conference on Information and Systems, Department of Electrical Engineering and Computer Science, Princeton University, Princeton, NJ, 1982.

[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] W. CULVER, *On the existence and uniqueness of the real logarithm of a matrix*, Proc. Amer. Math. Soc., 17 (1966), pp. 1146-1151.

[10] G. DAHLQUIST AND A. BJÖRCK, *Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[11] C. DEPRIMA AND C. JOHNSON, *The range of $A^{-1}A^*$ in* GL $(n, \mathbb{C})$, Linear Algebra Appl., 9 (1974), pp. 202-222.

[12] N. DUNFORD AND J. SCHWARTZ, *Linear Operators, Part* I: *General Theory*, 4th ed., John Wiley, New York, 1967.

[13] F. GANTMACHER, *The Theory of Matrices, Vol.* I, Chelsea, New York, 1959.

[14] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.

[15] A. GRAHAM, *Kronecker Products and Matrix Calculus with Applications*, John Wiley, New York, 1981.

[16] R. GREGORY AND D. KARNEY, *A Collection of Matrices for Testing Computational Algorithms*, John Wiley, New York, 1969.

[17] F. HAUSDORFF, *Die Symbolische Exponential formel in der Gruppentheorie*, Berichte der Sächsischen Akademie der Wissenschaften (Math. Phys. Klasse), Leipzig, Vol. 58, 1906, pp. 19-48.

[18] B. HELTON, *Logarithms of matrices*, Proc. Amer. Math. Soc., 19 (1968), pp. 733-738.

[19] N. J. HIGHAM, *Computing real square roots of a real matrix*, Linear Algebra Appl., 88/89 (1987), pp. 405-430.

[20] C. KENNEY AND A. J. LAUB, *Padé error estimates for the logarithm of a matrix*, Internat. J. Control, to appear, 1989.

[21] P. LANCASTER, *Explicit solutions of linear matrix equations*, SIAM Rev., 12 (1970), pp. 554-566.

[22] B. LEVINGER, *The square root of a* 2 × 2 *matrix*, Math. Mag., 53 (1980), pp. 222-224.

[23] W. MAGNUS, *On the exponential solution of differential equations for a linear operator*, Comm. Pure and Appl. Math., 7 (1954), pp. 649-673.

[24] C. B. MOLER AND C. VAN LOAN, *Nineteen dubious ways to compute the exponential of a matrix*, SIAM Rev., 20 (1978), pp. 801-836.

[25] B. C. MOORE AND A. J. LAUB, *Computation of supremal (A, B)-invariant and controllability subspaces*, IEEE Trans. Automat. Control, 23 (1978), pp. 783-792.

[26] P. PETKOV, N. D. CHRISTOV, AND M. M. KONSTANTINOV, *Computational methods for linear control systems—some open questions*, in Proc. 26th Conference on Decision and Control, Los Angeles, CA, 1987, pp. 818-823.

[27] S. PUTHENPURA AND N. SINHA, *Transformation of continuous-time model of a linear multivariable system from its discrete-time model*, Electronics Letters, 20 (1984), pp. 737-738.

[28] J. RICE, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287-310.

[29] N. SINHA AND G. LASTMAN, *Transformation algorithm for identification of continuous-time multivariable systems from discrete data*, Electronics Letters, 17 (1981), pp. 779-780.

[30] B. T. SMITH, J. M. BOYLE, J. J. DONGARRA, B. S. GARBOW, Y. IKEBE, V. C. KLEMA, AND C. B. MOLER, *Matrix Eigensystem Routines*—EISPACK *Guide*, 2nd ed., Lecture Notes in Computer Science 6, Springer-Verlag, New York, 1976.

[31] C. VAN LOAN, *The sensitivity of the matrix exponential*, SIAM J. Numer. Anal., 14 (1977), pp. 971-981.

[32] R. C. WARD, *Numerical computation of the matrix exponential with accuracy estimate*, SIAM J. Numer. Anal., 14 (1977), pp. 600-610.

[33] A. WOUK, *Integral representation of the logarithm of matrices and operators*, J. Math. Anal. Appl., 11 (1965), pp. 131-138.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 2, pp. 210–218, April 1989

006

CONSISTENCY AND CONVERGENCE OF THE PARALLEL MULTISPLITTING METHOD FOR SINGULAR M-MATRICES*

J. PHILIP KAVANAGH† AND MICHAEL NEUMANN‡

Abstract. O'Leary and White have suggested a parallel multisplitting iteration scheme for solving a nonsingular linear system $Ax = b$. Among other things they have shown that when A has a nonnegative inverse and the multisplitting is weak regular, then the iteration converges to the solution from any initial vector. The extension of this result to the case where A is a singular M-matrix is discussed. Problems of solvability, consistency, and convergence arise and their resolution is considered.

Key words. multisplittings, iterative methods, nonnegative matrices, matrix graph theory

AMS(MOS) subject classification. 65F10

1. Introduction.

Consider the linear system of equations

$$Ax=b, \tag{1.1}$$

where A is an $n \times n$ matrix and b is an n-vector. For the case when A is nonsingular, O'Leary and White [1985] have introduced the parallel multisplitting iteration method for obtaining the solution to (1.1), which they formulated as follows. Split A into

$$A=M_l-N_l \quad \text{with } \det(M_l)\neq 0, \qquad l=1,\cdots,k, \tag{1.2}$$

and, beginning with the initial vector x_0, perform the iteration

$$x_{i+1}=\sum_{l=1}^{k} E_lM_l^{-1}N_lx_i+\sum_{l=1}^{k} E_lM_l^{-1}b, \qquad i=0,1,\cdots. \tag{1.3}$$

Here E_l, $l = 1, \cdots, k$, are nonnegative diagonal matrices such that

$$\sum_{l=1}^{k} E_l=I. \tag{1.4}$$

O'Leary and White's idea was to implement each step of the iteration (1.3) using a parallel machine with perhaps k processors in the following manner. The lth processor receives the approximation x_i from the central processor or a shared memory and computes the vector

$$E_lM_l^{-1}N_lx_i+E_lM_l^{-1}b. \tag{1.5}$$

An essential observation is that the lth processor needs only to compute those entries of (1.5) that correspond to the nonzero diagonal entries of E_l.

O'Leary and White have investigated the convergence of the multisplitting iteration method for systems (1.1) where the coefficient matrix A is an inverse positive (e.g., a

* Received by the editors June 22, 1988; accepted for publication (in revised form) September 20, 1988. The research of the first author was performed under Air Force Office of Scientific Research grant 88-0047 while the author was a visiting research specialist at the University of Connecticut. The research of the second author was partly supported by Air Force Office of Scientific Research grant 88-0047 and National Science Foundation grant DMS-870064.

† Department of Mathematics, University of Connecticut, Storrs, Connecticut 06268. Present address, Department of Mathematics, Illinois Wesleyan University, Bloomington, Illinois 61701.

‡ Department of Mathematics, University of Connecticut, Storrs, Connecticut 06268 (NEUMANN@UCONNVM.BITNET).

nonsingular M-matrix) or a positive definite matrix. In particular, they have shown that if each of the splittings (1.2) is weak regular, that is,

$$(1.6) \qquad M_l^{-1} \geqq 0 \quad \text{and} \quad M_l^{-1}N_l \geqq 0,$$

then the iteration (1.3) converges to the solution to (1.1) for all x_0, i.e.,

$$(1.7) \qquad \rho\left(\sum_{l=1}^{k} E_l M_l^{-1} N_l\right) < 1$$

if and only if

$$(1.8) \qquad A^{-1} \geqq 0.$$

Here $\rho(\cdot)$ denote the spectral radius.

Put

$$(1.9) \qquad H := \sum_{l=1}^{k} E_l M_l^{-1} N_l \quad \text{and} \quad c := \sum_{l=1}^{k} E_l M_l^{-1} b,$$

observe that with

$$(1.10) \qquad P := \sum_{l=1}^{k} E_l M_l^{-1}$$

we have that

$$(1.11) \qquad I - H = \left(\sum_{l=1}^{k} E_l M_l^{-1}\right) A = PA$$

and consider the system

$$(1.12) \qquad (I-H)x = c.$$

Under the conditions (1.6) and (1.8) on A and the splittings in (1.2), respectively, both system (1.1) and (1.12) have a unique solution. Moreover, since any solution to (1.1) is, by (1.11), necessarily, a solution to (1.12), it follows that the unique limit point of the multisplitting iteration scheme *must* also be the solution to (1.1). In this case, in the language of Young [1971], the iteration (1.3) is *consistent* with the system (1.1).

Singular systems of equations (1.1) in which the coefficient matrix is an M-matrix arise, for example, in the computation of steady state solutions for Markov processes and in the discrete solution to elliptic partial differential equations subject to the Neumann boundary value condition (see, e.g., Berman and Plemmons [1979, Chap. 8], Plemmons [1976]). The question naturally arises about the applicability of the multisplitting technique to the solution of such systems. As we shall see the extension of this technique to the singular case raises not only problems with convergence, as indeed is the case for conventional splittings (i.e., when $k = 1$), but also questions concerning the consistency of the iteration (1.3) with the system (1.1).

The following example illustrates the nature of the problem of consistency in the case that A is singular.

Example 1.1. Let

$$A = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}}_{M_1} - \underbrace{\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}}_{N_1} = \underbrace{\begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}}_{M_2} - \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{N_2},$$

and let

$$E_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = I - E_2.$$

Then $H = E_1 M_1^{-1} N_1 + E_2 M_2^{-1} N_2 = I$, $P = (\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix})$ and (1.3) becomes

$$x_{i+1} = x_i + Pb. \tag{1.13}$$

If $b \in R(A)$, the *range* of A, then $Pb = 0$ and we see that (1.12) converges trivially (to x_0) from each initial vector x_0. But of course, *not* every vector x_0 is a solution to (1.1). We further note that if $b \notin R(A)$, then $Pb \neq 0$ and (1.13) diverges for each choice of initial vector x_0.

In this paper, for systems (1.1), where A is an M-matrix with property-c and each of the splittings in (1.2) is graph compatible weak regular (see § 2 for definition), we show (Theorem 3.2) that the system (1.1) is solvable if and only if the system (1.12) is solvable. We further show (Theorem 3.3) that in this case every solution to (1.12) is also a solution to (1.1) or, equivalently, the iteration scheme (1.3) is consistent with the system (1.1) if and only if the matrix P is nonsingular. We subsequently provide a graph-theoretic criterion for P to be nonsingular.

The outline of our paper is as follows. In § 2 we define the notion of a graph compatible multisplitting as well as describe other graph-theoretic properties that we then use in § 3 to prove our main results. In § 3 we also provide examples to show that our results do not necessarily hold when the M-matrix A does not have property-c. A byproduct of our proof to Theorem 3.3 is that the matrix $I - H$ is an M-matrix with property-c. However, the iteration scheme (1.3), which is now consistent with the system (1.1), does not necessarily converge from every initial vector x_0 since H may have eigenvalues other than one on the unit circle. We can remedy the situation as follows. If each of the splittings (1.2) is graph compatible weak regular, then for any $\varepsilon \in (0, 1)$ the splittings

$$A = (1+\varepsilon)M_l - [\varepsilon M_l + N_l], \qquad l = 1, \cdots, k,$$

are again graph compatible weak regular. These splittings induce the (extrapolated) multisplitting iteration scheme

$$\begin{aligned} z_{i+1} &= \sum_{l=1}^{k} E_l \left[\frac{\varepsilon}{1+\varepsilon} I + \frac{1}{1+\varepsilon} M_l^{-1} N_l \right] z_i + \frac{1}{1+\varepsilon} \sum_{l=1}^{k} E_l M_l^{-1} b \\ &= \left[\frac{\varepsilon}{1+\varepsilon} I + \frac{1}{1+\varepsilon} H \right] z_i + \frac{1}{1+\varepsilon} \sum_{l=1}^{k} E_l M_l^{-1} b. \end{aligned}$$

This scheme converges from every initial point z_0 as all the eigenvalues of $\varepsilon/(1 + \varepsilon)I + 1/(1 + \varepsilon)H$ other than one lie in the interior of the unit circle.

2. Preliminaries. Many of the following notation and definitions are standard. We frequently borrow from the language of Schneider as used in his papers [1984] and [1986].

A matrix B is called *nonnegative*, $B \geqq 0$, if each of its entries is nonnegative. We call B *semipositive*, $B > 0$, if $B \geqq 0$ and $B \neq 0$. We call B *strictly positive*, $B \gg 0$, if each of its entries is positive. A square matrix A is a *Z-matrix* if it has the form $A = sI - B$, where $B \geqq 0$. If in addition $s \geqq \rho(B)$, we call A an *M-matrix*.

Let A be an $n \times n$ matrix. The *directed graph* of A, $\Gamma(A)$, is the graph with vertices $\langle n \rangle := \{1, 2, \cdots, n\}$ and edges $\{(i, j) \mid a_{ij} \neq 0\}$. We identify $\Gamma(A)$ with its edge set. Denote by $\overline{\Gamma(A)}$ the reflexive transitive closure of $\Gamma(A)$. For S, $T \subseteq \langle n \rangle$ we say that S has *access* to T if there is a path in $\Gamma(A)$ from a member of S to a member of T. A *class* of A is the vertex set of a strongly connected component of $\Gamma(A)$. We denote the classes of A by $\alpha_1, \cdots, \alpha_P$. If $p = 1$, then A is called *irreducible*. For $T \subseteq \langle n \rangle$ we denote by $A[T]$ the principal submatrix of A whose rows and column are indexed by T. Similarly, if x is an n-dimensional column vector, we denote by $x[T]$ the subvector of x whose entries are indexed by T. If A is reducible, then A is permutation similar to the block upper triangular *Frobenius normal form* whose diagonal blocks are the irreducible matrices $A[\alpha_i]$, $i = 1, \cdots, p$. We call a class α of A *(non) singular* if $A[\alpha]$ is (non) singular. A class α of A is called *final* if it has access to no other class. The *reduced graph* of A is the set $\{\alpha_1, \cdots, \alpha_p\}$ together with the partial order of access that is induced from $\Gamma(A)$. A *chain of classes* is a sequence $(\alpha_{i_1}, \cdots, \alpha_{i_t})$ such that α_{i_j} has access to $\alpha_{i_{j+1}}, j = 1, \cdots, t-1$.

Given an $n \times n$ matrix A in Frobenius normal form and an n-dimensional vector b, we partition b conformably with the block structure of A, and define the *support* of b to be the set of classes α of A for which $b[\alpha] \neq 0$. We next state a result of Carlson which is needed in § 3.

THEOREM 2.1 (Carlson [1963, Thm. 1]).[1] *Let A be an M-Matrix and let b be a nonnegative vector. Then there exists a nonnegative vector x such that $Ax = b$ if and only if no singular class of A has access to the support of b.*

Given an $n \times n$ matrix A, we denote by $\operatorname{mult}_0(A)$ the algebraic multiplicity of zero as an eigenvalue of A. We further denote by $\operatorname{index}_0(A)$ the size of the largest Jordan block of A corresponding to zero. The following result of Rothblum relates the index of an M-matrix with its reduced graph.

THEOREM 2.2 (Rothblum [1975, Thm. 3.1(ii)]). *Let A be an M-Matrix. Then* $\operatorname{index}_0(A)$ *is the largest number of singular classes in any chain in the reduced graph of A.*

An M-matrix A is said to have *property-c* if $\operatorname{index}_0(A) \leqq 1$. It follows from Theorem 2.2 that an M-matrix A has property-c if and only if no singular class of A has access to any other singular class of A, a result first observed in Schneider [1956, Thm. 3]. The next result is essentially due to Schneider [1956, Thm. 3].

THEOREM 2.3. *Let A be a Z-matrix. If there exists a vector $x \gg 0$ such that $Ax = 0$, then A is an M-matrix with property-c and the set of singular classes of A coincides with the set of final classes.*

We next come to splittings.

DEFINITION 2.4. Let A be a real square matrix. *A splitting* of A is a pair of matrices M, N such that M is nonsingular and $A = M - N$. The splitting is called:

(i) *Weak regular* (Ortega and Rheinboldt [1967]) if $M^{-1} \geqq 0$ and $M^{-1}N \geqq 0$;

(ii) *Regular* (Varga [1962]) if $M^{-1} \geqq 0$ and $N \geqq 0$;

(iii) *Graph compatible* (Schneider [1984]) if $\Gamma(M) \subseteq \overline{\Gamma(A)}$.

We remark that for a graph compatible splitting any access in the graphs of M and N and hence of $M^{-1}N$ also occurs in the graph of A. In particular, if A is in Frobenius normal form, then each of these matrices is also block upper triangular when partitioned conformably with A.

DEFINITION 2.5. Let A be a real square matrix. A *(k-fold) multisplitting* of A is a set of matrices $\{M_l, N_l, E_l\}_{l=1}^{k}$ such that

$$\begin{array}{ll} A = M_l - N_l \quad \text{is a splitting} & \\ 0 \leqq E_l \leqq I, & l = 1, \cdots, k, \\ \sum_{l=1}^{k} E_l = I. & \end{array} \tag{2.1}$$

[1] This result is restated in Schneider's survey paper [1986, Thm. 4.3(i)].

The multisplitting is called (*weak*) *regular*, respectively *graph compatible*, if each of the splittings (1.2) is (weak) regular or graph compatible.

Finally, we denote the column space of a matrix A by $R(A)$ and we denote by E_{ij} the square matrix whose (i, j)-entry is one and whose remaining entries are zero.

3. Consistency and convergence. Let (2.1) be a graph compatible weak regular multisplitting of an M-matrix A in Frobenius normal form. In the following all matrices and all vectors are assumed to be partitioned conformably with the block structure of A. Let $T = \{\alpha_j, \cdots, \alpha_{j+\nu}\}$ be any set of contiguous classes of A. Then for each $l = 1, \cdots, k$, $M_l[T]$ is a nonsingular block upper triangular matrix, whose block structure conforms with that of $A[T]$. Thus, by (1.9)–(1.12),

$$H[T] = \sum_{l=1}^{k} E_l[T](M_l[T])^{-1}N_l[T] = I - P[T]A[T].$$

In particular, if T consists of a single class of A, then $H[T]$ is the iteration matrix of the corresponding multisplitting of the irreducible M-matrix $A[T]$. In view of this we first investigate the properties of multisplittings of irreducible singular M-matrices.

LEMMA 3.1. *Let A be an $n \times n$ irreducible singular M-matrix, let* (2.1) *be a weak regular multisplitting, and let P be defined as in* (1.10). *Then we have the following*:

(i) *PA is an M-matrix with property-c and the set of singular classes of PA coincides with the set of final classes of PA.*

(ii) *If $b \in R(A)$, then the systems* (1.1) *and* (1.12) *have the same solution set if and only if P is a nonsingular matrix.*

Proof. (i) Evidently $PA = I - H$ is a Z-matrix and the strictly positive null vector of A is also a null vector of PA. The result now follows from the results of Schneider, Theorem 2.3.

(ii) The proof of the "sufficiency" part is trivial.

Suppose that the solution sets coincide. Then the nullspaces of A and PA are identical. Since $\text{index}_0(PA) = \text{index}_0(A) = 1$ we now must have that $\text{mult}_0(PA) = \text{mult}_0(A) = 1$. Suppose y is a vector such that $y^TP = 0$. Then as $y^T(PA) = 0$ and $\text{mult}_0(PA) = 1$, $y = \alpha x$, where $x > 0$ is a left nullvector of PA. Since P is a nonnegative matrix with all rows nonzero, it is not possible for $\alpha x^TP = 0$, unless $\alpha = 0$. Hence P is nonsingular. □

It is immediate from (1.2)–(1.4) that if the system (1.1) is solvable, then so is the system (1.12). Interestingly, for M-matrices with property-c we have the following partial converse which does not require the nonsingularity of P.

THEOREM 3.2. *Let A be an $n \times n$ M-matrix with property-c and let* (2.1) *be a graph compatible weak regular multisplitting. Then the system* (1.1) *is solvable if and only if the system* (1.12) *is solvable.*

Proof. We need only show sufficiency. Suppose for the sake of contradiction that the vector x satisfies the system (1.12), so that $PAx = Pb$, but $b \notin R(A)$, where P is given in (1.10). Since $\text{index}_0(A) = 1$, b admits the decomposition into

$$b = u + v, \quad 0 \neq u \in N(A), \quad v = Aw \in R(A).$$

Thus

$$PA(x - w) = Pu \quad \text{and} \quad u \notin R(A).$$

Let $u[\alpha_i]$ be the bottom nonzero component of u. Since A is block triangular we have that

$$A[\alpha_i]u[\alpha_i] = (Au)[\alpha_i] = 0.$$

Hence $u[\alpha_i]$ is a nonzero null vector of the irreducible M-matrix $A[\alpha_i]$ and we may assume $u[\alpha_i] \gg 0$ (otherwise we can replace the vectors x and b by their negatives). Let $T = \cup_{j=i}^{p} \alpha_j$. Since PA is block upper triangular we have

$$\begin{aligned}(PA[T])(x-w)[T] &= (PA(x-w))[T]\\ &= (Pu)[T]\\ &= P[T]u[T] > 0.\end{aligned}$$

Since $PA[T]$ is an M-matrix with property-c, by Berman and Plemmons [1979, Thm. 6.14.12, condition E_{12}], there exists a vector $y > 0$ such that

$$PA[T]y = P[T]u[T] > 0.$$

Note that the leading block of $PA[T]$, namely $PA[\alpha_i]$, is singular, while the first component of $P[T]u[T]$, namely $P[\alpha_i]u[\alpha_i]$, is strictly positive. Thus a singular class of $PA[T]$ has access to the support of $(Pu)[T]$. This contradicts Carlson's result, Theorem 2.1. □

Note that in the proof of Theorem 3.2 the hypothesis of graph compatibility allows us to restrict the multisplitting to certain principle submatrices. In the proof of our second main result we again exploit this technique to extend Lemma 3.1(ii).

THEOREM 3.3. *Let A be an M-matrix with property-c and let* (2.1) *be a graph compatible weak regular multisplitting. If $b \in R(A)$, then the systems* (1.1) *and* (1.12) *have the same solution set if and only if the matrix P defined in* (1.10) *is nonsingular.*

Proof. The "sufficiency" part of the statement is trivial. Suppose then that the systems (1.1) and (1.12) have the same (nonempty) solution set.

We first claim that PA is an M-matrix. To see this consider a single block $A[\alpha]$ in the Frobenius normal form of A. Since the multisplitting is graph compatible, the corresponding diagonal block of PA arises from a multisplitting of $A[\alpha]$. If $A[\alpha]$ is nonsingular, then it follows from (1.6)–(1.11) that $PA[\alpha] = P[\alpha]A[\alpha]$ is a nonsingular M-matrix. Alternatively, if $A[\alpha]$ is singular, then, by Lemma 3.1(i), $PA[\alpha]$ is an M-matrix with property-c. Thus the matrix PA is a block triangular Z-matrix whose diagonal blocks are M-matrices, and hence is itself an M-matrix.

Next, we show that PA has property-c by showing that no singular class of PA has access to any other singular class of PA. Let β_1 and β_2 be singular classes of PA. Then there exist singular classes α_1 and α_2 of A with $\beta_1 \subseteq \alpha_1$ and $\beta_2 \subseteq \alpha_2$. If $\alpha_1 \neq \alpha_2$, then since $\Gamma(PA) \subseteq \overline{\Gamma(A)}$ and A has property-c it follows that β_1 has no access to β_2. On the other hand, if $\alpha_1 = \alpha_2$, then as $PA[\alpha_1]$ has property-c, and we again conclude that β_1 has no access to β_2.

Finally, as (1.1) and (1.2) have the same (nonempty) solution set, A and PA have the same nullspace. Now, since index (A) = index $(PA) \leqq 1$, we see that

$$\text{mult}_0(PA) = \text{mult}_0(A).$$

Hence each class α of A contains at most one singular class of PA and so $P[\alpha]$ is nonsingular as in the proof of Lemma 3.1(ii). □

While the M-matrices that arise in the applications mentioned in § 1 have property-c, much of the proof of Theorem 3.3 applies to general M-matrices. In particular we have the following results.

COROLLARY 3.4. *Let A be an M-matrix and let* (2.1) *be a graph compatible weak regular multisplitting. Then we have the following*:

(i) *PA is an M-matrix.*

(ii) $\text{index}_0(PA) \leqq \text{index}_0(A)$.

(iii) $\operatorname{mult}_0(PA) \geqq \operatorname{mult}_0(A)$ *and equality holds if and only if P is nonsingular.*

Proof. Statements (i) and (iii) follow from the proof of Theorem 3.3.

By Rothblum's result, Theorem 2.2, it suffices to show that corresponding to any chain of classes of PA, A has a chain with the same number of singular classes. Let β_1 and β_2 be singular classes of PA. Then there exist singular classes α_1 and α_2 of A such that $\beta_1 \subseteq \alpha_1$ and $\beta_2 \subseteq \alpha_2$. If $\alpha_1 = \alpha_2$, then likewise to the proof of the theorem, β_1 has no access to β_2 in the reduced graph of PA. Thus if β_1 has access to β_2, α_1, and α_2 must be distinct and, since $\Gamma(PA) \subseteq \overline{\Gamma(A)}$, α_1 must have access to α_2 in the reduced graph of A. The result now follows. □

In the case when $\operatorname{index}_0(A) > 1$, Corollary 3.4 does not allow us to draw the same conclusion as in Theorem 3.3. This is because the inequality $\operatorname{mult}_0(PA) > \operatorname{mult}_0(A)$ does not necessarily imply that nullity $(PA) >$ nullity (A). We illustrate this with the following example, which also serves to show that Theorem 3.2 does not extend to general M-matrices.

Example 3.5. Let

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

and let

$$M_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

Let $N_i = M_i - A$, $i = 1, 2$, so that $A = M_i - N_i$, $i = 1, 2$, are both graph compatible weak regular splittings. Let

$$E_1 = I - E_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Then

$$P = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \text{and} \quad PA = \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Observe that P is singular, but nullity (A) = nullity $(PA) = 2$. Thus for any vector $b \in R(A)$, the systems (1.1) and (1.12) have the same solution set. On the other hand, if $0 \neq b \in N(P)$, e.g., $b = (0\ 0\ 1\ {-1})^T$, then $b \notin R(A)$, but $Pb \in R(PA)$.

We conclude the paper with a series of remarks.

Remark 3.6. The hypothesis of graph compatibility may not be dropped in Theorems 3.2 and 3.3. Indeed, even the case of a single splitting, i.e., $k = 1$, PA need not be an M-matrix (see, e.g., Neumann and Plemmons [1978, Remark 1]).

Remark 3.7. For a weak regular multisplitting of a singular irreducible M-matrix A, Lemma 3.1 gives us a graph-theoretic criterion for the nonsingularity of P, namely, P is nonsingular if and only if the graph of PA has precisely one final class. More generally, for a graph compatible weak regular multisplitting of an M-matrix with property-c, P is nonsingular if and only if for each singular class α of A, $PA[\alpha]$ has exactly one final class. This observation together with Theorem 3.3 yield graph-theoretic means of deciding when the systems (1.1) and (1.12) have the same solution set.

Remark 3.8. Given an M-matrix with property-c and a set of k graph compatible weak regular splittings, the problem arises of how to choose the weighting matrices E_l, $l = 1, \cdots, k$, so that matrix P is nonsingular and Theorem 3.3 applies. Recall that P is nonsingular if and only if $P[\alpha]$ is nonsingular for each class α of A. Further, if α is a nonsingular class of A, then $P[\alpha]$ is nonsingular by the results of O'Leary and White described in the Introduction. Suppose then that α is a singular class of A. If, for instance, there is some weighting matrix E_{l_0}, $1 \leqq l_0 \leqq k$, such that $E_{l_0}[\alpha]$ is nonsingular, then the graph of $PA[\alpha]$ contains that of $M_{l_0}^{-1}A[\alpha]$, which has a unique final class. Hence $PA[\alpha]$ itself has a unique final class and so $P[\alpha]$ is nonsingular by the previous remark. It follows that a sufficient condition for the nonsingularity of P is that for each singular class α of A, there exists l_α, $1 \leqq l_\alpha \leqq k$, such that $E_{l_\alpha}[\alpha]$ is nonsingular. The above condition, however, may not be particularly useful in practice since it may not be desirable to have a large number of nonzero entries in any of the weighting matrices. This, in turn, is because, as pointed out in § 1, the number of nonzero entries of E_l affects the amount of computation per iteration of the lth processor.

The above sufficient condition for the nonsingularity of P restricts only the choice of the weighting matrices and does not depend on the set of the k (graph compatible weak regular) splittings. An alternative approach to the problem is to impose conditions on the set of splittings chosen that ensure that P is nonsingular for any choice of the weighting matrices. An example of such a condition is the following. Assume A is irreducible and the multisplitting is *regular*. Suppose there exist indices $1 \leqq i, j \leqq n$ such that the (i, j)-entry of N_l is nonzero for each $l = 1, \cdots, k$. Then for sufficiently small $\varepsilon > 0$, the splittings

$$A + \varepsilon E_{ij} = M_l - (N_l - \varepsilon E_{ij}), \qquad l = 1, \cdots, k$$

are regular splittings of a nonsingular M-matrix. Thus $P = \sum_{l=1}^{k} E_l M_l^{-1}$ is nonsingular regardless of the choice of the weighting matrices. It now follows that for a graph compatible regular multisplitting of an M-matrix A, the matrix P is nonsingular provided that for each singular class α of A,

$$\bigcap_{l=1}^{k} \Gamma(N_l[\alpha]) \neq \varnothing. \tag{3.1}$$

An example of a situation where condition (3.1) is satisfied is the following Gauss–Seidel type of multisplitting. Let

$$A = D - L - U$$

be an M-matrix, where D, L, and U are, respectively, nonsingular diagonal, strictly lower triangular, and strictly upper triangular matrices. For $l = 1, \cdots, k$ let

$$0 \leqq L_l \leqq L \quad \text{and} \quad M_l = D - L_l$$

so that

$$N_l = U + L - L_l \geqq U \geqq 0.$$

Since A has only positive diagonal entries, any singular block of A has size at least 2 × 2, and hence shares a nonzero entry with U.

Remark 3.9. Let A be an M-matrix with property-c and let (2.1) be a graph compatible weak regular multisplitting. If $b \in R(A)$, then by Theorem 3.3 every limit point of the multisplitting iteration scheme (1.3) is a solution to (1.1) if and only if the matrix P is nonsingular. We observe that in this case the multisplitting iteration matrix H given in (1.9) can be obtained from the (single) graph compatible weak regular splitting

$$A = P^{-1} - P^{-1}H. \tag{3.2}$$

For a weak regular multisplitting of a nonsingular M-matrix A, the matrix P must be nonsingular, and hence (3.2) is a weak regular splitting of A, a fact observed by Elsner [1988]. It is worthwhile noting that every weak regular splitting of a nonsingular M-matrix is graph compatible (see Kavanagh [1988]).

REFERENCES

A. Berman and R. J. Plemmons [1979], *Nonnegative Matrices in the Mathematical Science*, Academic Press, New York.

D. H. Carlson [1963], *A note on M-matrix equations*, J. Soc. Industrial Appl. Math., 11, pp. 1027–1033.

L. Elsner [1988], private communication.

J. P. Kavanagh [1988], *Splittings of M-matrices*, Ph.D. thesis, University of Wisconsin, Madison, WI.

M. Neumann and R. J. Plemmons [1978], *Convergent nonnegative matrices and iterative methods for consistent linear systems*, Numer. Math., 31, pp. 265–279.

D. P. O'Leary and R. E. White [1985], *Multi-splittings of matrices and parallel solution of linear systems*, SIAM J. Algebraic Discrete Meth., 6, pp. 630–640.

J. M. Ortega and W. C. Rheinboldt [1967], *Monotone iterations for nonlinear equations with applications to Gauss–Seidel methods*, SIAM J. Numer. Anal., 4, pp. 171–190.

R. J. Plemmons [1976], *Regular splittings and the discrete Neumann problem*, Numer. Math., 25, pp. 153–161.

U. G. Rothblum [1975], *Algebraic eigenspaces of nonnegative matrices*, Linear Algebra Appl., 12, pp. 281–292.

H. Schneider [1956], *The elementary divisors, associated with* 0, *of a singular M-matrix*, Proc. Edinburgh Math. Soc., 10, pp. 108–122.

———, [1984], *Theorems on M-splittings of a singular M-matrix which depend on graph structure*, Linear Algebra Appl., 58, pp. 407–424.

———, [1986], *The influence of the marked reduced graph of a nonnegative matrix on the Jordan form and on related properties: a survey*, Linear Algebra Appl., 84, pp. 161–189.

R. S. Varga [1962], *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ.

D. M. Young [1971], *Iterative Solutions of Large Linear Systems*, Academic Press, New York.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 2, pp. 219-228, April 1989

007

ALGORITHMS FOR THE RECONSTRUCTION OF SPECIAL JACOBI MATRICES FROM THEIR EIGENVALUES*

M. HEGLAND† AND J. T. MARTI†

Dedicated to the memory of Peter Henrici

Abstract. Two algorithms for the reconstruction of symmetric tridiagonal (not necessarily persymmetric) matrices J with subdiagonal entries equal to one from their eigenvalues are established. The first algorithm is an iteration method using orthogonal similarity transformations in the sense of an inverted Jacobi algorithm and is shown to be locally convergent. Since reconstruction problems are often rather ill-conditioned, the algorithm may be slow, but it gives good approximations J' to J. J' may be used as a starting value for the second algorithm, a Newton method iterating the characteristic polynomial of J'. Numerical examples demonstrate the convergence behavior, also for nonpersymmetric matrices J.

Key words. reconstruction algorithms, Jacobi matrices, Jacobi algorithm, inverse matrix eigenvalue problem

AMS(MOS) subject classifications. 65F30, 65F15

1. Introduction. *Special Jacobi matrices* are tridiagonal matrices of the following form:

$$J_q = \begin{bmatrix} q_1 & 1 & 0 & 0 & \cdot \\ 1 & q_2 & 1 & . & 0 \\ 0 & 1 & . & 1 & 0 \\ 0 & \cdot & 1 & q_{n-1} & 1 \\ \cdot & 0 & 0 & 1 & q_n \end{bmatrix}. \tag{1.1}$$

Such matrices occur in the discretized version

$$J_q x = \mu x \tag{1.2}$$

of the Sturm–Liouville problem

$$-u'' + au = \lambda u, \qquad u(0) = u(1) = 0, \tag{1.3}$$

where J_q has diagonal elements

$$q_i := -h^2 a(x_i), \quad x_i := ih, \quad 1 \leqq i \leqq n, \quad h := (n+1)^{-1}, \tag{1.4}$$

the symbol a denotes the potential function in (1.3), q is a vector in $\mathbf{R}^n$, and the first eigenvalues μ_i of J_q approximate $2 - h^2\lambda_i$, where $\lambda_1 \leqq \lambda_2 \leqq \cdots$ are the eigenvalues λ of (1.3) and $\mu_1 \geqq \mu_2 \geqq \cdots \geqq \mu_n$.

The inverse Sturm–Liouville problem related to (1.3) is the problem of the reconstruction of potentials a from the set of eigenvalues of (1.3). Strongly related to this problem is the problem of reconstructing $q \in \mathbf{R}^n$, and hence the special Jacobi matrix J_q from the spectrum $\sigma(J_q) = \{\mu_1, \cdots, \mu_n\}$ of J_q.

* Received by the editors December 28, 1987; accepted for publication (in revised form) September 21, 1988.

† Institute of Applied Mathematics, Eidgenössische Technische Hochschule Zurich, CH-8092 Zürich, Switzerland (hegland@math.ethz.ch and marti@math.ethz.ch).

It is well known [5] that J_q has n different eigenvalues and that the *inverse matrix eigenvalue problem* of reconstructing J_q from $\sigma(J_q)$ may have more than one solution. If P is the reversion matrix given by

$$P=\begin{bmatrix} \cdot & 0 & 0 & 0 & 1 \\ 0 & \cdot & 0 & 1 & 0 \\ 0 & 0 & \cdot & 0 & 0 \\ 0 & 1 & 0 & \cdot & 0 \\ 1 & 0 & 0 & 0 & \cdot \end{bmatrix},$$

then a symmetric matrix A is called *persymmetric* if $A = PAP$ or, equivalently, if $PA = AP$. In [5] it is also shown that the above problem for persymmetric matrices J_q has a unique solution. Algorithms for the solution of such inverse eigenvalue problems have been given by de Boor and Golub [2], Biegler-König [1], and Gragg and Harrod [4]. In the above situation, the first algorithm applies to persymmetric matrices J_q only, whereas the second algorithm, as a Newton method, converges rapidly, also for nonpersymmetric matrices, but depends on starting matrices close enough to J_q. The Lanczos and Rutishauser algorithms treated in the last paper are extremely fast and stable and reconstruct Jacobi matrices with arbitrary (positive) subdiagonal elements.

In this paper we describe an algorithm working with orthogonal similarity transformations using plane rotations $R(j, k, t)$ that rotate the (j, k) plane in $\mathbf{R}^n$ through an angle $\theta = \arctan t$,

$$R(j,k,t):=\begin{array}{cc} \left[\begin{array}{ccccc} 1 & \vdots & & \vdots & \\ \cdots & c & \cdots & s & \cdots \\ & \vdots & & \vdots & \\ \cdots & -s & \cdots & c & \cdots \\ & \vdots & & \vdots & 1 \end{array}\right] & \begin{array}{c} \\ j \\ \\ k, \\ \\ \end{array} \\ \begin{array}{ccccc} & j & & k & \end{array} & \end{array} \qquad t\in\mathbf{R}, \quad 1\leqq j<k\leqq n,$$

where $c = \cos\theta$ and $s = \sin\theta$, starting with the diagonal matrix $A_0 := \operatorname{diag}(\mu_n, \cdots, \mu_1)$ to obtain the iteration sequence $\{A_m\}$ given by

$$(1.5) \qquad A_m := Q(A_{m-1})^T A_{m-1} Q(A_{m-1}), \qquad m\in\mathbf{N}.$$

The matrices $Q(A_{m-1})$ are defined by

$$(1.6) \qquad Q(A_{m-1}) := \prod_{j=1}^{n-1} \prod_{k=j+1}^{n} R(j,k,s_{jk}),$$

where the matrices S of tangents s_{jk} of rotation angles depend on A_{m-1} and are chosen such that after each similarity transformation with a plane rotation the iterated (symmetric) matrix A_m has off-diagonal elements that are closer to one for subdiagonal and to zero for other off-diagonal elements. The details are given in § 2. The algorithm for the reconstruction of special Jacobi matrices J_q, $q \in \mathbf{R}^n$ is stated explicitly in § 3. In § 4 we present sufficient conditions on the potential vector q for the local convergence of the algorithm. Since the inverse eigenvalue problem of reconstructing J_q from $\sigma(J_q)$ is often rather ill-conditioned it is not surprising that numerical evidence shows a rather slow convergence of the algorithm. However, locally fast converging Newton methods require initial potential vectors that are relatively close to q. These starting vectors are now available as elements of the iteration sequence $\{A_m\}$ given by (1.5). A Newton algorithm iterating the characteristic polynomial of special Jacobi matrices J_p such that p converges to q is given in § 5. This method is much simpler and faster but probably less stable than the Newton iteration proposed in [1]. Numerical examples for persymmetric, nonpersymmetric and for strongly varying potential vectors q in J_q are given in § 6.

2. The choice of the rotation angles. The tangents of the rotation angles, i.e., the superdiagonal entries $s_{jk}(1 \leqq j < k \leqq n)$ of the "sweep matrix" S are chosen such that $\text{off}(B - J_0) \leqq \text{off}(A - J_0)$, where

$$\text{off}(C) := \|C\|_F^2 - \sum_{i=1}^{n} c_{ii}^2 \tag{2.1}$$

and $\|C\|_F$ is the Frobenius norm of $C \in \mathbf{R}^{n \times n}$ and where B is the iterated matrix obtained by an orthogonal similarity transform of A,

$$B := R(j,k,s_{jk})^T A R(j,k,s_{jk}).$$

To be more specific, we have by an elementary computation

$$\begin{aligned} \text{off}(B - J_0) = {} & 2[zsc + (c^2 - s^2)a_{jk}]^2 + 4x(1-c) + 4ys \\ & - 4\delta_{j+1,k}[zsc + a_{jk}(c^2 - s^2)] + \text{off}(A - J_0) - 2a_{jk}^2 - 4\delta_{j+1,k}a_{jk}, \end{aligned} \tag{2.2}$$

where $c := (1 + s_{jk}^2)^{-1/2}$, $s := s_{jk}c$, δ_{jk} is the Kronecker symbol,

$$x := a_{j-1,j} + a_{k,k+1} + (1 - \delta_{j+1,k})[a_{j,j+1} + a_{k-1,k}], \tag{2.3}$$

$$y := a_{j-1,k} - a_{j,k+1} + (1 - \delta_{j+1,k})[a_{j+1,k} - a_{j,k-1}], \tag{2.4}$$

and

$$z := a_{jj} - a_{kk}. \tag{2.5}$$

A choice of s_{jk} such that $\text{off}(B - J_0)$ is minimal would be optimal, but one step of a Newton iteration for a zero of the first derivative of $\text{off}(B - J_0)$ with respect to s_{jk} turns out to be a workable practical compromise and asymptotically sufficient. Again by an elementary computation, this leads to the formula

$$s_{jk} = \frac{z(a_{jk} - \delta_{j+1,k}) + y}{4a_{jk}^2 - z^2 - 4\delta_{j+1,k}a_{jk} - x}. \tag{2.6}$$

3. An algorithm for the reconstruction of special Jacobi matrices from the eigenvalues. The following algorithm starts with a diagonal matrix A_0 with the eigenvalues $\mu_n < \mu_{n-1} < \cdots < \mu_1$ of J_q as diagonal elements. The elements of A_{m-1} are overwritten by the elements of A_m.

ALGORITHM 3.1 (one sweep for iteration (1.5)).

```
For $j = 1, \cdots, n-1$
    For $k = j+1, \cdots, n$
    computation of $s$, $c$, and $z$ by (2.3)-(2.6)
        For $l = 1, \cdots, n \ (l \neq j,k)$
        $a := a_{jl}$
        $a_{jl} := a_{lj} := ca - sa_{kl}$
        $a_{kl} := a_{lk} := sa + ca_{kl}$
    $a := a_{jj}$, $b := a_{kk}$, $d := a_{jk}$
    $a_{jj} := c^2a - 2scd + s^2b$
    $a_{kk} := s^2a + 2scd + c^2b$
    $a_{jk} := a_{kj} := scz + (c^2 - s^2)d$
```

4. Convergence proof. For $q \in \mathbf{R}^n$ the manifold

$$I_q := \{A \in \mathbf{R}^{n\times n} : A^T = A, \sigma(A) = \sigma(J_q)\}$$

of $\mathbf{R}^{n\times n}$ is called the *isospectral set* of J_q. Since the algorithm (1.5) works with orthogonal similarity transforms, we introduce the mapping $S_q : \mathrm{SO}(n) \to I_q$ given by

$$S_q(U) := U^T J_q U, \qquad U \in \mathrm{SO}(n),$$

where SO (n) is the group of orthogonal matrices in $\mathbf{R}^n$ with determinant one. The following proposition shows the connection of the two manifolds I_q and $S_q(\mathrm{SO}(n))$, i.e., that the isospectral set I_q of J_q is the orbit $S_q(\mathrm{SO}(n))$ of J_q under the group of special orthogonal similarity transformations.

PROPOSITION 1. *For every $q \in \mathbf{R}^n$, $I_q = S_q(\mathrm{SO}(n))$.*

Proof. It is clear that $I_q \supset S_q(\mathrm{SO}(n))$. It is well known (see, e.g., [3, Thm. 8.1.1] or [6, Thm. 1.4]) that for $A \in I_q$ there is an orthogonal matrix $U = [u_1 \cdots u_n]$ such that

$$A = U \Lambda_q U^T,$$

where $\Lambda_q := \operatorname{diag}(\mu_1, \cdots, \mu_n)$ and $\mu_1, \cdots, \mu_n$ are the eigenvalues of J_q. Next, let $U_- := [u_1 \cdots u_{n-1} \; -u_n]$. Since we also have

$$A = U_- \Lambda_q U_-^T$$

and U_- is orthogonal with determinant -1 we may, without loss of generality, assume that $U \in \mathrm{SO}(n)$. Therefore, we also have $J_q = V\Lambda_q V^T$ and thus $A = UV^T J_q V U^T$ for some $V \in \mathrm{SO}(n)$, which finally implies that $A \in S_q(\mathrm{SO}(n))$. □

Next, let $\{f_A\}$ be the family of real functions on SO (n) given by

$$f_A(U) := \operatorname{off}(U^T A U - J_0), \qquad U \in \mathrm{SO}(n), \quad A \in \mathbf{R}^{n\times n}.$$

It follows that

$$f_{J_q}(U) = \operatorname{off}(S_q(U) - J_0), \qquad U \in \mathrm{SO}(n).$$

Moreover, let $Q : \mathbf{R}^{n\times n} \to \mathrm{SO}(n)$ be a continuous mapping, here called an *iteration mapping*, such that $Q(J_q) = I$, $q \in \mathbf{R}^n$. Then given $q \in \mathbf{R}^n$ and $A_0 \in I_q$, Q defines an *iteration sequence* $\{A_m\} \subset I_q$ by

$$A_m := Q(A_{m-1})^T A_{m-1} Q(A_{m-1}), \qquad m \in N.$$

This implies that

$$f_{A_m}(I) = f_{A_{m-1}}(Q(A_{m-1})), \qquad m \in \mathbf{N}.$$

DEFINITION 1. Let Y_Q be the set of all $q \in \mathbf{R}^n$ such that there is a compact neighbourhood W of I in SO (n) satisfying

(i) $Q(A)^T A Q(A) \in S_q(W), A \in S_q(W)$;

(ii) $f_A(Q(A)) < f_A(I)$ if $J_q \neq A \in S_q(W)$.

THEOREM 1. *For every $q \in Y_Q$ there is a neighbourhood W of I in* SO (n) *such that for $A_0 \in S_q(W)$ the iteration sequence $\{A_m\}$ defined by Q converges to J_q in I_q.*

Proof. If q is in Y_Q there is a neighbourhood W obeying properties (i) and (ii) of Definition 1. Since S_q is continuous on the compact set W in SO (n) there is at least one limit point A of $\{A_m\}$ in $S_q(W)$. If $A \neq J_q$ the continuous dependence of $f_A(U)$ on A and U, the continuity of Q and (ii) then imply the existence of an m in $\mathbf{N}$ such that

$$f_{A_m}(Q(A_m)) < f_A(I).$$

This contradiction to the fact that $\{f_{A_m}(I)\}$ is a monotone sequence finally yields $\lim_m A_m = J_q$. □

To prove the following lemma we now introduce a mapping g of $\mathbf{R}^\nu$, where $\nu := n(n-1)/2$, into SO (n) such that g is a homeomorphism of a zero neighbourhood in $\mathbf{R}^\nu$ into a neighbourhood of the identity I of SO (n). For this purpose, let $g:\mathbf{R}^\nu \to$ SO (n) be the Cayley transform

$$g(x) := (I-Sx)(I+Sx)^{-1}, \qquad x\in\mathbf{R}^\nu,$$

where S is the one-to-one linear transformation of $\mathbf{R}^\nu$ into the subspace $\mathbf{S}^\nu$ of all skew symmetric matrices in $\mathbf{R}^{n\times n}$, given by

$$Sx=\begin{bmatrix} 0 & x_1 & & \cdots & x_{n-1} \\ -x_1 & 0 & x_n & \cdots & x_{2n-3} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ & & & & x_\nu \\ -x_{n-1} & & & -x_\nu & 0 \end{bmatrix}, \qquad x\in\mathbf{R}^\nu.$$

Let $d_x g$ denote the first derivative of g at x and $d_x^2 g$ the corresponding second derivative (i.e., the Hessian bilinear form). An elementary calculation then shows that

$$d_0 g(x) = -2Sx \quad \text{and} \quad d_0^2 g(x,x) = 4Sx^2, \quad x\in\mathbf{R}^\nu.$$

Similarly, let $d_U^p f_{J_q}$ $(p = 1, 2)$ be the corresponding derivatives of f_{J_q} at $U\in$ SO (n).

LEMMA. *If $q\in\mathbf{R}^n$ is such that the bilinear form $d_I^2 f_{J_q}$ is positive definite on* $\mathbf{S}^\nu\times\mathbf{S}^\nu$ *there is a compact neighbourhood V of I in* SO (n) *such that*

(*) *I is the only critical point of f_{J_q} in V;*

(**) *For every compact neighbourhood V' of I in V there is a compact neighbourhood W of I in V' such that*

$$W = \{U\in V: f_{J_q}(U) \leq \inf\{f_{J_q}(U'): U'\in\partial V'\}\}.$$

Proof. We use the facts that f_{J_q} is smooth on SO (n), that $f_{J_q}(I) = 0$, and that $f_{J_q} \geq 0$. Thus, I is an absolute minimum of f_{J_q} and a critical point, i.e., $d_I f_{J_q} = 0$. Let X be the closed unit ball of $\mathbf{R}^\nu$. Then for $s\in(0, 1)$ and any $x\in\partial X$ by Taylor's Theorem and the chain rule we obtain

$$\begin{aligned} d_{sx}(f_{J_q}\circ g)(x) &= \int_0^1 d_{stx}^2(f_{J_q}\circ g)(sx,x)\, dt \\ &= \int_0^1 [d_{g(stx)}^2 f_{J_q}(d_{stx}g(sx), d_{stx}g(x)) + d_{g(stx)}f_{J_q}(d_{stx}^2 g(sx,x))]\, dt. \end{aligned}$$

Since ∂X is compact we get

$$a := \inf\{d_I^2 f_{J_q}(Sx,Sx): x\in\partial X\} > 0,$$

and thus

$$\begin{aligned} \lim_{s\to 0} s^{-1} d_{sx}(f_{J_q}\circ g)(x) &= \lim_{s\to 0} 4\int_0^1 [d_I^2 f_{J_q}(Sx,Sx) + d_{g(stx)} f_{J_q}(Sx^2)]\, dt \\ &= 4\, d_I^2 f_{J_q}(Sx,Sx) \\ &\geq 4a, \qquad x\in\partial X. \end{aligned}$$

This implies the existence of an $\varepsilon > 0$ such that

$$s^{-1} d_{sx}(f_{J_q} \circ g)(x) \geqq 2a, \qquad s \in (0, \varepsilon], \quad x \in \partial X. \tag{4.1}$$

Hence $d_U f_{J_q} \neq 0$ for $U \neq I$ on the compact neighbourhood $V := g(\varepsilon X)$ of I in SO (n). This shows (*).

Moreover, by Taylor's Theorem,

$$\begin{aligned} f_{J_q} \circ g(sx) &= \int_0^1 d_{stx}(f_{J_q} \circ g)(sx)\, dt \\ &= \int_0^1 s d_{stx}(f_{J_q} \circ g)(x)\, dt \\ &\geqq \int_0^1 2s^2 ta\, dt = s^2 a, \qquad s \in (0, \varepsilon], \quad x \in \partial X. \end{aligned}$$

Therefore, $f_{J_q}(U) > 0$, $I \neq U \in V$. Next, let V' be any compact neighbourhood of I in V. Since f_{J_q} attains its absolute minimum on $\partial V'$, we then have

$$b := \inf \{ f_{J_q}(U) : U \in \partial V' \} > 0.$$

By the continuity of f_{J_q},

$$W := \{ U \in V : f_{J_q}(U) \leqq b \}$$

is a compact neighbourhood of I in V. Finally, we can show that $W \subset V'$. For this purpose, let $x \in \mathbf{R}^\nu$ be such that $g(x) \in W \setminus V'$ and let

$$r := \sup \{ t \in (0, 1) : g(tx) \in V' \}.$$

Obviously, $r \in (0, 1)$. Again by Taylor's Theorem, the substitution $s = rt$ and (4.1), we then obtain

$$\begin{aligned} f \circ g(x) &= \int_0^1 d_{sx}(f_{J_q} \circ g)(x)\, ds \\ &= \int_0^1 d_{trx}(f_{J_q} \circ g)(rx)\, dt + \int_r^1 d_{sx}(f_{J_q} \circ g)(x)\, ds \\ &\geqq f_{J_q} \circ g(rx) + \int_r^1 2as\, ds \\ &\geqq b + a - ar^2 > b. \end{aligned}$$

This proves (**). □

THEOREM 2. *If Q is an iteration mapping, then the following conditions are sufficient for $q \in \mathbf{R}^n$ to be an element of Y_Q:*

(1) *$d_I^2 f_{J_q}$ is positive definite on $\mathbf{S}^\nu \times \mathbf{S}^\nu$;*

(2) *There is a neighbourhood V of I in* SO (n) *such that*

(a) $f_{S_q(U)}(Q \circ S_q(U)) \leqq f_{J_q}(U)$, $U \in V$,

(b) $f_{S_q(U)}(Q \circ S_q(U)) < f_{J_q}(U)$, *if $I \neq U \in V$ and U is not a critical point of f_{J_q}.*

Proof. Let $d_I^2 f_{J_q}$ be positive definite for some $q \in \mathrm{R}^n$. By the above lemma we may assume that V coincides with the neighbourhood V of the lemma. Since Q and S_q are continuous there is a compact neighbourhood V' of I in V such that

$$UQ \circ S_q(U) \in V, \qquad U \in V'. \tag{4.2}$$

Again by the above lemma there is a compact neighbourhood W of I in V' such that

$$W=\{U\in V{:}f_{J_q}(U)\leqq\inf\{f_{J_q}(U'){:}\,U'\in\partial V'\}\}. \tag{4.3}$$

In view of (4.2) we have

$$Q\circ S_q(U)^TS_q(U)Q\circ S_q(U)\in S_q(V), \qquad U\in W.$$

Thus for every $U\in W$ there is an $U'\in V$ such that

$$Q\circ S_q(U)^TS_q(U)Q\circ S_q(U)=S_q(U'), \tag{4.4}$$

and thus

$$f_{S_q(U)}(Q\circ S_q(U))=f_{J_q}(U').$$

By Theorem 2(2)(a) this yields

$$f_{J_q}(U')\leqq f_{J_q}(U),$$

which by (4.3) implies that $U'\in W$. Hence $A:=S_q(U)(U\in W)$ by (4.4) satisfies

$$Q(A)^TAQ(A)\in S_q(W), \qquad A\in S_q(W),$$

which is (i) of Definition 1.

To show (ii) we use the fact that I is the only critical point of f_{J_q} in W. Then the relation $A=S_q(U)$ implies that the condition (ii) is an immediate consequence of condition (2)(b). □

Let us now compute the bilinear form $d_I^2f_{J_q}$ on $\mathbf{S}^\nu\times\mathbf{S}^\nu$. If $E_{jk}\in\mathbf{S}^\nu$ is given by

$$(E_{jk})_{lm}:=\delta_{lj}\delta_{mk}-\delta_{lk}\delta_{mj}, \qquad 1\leqq j<k\leqq n, \qquad 1\leqq l<m\leqq n,$$

it is clear that the set $\{E_{jk}:1\leqq j<k\leqq n\}$ is a basis for $\mathbf{S}^\nu$, that $\mathbf{S}^\nu=S(\mathbf{R}^\nu)$, and that the elements $S^{-1}E_{jk}$ are the column vectors of the unit matrix of order ν. Using the definition of $d_U^pf_{J_q}$, $U\in\mathrm{SO}(n)$ $(p=1,2)$, we obtain by an elementary computation

$$d_Uf_{J_q}(E_{jk})=2\sum_{1\leqq r<r'\leqq n}(U^TJ_qU-J_0)_{rr'}(E_{jk}^TJ_qU+U^TJ_qE_{jk})_{rr'}$$

and

$$d_I^2f_{J_q}(E_{jk},E_{lm})=2\sum_{1\leqq r<r'\leqq n}(E_{jk}^TJ_q+J_qE_{jk})_{rr'}(E_{lm}^TJ_q+J_qE_{lm})_{rr'}.$$

Since for $1\leqq j<k\leqq n$ and $q_{jk}:=q_j-q_k$

$$E_{jk}J_q+J_qE_{jk}^T=\begin{array}{cc}\left[\begin{array}{ccccccc}\cdot & & & & -1 & & \\ & \cdot & & & 1-q_{jk} & 1 & \\ & & \cdot & & -1 & & \\ & 1 & & \cdot & & & \\ -1 & q_{jk}-1 & & & \cdot & & \\ & 1 & & & & \cdot & \\ & & & & & & \cdot\end{array}\right] & \begin{array}{c} \\ j \\ \\ \\ k \\ \\ \\ \end{array} \\ \begin{array}{ccccccc} & j & & & \phantom{1-q_{jk}} k & & \end{array} & \end{array}$$

the mapping $(x,y)\rightarrow d_I^2f_{J_q}(Sx,Sy)$, $x,y\in\mathbf{R}^\nu$ defines a bilinear form. The (Hessian) matrix of this form is $\frac{1}{2}$ times a similarity transform with a permutation matrix of a symmetric pentadiagonal block matrix H_q with diagonal blocks $B_k(1\leqq k<n)$ and

subdiagonal blocks $C_k (1 < k < n)$ and $D_k (2 < k < n)$. The block $B_1 \in \mathbf{R}^{(n-1)\times(n-1)}$ is tridiagonal with diagonal entries

$$1+q_{12}^2, 2+q_{23}^2, 2+q_{34}^2, \cdots, 2+q_{n-2,n-1}^2, 1+q_{n-1,n}^2$$

and subdiagonal entries -1. The matrices $B_k \in \mathbf{R}^{(n-k)\times(n-k)} (1 < k < n-1)$ are given by

$$B_k = \operatorname{diag}\,(3+q_{1,k+1}^2, 4+q_{2,k+2}^2, 4q_{3,k+3}^2, \cdots, 4q_{n-k-1,n-1}^2, 3+q_{n-k,n}^2)$$

and $B_{n-1} := [2+q_{1n}^2]$. The blocks $C_k \in \mathbf{R}^{(n-k)\times(n-k+1)}$ for $1 < k < n$ are matrices with vanishing elements, except the diagonal elements

$$2q_i - q_{i+k-1} - q_{i+k} (1 \leqq i \leqq n-k)$$

and the superdiagonal elements

$$-q_i - q_{i+1} + 2q_{i+k} (1 \leqq i \leqq n-k).$$

Finally, $D_k \in \mathbf{R}^{(n-k)\times(n-k+2)}$ are matrices with vanishing elements except the diagonal and the first two superdiagonal elements which have constant values 1, -2, and 1, respectively.

The following example shows the existence of potential vectors $q \in \mathbf{R}^n$ such that H_q is singular and hence not positive definite. For $n = 3$ we obtain

$$\det H_q = \det \begin{bmatrix} 1+q_{12}^2 & -1 & q_{23}+q_{31} \\ -1 & 1+q_{23}^2 & q_{13}+q_{23} \\ q_{21}+q_{31} & q_{13}+q_{23} & 2+q_{13}^2 \end{bmatrix}$$
$$= q_{13}^2$$

and the last quantity obviously vanishes if and only if J_q is persymmetric.

THEOREM 3. *If Q is defined by* (1.6) *and Algorithm* 3.1 *then the set Y_Q is dense in $\mathbf{R}^n$ and for every $q \in Y_Q$ there is a neighbourhood W of I in* $\mathrm{SO}(n)$ *such that* $\lim_m A_m = J_q$ *for any starting matrix* $A_0 \in S_q(W)$.

Proof. For every $q \in \mathbf{R}^n$ the matrix H_q is positive semidefinite. The determinants of the leading principal submatrices of H_q are polynomials in q, and hence do not vanish on a dense set of points in $\mathbf{R}^n$. Thus it is clear that the set of q's where H_q is positive definite is again dense in $\mathbf{R}^n$. Since Q given by (1.6) and Algorithm 3.1 satisfies Theorem 2(2) it follows that Y_Q is dense in $\mathbf{R}^n$. The rest of the proof follows as a direct consequence of Theorems 2 and 1. □

5. A Newton method for the reconstruction of special Jacobi matrices. A well-known recursion relation (see, e.g., [2]) for the characteristic polynomial $\varphi_n(\lambda)$ of $J_p (p \in \mathbf{R}^n)$ is

$$\varphi_k(\lambda) = (\lambda - p_k)\varphi_{k-1}(\lambda) - \varphi_{k-2}(\lambda), \qquad 1 \leqq k \leqq n, \tag{5.1}$$

where $\varphi_{-1}(\lambda) := 0$ and $\varphi_0(\lambda) := 1$, $\lambda \in \mathbf{C}$. We write $\varphi_n(\lambda) = a_n^T z + \lambda^n$ and

$$\varphi_k(\lambda) = a_k^T z, \qquad -1 \leqq k < n,$$

where a_k and $z := (1, \lambda, \lambda^2, \cdots, \lambda^{n-1})^T \in \mathbf{C}^n$. Equation (5.1) is then equivalent to

$$a_k = (E - p_k I) a_{k-1} - a_{k-2}, \qquad 1 \leqq k \leqq n,$$

where $\{e_1, \cdots, e_n\}$ is the standard basis of $\mathbf{R}^n$, E is the shift matrix $[e_2 e_3 \cdots e_n 0]$, $a_{-1} := 0$, and $a_0 := e_1$. Now let $B = [b_1 \cdots b_n] := [\partial a_n / \partial p_1 \cdots \partial a_n / \partial p_n]$ be the Jacobian matrix of the vector a_n (both depending on p) with respect to the potential vector p of J_p and let $c \in \mathbf{R}^n$ be a vector such that the characteristic polynomial φ of J_q is given by

$$\varphi(\lambda)=\prod_{i=1}^{n}(\lambda-\mu_i)=c^Tz+\lambda^n, \qquad \lambda\in\mathbf{C}.$$

One step of Newton's method for the computation of the zero q of $a_n - c$ iterating an approximation p for (the potential vector) q then is

$$p:=p-x, \quad x=\text{solution of } Bx=a_n-c.$$

The following algorithm computes the Jacobian matrix B of a_n and the right-hand side of the linear system $Bx = c - a_n$, where for the sake of compactness we use the notation b_0 for a_n.

ALGORITHM 5.1 (computes the Jacobian matrix B and the vectors a_n ($=b_0$) and c for the proposed Newton method).

```
$b_0 := c := e_1$
For $k=0, \cdots, n$
    $s := 0$
    For $j=k+1, \cdots, n$
        If $k=0$ then $[b_j = b_0, c := Ec-\mu_j c]$
        $t:=s, s:=b_k, b_k:=Es-p_j s-t$
```

6. Numerical examples. The following numerical examples are concerned with the reconstruction of special Jacobi matrices $J_q \in \mathbf{R}^{n\times n}$ for $n = 15$ with $q \in \mathbf{R}^n$ given by (1.4), where $a : (0, 1) \to \mathbf{R}$ is given by

$$h^{-2}a(x):=\begin{cases}\sin(\pi x), & \\ \text{sqrt}(x), & x\in(0,1). \\ \text{sign}(x-\frac{1}{2}), & \end{cases}$$

The data for the reconstruction are the spectra $\sigma(J_q)$ of J_q that are computed here by the well-known QR algorithm for tridiagonal matrices. The numerical results have been obtained by using 512 iterations of the type (1.5) based on Algorithm 3.1. Let $p^{(m)} \in \mathbf{R}^n$ be the resulting vectors of diagonal elements of A_m. $p^{(512)}$ then has been used as a starting vector for $k = 2, 4, 8, \cdots, 64$ steps of the Newton algorithm, Algorithm 5.1, producing approximations $q^{(k)}$ in $\mathbf{R}^n$ for the potential vector q of J_q. Since the

TABLE 6.1

Example		1	2	3
$m =$	8	1.0 − 0	1.0 − 0	1.0 − 0
	32	5.7 − 1	5.7 − 1	5.7 − 1
	128	3.7 − 1	3.7 − 1	3.7 − 1
	512	1.8 − 1	1.8 − 1	1.8 − 1
$k =$	2	6.0 − 2	5.7 − 2	5.5 − 2
	4	1.5 − 2	1.4 − 2	1.2 − 2
	8	2.3 − 3	1.5 − 3	7.7 − 4
	16	9.7 − 4	3.7 − 4	6.0 − 5
	32	4.7 − 4	5.8 − 5	6.6 − 7
	64	2.4 − 4	2.2 − 6	1.7 − 8

condition numbers of the Jacobian matrices B are strongly increasing with k, the parallel-chord Newton method has been applied for $k > 6$, where the matrix B of the case $k = 6$ has been kept for the corresponding B's for $k > 6$. Slightly better results could be achieved using the more elaborate singular value decomposition instead of the Gauss algorithm with partial pivoting for solving the system $Bx = c - a_n$.

Table 6.1 states the error in the mean for $p^{(m)}$ and $q^{(k)}$ given by the Euclidean norm of the error vectors $p^{(m)} - q$ and $q^{(k)} - q$, respectively, divided by $n^{1/2}$.

REFERENCES

[1] F. W. BIEGLER-KÖNIG, *A Newton iteration process for inverse eigenvalue problems*, Numer. Math., 37 (1981), pp. 349-354.

[2] 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.

[3] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[4] W. B. GRAGG AND W. J. HARROD, *The numerically stable reconstruction of Jacobi matrices from spectral data*, Numer. Math., 44 (1984), pp. 317-335.

[5] H. HOCHSTADT, *On the construction of Jacobi matrices from spectral data*, Linear Algebra Appl., 8 (1974), pp. 435-446.

[6] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 2, pp. 229-232, April 1989

GEOMETRIC PROPERTIES OF HIDDEN MINKOWSKI MATRICES*

WALTER D. MORRIS, JR.† AND JIM LAWRENCE†

Abstract. It is known that a vector satisfying a certain exponentially large set of inequalities related to a matrix M will allow the solution of linear complementarity problems with matrix M in n steps. It is shown that the set of such vectors, if nonempty, is the interior of a simplicial cone. The defining inequalities for this cone show that M^T is hidden Minkowski if such a vector exists.

Key words. hidden Minkowski matrices, linear complementarity problem

AMS(MOS) subject classifications. 15A39, 90C99

1. Introduction. Let $m_1, \cdots, m_n$ be a set of n points in $\mathbb{R}^n$, and let $e_1, \cdots, e_n$ be the standard basis for $\mathbb{R}^n$. A subset of $S = \{e_1, \cdots, e_n, m_1, \cdots, m_n\}$ is called *complementary* if it contains no two points with the same subscript. If $M \in \mathbb{R}^{n \times n}$ and $q \in \mathbb{R}^n$, the linear complementarity problem (M, q) is to find a complementary subset of S, with $m_i = -M_i$, the ith column of $-M$, for $i = 1, \cdots, n$, so that the cone generated by the subset contains q. It is known (see [PC]) that Lemke's algorithm for the linear complementarity problem will find a solution quickly if we can produce a positive vector p satisfying, for any index set I contained in $\{1, \cdots, n\}$,

$$(*) \qquad M_{II}^{-1} p_I > 0.$$

A set of conditions similar to $(*)$, with nonstrict inequalities replacing strict inequalities, was obtained by Cottle [C] to characterize matrices for which a certain monotonicity property of the parametric linear complementarity problem holds. These considerations have led to the study of properties of matrices for which such a p can be found.

It has been conjectured (see [KMW]) that the set of $p > 0$ satisfying $(*)$, if nonempty, is the interior of an n-dimensional simplicial cone in $\mathbb{R}^n$ when M is symmetric and positive definite. One purpose of this note is to prove that this is true when M is a P-matrix. The defining inequalities for this simplicial cone will show that if a $p > 0$ satisfying $(*)$ exists, then M^T must be a hidden Minkowski matrix. It is already known that a $p > 0$ satisfying $(*)$ exists when M^T is hidden Minkowski (see [PC]), so this result shows that the two matrix classes are the same. This characterization is very useful, because [P] has given a polynomial time algorithm to check if M^T is hidden Minkowski and [PC] finds a $p > 0$ satisfying $(*)$ in polynomial time if M^T is indeed hidden Minkowski.

DEFINITIONS. Let $S = \{e_1, \cdots, e_n, m_1, \cdots, m_n\}$, with $m_1, \cdots, m_n$ arbitrary and $e_1, \cdots, e_n$ the standard basis for $\mathbb{R}^n$. S is called *nondegenerate* if the points in any complementary subset of S are linearly independent. A complementary subset of S with m elements is called an m-complementary (m-c.) subset of S. A matrix $M \in \mathbb{R}^{n \times n}$ is called a P-matrix if one of the following equivalent (see [Mu], [STW]) conditions is satisfied:

(P1) The principal minors of M are all positive.

(P2) For every $0 \neq x \in \mathbb{R}^n$, there is an index i so that $x_i(Mx)_i > 0$.

* Received by the editors July 5, 1988; accepted for publication (in revised form) October 18, 1988. The research of the second author was supported by National Science Foundation grant DMS-8711581.

† Department of Mathematical Sciences, George Mason University, Fairfax, Virginia 22030 (wmorris @gmuvax).

(P3) When we let $m_1, \cdots, m_n$ be given by the columns of M, S is nondegenerate and for every index i the points e_i and m_i are on the same side of any hyperplane containing an $(n-1)$-c. subset of $S \setminus \{e_i, m_i\}$.

(P4) There is a unique solution to the linear complementary problem (M, q) for any $q \in \mathbb{R}^n$.

2. A geometric property.

LEMMA 1. *Let S be nondegenerate and let the columns of M be given by $m_1, \cdots, m_n$. M is a P-matrix and there is a $p > 0$ satisfying $(*)$ if and only if the 2^n cones generated by n-c. subsets of S have nonempty intersection of dimension n.*

Proof. Suppose that $p > 0$ satisfies $(*)$ and that M is a P-matrix. Then we would like to show that for any n-c. set $\{e_j, j \in J\} \cup \{m_i, i \in I\}$ there exist $\lambda_i > 0$, $i = 1, \cdots, n$, so that

$$p = \sum_{j \in J} \lambda_j e_j + \sum_{i \in I} \lambda_i m_i.$$

Consider an n-c. subset $\{e_j, j \in J\} \cup \{m_i, i \in I\}$. Let

$$p = \sum_{j \in J} \lambda_j e_j + \sum_{i \in I} \lambda_i m_i \tag{1}$$

be the expression of p in terms of $\{e_j, j \in J\} \cup \{m_i, i \in I\}$. The assumption $(*)$ implies that $\lambda_i > 0$ for $i \in I$. In particular, this implies that p is in the interior of the cone generated by $\{m_i, i = 1, \cdots, n\}$. If $J \neq \varnothing$, let $k \in J$ and let

$$p = \sum_{j \in J \setminus k} \mu_j e_j + \mu_k m_k + \sum_{i \in I} \mu_i m_i \tag{2}$$

be the unique representation of p in terms of $\{e_j, j \in J \setminus k\} \cup \{m_i, i \in I \cup k\}$. Subtracting line (2) from line (1), we obtain

$$\mu_k m_k + \sum_{i \in I} (\mu_i - \lambda_i) m_i = \lambda_k e_k + \sum_{j \in J \setminus k} (\lambda_j - \mu_j) e_j. \tag{3}$$

Define $x \in \mathbb{R}^n$ by $x_k = \mu_k$, $x_i = \mu_i - \lambda_i$, $i \in I$, $x_j = 0$ otherwise. Then $(Mx)_j = \lambda_j - \mu_j$, $j \in J \setminus k$, $(Mx)_k = \lambda_k$, $(Mx)_i = 0$, $i \in I$. Thus by (P2), $x_k(Mx)_k = \lambda_k \mu_k > 0$, implying that $\lambda_k > 0$. Thus $\lambda_j > 0$, $j \in J$, and p is in the interior of the cone generated by $\{e_j, j \in J\} \cup \{m_i, i \in I\}$.

Conversely, if the 2^n cones have nonempty intersection of dimension n, let p be a point in this intersection. Then clearly $(*)$ holds, and (P3) must also hold. □

By Lemma 1, the set of $p > 0$ satisfying $(*)$ is the interior of an n-dimensional polyhedral cone. The facets of this cone are contained in cones generated by $(n-1)$-c. subsets of S, because the cones generated by the $(n-1)$-c. subsets of S are the facets of the cones generated by the n-c. subsets of S.

LEMMA 2. *With the assumptions of Lemma 1, suppose M is a P-matrix and that the set of $p > 0$ satisfying $(*)$ is nonempty. Let C be the closure of this set. Then C is a closed polyhedral cone. Let F be a facet of C. Without loss of generality, assume F is contained in the cone generated by $\{e_1, \cdots, e_{n-1}\}$. Then the points e_n and m_n are in one of the open halfspaces created by the hyperplane H containing the points $\{e_1, \cdots, e_{n-1}\}$ and the points $\{e_1, \cdots, e_{n-1}, m_1, \cdots, m_{n-1}\}$ are in the complementary closed halfspace.*

Proof. By (P3), e_n and m_n are on the same side of H. Suppose m_{n-1} is in the open halfspace containing e_n and m_n. Let q be a point in the relative interior of F. Then $q +$

εe_n is in the interior of C, for $\varepsilon > 0$ sufficiently small. Thus, for sufficiently small $\varepsilon > 0$, there exist $\lambda_i > 0$, $i = 1, \cdots, n$ and $\mu_i > 0$, $i = 1, \cdots, n$, so that

$$q + \varepsilon e_n = \lambda_1 e_1 + \cdots + \lambda_n e_n, \tag{4}$$

$$q + \varepsilon e_n = \mu_1 e_1 + \cdots + \mu_{n-2} e_{n-2} + \mu_{n-1} m_{n-1} + \mu_n e_n. \tag{5}$$

Furthermore, $\lambda_n = \varepsilon$. Subtracting (4) from (5), we obtain

$$\mu_{n-1} m_{n-1} + (\mu_n - \varepsilon) e_n = (\lambda_1 - \mu_1) e_1 + \cdots + (\lambda_{n-2} - \mu_{n-2}) e_{n-2} + \lambda_{n-1} e_{n-1}. \tag{6}$$

Now if $\mu_n - \varepsilon > 0$, then e_n and m_{n-1} are on opposite sides of H. To show this, let $q = \gamma_1 e_1 + \cdots + \gamma_{n-2} e_{n-2} + \gamma_{n-1} m_{n-1} + \gamma_n e_n$ be the representation of q in terms of $\{e_1, \cdots, e_{n-2}, m_{n-1}, e_n\}$. Then $\gamma_n > 0$ because m_{n-1} is not on H. For ε small enough, then, μ_n will be close to γ_n and thus $\mu_n - \varepsilon > 0$. □

THEOREM 1. *If the set of $p > 0$ satisfying $(*)$ is nonempty, then the cone C defined in Lemma 2 is simplicial.*

Proof. From Lemma 2, it is clear that the intersection of any $(n-1)$-c. cone generated by points with subscripts less than n with the cone generated by $\{e_1, \cdots, e_n\}$ (and thus with C) must be contained in the hyperplane H. Thus F is the only facet of C contained in a cone generated by points with subscripts less than n. Since n was arbitrarily chosen as the missing subscript for F, there must be at most n facets, one for each subscript. However, an n-dimensional cone must have at least n facets, so C is a simplicial cone. □

Remark. The existence of a $p > 0$ satisfying $(*)$ is related (see [KMW]) to the existence of a "CP-point." Let D be a nonsingular matrix in $\mathbb{R}^{n \times n}$. A vector $b \in \mathbb{R}^n$ is a CP-point for the cone generated by the columns of D if it is in the interior of this cone and the projection of b onto the linear span of any face of the cone is in the relative interior of that face. Let $M = D^T D$. It is proved in [KMW] that $p = D^T b$ is positive and satisfies $(*)$ if and only if b is a CP-point. The set of $p > 0$ satisfying $(*)$, if nonempty, is the interior of an n-dimensional simplicial cone. Thus the set of CP-points, which is the set of $(D^{-1})^T p$ for such p, must be the interior of an n-dimensional simplicial cone if nonempty because $(D^{-1})^T$ is nonsingular.

3. Relationship to hidden Minkowski matrices. There are many examples known of P-matrices (even positive-definite matrices) for which there is no $p > 0$ satisfying $(*)$. The next result shows that even if there is no such p, we can find hyperplanes H^i as in Lemma 2.

Define the matrix M^i by

$$M^i_{jk} = \begin{cases} M_{ii} & \text{if } j = i = k, \\ -M_{kj} & \text{if } i = j \neq k \text{ or if } i = k \neq j, \\ M_{kj} & \text{otherwise.} \end{cases}$$

(M^i is obtained from M^T by first negating column i and then negating row i. This leaves M_{ii} unchanged.)

LEMMA 3. *M^i is a P-matrix if and only if M is.*

Proof. The proof is immediate from (P1).

LEMMA 4. *Let M be a P-matrix. The vector (w, z) such that $w = M^i_z + M^i_i$ solves the linear complementarity problem (M^i, M^i_i) if and only if $z^{*i} = z - e_i$ satisfies*

$$z^{*i}_i = -1, \quad (z^{*i})^T M_i < 0, \quad (z^{*i})^T M_j \geqq 0, \quad z^{*i}_j \geqq 0, \quad ((z^{*i})^T M_j) z^{*i}_j = 0, \quad j \neq i. \tag{7}$$

Proof. If M^i is a P-matrix, by nondegeneracy of M^i any solution (w, z) to the linear complementarity problem (M^i, M^i_i) must satisfy $z_i = 0$ and $w_i > 0$. Now (w, z) solves the linear complementarity problem (M^i, M^i_i) if and only if for $\bar{z} = z + e_i$ we have $w = M^i\bar{z}$ and

$$\bar{z}_i = 1, \quad (M^i\bar{z})_i > 0, \quad (M^i\bar{z})_j \geqq 0, \quad \bar{z}_j \geqq 0, \quad (M^i\bar{z})_j\bar{z}_j = 0, \quad j \neq i. \tag{8}$$

From the definition of M^i, we get that $\bar{z}$ satisfies (8) if and only if z^{*i} satisfies (7). □

THEOREM 2. *If M is a P-matrix, then for each $i = 1, \cdots, n$ there is a unique hyperplane H^i containing an $(n-1)$-c. set of $S \setminus \{e_i, m_i\}$ and such that e_i and m_i are in one of the open halfspaces defined by H^i, and all of the other points of S are in the complementary closed halfspace.*

Proof. Such an H^i is defined by a z^{*i} satisfying (7). z^{*i} is unique because the solution to the linear complementarity problem (M^i, M^i_i) is unique, by (P4).

Theorem 2 implies that if we know that M is a P-matrix, and we know the solutions to the n linear complementarity problems (M^i, M^i_i), then we can find a p satisfying $(*)$, if it exists, by solving a linear program.

A square matrix is called a Z-matrix if all of its off-diagonal elements are nonpositive. A P-matrix that is also a Z-matrix is called a Minkowski matrix. A P-matrix M is called a hidden Minkowski (see [CP], [Ma], [P]) matrix if there exist Z-matrices X and Y so that $Y = MX$ and there exist $r, s \in \mathbb{R}^n$, $(r, s) \geqq 0$, $r^TX + s^TY > 0$.

THEOREM 3. *Let M be a P-matrix and let $p > 0$ satisfy $(*)$. Then M^T is a hidden Minkowski matrix.*

Proof. We are not aware of any previous proofs of Theorem 3, even though its converse is well known (see [PC]). Define the matrix X by letting the ith column of X be $-z^{*i}$, $i = 1, \cdots, n$. Then X and $Y = M^TX$ are Z-matrices, by (7). Suppose $p \in \operatorname{int}(C)$. Then we have $p^T(-z^{*i}) > 0$, $i = 1, \cdots, n$, since $e_i^T(-z^{*i}) > 0$, $i = 1, \cdots, n$ and p must be on the same side of H^i as e_i, for $i = 1, \cdots, n$. Thus, $p^TX > 0$ and M^T is hidden Minkowski. □

Pang [PC] gives an algorithm to find a p satisfying $(*)$ in polynomial time if M^T is a hidden Minkowski matrix.

REFERENCES

[C] R. W. COTTLE, *Monotone solutions of the parametric linear complementarity problem*, Math. Programming, 3 (1972), pp. 210–224.

[CP] R. W. COTTLE AND J. S. PANG, *On solving linear complementarity problems as linear programs*, Math. Programming Stud., 7 (1978), pp. 88–107.

[KMW] L. M. KELLY, K. G. MURTY, AND L. T. WATSON, "CP-*rays in simplicial cones*, Technical Report No. 87-17, Department of Industrial and Operations Engineering, The University of Michigan, Ann Arbor, MI, 1987.

[Ma] O. L. MANGASARIAN, *Linear complementarity problems solvable by a single linear program*, Math. Programming, 10 (1976), pp. 263–270.

[Mu] K. G. MURTY, *On the number of solutions to the complementarity problem and spanning properties of complementary cones*, Linear Algebra Appl., 5 (1972), pp. 65–108.

[P] J. S. PANG, *On discovering hidden Z-matrices*, in Constructive Approaches to Mathematical Models, C. V. Coffman and G. J. Fix, eds., Academic Press, New York, 1979, pp. 231–241.

[PC] J. S. PANG AND R. CHANDRASEKARAN, *Linear complementarity problems solvable by a polynomially bounded pivoting algorithm*, Math. Programming Stud., 25 (1985), pp. 13–27.

[STW] H. SAMELSON, R. M. THRALL, AND O. WESLER, *A partition theorem for Euclidean n-space*, Proc. Amer. Math. Soc., 9 (1958), pp. 805–807.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 2, pp. 233–243, April 1989

009

A CANONICAL FORM FOR HERMITIAN MATRICES UNDER COMPLEX ORTHOGONAL CONGRUENCE*

YOOPYO HONG†

Abstract. It is shown that a Hermitian matrix can be reduced to a Hermitian canonical form by a complex orthogonal congruence. As a consequence, a short proof is given showing that a nonsingular symmetric matrix and a Hermitian matrix can be simultaneously reduced to the identity matrix and a Hermitian canonical form by complex orthogonal *T*- and $*$-congruences, respectively.

Key words. orthogonal congruence, Hermitian, simultaneous reduction

AMS(MOS) subject classification. 15A21

1. Notation and introduction. We denote the set of all m-by-n complex matrices by $M_{m,n}$; $M_n \equiv M_{n,n}$. For $A \in M_n$, we denote the transpose of A by A^T, the complex conjugate by $\bar{A}$, and the Hermitian adjoint by $A^* = \bar{A}^T$. We say that two matrices $A, B \in M_n$ are *consimilar*, T-*congruent*, or $*$-*congruent*, respectively, if there is a nonsingular $P \in M_n$ such that $P^{-1}A\bar{P} = B$, $P^TAP = B$, or $P^*AP = B$, respectively.

If $A \in M_n$ is symmetric, then any $B \in M_n$ that is *T*-congruent to A must also be symmetric. Similarly, if $A \in M_n$ is Hermitian, then any $B \in M_n$ that is $*$-congruent to A must also be Hermitian. Therefore, symmetry and Hermiticy are invariant under *T*-congruence and $*$-congruence, respectively. A nonsingular $Q \in M_n$ is called *orthogonal* if $QQ^T = I$, i.e., $Q^{-1} = Q^T$. Two matrices $A, B \in M_n$ are *orthogonally* T-*congruent* if there is an orthogonal $Q \in M_n$ such that $Q^TAQ = B$. Similarly, two matrices $A, B \in M_n$ are *orthogonally*$*$-*congruent* if there is an orthogonal $Q \in M_n$ such that $Q^*AQ = B$. Note that an orthogonal *T*-congruence is an orthogonal similarity and an orthogonal $*$-congruence is an orthogonal consimilarity.

We denote the spectrum (set of eigenvalues, counting multiplicities) of a given $A \in M_n$ by $\sigma(A)$. We denote a k-by-k identity matrix by I_k. For a Hermitian $A \in M_n$, we denote the *inertia* of A by $i(A) = (i_+(A), i_-(A), i_0(A))$, where $i_+(A)$, $i_-(A)$, and $i_0(A)$ indicate the number of positive, negative, and zero eigenvalues of A (all counting multiplicities), respectively. The *inertia matrix* of a given Hermitian $A \in M_n$ is a diagonal matrix $I(A) = I_r \oplus -I_S \oplus 0_0$, where 0_0 is a square zero matrix of dimension $i_0(A)$, $r = i_+(A)$, and $s = i_-(A)$. Sylvester's inertia theorem guarantees that every Hermitian A is $*$-congruent to $I(A)$.

We say that $A \in M_n$ is *condiagonalizable* if A is diagonalizable by consimilarity, i.e., there exists a nonsingular P such that $P^{-1}A\bar{P} = \Lambda$, where $\Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$. There is no loss of generality when we assume that $\Lambda = \operatorname{diag}(|\lambda_1|, \cdots, |\lambda_n|)$; the nonnegative diagonal entries of Λ are called *coneigenvalues* of A [4]. A Jordan block $J_k(\lambda) \in M_k$ has the form $J_k(\lambda) = \lambda I_k + N_k$, where $N_k \in M_k$ is a nilpotent matrix with all entries zero except for ones on the first superdiagonal.

The following facts are well known.

LEMMA 1.1. *Each $A \in M_n$ is similar to a symmetric Jordan canonical form $J_S(A) = S_{n_1}(\lambda_1) \oplus \cdots \oplus S_{n_p}(\lambda_p)$, where each $S_k(\lambda) \in M_k$ is symmetric and similar to a Jordan block $J_k(\lambda)$ of A.*

* Received by the editors July 13, 1988; accepted for publication in revised form October 28, 1988.

† Department of Mathematical Sciences, Northern Illinois University, DeKalb, Illinois 60115.

Proof. See § 4.4 of [7] for a proof and a standard form for $S_k(\lambda)$.

LEMMA 1.2. *Two symmetric matrices are similar if and only if they are orthogonally similar.*

Proof. See Chapter 1 of [2] for the proof.

By Lemmata 1.1 and 1.2, a symmetric matrix is orthogonally similar to a symmetric Jordan canonical form. Thus, the following lemma is immediate.

LEMMA 1.3. *Let $A \in M_n$ be a given symmetric matrix with k negative eigenvalues, $0 \leqq k \leqq n$. Then A is orthogonally similar to a direct sum $B \oplus C$ where $B \in M_{n-k}$ is symmetric and has no negative eigenvalues, and all the eigenvalues of the symmetric $C \in M_k$ are negative.*

LEMMA 1.4. *Let $A \in M_n$, $B \in M_m$ be given and suppose $\sigma(A) \cap \sigma(B) = \emptyset$. Then $AC - CB = 0$ for some $C \in M_{n,m}$ if and only if $C = 0$.*

Proof. See Chapter 1.4 of [2] for the proof. In particular, if $\sigma(A) \cap \sigma(\bar{B}) = \emptyset$ then $AC - C\bar{B} = 0$ implies $C = 0$.

DePrima and Johnson [1] have shown that if $T \in M_n$ is nonsingular with no negative eigenvalues, then there is a unique $T_1 \in M_n$ such that (a) $T_1^2 = T$, (b) $\sigma(T_1)$ lies in the open right half plane, and (c) $T_1C = CT_1$ for every $C \in M_n$ such that $TC = CT$. We adopt their method and extend the result.

Let $T \in M_n$ be a given nonsingular matrix with no negative eigenvalues. Let Γ be the oriented Jordan curve consisting of circular arcs and line segments of the sort illustrated in Fig. 1, whose interior Δ contains $\sigma(T)$. Since Γ is symmetric with respect to the real axis, $\sigma(\bar{T}) \in \Delta$ as well.

Set $T_1 \equiv 1/2\pi i \int_\Gamma z^{1/2}(zI - T)^{-1}dz$. It is a simple matter to verify the following.

LEMMA 1.5. *Let $T \in M_n$ be a given nonsingular matrix with no negative eigenvalues. There is a unique $T_1 \in M_n$ such that* (a) $T_1^2 = T$, (b) $\sigma(T_1)$ *lies in the open right half plane, and* (c) $T_1A = AT_1$ *for all $A \in M_n$ such that $TA = AT$. Moreover, for this matrix T_1 we have:* (d) T_1 *is symmetric whenever T is symmetric, and* (e) $T_1A = A\bar{T}_1$ *for all $A \in M_n$ such that $TA = A\bar{T}$.*

Proof. The existence of a unique T_1 that satisfies (a), (b), and (c) is shown in [1]. If T is symmetric then T_1 is clearly symmetric by its construction. If $TA = A\bar{T}$ for some $A \in M_n$, then $(zI - T)^{-1}A = A(zI - \bar{T})^{-1}$ for all $z \in \Gamma$. Therefore,

$$T_1A = \frac{1}{2\pi i}\int_\Gamma z^{1/2}(zI - T)^{-1}dzA = \frac{1}{2\pi i}\int_\Gamma z^{1/2}A(zI - \bar{T})^{-1}dz = A\bar{T}_1. \qquad \square$$

The following is a fundamental theorem about consimilarity [4].

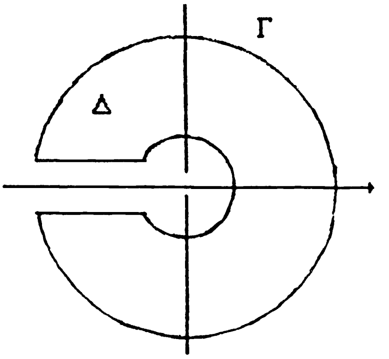


FIG. 1.

THEOREM 1.6. *Let A, $B \in M_n$. Then A is consimilar to B if and only if $A\bar{A}$ is similar to $B\bar{B}$ and the following alternating product rank condition is satisfied*: rank $[(A\bar{A})^k A]$ = rank $[(B\bar{B})^k B]$, $k = 0, \cdots, [n/2]$. *In particular, A is condiagonalizable if and only if $A\bar{A}$ is diagonalizable and has nonnegative eigenvalues and* rank (A) = rank $(A\bar{A})$.

There is an analogue of Lemma 1.1 for consimilarity: every $A \in M_n$ is consimilar to a canonical form that is Hermitian [3]. We now describe the structure of this form.

Let $A \in M_n$ be given. The Hermitian canonical form of A under consimilarity, denoted by $J_H(A)$, is a direct sum of three Hermitian matrices:

(1.7) $$J_H(A) \equiv H_p(A) \oplus K_N(A) \oplus K_C(A),$$

(1.8) $H_p(A) = H_{m_1}(\lambda_1) \oplus \cdots \oplus H_{m_p}(\lambda_p)$, where all $\lambda_i \geqq 0$ and λ_i^2 are the nonnegative eigenvalues of $A\bar{A}$,

(1.9) $K_N(A) = K_{2n_1}(\mu_1) \oplus \cdots \oplus K_{2n_r}(\mu_r)$, where all $\mu_i > 0$ and $-\mu_i^2$ are the negative eigenvalues of $A\bar{A}$,

$$K_{2n_i}(\mu_i) = \left[\begin{array}{c|c} 0 & -iH_{N_i}(\mu_i) \\ \hline iH_{n_i}(\mu_i) & 0 \end{array}\right] \in M_{2k},$$

and

(1.10) $K_C(A) = L_{2k_1}(\xi_1) \oplus \cdots \oplus L_{2k_S}(\xi_S)$, where all $\xi_i \in C$ are nonreal and ξ_i^2 are the nonreal eigenvalues of $A\bar{A}$,

$$L_{2k_i}(\xi_i) = \left[\begin{array}{c|c} 0 & H_{k_i}(\xi_i) \\ \hline H_{k_i}(\xi_i)^* & 0 \end{array}\right] \in M_{2k_i},$$

where

$$H_m(\lambda) = \tfrac{1}{2}\left(\begin{bmatrix} 0 & & 1 & 2\lambda \\ & \cdots & \ddots & 1 \\ 1 & & \cdots & \\ 2\lambda & 1 & & 0 \end{bmatrix} + i\begin{bmatrix} 0 & 1 & & 0 \\ -1 & & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & -1 & 0 \end{bmatrix}\right) \in M_m, \qquad \lambda \in C.$$

Note that $H_m(\lambda)$ is Hermitian if λ is real. The Hermitian blocks $H_{m_i}(\lambda_i)$, $K_{2n_i}(\mu_i)$, and $L_{2k_i}(\xi_i)$ are derived in an explicit way from the Jordan blocks and the quasi-Jordan blocks of the concanonical form of A.

The formal result about consimilarity and Hermitian canonical forms is the following [3].

THEOREM 1.11. *For each $A \in M_n$ there is a nonsingular $P \in M_n$ such that $P^{-1}A\bar{P} = J_H(A)$. Moreover, $J_H(A)$ is unique up to a permutation of its diagonal blocks.*

Note that if $P^{-1}A\bar{P} = J_H(A)$ then

$$(e^{-i\theta/2}P)^{-1}A(\overline{e^{-i\theta/2}P}) = e^{i\theta}J_H(A)$$

and $(e^{-i\theta/2}P)^{-1}(e^{-i\theta}A)(\overline{e^{-i\theta/2}P}) = J_H(A)$ for all $\theta \in \mathbb{R}$. Thus, $J_H(A) = J_H(e^{-i\theta}A)$ for all $\theta \in \mathbb{R}$ [3]. In particular, if $\theta = \pi/2$ then $(e^{-i\pi/4}P)^{-1}A(\overline{e^{-i\pi/4}P}) = iJ_H(A)$ and that $iJ_H(A)$ is a skew Hermitian matrix. Thus, $iJ_H(A)$ is a skew Hermitian canonical form.

COROLLARY 1.12. *Let $A \in M_n$. Then for all $\theta \in \mathbb{R}$, $J_H(A) = J_H(e^{-i\theta}A)$ and there is a nonsingular $P \in M_n$ such that $P^{-1}A\bar{P} = e^{i\theta}J_H(A)$. In particular, $J_H(A) = J_H(-iA)$ and there is a nonsingular $P \in M_n$ such that $P^{-1}A\bar{P} = iJ_H(A)$, where $iJ_H(A)$ is a skew Hermitian canonical form. Moreover, $iJ_H(A)$ is unique up to a permutation of its diagonal blocks.*

2. Main results. If two symmetric matrices are similar, then they are orthogonally similar. An analogous theorem holds for two Hermitian matrices.

PROPOSITION 2.1. *Let A, $B \in M_n$. There is an orthogonal $Q \in M_n$ such that $Q^*AQ = B$ if and only if there is a nonsingular $P \in M_n$ such that PP^T has no negative eigenvalues, $\bar{P}^{-1}AP = B$, and $\bar{P}^{-1}A^*P = B^*$.*

Proof. Suppose $Q^*AQ = B$, $Q^T = Q^{-1}$, $Q \in M_n$. Then $QQ^T = I$ has no negative eigenvalues, $Q^*AQ = \bar{Q}^{-1}AQ = B$, and $B^* = (Q^*AQ)^* = Q^*A^*Q$.

Conversely, suppose there is a nonsingular $P \in M_n$ such that $\bar{P}^{-1}AP = B$, $\bar{P}^{-1}A^*P = B^*$, and PP^T has no negative eigenvalues. Then $B = \bar{P}^{-1}AP$ and $B = (B^*)^* = (\bar{P}^{-1}A^*P)^* = P^*AP^{T-1}$. Thus, $P^*AP^{T-1} = \bar{P}^{-1}AP$, or $A(PP^T) = (\overline{PP^T})A$. Set $PP^T = S$. Then $S^T = S$ and $AS = \bar{S}A$. Since S has no negative eigenvalues, by Lemma 1.5 there is a unique (necessarily nonsingular) $S_1 \in M_n$ with spectrum in the open right half plane such that $S_1^2 = S$ and $AS_1 = \bar{S}_1A$, or $\bar{S}_1^{-1}AS_1 = A$. If we set $Q = S_1^{-1}P$ then $QQ^T = S_1^{-1}PP^TS_1^{-1} = S_1^{-1}(S_1^2)S_1^{-1} = I$, i.e., Q is an orthogonal matrix. Then $P = S_1Q$ and $B = \bar{P}^{-1}AP = Q^*\bar{S}_1^{-1}AS_1Q = Q^*AQ$, as desired. □

COROLLARY 2.2. *Let A, $B \in M_n$. There is an orthogonal $Q \in M_n$ such that $Q^*AQ = -B$ if and only if there is a nonsingular $P \in M_n$ such that all the eigenvalues of PP^T are negative, $\bar{P}^{-1}AP = B$, and $\bar{P}^{-1}A^*P = B^*$.*

Proof. Suppose there is an orthogonal $Q \in M_n$ such that $Q^*AQ = -B$. Then, $iQ^*AiQ = B$ and $iQ^*A^*iQ = B^*$. Set $P = iQ$. Then $PP^T = -I$ and $P^{-1} = -iQ^T$ and hence $\bar{P}^{-1} = iQ^*$. Thus, there is a nonsingular $P \in M_n$ such that $\bar{P}^{-1}AP = B$, $\bar{P}^{-1}A^*P = B^*$, and all the eigenvalues of PP^T are negative.

Conversely, suppose there is a nonsingular $P \in M_n$ such that $\bar{P}^{-1}AP = B$, $\bar{P}^{-1}A^*P = B^*$ and all the eigenvalues of PP^T are negative. Set $E = iI$. If we set $R = PE$ then $\sigma(RR^T) = \sigma(-PP^T)$, and hence all the eigenvalues of RR^T are positive, $\bar{R}^{-1}AR = EBE = -B$, and $\bar{R}^{-1}A^*R = (-B)^*$. Therefore, by Proposition 2.1 there is an orthogonal $Q \in M_n$ such that $Q^*AQ = -B$. □

Suppose $A, B \in M_n$ is a pair of Hermitian, skew Hermitian, or orthogonal matrices. If there is a nonsingular $P \in M_n$ such that $\bar{P}^{-1}AP = B$, then it follows that $\bar{P}^{-1}A^*P = B^*$. Thus, following result is immediate.

COROLLARY 2.3. *Let A, $B \in M_n$ be a pair of Hermitian, skew Hermitian, or orthogonal matrices. Then there is an orthogonal $Q \in M_n$ such that $Q^*AQ = B$ if and only if there is a nonsingular $P \in M_n$ such that PP^T has no negative eigenvalues and $\bar{P}^{-1}AP = B$. There is an orthogonal $Q \in M_n$ such that $Q^*AQ = -B$ if and only if there is a nonsingular $P \in M_n$ such that all the eigenvalues of PP^T are negative and $\bar{P}^{-1}AP = B$.*

We have a more general result than the preceding corollary. Suppose there exists a complex polynomial F in $\bar{x}x$, i.e., $F(x) = a_0 + a_1(\bar{x}x) + a_2(\bar{x}x)^2 + \cdots + a_m(\bar{x}x)^m$, such that $A^* = AF(A)$ and $B^* = BF(B)$. Then $\bar{P}^{-1}AP = B$ implies $\bar{P}^{-1}A^*P = B^*$. Indeed, if $\bar{P}^{-1}AP = B$ for some nonsingular P, then

$$\begin{aligned} B^* = BF(B) &= \bar{P}^{-1}APF(\bar{P}^{-1}AP) \\ &= \bar{P}^{-1}AP[a_0 + a_1(P^{-1}\bar{A}\bar{P}\bar{P}^{-1}AP) + a_2(P^{-1}\bar{A}\bar{P}\bar{P}^{-1}AP)^2 + \cdots] \\ &= \bar{P}^{-1}a_0AP + \bar{P}^{-1}a_1A(\bar{A}A)P + \bar{P}^{-1}a_2A(\bar{A}A)^2P + \cdots = \bar{P}^{-1}(AF(A))P \\ &= \bar{P}^{-1}A^*P. \end{aligned}$$

COROLLARY 2.4. *Let $A, B \in M_n$. Suppose $A^* = AF(A)$ and $B^* = BF(B)$ for some complex polynomial F in $\bar{x}x$. Then, there is an orthogonal $Q \in M_n$ such that $Q^*AQ = B$ if and only if there is a nonsingular $P \in M_n$ such that $\bar{P}^{-1}AP = B$ and PP^T has no negative eigenvalues.*

LEMMA 2.5. *Let $A \in M_n$ be Hermitian. Suppose there is a nonsingular $P \in M_n$ such that $\bar{P}^{-1}AP$ is Hermitian, and suppose PP^T has r negative eigenvalues (counting multiplicities). Then there is an orthogonal $Q \in M_n$ such that $Q^*AQ = A_1 \oplus A_2$, where $A_1 \in M_{n-r}$ and $A_2 \in M_r$ are Hermitian.*

Proof. Since A and $\bar{P}^{-1}AP$ are Hermitian, $\bar{P}^{-1}AP = P^*A(P^T)^{-1}$. Thus, $A = (\overline{PP^T})A(PP^T)^{-1}$. If we set $S = PP^T$, then $\bar{S}AS^{-1} = A$. Since S is a symmetric matrix with r negative eigenvalues, by Lemma 1.3 there is an orthogonal $Q \in M_n$ such that

$$Q^TSQ = \left[\begin{array}{c|c} T_1 & 0 \\ \hline 0 & T_2 \end{array}\right],$$

where $T_1 \in M_{n-r}$ has no negative eigenvalues and $T_2 \in M_r$ has only negative eigenvalues. Since $\bar{S}AS^{-1} = A$,

$$Q^*AQ = Q^*\bar{S}AS^{-1}Q = Q^*\bar{S}\bar{Q}Q^*AQQ^TS^{-1}Q = (\overline{Q^TSQ})(Q^*AQ)(Q^TSQ)^{-1}.$$

Thus, if we set $Q^*AQ = A_1$ and $Q^TSQ = S_1$ then $\bar{S}_1A_1S_1^{-1} = A_1$, or $\bar{S}_1A_1 = A_1S_1$.

Now write A_1 in block form as

$$A_1 = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right], \qquad A_{11} \in M_{n-r}, \quad A_{22} \in M_r,$$

and note that $\bar{S}_1A_1 = A_1S_1$ implies that $\bar{T}_1A_{12} - A_{12}T_2 = 0$ and $\bar{T}_2A_{21} - A_{21}T_1 = 0$. Thus, $A_{12} = A_{21} = 0$ by Lemma 1.4 since $\sigma(\bar{T}_1) \cap \sigma(T_2) = \varnothing$. □

Now, let $J_H(A)$ be a Hermitian canonical form of $A \in M_n$, $J_H(A) = H_P(A) \oplus K_N(A) \oplus K_C(A)$ where $H_P(A)$, $K_N(A)$, and $K_C(A)$ are defined in (1.8), (1.9), and (1.10), respectively.

LEMMA 2.6. *$K_n(A)$ and $K_C(A)$ are always orthogonally $*$-congruent to $-K_N(A)$ and $-K_C(A)$, respectively.*

Proof. Since $K_N(A) = K_{2n_1}(\mu_1) \oplus \cdots \oplus K_{2n_r}(\mu_r)$ and $K_C(A) = L_{2k_1}(\xi_1) \oplus \cdots \oplus L_{2k_S}(\xi_S)$, it is sufficient to show that $K_{2n_j}(\mu_j)$ and $L_{2k_j}(\xi_j)$ are orthogonally $*$-congruent to $-K_{2n_j}(\mu_j)$ and $-L_{2k_j}(\xi_j)$, respectively. Since

$$K_{2n}(\mu) = \begin{bmatrix} 0 & -iH_n(\mu) \\ iH_n(\mu)^* & 0 \end{bmatrix} \quad \text{and} \quad L_{2k}(\xi) = \begin{bmatrix} 0 & H_k(\xi) \\ H_k(\xi)^* & 0 \end{bmatrix},$$

the result follows from the consimilarities

$$\begin{bmatrix} I_n & 0 \\ 0 & -I_n \end{bmatrix} \begin{bmatrix} 0 & -iH_n(\mu) \\ iH_n(\mu)^* & 0 \end{bmatrix} \begin{bmatrix} I_n & 0 \\ 0 & -I_n \end{bmatrix} = \begin{bmatrix} 0 & iH_n(\mu) \\ -iH_n(\mu)^* & 0 \end{bmatrix},$$

$$\begin{bmatrix} I_k & 0 \\ 0 & -I_k \end{bmatrix} \begin{bmatrix} 0 & H_k(\xi) \\ H_k(\xi)^* & 0 \end{bmatrix} \begin{bmatrix} I_k & 0 \\ 0 & -I_k \end{bmatrix} = \begin{bmatrix} 0 & -H_k(\xi) \\ -H_k(\xi)^* & 0 \end{bmatrix}. \qquad \square$$

Let $J_H(A) \in M_n$ be a given Hermitian canonical form of $A \in M_n$, $J_H(A) \equiv H_P(A) \oplus K_N(A) \oplus K_C(A)$, and let $\mathbf{e} \equiv [\varepsilon_1 \cdots \varepsilon_p]^T$ be a given p-vector with each $\varepsilon_j = \pm 1$.

Define $J^{\mathbf{e}}_H(A) \equiv H^{\mathbf{e}}_P(A) \oplus K_N(A) \oplus K_C(A)$, where $H^{\mathbf{e}}_P(A) \equiv \varepsilon_1 H_{n_1}(\lambda_1) \oplus \cdots \oplus \varepsilon_p H_{n_p}(\lambda_p)$, $\lambda_i \geqq 0$.

We can now give the main result of the section.

THEOREM 2.7. *Let $A \in M_n$ be given. Then $A = A^*$ if and only if there is a vector $\mathbf{e} = [\varepsilon_1 \cdots \varepsilon_p]^T$ with each $\varepsilon_i = \pm 1$ and an orthogonal $Q \in M_n$ such that $Q^*AQ = J^{\mathbf{e}}_H(A)$, where $J_H(A)$ is a Hermitian canonical form of A.*

Proof. Suppose $A \in M_n$ is Hermitian. There is a nonsingular $P \in M_n$ such that $\bar{P}^{-1}AP = J_H(A)$, a Hermitian canonical form of A. Let r be the number of negative eigenvalues of PP^T. If r is zero, then by Corollary 2.3 we are done since there is an orthogonal $Q \in M_n$ such that $Q^*AQ = J_H(A) = J^{\mathbf{e}}_H(A)$ where all $\varepsilon_i = 1$. Similarly, if $r = n$ then by Corollary 2.3 there is an orthogonal $Q \in M_n$ such that $Q^*AQ = -J_H(A)$. By Lemma 2.6 there is an orthogonal $Q_1 \in M_n$ such that $Q_1^*Q^*AQQ_1 = Q_1^*(-J_H(A))Q_1 = J^{\mathbf{e}}_H(A)$ where all $\varepsilon_i = -1$.

Finally, suppose $0 < r < n$ and let $n_1 = r$. By Lemma 2.5 there is an orthogonal $Q_2 \in M_n$ such that

$$Q_2^*AQ_2 = \left[\begin{array}{c|c} A_1 & 0 \\ \hline 0 & A_2 \end{array}\right],$$

where $A_1 = A_1^* \in M_{n_1}$ and $A_2 = A_2^* \in M_{n_2}$, $n_1 + n_2 = n$. Now, let P_i be nonsingular and such that $\bar{P}_i^{-1}A_iP_i = J_H(A_i) \in M_{n_i}$ for $i = 1, 2$, where $J_H(A_i)$ is a Hermitian canonical form of A_i. Note that $J_H(A_1) \oplus J_H(A_2)$ is a Hermitian canonical form of A. Now, for each $i = 1, 2$, if the number of negative eigenvalues of $P_iP_i^T$ is either zero or n_i then A_i is orthogonally $*$-congruent to either $J_H(A_i)$ or $-J_H(A_i)$, respectively, and hence we are done as before. Otherwise, reduce A_i to a direct sum of Hermitian matrices of lesser dimension under orthogonal $*$-congruence by using Lemma 2.5 again. Since this reduction process must end after at most n steps, we obtain an orthogonal $*$-congruence that reduces A to a direct sum of k Hermitian matrices, $1 < k \leqq n$,

$$Q^*AQ = \begin{bmatrix} B_1 & & & 0 \\ & B_2 & & \\ & & \ddots & \\ 0 & & & B_k \end{bmatrix},$$

$B_i = B_i^* \in M_{n_i}$, $n_1 + n_2 + \cdots + n_k = n$. Moreover, for each $i = 1, \cdots, k$ there is a nonsingular $P_i \in M_n$ such that $\bar{P}_i^{-1}B_iP_i = J_H(B_i)$ and the number of negative eigenvalues of $P_iP_i^T$ is either zero or n_i. Therefore, by Corollary 2.3 each B_i is orthogonally $*$-congruent to either $J_H(B_i)$ or $-J_H(B_i)$. By the uniqueness of $J_H(A)$ and Lemma 2.6 are done.

The converse is immediate: since $J^{\mathbf{e}}_H(A)$ is Hermitian, $A = \bar{Q}J^{\mathbf{e}}_H(A)Q^T$ is Hermitian. □

COROLLARY 2.8. *Let $A \in M_n$. Then $A = -A^*$ if and only if there is a vector $\mathbf{e} = [\varepsilon_1 \cdots \varepsilon_p]^T$ with each $\varepsilon_i = \pm 1$ and an orthogonal $Q \in M_n$ such that $Q^*AQ = iJ^{\mathbf{e}}_H(A)$ where $iJ_H(A)$ is a skew Hermitian canonical form of A.*

Proof. If A is skew Hermitian, then $-iA$ is Hermitian. Thus, by Theorem 2.7 there is an orthogonal $Q \in M_n$ such that $Q^*(-iA)Q = J^{\mathbf{e}}_H(-iA)$ where $J_H(-iA) = J_H(A)$ is a Hermitian canonical form of A by Corollary 1.12. Therefore, $iQ^*(-iA)Q = Q^*AQ = iJ^{\mathbf{e}}_H(-iA) = iJ^{\mathbf{e}}_H(A)$ where $iJ_H(A)$ is a skew Hermitian canonical form of A.

The converse is immediate since $iJ^{\mathbf{e}}_H(A)$ is a skew Hermitian matrix and skew Hermiticy is invariant under $*$-congruence. □

By assuming more than Hermiticy of $A \in M_n$, we can obtain a special Hermitian canonical form by an orthogonal $*$-congruence. In the context of orthogonal $*$-congruence, the following result gives a natural analogue of the canonical form of a conjugate normal matrix ($AA^* = \overline{A^*A}$) [9] under unitary congruence.

COROLLARY 2.9. *Let $A \in M_n$ be Hermitian and assume that* (a) *$A\bar{A}$ is diagonalizable and* (b) rank (A) = rank ($A\bar{A}$). *Then there is a vector $\mathbf{e} = [\varepsilon_1 \cdots \varepsilon_p]^T$ with each $\varepsilon_i = \pm 1$ and an orthogonal $Q \in M_n$ such that*

$$(2.10) \qquad Q^*AQ = J_H^{\varepsilon}(A) = \begin{bmatrix} M^{\varepsilon} & 0 \\ 0 & \Sigma \end{bmatrix}$$

where $\Sigma \in M_{2k}$, $M^{\varepsilon} \in M_{n-2k}$, and $0 \leqq k \leqq [n/2]$. The matrix $M^{\varepsilon} \equiv M\mathscr{E}$, where

$$(2.11) \qquad M = \begin{bmatrix} \mu_1 & & & 0 \\ & \mu_2 & & \\ & & \ddots & \\ 0 & & & \mu_{n-2k} \end{bmatrix} \in M_{n-2k}, \qquad \mu_i \geqq 0,$$

μ_i^2 are the nonnegative eigenvalues of $A\bar{A}$, and

$$(2.12) \quad \mathscr{E} = \begin{bmatrix} \varepsilon_1 & & & 0 \\ & \varepsilon_2 & & \\ & & \ddots & \\ 0 & & & \varepsilon_{n-2k} \end{bmatrix} \in M_{n-2k}, \qquad \varepsilon_i = \pm 1 \quad \text{for } 1 \leqq i \leqq n-2k.$$

The block diagonal Hermitian matrix Σ has the form

$$(2.13) \qquad \Sigma = \begin{bmatrix} \Sigma_1 & & & 0 \\ & \Sigma_2 & & \\ & & \ddots & \\ 0 & & & \Sigma_k \end{bmatrix} \in M_{2k}, \qquad \Sigma_i \in M_2.$$

The two-by-two Hermitian matrices Σ_j have two possible forms:

$$\Sigma_j = \begin{bmatrix} 0 & i\sigma \\ -i\sigma & 0 \end{bmatrix}, \qquad \sigma > 0$$

correspond to the set of pairs of equal negative eigenvalues $\{-\sigma^2, -\sigma^2\}$ of $A\bar{A}$;

$$\Sigma_j = \begin{bmatrix} 0 & \xi \\ \bar{\xi} & 0 \end{bmatrix}, \qquad \xi \notin \mathbb{R}$$

correspond to the set of conjugate pairs of nonreal eigenvalues $\{\xi^2, \bar{\xi}^2\}$ of $A\bar{A}$.

Proof. Let $B \equiv \left[\begin{smallmatrix} M & 0 \\ 0 & \Sigma \end{smallmatrix}\right]$, where $M \in M_{n-2k}$ is as in (2.11), and note that B is a Hermitian canonical form. Note also that $B\bar{B}$ is a diagonal matrix with the eigenvalues of $A\bar{A}$ as its diagonal entries. Since $A\bar{A}$ is a diagonalizable matrix that has exactly the same eigenvalues as $B\bar{B}$, the matrices $A\bar{A}$ and $B\bar{B}$ are similar. Since rank (A) = rank ($A\bar{A}$) = rank ($B\bar{B}$) = rank (B), it is easy to check that the alternating product rank condition holds for A and B. Thus, by Theorem 1.6 A is consimilar to B. By the uniqueness of a Hermitian canonical form, B must be a Hermitian canonical form of A; $B = J_H(A)$. Thus, A is orthogonally $*$-congruent to $J_H^{\varepsilon}(A)$ by Theorem 2.7. □

Since a permutation similarity is a real orthogonal congruence, we can rearrange the diagonal entries of the matrix M^{ε} (2.11)–(2.12) by a permutation similarity so that the diagonal entries with the same signs are grouped together, i.e., there is a permutation matrix $P \in M_n(\mathbb{R})$ such that $P^T M^{\varepsilon} P = (P^T M P)(P^T \mathscr{E} P) = M' \mathscr{E}'$ where

$$(2.14) \quad M' = \begin{bmatrix} \mu'_1 & & & & & & & \\ & \ddots & & & & & 0 & \\ & & \mu'_r & & & & & \\ & & & \mu'_{r+1} & & & & \\ & & & & \ddots & & & \\ & & & & & \mu'_{r+s} & & \\ & 0 & & & & & 0 & \\ & & & & & & & \ddots \\ & & & & & & & & 0 \end{bmatrix} \in M_{n-2k}, \qquad \mu'_i > 0,$$

is a nonnegative diagonal matrix that corresponds to the positive or zero eigenvalues μ'^2_i of $A\bar{A}$ and

$$(2.15) \quad \mathscr{E}' = \begin{bmatrix} \varepsilon_1 & & & & & 0 \\ & \varepsilon_2 & & & & \\ & & \ddots & & & \\ & & & \varepsilon_{r+s} & & \\ & 0 & & & 0 & \\ & & & & & \ddots \\ & & & & & & 0 \end{bmatrix} \in M_{n-2k},$$

$$\varepsilon_i = \begin{cases} +1 & \text{if } 1 \leqq i \leqq r, \\ -1 & \text{if } r+1 \leqq i \leqq r+s. \end{cases}$$

If we further assume that a given Hermitian matrix is actually condiagonalizable, then all the eigenvalues of $A\bar{A}$ are nonnegative and the block Σ is absent from (2.10). Since an orthogonal $*$-congruence of a Hermitian matrix preserves its inertia, the matrix $\mathscr{E}'$ in (2.15) must be an inertia matrix of A. We summarize these observations as follows:

COROLLARY 2.16. *Let $A \in M_n$ be Hermitian and have inertia $i_+(A) = r$, $i_-(A) = s$. Then A is condiagonalizable if and only if there is an orthogonal $Q \in M_n$ and positive real numbers $\mu_1, \cdots, \mu_{r+s}$ such that* $Q^*AQ = \operatorname{diag}(\mu_1, \cdots, \mu_r, -\mu_{r+1}, \cdots, -\mu_{r+s}, 0, \cdots, 0) = MI(A)$.

If $A \in M_n$ is a positive definite Hermitian matrix, then AA has full rank and is similar to a positive diagonal matrix [7, Thm. (7.6.3)], and hence A is condiagonalizable by Theorem 1.6. But since A is positive definite, $i_+(A) = n$, or $\mathscr{E}' = I \in M_n$ (2.15). Thus a positive-definite Hermitian matrix is orthogonally $*$-congruent to a positive-diagonal matrix.

Unfortunately, not every positive-semidefinite matrix is condiagonalizable, as the example

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

shows: A is positive semidefinite and $1 = \operatorname{rank}(A) \neq \operatorname{rank}(A\bar{A}) = 0$. However, this is the only way a positive-semidefinite matrix can fail to be condiagonalizable. The following has been shown [3].

LEMMA 2.17. *Let $A \in M_n$ be positive semidefinite and suppose that* rank (A) − rank $(A\bar{A}) = r$. *Then,* $0 \leqq r \leqq [n/2]$ *and the Hermitian canonical form* $J_H(A)$ *is an almost diagonal matrix of the form* $\Lambda \oplus \Gamma$,

$$(2.18) \qquad \Lambda = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_{n-2r} \end{bmatrix} \in M_{n-2r},$$

where each $\lambda_i \geqq 0$ *and* λ_i^2 *is a nonnegative eigenvalue of* $A\bar{A}$, *and*

$$\Gamma = \begin{bmatrix} H_2(0) & & & 0 \\ & H_2(0) & & \\ & & \ddots & \\ 0 & & & H_2(0) \end{bmatrix} \in M_{2r},$$

where $H_2(0) = \frac{1}{2}\left[\begin{smallmatrix} 1 & -i \\ i & 1 \end{smallmatrix}\right]$.

Proof. See Corollary (2.7) of [3] for the proof.

Note that the Hermitian canonical form $J_H(A)$ of a positive-semidefinite $A \in M_n$ is positive semidefinite. Thus, by Theorem 2.7 a positive-semidefinite $A \in M_n$ is actually orthogonally $*$-congruent to a Hermitian canonical form in (2.18) since the inertia of a Hermitian matrix is preserved under orthogonal $*$-congruence.

COROLLARY 2.19. *Let $A \in M_n$ be positive semidefinite and suppose* rank (A) − rank $(A\bar{A}) = r$. *Then,* $0 \leqq r \leqq [n/2]$ *and A is orthogonally $*$-congruent to a Hermitian canonical form* $J_H(A)$ *in* (2.18).

The following is an immediate consequence of Corollary 2.19.

COROLLARY 2.20. *Let $A \in M_n$ be positive semidefinite. Then A is diagonalizable by an orthogonal $*$-congruence if and only if* rank (A) = rank $(A\bar{A})$.

Another interesting special class of Hermitian matrices is the Hermitian coninvolutory matrices: $E \in M_n$ is coninvolutory if $E\bar{E} = I$. It follows easily from Theorem 1.6 that $E \in M_n$ is coninvolutory if and only if $E = \bar{P}^{-1}P$ for some nonsingular $P \in M_n$. Thus, a coninvolutory matrix is condiagonalizable and all its coneigenvalues are equal to one. If a coninvolutory $E \in M_n$ is Hermitian, then it follows from Corollary 2.16 that it must be orthogonally $*$-congruent to the inertia matrix of E.

COROLLARY 2.21. *Let $E \in M_n$ be given. Then E is coninvolutory and Hermitian if and only if there is an orthogonal $Q \in M_n$ such that Q^*EQ is a diagonal matrix with entries* ±1.

COROLLARY 2.22. *Let $E \in M_n$. Then E is coninvolutory and positive definite if and only if $E = Q^*Q = \bar{Q}^{-1}Q$ for some orthogonal $Q \in M_n$.*

Proof. If E is coninvolutory and positive definite there is an orthogonal $Q_1 \in M_n$ such that $Q_1^* E Q_1 = I(E) = I$, or $E = Q^*Q$ for $Q^T = Q_1$.

The converse is immediate. □

3. A pair of Hermitian and symmetric matrices. Simultaneous reduction of a pair of Hermitian or symmetric matrices to standard form by $*$- or T-congruence, respectively, is a classical problem [6], [8]. Using Theorem 2.7, we give a simple proof for the mixed case when a symmetric matrix is nonsingular.

THEOREM 3.1. *Let $A, B \in M_n$ with B nonsingular and symmetric and A either Hermitian or skew Hermitian. Then there is a vector* $\mathbf{e} = [\varepsilon_1 \cdots \varepsilon_p]^T$ *with each* $\varepsilon_i = \pm 1$

and a nonsingular $P \in M_n$ such that $P^TBP = I$ and $P^*AP = J^\varepsilon_H(AB^{-1})$ if A is Hermitian, or $P^*AP = iJ^\varepsilon_H(AB^{-1})$ if A is skew Hermitian, where $J_H(AB^{-1})$ is a Hermitian canonical form of AB^{-1}.

Proof. First suppose A is Hermitian. Since $B = B^T$ is nonsingular, there is a nonsingular $R \in M_n$ such that $R^TBR = I$ [7, Cor. (4.4.6)]; then $A_1 \equiv R^*AR$ is Hermitian and $B^{-1} = RR^T$. Thus, $A_1 = R^*AR = R^*ARR^T(R^T)^{-1} = R^*(ARR^T)(R^T)^{-1} = R^*(AB^{-1})(R^T)^{-1} = R^*(AB^{-1})(\bar{R}^*)^{-1}$. Therefore, the Hermitian matrix A_1 is consimilar to AB^{-1}, and hence they have the same Hermitian canonical form. There is an orthogonal $Q \in M_n$ such that $Q^*A_1Q = J^\varepsilon_H(A_1) = J^\varepsilon_H(AB^{-1})$, and $Q^*A_1Q = Q^*(R^*AR)Q = (RQ)^*A(RQ)$. Also, $(RQ)^TB(RQ) = Q^TR^TBRQ = Q^TIQ = I$, so $P \equiv RQ$ accomplishes the desired simultaneous reduction of A to Hermitian canonical form and B to identity matrix. If A is skew Hermitian, the result follows in the same way from Corollary 2.8. □

If at least one is nonsingular, necessary and sufficient conditions for a pair of matrices A and B (where A and B are both symmetric, or both Hermitian, or one of each) to be simultaneously diagonalized by congruence are known [6]. As a consequence of Theorem 3.1, we can give a necessary and sufficient condition for simultaneous reduction of a symmetric and Hermitian pair by respective congruences to their inertia matrices.

COROLLARY 3.2. *Let $A, B \in M_n$ with A Hermitian and B nonsingular and symmetric. There is a nonsingular $P \in M_n$ such that $P^*AP = MI(A)$ and $P^TBP = I$, where $M =$* diag $(\lambda_1, \cdots, \lambda_n)$ *and $\lambda_i \geqq 0$ are coneigenvalues of AB^{-1}, if and only if AB^{-1} is condiagonalizable.*

Proof. If $A = (P^*)^{-1}MI(A)P^{-1}$ and $B = (P^T)^{-1}IP^{-1}$ then

$$AB^{-1} = (P^*)^{-1}MI(A)P^{-1}PP^T = (P^*)^{-1}MI(A)P^T,$$

so AB^{-1} is condiagonalizable.

Conversely, suppose AB^{-1} is condiagonalizable. Since $B = B^T$ is nonsingular, there is a nonsingular $R \in M_n$ such that $R^TBR = I$. Let $A_1 \equiv R^*AR$, which is Hermitian. Since A is $*$-congruent to A_1, they have the same inertia, i.e., $I(A_1) = I(A)$. Now, $A_1 = R^*AR = R^*ARR^T(R^T)^{-1} = R^*(ARR^T)(\bar{R}^*)^{-1} = R^*(AB^{-1})(\bar{R}^*)^{-1}$. Thus, A_1 is consimilar to the condiagonalizable matrix AB^{-1} and, therefore, A_1 is a condiagonalizable Hermitian matrix. By Corollary 2.16 there is an orthogonal $Q \in M_n$ such that $Q^*R^*ARQ = Q^*A_1Q = MI(A_1) = MI(A)$ and $Q^TR^TBRQ = I$. Then $P \equiv RQ$ accomplishes the desired reduction. □

If $A \in M_n$ is positive definite and $B \in M_n$ is nonsingular and symmetric, let $B = P^TP$ for some nonsingular $P \in M_n$. Then $AB = AP^TP$, and hence $\bar{P}ABP^{-1} = \bar{P}AP^T$, so AB is consimilar to the positive-definite matrix $\bar{P}AP^T$, which is always condiagonalizable [5]. The corollary follows easily.

COROLLARY 3.3. *Let A, $B \in M_n$ with A positive definite and B nonsingular and symmetric. Then there is a nonsingular $P \in M_n$ such that $P^*AP = MI(A)$ and $P^TBP = I$, where $M =$* diag $(\lambda_1, \cdots, \lambda_n)$, *and $\lambda_i \geqq 0$ are coneigenvalues of AB^{-1}.*

REFERENCES

[1] C. R. DePrima and C. R. Johnson, *The range of $A^{-1}A^*$ in* Gl (n, C), Linear Algebra Appl., 9 (1974), pp. 209–222.

[2] F. R. Gantmacher, *Applications of the Theory of Matrices*, Interscience, New York, 1959.

[3] Y. P. Hong, *A Hermitian canonical form for complex matrices*, Linear Algebra Appl., to appear.

[4] Y. P. HONG AND R. A. HORN, *A canonical form for matrices under consimilarity*, Linear Algebra Appl., 102 (1988), pp. 143–168.

[5] ——, *On the reduction of a matrix to triangular or diagonal form by consimilarity*, SIAM J. Algebraic Discrete Methods, 7 (1986), pp. 80–88.

[6] Y. P. HONG, R. A. HORN, AND C. R. JOHNSON, *On the reduction of pairs of Hermitian or symmetric matrices to diagonal form by congruence*, Linear Algebra Appl., 72 (1986), pp. 213–226.

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

[8] L.-K. HUA, *Orthogonal classification of Hermitian matrices*, Trans. Amer. Math. Soc., 59 (1946), pp. 508–523.

[9] M. VUJICIC, F. HERBUT, AND G. VUJICIC, *Canonical form for matrices under unitary congruence transformations* I: *conjugate-normal matrices*, SIAM J. Appl. Math., 23 (1972), pp. 225–238.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 2, pp. 244-258, April 1989

010

SELF-EQUIVALENT FLOWS ASSOCIATED WITH THE SINGULAR VALUE DECOMPOSITION *

D. S. WATKINS † AND L. ELSNER ‡

Abstract. A family of flows which are continuous analogues of the constant and variable shift QR algorithms for the singular value decomposition problem is presented, and it is shown that certain of these flows interpolate the QR algorithm exactly. Here attention is not restricted to bidiagonal matrices; arbitrary rectangular matrices are considered.

Key words. singular value decomposition, QR algorithm, unitary equivalence, flow

AMS(MOS) subject classifications. 15A18, 15A23, 58F19, 58F25, 65F15

1. Introduction. In recent years there has been considerable interest in continuous analogues of the QR algorithm and other algorithms for calculating eigenvalues of matrices. See, for example, the works of Symes [14]; Deift, Nanda, and Tomei [4]; Nanda [7], [8]; Chu [1]; Watkins [16]; and Watkins and Elsner [17], all of which have appeared since 1982. See also the work of Rutishauser [12],[13] from the 1950's, which has been overlooked until recently. Given a matrix $\hat{A}$ whose eigenvalues are desired, the QR algorithm produces a sequence $A_0, A_1, A_2, \cdots$ such that each member of the sequence is similar to $\hat{A}$, and the matrices tend to upper triangular form. A continuous analogue of the QR algorithm produces a smooth, matrix-valued function or flow $B(t)$, such that, for all t, $B(t)$ is similar to $\hat{A}$, and $B(m) = A_m$ for $m = 0, 1, 2, \cdots$. That is, the flow interpolates the QR algorithm. More generally we may have $B(t)$ similar to $\hat{B} = g(\hat{A})$ and $B(m) = g(A_m)$ for some specified function g. Such a flow must satisfy

$$B(t) = F(t)^{-1}\hat{B}F(t) \tag{1}$$

for some nonsingular matrix function $F(t)$. In [17] we studied functions of the type (1), which we called *self-similar flows.*

When studying eigenvalues it is natural to employ similarity transformations, since they preserve eigenvalues. For certain other problems, such as the generalized eigenvalue problem and the singular value problem, it is more natural to consider equivalences. Recall that two matrices $A, \tilde{A} \in C^{n\times m}$ are *equivalent* if there exist nonsingular matrices $F \in C^{n\times n}$ and $Z \in C^{m\times m}$ such that $\tilde{A} = FAZ$. If F and Z are unitary, A and $\tilde{A}$ are *unitarily equivalent.* A matrix-valued function $B(t)$ defined on some interval is called a *self-equivalent flow* if there exist smooth, nonsingular, matrix-valued functions $F(t) \in C^{n\times n}$ and $Z(t) \in C^{m\times m}$, and $\hat{B} \in C^{n\times m}$, such that $B(t) = F(t)\hat{B}Z(t)$. If $F(t)$ and $Z(t)$ are unitary for all t, the flow is *unitarily self-equivalent.* In this paper we will develop unitarily self-equivalent flows associated with the singular value decomposition (SVD). We presented self-equivalent flows associated with the generalized eigenvalue problem in [18].

In [2] Chu presented a flow which is a continuous analogue of the QR algorithm for the SVD. The present paper constructs a large family of flows of which the flow

* Received by the editors November 23, 1987; accepted for publication (in revised form) September 20, 1988.

† Department of Pure and Applied Mathematics, Washington State University, Pullman, Washington 99164-2930 (watkins@wsumath.bitnet).

‡ Fakultät für Mathematik, Universität Bielefeld, Postfach 86 40, D-4800 Bielefeld 1, Federal Republic of Germany.

of [2] is a single example. Where Chu restricted his attention to bidiagonal matrices, we consider arbitrary (full or banded) rectangular matrices.

Our presentation begins with the introduction of an explicit version of the QR algorithm which can be used to find the SVD of an (almost) arbitrary rectangular matrix. By contrast, the implicit version of the QR algorithm which is usually used can be applied only to unreduced bidiagonal matrices. Our explicit version is not recommended for practical use. It is important because it adds to our understanding of the QR algorithm and its relationship to self-equivalent flows. For simplicity we consider the QR algorithm with a constant shift at first. We show that the algorithm converges to the SVD for almost all starting matrices.

In §3 we show that every self-equivalent flow must satisfy a differential equation of the general form $\dot{B} = CB+BD$. Conversely, every solution of a differential equation of this form must be a self-equivalent flow. This is a slight generalization of observations made in [2], [3].

In §4 we introduce a family of unitarily self-equivalent flows associated with the QR algorithm and present theorems which show that under mild assumptions the flows converge to the SVD of the initial matrix $\hat{B}$. One member of the family interpolates the constant shift QR algorithm.

We then consider shifted and generalized QR algorithms and a family of analogous flows. These flows differ from those considered earlier only in that the differential equations they satisfy are nonautonomous. Given any shift strategy for the QR algorithm for which none of the shifts is an eigenvalue of $\hat{A}^*\hat{A}$ or $\hat{A}\hat{A}^*$, we show how to construct numerous flows which interpolate the shifted algorithm. Of course, almost all shift strategies satisfy this condition.

In the final section of the paper we show that all of the flows which we have discussed preserve banded forms. That is, if the initial matrix $\hat{B}$ is banded, then $B(t)$ has the same band structure for all $t > 0$.

2. The QR algorithm for the SVD. Let $\hat{A} \in C^{n\times m}$. The most common way of calculating the singular value decomposition of $\hat{A}$ is to apply a variant of the implicit QR algorithm due to Golub and Kahan (see [6]). This requires that $\hat{A}$ be reduced to bidiagonal form before the QR iterations are begun. We will discuss an explicit variant which does not require the preliminary reduction to bidiagonal form. While this variant is not recommended for practical use, it is useful for our development. To keep matters simple we restrict our attention to the constant shift case at first. Let μ be a fixed positive number. Setting $A_0 = \hat{A}$, we create a sequence of unitarily equivalent matrices $A_0, A_1, A_2, \cdots$ as follows: Given A_{i-1}, perform QR decompositions of both $A_{i-1}^*A_{i-1} + \mu I_m$ and $A_{i-1}A_{i-1}^* + \mu I_n$:

$$(2) \qquad A_{i-1}^*A_{i-1} + \mu I_m = \bar{Q}_i\bar{R}_i, \qquad A_{i-1}A_{i-1}^* + \mu I_n = \bar{P}_i\bar{S}_i,$$

where $\bar{Q}_i \in C^{m\times m}$ and $\bar{P}_i \in C^{n\times n}$ are unitary, and $\bar{R}_i \in C^{m\times m}$ and $\bar{S}_i \in C^{n\times n}$ are upper triangular with real, positive main diagonal entries. Now define A_i by

$$(3) \qquad A_i = \bar{P}_i^* A_{i-1}\bar{Q}_i.$$

The reason for using the positive shift μ is that it guarantees that $A_{i-1}^*A_{i-1}+\mu I_m$ and $A_{i-1}A_{i-1}^*+\mu I_n$ are nonsingular. Therefore the factors in the QR decompositions in (2) are uniquely determined, and so is A_i via (3). If $\hat{A}$ is square and nonsingular, we can take $\mu = 0$ and get uniquely determined A_i. (If $\hat{A}$ is singular or nonsquare, one can still carry out the steps (2) and (3) with $\mu = 0$, but not in a unique manner.)

Since obviously

$$A_i^*A_i = \bar{Q}_i^*A_{i-1}^*A_{i-1}\bar{Q}_i = \bar{R}_i\bar{Q}_i - \mu I_m \tag{4}$$

and

$$A_iA_i^* = \bar{P}_i^*A_{i-1}A_{i-1}^*\bar{P}_i = \bar{S}_i\bar{P}_i - \mu I_n, \tag{5}$$

we see that the transformations $A_{i-1}^*A_{i-1} \to A_i^*A_i$ and $A_{i-1}A_{i-1}^* \to A_iA_i^*$ amount to QR steps. Therefore by standard results (see, e.g., [19]), the sequences $(A_i^*A_i)$ and $(A_iA_i^*)$ converge to diagonal form.

For $i = 0, 1, 2, \cdots$ let

$$Q_i = \bar{Q}_1\bar{Q}_2\cdots\bar{Q}_i, \qquad P_i = \bar{P}_1\bar{P}_2\cdots\bar{P}_i,$$

$$R_i = \bar{R}_i\bar{R}_{i-1}\cdots\bar{R}_1, \qquad S_i = \bar{S}_i\bar{S}_{i-1}\cdots\bar{S}_1.$$

Then

$$A_i = P_i^*\hat{A}Q_i, \tag{6}$$

$$A_i^*A_i = Q_i^*\hat{A}^*\hat{A}Q_i, \tag{7}$$

$$A_iA_i^* = P_i^*\hat{A}\hat{A}^*P_i, \tag{8}$$

and by induction

$$(\hat{A}^*\hat{A} + \mu I_m)^i = Q_iR_i, \qquad (\hat{A}\hat{A}^* + \mu I_n)^i = P_iS_i. \tag{9}$$

These are QR decompositions.

In addition, it is not hard to show that

$$A_i^*A_i = \bar{R}_iA_{i-1}^*A_{i-1}\bar{R}_i^{-1} = R_i\hat{A}^*\hat{A}R_i^{-1}, \tag{10}$$

$$A_iA_i^* = \bar{S}_iA_{i-1}A_{i-1}^*\bar{S}_i^{-1} = S_i\hat{A}\hat{A}^*S_i^{-1} \tag{11}$$

and

$$A_i = \bar{S}_iA_{i-1}\bar{R}_i^{-1} = S_i\hat{A}R_i^{-1}, \tag{12}$$

$$A_i^* = \bar{R}_iA_{i-1}^*\bar{S}_i^{-1} = R_i\hat{A}^*S_i^{-1}. \tag{13}$$

Only in (12) and (13) do we use the fact that in (2) the same μ is used in both decompositions.

In order to get some idea of how this algorithm relates to the usual implicit QR algorithm for the SVD, suppose $\hat{A}$ is square, upper triangular, and nonsingular, with real, positive main diagonal entries. There is no loss of generality in making these assumptions, for there exist procedures [5] for reducing an arbitrary problem to problems for which the matrix has this form. Then by (12) all A_i will be upper triangular with positive main diagonal entries. By (3) we have

$$\bar{P}_iA_i = A_{i-1}\bar{Q}_i,$$

which shows that we can get A_i by computing the QR decomposition of $A_{i-1}\bar{Q}_i$. Thus it is enough to find $\bar{Q}_i$. If $\hat{A}$ is bidiagonal and unreduced, one can determine $\bar{Q}_i$ implicitly, without forming $A_{i-1}^*A_{i-1}$. This is documented in [6], for example.

2.1. Convergence of the QR algorithm for the SVD. We have already noted that $A_i^*A_i$ and $A_iA_i^*$ converge to diagonal form. If all A_i are upper triangular with positive main diagonal entries, convergence of the A_i to diagonal form can be inferred from convergence of the $A_i^*A_i$. For in this case A_i is the upper Cholesky factor of $A_i^*A_i$. By continuity of the Cholesky decomposition, convergence of $A_i^*A_i$ to diagonal form implies the same for A_i. The main diagonal entries of A_i converge to the singular values of $\hat{A}$. The columns of P_i and Q_i converge to the left and right singular vectors, respectively.

While the upper triangular case is the most important, it is nevertheless interesting to study the convergence of (A_i) in general. The following examples show that convergence of $A_i^*A_i$ and $A_iA_i^*$ does not, in general, imply convergence of A_i to diagonal form.

Example 1. Let

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

which has singular values $\sigma_1 = 1$ and $\sigma_2 = 0$. Then

$$\hat{A}^*\hat{A} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad \hat{A}\hat{A}^* = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

so $\bar{Q}_1 = I_2$ and $\bar{P}_1 = I_2$ in (2) and, from (3), $A_1 = \hat{A}$. Thus $A_i = \hat{A}$ for all i.

Example 2. Let $\hat{A} \in C^{n\times n}$ be any unitary matrix. Then $\hat{A}^*\hat{A} = I_n$ and $\hat{A}\hat{A}^* = I_n$. Again $A_i = \hat{A}$ for all i.

These examples notwithstanding, the sequence (A_i) usually does converge to diagonal form, as we shall now show. Our approach can also be used to prove the convergence of the flows. We have the choice of a geometric proof in the spirit of [11] and [15] or a proof along classical lines [19, p. 517]. In this case we opt for the latter because it is shorter.

In the statements and proofs of the convergence theorems we suppose $\hat{A} \in C^{n\times m}$ with $\operatorname{rank}(\hat{A}) = r$. Let $\hat{A} = U\Sigma V^*$ be the SVD of $\hat{A}$. Then $U = [u_1, \cdots, u_n] \in C^{n\times n}$, $V = [v_1, \cdots, v_m] \in C^{m\times m}$, and $\Sigma = \operatorname{diag}\{\sigma_1, \cdots, \sigma_r\} \in C^{n\times m}$, where $u_1, \cdots, u_n$ (the left singular vectors of $\hat{A}$) are orthonormal eigenvectors of $\hat{A}\hat{A}^*$, $v_1, \cdots, v_m$ (the left singular vectors of $\hat{A}$) are orthonormal eigenvectors of $\hat{A}^*\hat{A}$, and $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are the nonzero singular values of $\hat{A}$. The common eigenvalues of $\hat{A}^*\hat{A} + \mu I_m$ and $\hat{A}\hat{A}^* + \mu I_n$ which are greater than μ are $\lambda_i = \sigma_i^2 + \mu$, $i = 1, 2, \cdots, r$. Any additional eigenvalues are equal to μ.

Let $e_1, \cdots, e_j$ denote the canonical basis vectors for C^j, where the value of j depends on the context. Given vectors $w_1, \cdots, w_k \in C^j$, let $\langle w_1, \cdots, w_k\rangle$ denote the subspace of C^j spanned by $w_1, \cdots, w_k$.

THEOREM 2.1. *Suppose $\sigma_k > \sigma_{k+1}$ for some k, and*

$$\langle v_1, \cdots, v_k\rangle \cap \langle e_{k+1}, \cdots, e_m\rangle = \{0\} \qquad (\text{in } C^m), \tag{14}$$

$$\langle u_1, \cdots, u_k\rangle \cap \langle e_{k+1}, \cdots, e_n\rangle = \{0\} \qquad (\text{in } C^n). \tag{15}$$

Let $\{A_i\}$ be the sequence defined by (3). *Partition each A_i as*

$$A_i = \begin{bmatrix} A_{11}^{(i)} & A_{12}^{(i)} \\ A_{21}^{(i)} & A_{22}^{(i)} \end{bmatrix},$$

with $A_{11}^{(i)} \in C^{k\times k}$. *Then* $A_{21}^{(i)} \to 0$ *and* $A_{12}^{(i)} \to 0$ *as* $i \to \infty$. *The singular values of* $A_{11}^{(i)}$ *and* $A_{22}^{(i)}$ *converge to* $\{\sigma_1, \cdots, \sigma_k\}$ *and* $\{\sigma_{k+1}, \cdots\}$, *respectively. The convergence is linear with contraction number* λ_{k+1}/λ_k.

Remarks. (1) The subspace conditions (14) and (15) are satisfied for almost all choices of $\hat{A}$ [15, pp. 429-430]. However, (14) is violated by the matrix in Example 1, since in that case $v_1 = e_2$.

(2) The matrix of Example 2 does not satisfy $\sigma_k > \sigma_{k+1}$ for any k because $\sigma_1 = \sigma_2 = \cdots = \sigma_n = 1$. The only matrices for which all singular values are equal are multiples of unitary matrices.

(3) The assumption that the shift μ is positive simplifies the statement and proof of the theorem but is not crucial to our arguments. All that is really needed is that $-\mu$ is not an eigenvalue of $\hat{A}^*\hat{A}$ and $\hat{A}\hat{A}^*$, and $\hat{A}^*\hat{A} + \mu I_m$ and $\hat{A}\hat{A}^* + \mu I_n$ do not have any eigenvalues of equal magnitude and opposite sign.

Proof. Define $\Lambda \in C^{m\times m}$ by $\Lambda = \operatorname{diag}\{\lambda_1, \cdots, \lambda_r\}$. Then $\hat{A}^*\hat{A} + \mu I_m = V\Lambda V^*$. By the first equation of (9) we have

$$Q_iR_i = (\hat{A}^*\hat{A} + \mu I_m)^i = V\Lambda^i V^*.$$

The subspace condition (14) guarantees that V^* has a block LU decomposition

$$V^* = LX = \begin{bmatrix} I_k & 0 \\ L_{21} & I_{m-k} \end{bmatrix} \begin{bmatrix} X_{11} & X_{12} \\ 0 & X_{22} \end{bmatrix}.$$

Clearly

$$Q_iR_i = V(\Lambda^i L \Lambda^{-i})\Lambda^i X.$$

Define $\Lambda_1 \in C^{k\times k}$ and $\Lambda_2 \in C^{(m-k)\times(m-k)}$ by $\Lambda_1 = \operatorname{diag}\{\lambda_1, \cdots, \lambda_k\}$ and $\Lambda_2 = \operatorname{diag}\{\lambda_{k+1}, \cdots\}$. Then $\Lambda = \operatorname{diag}\{\Lambda_1, \Lambda_2\}$, and

$$\Lambda^i L \Lambda^{-i} = \begin{bmatrix} I_k & 0 \\ \Lambda_2^i L_{21} \Lambda_1^{-i} & I_{m-k} \end{bmatrix}.$$

Since $\lambda_k > \lambda_{k+1}$, $\Lambda^i L \Lambda^{-i} \to I_m$ linearly with contraction number λ_{k+1}/λ_k. Let $\tilde{Q}_i\tilde{R}_i$ be the QR decomposition of $\Lambda^i X$. Then since $\Lambda^i X$ is block upper triangular, $\tilde{Q}_i$ must have the block diagonal form $\tilde{Q}_i = \operatorname{diag}\{\tilde{Q}_1^{(i)}, \tilde{Q}_2^{(i)}\}$, where $\tilde{Q}_1^{(i)} \in C^{k\times k}$ and $\tilde{Q}_2^{(i)} \in C^{(m-k)\times(m-k)}$ are unitary. Now Q_iR_i can be written as

$$Q_iR_i = V\tilde{Q}_i(\tilde{Q}_i^*\Lambda^i L \Lambda^{-i}\tilde{Q}_i)\tilde{R}_i.$$

Let $\tilde{\tilde{Q}}_i\tilde{\tilde{R}}_i$ be the QR decomposition of $\tilde{Q}_i^*\Lambda^i L\Lambda^{-i}\tilde{Q}_i$. Since $\tilde{Q}_i^*\Lambda^i L\Lambda^{-i}\tilde{Q}_i \to I_m$ linearly with contraction number λ_{k+1}/λ_k, the same is true of $\tilde{\tilde{Q}}_i$. Since

$$Q_iR_i = (V\tilde{Q}_i\tilde{\tilde{Q}}_i)(\tilde{\tilde{R}}_i\tilde{R}_i),$$

and QR decompositions are unique, we have

$$Q_i = V\tilde{Q}_i\tilde{\tilde{Q}}_i.$$

Repeating this argument starting from the second equation of (9), we find that

$$P_i = U\tilde{P}_i\tilde{\tilde{P}}_i,$$

where $\tilde{P}_i$ and $\tilde{\tilde{P}}_i$ are unitary, $\tilde{P}_i = \text{diag}\{\tilde{P}_1^{(i)}, \tilde{P}_2^{(i)}\}$, with $\tilde{P}_1^{(i)} \in C^{k\times k}$, and $\tilde{\tilde{P}}_i \to I_n$ linearly with contraction number λ_{k+1}/λ_k.

By (6) $A_i = P_i^* \hat{A} Q_i$, so

$$A_i = \tilde{\tilde{P}}_i^* \tilde{P}_i^* (U^* \hat{A} V) \tilde{Q}_i \tilde{\tilde{Q}}_i = \tilde{\tilde{P}}_i^* (\tilde{P}_i^* \Sigma \tilde{Q}_i) \tilde{\tilde{Q}}_i.$$

Define $\Sigma_1 = \text{diag}\{\sigma_1, \cdots, \sigma_k\} \in C^{k\times k}$ and $\Sigma_2 = \text{diag}\{\sigma_{k+1}, \cdots\} \in C^{(n-k)\times(m-k)}$, and let $B_i = \tilde{P}_1^{(i)*} \Sigma_1 \tilde{Q}_1^{(i)}$ and $C_i = \tilde{P}_2^{(i)*} \Sigma_2 \tilde{Q}_2^{(i)}$. Then

$$A_i = \begin{bmatrix} A_{11}^{(i)} & A_{12}^{(i)} \\ A_{21}^{(i)} & A_{22}^{(i)} \end{bmatrix} = \tilde{\tilde{P}}_i^* \begin{bmatrix} B_i & 0 \\ 0 & C_i \end{bmatrix} \tilde{\tilde{Q}}_i.$$

Since $\tilde{\tilde{P}}_i^* \to I_n$ and $\tilde{\tilde{Q}}_i \to I_m$, we see that $A_{21}^{(i)} \to 0$ and $A_{12}^{(i)} \to 0$ at the claimed rate. Since B_i and C_i have singular values $\{\sigma_1, \cdots, \sigma_k\}$ and $\{\sigma_{k+1}, \cdots\}$, respectively, the singular values of $A_{11}^{(i)}$ and $A_{22}^{(i)}$ must converge to these sets at the stated rate. □

THEOREM 2.2. *Let $\tau_1 > \cdots > \tau_j$ be the distinct nonzero singular values of $\hat{A}$, and let $\nu_k = \tau_k^2 + \mu$, $k = 1, \cdots, j$, be the corresponding eigenvalues of $\hat{A}^* \hat{A} + \mu I_m$ and $\hat{A}\hat{A}^* + \mu I_n$. Let m_k denote the multiplicity of τ_k and ν_k, $k = 1, \cdots, j$. (Thus $m_1 + \cdots + m_j = r$.) Suppose the subspace conditions* (14) *and* (15) *hold for every k for which $\sigma_k > \sigma_{k+1}$. Then (A_i) converges to the block diagonal form*

$$\begin{bmatrix} \tau_1 W_1 & 0 & \cdots & 0 & 0 \\ 0 & \tau_2 W_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \tau_j W_j & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \tag{16}$$

where $W_k \in C^{m_k \times m_k}$ is unitary, $k = 1, \cdots, j$. Convergence of the kth main diagonal block is linear with contraction number $\rho_k = \max\{\nu_{k+1}/\nu_k, \nu_k/\nu_{k-1}\}$, where $\nu_0 = \infty$ and $\nu_{j+1} = \mu$. ($\nu_{j+1} = 0$ if $m = n = r$.)

Remarks. (1) If $\hat{A}$ is upper triangular, the blocks in (16) must be upper triangular. Since a matrix which is both upper triangular and unitary must be diagonal, we get convergence to diagonal form in this case, provided the subspace conditions (14) and (15) are satisfied.

(2) The nonzero singular values of most matrices are distinct. In this case, assuming that the subspace conditions are satisfied, $A_i \to \text{diag}\{\tilde{\sigma}_1, \cdots, \tilde{\sigma}_r\} \in C^{n\times m}$, where $|\tilde{\sigma}_k| = \sigma_k$, $k = 1, \cdots, r$. The columns of the cumulative transformation matrices Q_i and P_i converge to (multiples of unit modulus of) right and left singular vectors, respectively.

(3) Every problem can be reduced to one or more subproblems for which $\hat{A}$ is square and bidiagonal, with real, strictly positive entries on both the main diagonal and the superdiagonal. If $\hat{A}$ is of this form, then both $\hat{A}^* \hat{A} + \mu I$ and $\hat{A}\hat{A}^* + \mu I$ are unreduced tridiagonal matrices. It follows that the singular values are distinct [10], and the subspace conditions (14) and (15) are satisfied for all k [9], [15]. Thus convergence to diagonal form is guaranteed in this case.

Proof. It follows from Theorem 2.1 that the off-diagonal blocks tend to zero. Furthermore, the singular values of the kth main diagonal block tend to the multiple singular value τ_k at the stated rate. It remains only to show that the convergence of the singular values implies the convergence of the main diagonal blocks of (A_i). While this is not hard to do, we have found that it is just as easy to prove the theorem from scratch, using a variant of the argument which was used in the proof of Theorem 2.1. We will show that $P_i = U\tilde{P}_i$ and $Q_i = V\tilde{Q}_i$, where $\tilde{P}_i$ and $\tilde{Q}_i$ converge to specific block diagonal unitary matrices. It follows that (A_i) converges to the form (16).

Let $\Lambda = \operatorname{diag}\{\lambda_1, \lambda_2, \cdots\}$, as in the proof of Theorem 2.1. Under the present hypotheses Λ has the form $\Lambda = \operatorname{diag}\{\nu_1 I_{m_1}, \nu_2 I_{m_2}, \cdots, \nu_j I_{m_j}, 0\}$. As in the proof of Theorem 2.1 we have, from the first equation of (9), $Q_i R_i = V\Lambda^i V^*$. The subspace conditions (14) guarantee that V^* has a block LU decomposition

$$V^* = LX = \begin{bmatrix} L_{11} & 0 & \cdots & 0 \\ L_{21} & L_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ L_{j+1,1} & L_{j+1,2} & \cdots & L_{j+1,j+1} \end{bmatrix} \begin{bmatrix} I & X_{12} & \cdots & X_{1,j+1} \\ 0 & I & \cdots & X_{2,j+1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & I \end{bmatrix},$$

where $L_{kk} \in C^{m_k \times m_k}$, $k = 1, \cdots, j$. Noting that

$$Q_i R_i = V(\Lambda^i L \Lambda^{-i})\Lambda^i X,$$

we examine the product $\Lambda^i L \Lambda^{-i}$. Clearly

$$\Lambda^i L \Lambda^{-i} = \begin{bmatrix} M_{11} & 0 & 0 & \cdots \\ M_{21} & M_{22} & 0 & \cdots \\ M_{31} & M_{32} & M_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

where $M_{jk} = (\nu_j/\nu_k)^i L_{jk}$. Therefore

$$\Lambda^i L \Lambda^{-i} \to \operatorname{diag}\{L_{11}, L_{22}, \cdots, L_{j+1,j+1}\}.$$

Consider the QR factorization $\operatorname{diag}\{L_{11}, L_{22}, \cdots, L_{j+1,j+1}\} = \hat{Q}\hat{R}$. Obviously $\hat{Q}$ is block diagonal:

$$\hat{Q} = \operatorname{diag}\{\hat{Q}_1, \hat{Q}_2, \cdots, \hat{Q}_{j+1}\}.$$

Let $\tilde{Q}_i \tilde{R}_i$ be the QR decomposition of $\Lambda^i L \Lambda^{-i}$. Then $\tilde{Q}_i \to \hat{Q}$ as $i \to \infty$. Also

$$Q_i R_i = (V\tilde{Q}_i)(\tilde{R}_i \Lambda^i X).$$

Since $V\tilde{Q}_i$ is unitary and $\tilde{R}_i \Lambda^i X$ is upper triangular with positive main diagonal entries,

$$Q_i = V\tilde{Q}_i.$$

Repeating this argument starting from the second equation of (9), we find that

$$P_i = U\tilde{P}_i,$$

where $\tilde{P}_i \to \hat{P} = \operatorname{diag}\{\hat{P}_1, \hat{P}_2, \cdots, \hat{P}_{j+1}\}$.

It now follows easily that (A_i) converges to block diagonal form. For

$$A_i = P_i^* \hat{A} Q_i = \tilde{P}_i^* \Sigma \tilde{Q}_i \to \hat{P}^* \Sigma \hat{Q}.$$

By hypothesis, Σ has the form $\Sigma = \operatorname{diag}\{\tau_1 I_{m_1}, \cdots, \tau_j I_{m_j}, 0\}$, so

$$A_i \to \operatorname{diag}\{\tau_1 \hat{P}_1^* \hat{Q}_1, \cdots \tau_j \hat{P}_j^* \hat{Q}_j, 0\} = \operatorname{diag}\{\tau_1 W_1, \cdots, \tau_j W_j, 0\},$$

where $W_k = \hat{P}_k^* \hat{Q}_k$, $k = 1, \cdots, j$. □

3. The differential equation of a self-equivalent flow. Let $\hat{B} \in C^{n\times m}$, and consider the self-equivalent flow

$$B(t) = F(t)\hat{B}Z(t). \tag{17}$$

Then $B(t)$ satisfies a differential equation, which can be found by differentiating (17).

$$\begin{aligned} \dot{B} &= \dot{F}\hat{B}Z + F\hat{B}\dot{Z} \\ &= \dot{F}F^{-1}B + BZ^{-1}\dot{Z} \\ &= CB + BD, \end{aligned} \tag{18}$$

where $C = \dot{F}F^{-1}$ and $D = Z^{-1}\dot{Z}$. Conversely, suppose $B(t)$ is the unique solution of an initial value problem

$$\dot{B} = CB + BD, \qquad B(0) = \hat{B}. \tag{19}$$

Let F and G be the solutions of the initial value problems

$$\dot{F} = CF, \qquad F(0) = I,$$

$$\dot{Z} = ZD, \qquad Z(0) = I.$$

Then $B(t) = F(t)\hat{B}Z(t)$. That is, $B(t)$ is a self-equivalent flow. To prove this result, let $\tilde{B}(t) = F(t)\hat{B}Z(t)$. Differentiate $\tilde{B}$ as in (18) to find that $\tilde{B}$ satisfies the initial value problem (19). Since the solution of (19) is unique, $\tilde{B} = B$. This result is a slight generalization of theorems appearing in Chu [2],[3].

In (19) we have purposely left the form of C and D vague to show that the form is unimportant. C and D could be constant matrices or prespecified functions of t, but the most interesting instances of (19) are those for which C and D also depend on B, since (19) is then nonlinear.

It will sometimes be useful to write the self-equivalence relation in slightly different ways, such as $B(t) = S(t)\hat{B}R(t)^{-1}$. Using the equation $\frac{d}{dt}(R^{-1}) = -R^{-1}\dot{R}R^{-1}$, we find that $B(t) = S(t)\hat{B}R(t)^{-1}$ if and only if

$$\dot{B} = CB - BD, \qquad B(0) = \hat{B}, \tag{20}$$

where S and R satisfy

$$\dot{S} = CS, \qquad S(0) = I,$$

$$\dot{R} = DR, \qquad R(0) = I.$$

Similarly, the relationship $B(t) = P(t)^{-1}\hat{B}Q(t)$ holds if and only if

$$\dot{B} = BD - CB, \qquad B(0) = \hat{B}, \tag{21}$$

where P and Q satisfy

$$\dot{P} = PC, \qquad P(0) = I,$$

$$\dot{Q} = QD, \qquad Q(0) = I.$$

Finally we note that P (respectively, Q) is unitary for all t if and only if $C(t)$ (respectively, $D(t)$) is skew-Hermitian for all t.

4. *QR* flows for the SVD. Every matrix $C \in C^{k\times k}$ ($k = n$ or m) can be expressed uniquely as a sum

$$C = \rho(C) + \sigma(C), \tag{22}$$

where $\rho(C)$ is skew-Hermitian, and $\sigma(C)$ is upper triangular with real entries on the main diagonal. Let $\hat{B} \in C^{n\times m}$. Given any real-valued function f defined on the spectra of $\hat{B}^*\hat{B}$ and $\hat{B}\hat{B}^*$, consider the flow

$$\dot{B} = B\rho(f(B^*B)) - \rho(f(BB^*))B, \qquad B(0) = \hat{B}. \tag{23}$$

This has the form (21), so $B(t) = P(t)^{-1}\hat{B}Q(t)$, where

$$\dot{P} = P\rho(f(BB^*)), \qquad P(0) = I, \tag{24}$$

$$\dot{Q} = Q\rho(f(B^*B)), \qquad Q(0) = I. \tag{25}$$

Since $\rho(f(BB^*))$ and $\rho(f(B^*B))$ are skew-Hermitian, $P(t)$ and $Q(t)$ are unitary, and we have

$$B(t) = P(t)^*\hat{B}Q(t). \tag{26}$$

We get as a special case the flow of Chu [2] by taking $\hat{B}$ to be real, square, and bidiagonal, and taking $f(x) = x$.

Using (22) and the equation $f(BB^*)B = Bf(B^*B)$, we see that (23) can also be written as

$$\dot{B} = \sigma(f(BB^*))B - B\sigma(f(B^*B)), \qquad B(0) = \hat{B}. \tag{27}$$

This has the form (20), so

$$B(t) = S(t)\hat{B}R(t)^{-1}, \tag{28}$$

where

$$\begin{aligned} \dot{S} &= \sigma(f(BB^*))S, & S(0) &= I, \\ \dot{R} &= \sigma(f(B^*B))R, & R(0) &= I. \end{aligned}$$

Since $\sigma(f(BB^*))$ and $\sigma(f(B^*B))$ are upper triangular with real main diagonal entries, $S(t)$ and $R(t)$ must be upper triangular with positive main diagonal entries.

Taking the conjugate transpose of (23), we find that B^* satisfies the differential equations

$$\dot{B}^* = \left\{ \begin{array}{lcl} B^*\rho(f(BB^*)) & - & \rho(f(B^*B))B^* \\ \sigma(f(B^*B))B^* & - & B^*\sigma(f(BB^*)) \end{array} \right\}, \qquad B(0)^* = \hat{B}^*,$$

from which it follows that

$$B(t)^* = Q(t)^*\hat{B}^*P(t) = R(t)\hat{B}^*S(t)^{-1}, \tag{29}$$

where P, Q, R, and S are as defined above. (Of course the expression $B(t)^* = Q(t)^*\hat{B}^*P(t)$ is already obvious.) The matrices $B(t)^*B(t)$ and $B(t)B(t)^*$ also satisfy certain differential equations. Easy computations show that

$$\frac{d}{dt}(B^*B) = [B^*B, \rho(f(B^*B))] = [\sigma(f(B^*B)), B^*B], \tag{30}$$

$$\frac{d}{dt}(BB^*) = [BB^*, \rho(f(BB^*))] = [\sigma(f(BB^*)), BB^*], \tag{31}$$

where $[X, Y] = XY - YX$. Thus B^*B and BB^* are QR flows of the type described in [16],[17], and elsewhere. We also note that

$$B(t)^*B(t) = Q(t)^*\hat{B}^*\hat{B}Q(t) = R(t)\hat{B}^*\hat{B}R(t)^{-1}, \tag{32}$$

$$B(t)B(t)^* = P(t)^*\hat{B}\hat{B}^*P(t) = S(t)\hat{B}\hat{B}^*S(t)^{-1}. \tag{33}$$

Because these are QR flows, we have [16],[17]

$$\exp(f(\hat{B}^*\hat{B})t) = Q(t)R(t), \tag{34}$$

$$\exp(f(\hat{B}\hat{B}^*)t) = P(t)S(t). \tag{35}$$

These are QR decompositions.

4.1. The relationship between the QR flows and the QR algorithm for the SVD. For a special choice of f the QR flow interpolates the constant shift QR algorithm. Obviously $f(x) = \log(x + \mu)$, $\mu > 0$, is defined on the common spectrum of $\hat{B}^*\hat{B}$ and $\hat{B}\hat{B}^*$.

THEOREM 4.1. *The QR algorithm* (2), (3) *with initial matrix $\hat{A}$ and the QR flow* (23) *with $f(x) = \log(x + \mu)$ and initial matrix $\hat{B} = \hat{A}$ are related by $A_i = B(i)$, $i = 0, 1, 2, \cdots$. In other words, the QR flow with $f(x) = \log(x + \mu)$ interpolates the QR algorithm with constant shift μ.*

Proof. The assumptions imply that $\hat{A}^*\hat{A} + \mu I_m = \exp(f(\hat{B}^*\hat{B}))$ and $\hat{A}\hat{A}^* + \mu I_n = \exp(f(\hat{B}\hat{B}^*))$. Therefore (34) and (35), taken at $t = 0, 1, 2, \cdots$, can be rewritten as

$$\begin{array}{rcl} (\hat{A}^*\hat{A} + \mu I_m)^i & = & Q(i)R(i), \\ (\hat{A}\hat{A}^* + \mu I_n)^i & = & P(i)S(i), \end{array} \quad i = 0, 1, 2, \cdots. \tag{36}$$

Comparing these with the decompositions (9) and recalling that the QR decompositions are unique in the nonsingular case, we find that

$$\begin{array}{cc} Q(i) = Q_i, & R(i) = R_i, \\ P(i) = P_i, & S(i) = S_i, \end{array} \quad i = 0, 1, 2, \cdots. \tag{37}$$

Thus, by (6) and (26) we have

$$A_i = P_i^*\hat{A}Q_i = P(i)^*\hat{B}Q(i) = B(i), \qquad i = 0, 1, 2, \cdots.$$

The same conclusion can also be obtained using (12) and (28) instead of (6) and (26): $A_i = S_i\hat{A}R_i^{-1} = S(i)\hat{B}R(i)^{-1} = B(i)$. □

For choices of f other than $\log(x + \mu)$ we have the following weaker interpolation properties.

THEOREM 4.2. *The QR algorithm* (2,3) *with initial matrix $\hat{A}$ and the QR flow* (23) *have the following relationships:*

If $\hat{A}^*\hat{A} + \mu I_m = \exp(f(\hat{B}^*\hat{B}))$, then $A_i^*A_i + \mu I_m = \exp(f(B(i)^*B(i)))$ for $i = 0, 1, 2, \cdots$.

If $\hat{A}\hat{A}^* + \mu I_n = \exp(f(\hat{B}\hat{B}^*))$, then $A_iA_i^* + \mu I_n = \exp(f(B(i)B(i)^*))$ for $i = 0, 1, 2, \cdots$.

Proof. If $\hat{A}^*\hat{A} + \mu I_m = \exp(f(\hat{B}^*\hat{B}))$, then the equations in the first line of (36) and (37) hold. In particular, $Q(i) = Q_i$, $i = 0, 1, 2, \cdots$. Therefore, by (7) and (32),

$$\begin{aligned} A_i^* A_i + \mu I_m &= Q_i^*(\hat{A}^*\hat{A} + \mu I_m)Q_i \\ &= Q(i)^* \exp(f(\hat{B}^*\hat{B}))Q(i) \\ &= \exp(f(Q(i)^*\hat{B}^*\hat{B}Q(i))) \\ &= \exp(f(B(i)^*B(i))). \end{aligned}$$

The second assertion is proved similarly. □

4.2. Convergence of QR flows. The flows satisfy convergence theorems analogous to Theorems 2.1 and 2.2. Let $\hat{B} = U\Sigma V^*$ be the SVD of $\hat{B}$, with $U = [u_1, \cdots, u_n] \in C^{n\times n}$, $\Sigma = \operatorname{diag}\{\sigma_1, \cdots, \sigma_r\} \in C^{n\times m}$, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, and $V = [v_1, \cdots, v_m] \in C^{m\times m}$. The eigenvalues of $\exp(f(\hat{B}^*\hat{B}))$ and $\exp(f(\hat{B}\hat{B}^*))$ are $\lambda_i = \exp(f(\sigma_i^2))$, $i = 1, \cdots, r$. If $r < m$ (or $r < n$), $\exp(f(\hat{B}^*\hat{B}))$ (respectively, $\exp(f(\hat{B}\hat{B}^*))$) has the additional eigenvalue $\lambda_{r+1} = \exp(f(0))$ of multiplicity $m - r$ (respectively, $n - r$). For convenience we will assume that f is a strictly increasing function. This has the effect that the eigenvalues satisfy $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > \lambda_{r+1}$. In analogy with Theorem 2.1 we have Theorem 4.3.

THEOREM 4.3. *Let $B(t)$ be the solution of* (23), *where f is strictly increasing. Suppose $\sigma_k > \sigma_{k+1}$ for some k, and*

$$(38) \qquad \langle v_1, \cdots, v_k\rangle \cap \langle e_{k+1}, \cdots, e_m\rangle = \{0\} \qquad (\textit{in } C^m),$$

$$(39) \qquad \langle u_1, \cdots, u_k\rangle \cap \langle e_{k+1}, \cdots, e_n\rangle = \{0\} \qquad (\textit{in } C^n).$$

Partition $B(t)$ as

$$B(t) = \begin{bmatrix} B_{11}(t) & B_{12}(t) \\ B_{21}(t) & B_{22}(t) \end{bmatrix},$$

with $B_{11}(t) \in C^{k\times k}$. Then $B_{21}(t) \to 0$ and $B_{12}(t) \to 0$ as $t \to \infty$. The singular values of $B_{11}(t)$ and $B_{22}(t)$ converge to $\{\sigma_1, \cdots, \sigma_k\}$ and $\{\sigma_{k+1}, \cdots\}$, respectively. The convergence is linear with contraction number λ_{k+1}/λ_k.

Proof. The proof is identical to that of Theorem 2.1, except that $\exp(f(\hat{B}^*\hat{B}))$ and $\exp(f(\hat{B}\hat{B}^*))$ replace $\hat{A}^*\hat{A} + I_m$ and $\hat{A}\hat{A}^* + I_n$, and the continuous variable t replaces the discrete variable i. □

In analogy with Theorem 2.2 we have Theorem 4.4.

THEOREM 4.4. *Let $B(t)$ be the solution of* (23), *where f is strictly increasing. Let $\tau_1 > \cdots > \tau_j$ be the distinct nonzero singular values of $\hat{B}$, and let $\nu_k = \exp(f(\tau_k^2))$, $k = 1, \cdots, j$, be the corresponding eigenvalues of $\exp(f(\hat{B}^*\hat{B}))$ and $\exp(f(\hat{B}\hat{B}^*))$. Let m_k denote the multiplicity of τ_k and ν_k, $k = 1, \cdots, j$. (Thus $m_1 + \cdots + m_j = r$.) Suppose the subspace conditions* (38) *and* (39) *hold for every k for which $\sigma_k > \sigma_{k+1}$. Then $B(t)$ converges to the block diagonal form*

$$(40) \qquad \begin{bmatrix} \tau_1 W_1 & 0 & \cdots & 0 & 0 \\ 0 & \tau_2 W_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \tau_j W_j & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix},$$

where $W_k \in C^{m_k \times m_k}$ *is unitary,* $k = 1, \cdots, j$. *Convergence of the kth main diagonal block is linear with contraction number* $\rho_k = \max\{\nu_{k+1}/\nu_k, \nu_k/\nu_{k-1}\}$, *where* $\nu_0 = \infty$ *and* $\nu_{j+1} = \exp(f(0))$. ($\nu_{j+1} = 0$ *if* $m = n = r$.)

Proof. The proof is analogous to that of Theorem 2.2. □

Remarks. (1) If $\hat{B}$ is upper triangular, then $B(t)$ is upper triangular for all t by (28). Therefore each of the main diagonal blocks in (40) must be both unitary and upper triangular, hence diagonal. Therefore $B(t)$ converges to diagonal form, provided the subspace conditions (38) and (39) are satisfied.

(2) If the nonzero singular values of $\hat{B}$ are distinct, and the subspace conditions are satisfied, $B(t) \to \operatorname{diag}\{\tilde{\sigma}_1, \cdots, \tilde{\sigma}_r\} \in C^{n \times m}$, where $|\tilde{\sigma}_k| = \sigma_k$, $k = 1, \cdots, r$. The columns of the transformation matrices $Q(t)$ and $P(t)$ converge to (multiples of unit modulus of) right and left singular vectors, respectively.

(3) Consider the important special case $f(x) = \log(x + \mu)$. If $\hat{B}$ is bidiagonal, with real, strictly positive entries on both the main diagonal and the superdiagonal, then both $\exp(f(\hat{B}^*\hat{B})) = \hat{B}^*\hat{B} + \mu I_m$ and $\exp(f(\hat{B}\hat{B}^*)) = \hat{B}\hat{B}^* + \mu I_n$ are unreduced tridiagonal matrices. Thus the singular values are distinct [10], the subspace conditions (38) and (39) are satisfied for all k [9], [15], and convergence to diagonal form is guaranteed.

(4) Both Theorem 4.3 and Theorem 4.4 can be extended to the case in which f is not monotone. In this case the ordering of the eigenvalues can differ from that of the singular values. The order of the blocks in (40) depends on the order of the eigenvalues, not the singular values. In particular the zero block on the main diagonal need not be at the end; it can be sandwiched between nonzero blocks. Note that this block is not necessarily square; it has dimensions $(n-r) \times (m-r)$. Both the statement and the proof of Theorem 4.3 become more delicate in this case.

5. Generalized QR algorithms and flows. The practical QR algorithm uses a different shift at each step to speed convergence. At step i a shift σ_i is chosen. Instead of (2) we have

$$(41) \qquad A_{i-1}^*A_{i-1} - \sigma_i I = \bar{Q}_i\bar{R}_i, \qquad A_{i-1}A_{i-1}^* - \sigma_i I = \bar{P}_i\bar{S}_i.$$

Then A_i is defined by

$$(42) \qquad A_i = \bar{P}_i^* A_{i-1}\bar{Q}_i,$$

as before. Equations (41) can be expressed more compactly as

$$(43) \qquad p_i(A_{i-1}^*A_{i-1}) = \bar{Q}_i\bar{R}_i, \qquad p_i(A_{i-1}A_{i-1}^*) = \bar{P}_i\bar{S}_i,$$

where $p_i(x) = x - \sigma_i$. More generally we can carry out the process (43), (42), where $p_1, p_2, p_3, \cdots$ is any sequence of functions defined on the spectra of $\hat{A}^*\hat{A}$ and $\hat{A}\hat{A}^*$. This is the *generalized QR algorithm* for the SVD problem.

Clearly $A_i^*A_i = \bar{Q}_i^*A_{i-1}^*A_{i-1}\bar{Q}_i$ and $A_iA_i^* = \bar{P}_i^*A_{i-1}A_{i-1}^*\bar{P}_i$, showing that the transformations $A_{i-1}^*A_{i-1} \to A_i^*A_i$ and $A_{i-1}A_{i-1}^* \to A_iA_i^*$ amount to shifted or generalized QR steps. Equations (6)–(8) continue to hold. Equations (9) are replaced by

$$(44) \qquad \prod_{j=1}^{i} p_j(\hat{A}^*\hat{A}) = Q_iR_i, \qquad \prod_{j=1}^{i} p_j(\hat{A}\hat{A}^*) = P_iS_i.$$

Notice that if $p_i(A_{i-1}^*A_{i-1})$ and $p_i(A_{i-1}A_{i-1}^*)$ are nonsingular, then both of the QR decompositions in (43) are unique. This is typically the case. For example, if $p_i(x) = x - \sigma_i$, where $\sigma_i \neq 0$ is not an eigenvalue of $\hat{A}^*\hat{A}$ and $\hat{A}\hat{A}^*$, then both $p_i(A_{i-1}^*A_{i-1})$ and $p_i(A_{i-1}A_{i-1}^*)$ are nonsingular.

If all of $p_i(\hat{A}^*\hat{A})$ and $p_i(\hat{A}\hat{A}^*)$, $i = 1,2,3,\cdots$ are nonsingular, then equations (10)–(13) all hold, and the QR decompositions in (44) are unique.

The algorithm can be shown to converge for various choices of $p_1, p_2, p_3, \cdots$. For example, if $p_i(x) = x - \sigma_i$, where (σ_i) converges to an eigenvalue, and the subspace conditions (38) and (39) are satisfied, the algorithm will converge. Because the shifts approach an eigenvalue, the block in the lower right-hand corner will converge rapidly.

5.1. Generalizing the QR flow. Given a generalized QR algorithm, we would like to find flows which interpolate the algorithm at integer times. To this end we consider nonautonomous flows satisfying differential equations of the form

$$\dot{B} = B\rho(f(t, B^*B)) - \rho(f(t, BB^*))B, \qquad B(0) = \hat{B}, \tag{45}$$

where f is piecewise continuous in t. For this type of flow the properties (24) through (33) all continue to hold, except that f now depends on t. In particular,

$$\frac{d}{dt}(B^*B) = [B^*B, \rho(f(t, B^*B))],$$

$$\frac{d}{dt}(BB^*) = [BB^*, \rho(f(t, BB^*))],$$

showing that B^*B and BB^* are nonautonomous QR flows of the type studied in §9 of [17]. Therefore by Theorem 9.1 of [17], we have

$$\exp\left\{\int_0^t f(s, \hat{B}^*\hat{B})ds\right\} = Q(t)R(t), \tag{46}$$

$$\exp\left\{\int_0^t f(s, \hat{B}\hat{B}^*)ds\right\} = P(t)S(t), \tag{47}$$

where Q, R, P, and S are the unique solutions of

$$\dot{Q} = Q\rho(f(t, B^*B)), \qquad Q(0) = I, \tag{48}$$

$$\dot{R} = \sigma(f(t, B^*B))R, \qquad R(0) = I, \tag{49}$$

$$\dot{P} = P\rho(f(t, BB^*)), \qquad P(0) = I, \tag{50}$$

$$\dot{S} = \sigma(f(t, BB^*))S, \qquad S(0) = I. \tag{51}$$

5.2. The connection between generalized QR algorithms and QR flows.

THEOREM 5.1. *Suppose*

$$\int_{j-1}^{j} f(s,x)ds = \log(p_j(x)), \qquad j = 1,2,3,\cdots. \tag{52}$$

Then the generalized QR algorithm based on $p_1, p_2, p_3, \cdots$, with initial matrix $\hat{A} \in C^{n\times m}$, and the generalized QR flow based on f, with initial matrix $\hat{B} = \hat{A}$, are related by $A_i = B(i)$, $i = 0,1,2,\cdots$.

Proof. Substituting $\hat{A}^*\hat{A}$ $(=\hat{B}^*\hat{B})$ for x in (52), summing j from 1 to i, and taking exponents, we find that for $i = 1,2,3,\cdots$,

$$\exp\left\{\int_0^i f(s,\hat{B}^*\hat{B})ds\right\} = \prod_{j=1}^{i} p_j(\hat{A}^*\hat{A}).$$

Then by (46) and (44), $Q(i)R(i) = Q_iR_i$, $i = 0,1,2,\cdots$. By uniqueness of the QR decomposition, $Q(i) = Q_i$ and $R(i) = R_i$, $i = 0,1,2,\cdots$. Performing the same steps with $\hat{A}\hat{A}^*$ in place of $\hat{A}^*\hat{A}$, we find that $P(i) = P_i$ and $S(i) = S_i$, $i = 0,1,2,\cdots$. Therefore by (6) and (26),

$$A_i = P_i^*\hat{A}Q_i = P(i)^*\hat{B}Q(i) = B(i)$$

for $i = 0,1,2,\cdots$. □

Remark. We could have drawn the same conclusion using R and S instead of Q and P.

Provided that $p_1, p_2, p_3, \cdots$ are chosen so that $\log(p_i(\hat{A}^*\hat{A}))$ and $\log(p_i(\hat{A}\hat{A}^*))$ are always meaningful, there are many ways to choose f so that the equations (52) are satisfied. Some examples are given in [17, Examples 9.4–9.7]. There is no need to repeat them here.

6. Preservation of band structure. A matrix $C = (c_{ij}) \in C^{n\times m}$ is said to be *lower k-banded* if $c_{ij} = 0$ whenever $i - j > k$. For example, upper triangular matrices are lower 0-banded. It is easy to show that the product of a lower k-banded matrix with an upper triangular matrix, in either order, is lower k-banded. A matrix is *upper k-banded* if its transpose is lower k-banded. A matrix that is both lower 0-banded and upper 1-banded is bidiagonal.

THEOREM 6.1. *Let $B(t)$ be a flow which satisfies an initial value problem of the form* (45). *If $\hat{B}$ is lower k-banded, then $B(t)$ is lower k-banded for all t. If $\hat{B}$ is upper j-banded, then $B(t)$ is upper j-banded for all t. In particular, if $\hat{B}$ is bidiagonal, then $B(t)$ is bidiagonal for all t.*

Proof. Suppose $\hat{B}$ is lower k-banded. By (28) $B(t) = S(t)\hat{B}R(t)^{-1}$, where both $S(t)$ and $R(t)^{-1}$ are upper triangular. Thus $B(t)$ is lower k-banded.

Now suppose $\hat{B}$ is upper j-banded. Then $\hat{B}^*$ is lower j-banded. By (29) $B(t)^* = R(t)\hat{B}^*S(t)^{-1}$, where $R(t)$ and $S(t)^{-1}$ are both upper triangular. Therefore $B(t)^*$ is lower j-banded for all t; that is, $B(t)$ is upper j-banded for all t. □

REFERENCES

[1] M. CHU, *The generalized Toda flow, the QR algorithm, and the centre manifold theory*, SIAM J. Algebraic Discrete Meth., 5 (1984), pp. 187-201.

[2] M. CHU, *A differential equation approach to the singular value decomposition of bidiagonal matrices,* Linear Algebra Appl. 80, (1986), pp. 71-80.

[3] ———, *A continuous approximation to the generalized Schur decomposition,* Linear Algebra Appl., 78 (1986), pp. 119-132.

[4] P. DEIFT, T. NANDA, AND C. TOMEI, *Differential equations for the symmetric eigenvalue problem,* SIAM J. Numer. Anal., 20 (1983), pp. 1-22.

[5] 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.

[6] G. GOLUB AND C. VAN LOAN, *Matrix Computations,* The Johns Hopkins University Press, Baltimore, MD, 1983.

[7] T. NANDA, *Isospectral flows on band matrices,* Ph.D. thesis, Courant Institute of Mathematical Sciences, New York University, New York, NY, 1982.

[8] ———, *Differential equations and the QR algorithm,* SIAM J. Numer. Anal., 22 (1985), pp. 310-321.

[9] B. N. PARLETT, *Global convergence of the basic QR algorithm on Hessenberg matrices,* Math. Comp., 22 (1968), pp. 803-817.

[10] ———, *The Symmetric Eigenvalue Problem,* Prentice-Hall, Englewood Cliffs, NJ, 1980.

[11] B. N. PARLETT AND W. G. POOLE, JR., *A geometric theory for the QR, LU, and power iterations,* SIAM J. Numer. Anal., 8 (1973), pp. 389-412.

[12] H. RUTISHAUSER, *Ein infinitesimales Analogon zum Quotienten-Differenzen-Algorithmus,* Arch. Math. (Basel), 5 (1954), pp. 132-137.

[13] ———, *Solution of Eigenvalue Problems with the LR-Transformation,* National Bureau of Standards Applied Mathematics Series 49, 1958, pp. 47-81.

[14] W. W. SYMES, *The QR algorithm and scattering for the finite nonperiodic Toda lattice,* Physica D, 4 (1982), pp. 275-280.

[15] D. S. WATKINS, *Understanding the QR algorithm,* SIAM Rev., 24 (1982), pp. 427-440.

[16] ———, *Isospectral flows,* SIAM Rev., 26 (1984), pp. 379-391.

[17] D. S. WATKINS AND L. ELSNER, *Self-similar flows,* Linear Algebra Appl., 110 (1988), pp. 213-242.

[18] ———, *Self-equivalent flows associated with the generalized eigenvalue problem,* Linear Algebra Appl., to appear.

[19] J. H. WILKINSON, *The Algebraic Eigenvalue Problem,* Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 2, pp. 259-277, April 1989

011

CONTINUOUS HOMOTOPIES FOR THE LINEAR COMPLEMENTARITY PROBLEM*

LAYNE T. WATSON[†], J. PATRICK BIXLER[†], AND AUBREY B. POORE[‡]

Abstract. There are various formulations of the linear complementarity problem as a Kakutani fixed point problem, a constrained optimization, or a nonlinear system of equations. These formulations have remained a curiosity since not many people seriously thought that a linear combinatorial problem should be converted to a nonlinear problem. Recent advances in homotopy theory and new mathematical software capabilities such as HOMPACK indicate that continuous nonlinear formulations of linear and combinatorial problems may not be farfetched. Several different types of continuous homotopies for the linear complementarity problem are presented and analyzed here, with some numerical results. The homotopies with the best theoretical properties (global convergence and no singularities along the zero curve) turn out to also be the best in practice.

Key words. homotopy algorithm, globally convergent, linear complementarity problem, fixed point, expanded Lagrangian, nonlinear equations

AMS(MOS) subject classifications. 65H10, 65L10, 65L60

1. Introduction. Given a real $n \times n$ matrix M and a real n-vector q, the linear complementarity problem (LCP), denoted by (q, M), is to find n-vectors w and z such that

$$w - Mz = q,$$
$$w \geqq 0, \quad z \geqq 0, \quad w^t z = 0.$$

The constraint $w^t z = 0$ is called the complementarity condition since for any i, $1 \leqq i \leqq n$, $z_i = 0$ if $w_i \neq 0$, and vice versa. A solution where some $z_i = w_i = 0$ is called degenerate. The linear complementarity problem arises in such diverse areas as economic modeling [15], [16], [59]; bimatrix games [29], [32]; mathematical programming [10], [19], [34]; mechanics [17]; lubrication [28]; and numerical analysis [9].

There are numerous algorithms for solving special classes of linear complementarity problems. Those based on pivoting or simplex-type processes include Lemke's complementary pivot algorithm [29]; Cottle and Dantzig's principal pivot method [6]; Bard-type algorithms [4], [45], [60]; and the n-cycle algorithm [62], [64]. There are also linear iterative techniques, similar to those for solving linear systems of equations, such as SOR [2], [3], [8], [35], [50], [51], [61] and various related fixed point iteration schemes. A very different algorithm is the simplicial homotopy algorithm of Merrill [37], applied to a Kakutani fixed point formulation (solution is a fixed point of a point-to-set mapping) of the linear complementarity problem.

* Received by the editors December 23, 1987; accepted for publication (in revised form) October 21, 1988.

[†] Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061. The work of L. T. Watson was supported in part by Air Force Office of Scientific Research grant 85–0250.

[‡] Department of Mathematics, Colorado State University, Fort Collins, Colorado 80523. The work of this author was supported in part by National Science Foundation grant DMS–8704679.

A more recent development was the formulation of the linear complementarity problem as a differentiable nonlinear system of equations [33], and the solution of this system of equations by a globally convergent homotopy method [66]. This approach has remained a curiosity because few people took seriously the formulation of a linear combinatorial problem (like the LCP) as a highly nonlinear problem. Recent advances in homotopy theory and mathematical software for nonlinear systems of equations [68]–[69], and new nonlinear formulations of linear, discrete, and combinatorial problems ([33], [53], [54], [66], [67]) suggest that nonlinear formulations of the linear complementarity problem should be investigated further.

The present paper proposes and analyzes several nonlinear homotopies for the linear complementarity problem. The existence theorems implied by the globally convergent homotopy theorems are as general as any derived by other methods. Section 2 defines some terminology, §§3–9 describe and analyze different homotopy maps, §10 describes some numerical experiments, and §11 summarizes.

2. Preliminaries. In this section we gather some terms and fundamental results about globally convergent homotopy methods. For additional background refer to [65], [68].

Let E^n denote n-dimensional, real Euclidean space and let $E^{n\times n}$ be the set of real $n \times n$ matrices. The ith component of a vector $v \in E^n$ is denoted by v_i, and for a matrix $A \in E^{n\times n}$, $A_{i\cdot}$ denotes the ith row and $A_{\cdot j}$ denotes the jth column. For subsets $\emptyset \neq I, J \subset \{1, \cdots, n\}$, A_{IJ} denotes the submatrix of A with rows indexed by I and columns indexed by J. Let $e \in E^n$ be the vector such that $e_i = 1$ for all i. For $v \in E^n$, $v+$ denotes the vector with components $(v+)_i = \max\{0, v_i\}$, and $v-$ denotes the vector with components $(v-)_i = \max\{0, -v_i\}$. The *support* of v, denoted by $S(v)$, is simply $\{i \mid v_i \neq 0\}$. We use the following notation when comparing a vector $a \in E^n$ to 0:

$$\begin{aligned} a &\geqq 0 \quad \text{if } a_i \geqq 0 \text{ for all } i,\\ a &\geq 0 \quad \text{if } a \geqq 0 \text{ and } a \neq 0,\\ a &> 0 \quad \text{if } a_i > 0 \text{ for all } i. \end{aligned}$$

Let $M \in E^{n\times n}$ be a real $n \times n$ matrix and let q be a real n-vector. M is *nonnegative* if each element of M is nonnegative, *copositive* if $x^tMx \geqq 0$ for all $x \geqq 0$, and *strictly copositive* if $x^tMx > 0$ for all $x \geq 0$. M is called *nondegenerate* if all of its principal minors are nonzero, and a *P-matrix* if all of its principal minors are positive. The vector q is *nondegenerate* with respect to M if q is not a linear combination of any $n-1$ columns of $(I, -M)$. Finally, M is *strictly semimonotone* if for each vector $x \geq 0$ there exists an index k such that $x_k(Mx)_k > 0$.

When $w \geqq 0$ and $z \geqq 0$ satisfy $w - Mz = q$, (w, z) is called a *feasible solution.* If $w^tz = 0$ also, (w, z) is called a *complementary feasible solution.*

A C^2 (twice continuously differentiable) function $F : E^n \rightarrow E^m$ is said to be *transversal to zero* if the $m \times n$ Jacobian matrix $DF(x)$ has rank m on $F^{-1}(0)$. The theoretical justification for modern probability-one homotopy methods rests on a result from differential geometry, known as a parameterized Sard's theorem [65]:

LEMMA 2.1. *Let $\rho : E^m \times [0,1) \times E^n \to E^n$ be a C^2 map which is transversal to zero, and define*

$$\rho_a(\lambda, z) = \rho(a, \lambda, z).$$

Then for almost all $a \in E^m$, the map ρ_a is also transversal to zero.

The significance of Lemma 2.1 is partially given by:

LEMMA 2.2. *In addition to the hypotheses of Lemma 2.1, suppose that for each $a \in E^m$ the system $\rho_a(0, z) = 0$ has a unique solution $z^{(0)}$. Then for almost all $a \in E^m$ there is a smooth zero curve $\gamma \subset [0,1) \times E^n$ of $\rho_a(\lambda, z)$, emanating from $(0, z^{(0)})$, along which the Jacobian matrix $D\rho_a(\lambda, z)$ has rank n. γ does not intersect itself or any other zero curves of ρ_a, does not bifurcate, has finite arc length in any compact subset of $[0,1) \times E^n$, and either goes to infinity or reaches the hyperplane $\lambda = 1$.*

LEMMA 2.3. *Under the hypotheses of Lemma 2.2, if the zero curve γ is bounded, then it has an accumulation point $(1, \bar{z})$. Furthermore, if rank $D\rho_a(1, \bar{z}) = n$, then γ has finite arc length.*

Conceptually, the algorithm for solving the nonlinear system of equations $F(z) = 0$ is simple. Using the lemmas above, just follow the zero curve γ, starting from some point $(0, z^{(0)})$ and ending at a point $(1, \bar{z})$, where $\bar{z}$ is a zero of $F(z)$. Computationally this may be nontrivial, but at least the idea is clear. A typical simple choice for the homotopy map is

$$\rho_a(\lambda, z) = \lambda F(z) + (1 - \lambda)(z - a).$$

Although this homotopy map has the same form as a standard continuation or embedding mapping, there are two important differences. First, in standard continuation the embedding parameter λ increases monotonically from 0 to 1 as the trivial problem $(z - a) = 0$ is continuously deformed to the given problem $F(z) = 0$. With the present homotopy method, turning points on γ cause no special difficulties and so λ can increase and decrease as the curve is being tracked. Secondly, the fact that the Jacobian matrix $D\rho_a$ has full rank along γ and the way in which the zero curve is tracked guarantee that there are never any "singular points" which afflict standard continuation methods.

3. The 1979 homotopy.

To provide a backdrop for the homotopies presented in the next few sections, we briefly review the homotopy map of [66]. Mangasarian [33] has shown that the linear complementarity problem (q, M) can be reformulated as a zero finding problem

$$H(z) = 0,$$

where $H(z)$ can be made as smooth as desired. Taking $\theta(t) = t^3$ in Mangasarian's Theorem 1 [33], we define $H(z)$ by

$$H_i(z) = -|M_{i\cdot}z + q_i - z_i|^3 + (M_{i\cdot}z + q_i)^3 + {z_i}^3$$

and

$$\rho_a(\lambda, z) = \lambda H(z) + (1 - \lambda)(z - a).$$

By noting the signs of each term in H, it is clear that $z \geqq 0$, $Mz + q \geqq 0$, and $(Mz + q)^t z = 0$ if and only if $H(z) = 0$. That is to say, z solves the LCP if and only if $H(z) = 0$. The following result from [66] gives conditions on the matrix M to insure that a zero curve of the homotopy map ρ_a can be tracked to obtain a zero of H.

THEOREM 3.1. *Let $M \in E^{n \times n}$ be either positive definite, a P-matrix, nondegenerate strictly copositive, or nondegenerate strictly semimonotone, and let $q \in E^n$ be nondegenerate with respect to M. Then there exists $\delta > 0$ such that for almost all $a \geqq 0$ with $\|a\|_\infty < \delta$ there is a zero curve γ of $\rho_a(\lambda, z)$, along which $D\rho_a(\lambda, z)$ has full rank, having finite arc length and connecting $(0, a)$ to $(1, \bar{z})$, where $\bar{z}$ is a zero of $H(z)$.*

Although it was not stated in [66], the proof of Theorem 3.1 there showed that if the nondegeneracy assumptions are removed, then the conclusion still holds, except that $(1, \bar{z})$ is only an accumulation point of the zero curve γ (of possibly unbounded variation). The map ρ_a above is the standard homotopy map. In the context of this paper we can view it as relaxing all of the solution requirements of the LCP while the zero curve is being tracked. Initially z is set to some arbitrary point a having nothing to do with the solution to (q, M). As λ gets closer to 1 we can say that, in some sense, z gets closer to such a solution. However, for any $\lambda < 1$, z and $w = Mz + q$ do not necessarily form a feasible solution or a complementary solution to (q, M). These conditions are imposed only at the end, when $\lambda = 1$, and then all at once. In the next few sections we present several homotopies, based on Mangasarian's function, that attempt to maintain at least feasibility or complementarity for a modified LCP right from the start. The hope is that the homotopy process is then more efficient.

4. Relaxation of M.

In this map, all of the continuation is applied to the matrix. We maintain a complementary feasible solution for some other matrix which is a convex combination of M and the identity. When $\lambda = 0$ the matrix is the identity, and when $\lambda = 1$ the matrix is M. We can view this map as relaxing only the matrix M as the zero curve is being tracked.

Define $\Lambda : [0, 1) \times E^n \to E^n$ by

$$\Lambda_i(\lambda, z) = -\left|[(1-\lambda)I + \lambda M]_{i\cdot} z + q_i - z_i\right|^3 + \left([(1-\lambda)I + \lambda M]_{i\cdot} z + q_i\right)^3 + z_i^3$$

for $i = 1, \cdots, n$.

Observe that since this is simply Mangasarian's map with a modified matrix for M, feasibility and complementarity are preserved wherever Λ is zero.

LEMMA 4.1. *Let P be any of the following properties:*

(a) *positive definite,*

(b) *P-matrix,*

(c) *nondegenerate strictly copositive,*

(d) *nondegenerate strictly semimonotone,*

and let $0 \leqq \lambda \leqq 1$. If a matrix $M \in E^{n \times n}$ has property P, then $(1-\lambda)I + \lambda M$ also has property P except possibly for finitely many values of λ.

Proof. (a) It follows from the definition of positive definite that

$$x^t[(1-\lambda)I + \lambda M]x = (1-\lambda)(x^t x) + \lambda(x^t M x) > 0$$

for all $x \neq 0$ whenever M is positive definite.

(b) It can be shown [13] that M is a P-matrix if and only if for all $x \neq 0$ there is an index k such that $x_k(Mx)_k > 0$. Let M be a P-matrix and let $x \neq 0$. Then

$$\begin{aligned} x_k\big([(1-\lambda)I + \lambda M]x\big)_k &= (1-\lambda)x_k(Ix)_k + \lambda x_k(Mx)_k \\ &= (1-\lambda)x_k^2 + \lambda x_k(Mx)_k \\ &> 0 \end{aligned}$$

for some index k.

Let M be nondegenerate and let $K \subset \{1,\cdots,n\}$. Because the determinant is multilinear we have

$$\det\big((1-\lambda)I + \lambda M\big)_{KK} = \sum_{J \subset K} (1-\lambda)^{|K|-|J|}\lambda^{|J|} \det M_{JJ},$$

which is simply a polynomial in λ. Notice that, since $\{(1-\lambda)^{k-j}\lambda^j \mid 0 \leqq j \leqq k\}$ forms a linearly independent set of polynomials, and $\det M_{JJ} \neq 0$ for any subset $J \subset \{1,\cdots,n\}$, this polynomial is not identically zero. (By convention, $\det M_{\phi\phi} = 1$.) This polynomial has only a finite number of zeros and so $(1-\lambda)I + \lambda M$ is nondegenerate except for finitely many values of λ.

(c) It follows from the definition of strictly copositive that

$$x^t[(1-\lambda)I + \lambda M]x = (1-\lambda)(x^t x) + \lambda(x^t Mx) > 0$$

for all $x \geq 0$ whenever M is strictly copositive.

(d) An argument similar to that for (b) holds if M is strictly semimonotone and $x \geq 0$. □

Lemma 4.1, Theorem 3.1, and the subsequent remark give us the following theorem.

THEOREM 4.1. *Let $M \in E^{n\times n}$ be positive definite or a P-matrix, and let $q \in E^n$. Then there exists a zero curve γ of Λ emanating from $(0, q-)$ and reaching a point $(1, \bar{z})$, where $\bar{z}$ solves the* LCP (q, M).

Note that Theorem 4.1 does not include strictly copositive or strictly semimonotone matrices, nor any reference to the rank of the Jacobian matrix along the zero curve γ. If M is nondegenerate strictly copositive or nondegenerate strictly semimonotone, there is a solution to the LCP $\big(q, [(1-\lambda)I + \lambda M]\big)$ for every $\lambda \in [0,1]$ by Theorem 3.1. However, there may be multiple solutions, and when the number of solutions changes at some $\bar{\lambda}$ some of the zero curves of Λ either "stop" or "start" at $\bar{\lambda}$. Thus there is no guarantee that a *single* zero curve of Λ will reach all the way from $\lambda = 0$ to $\lambda = 1$. For example, take

$$M = \begin{pmatrix} 3 & 2 & 10 \\ 0 & 1 & 10 \\ 1 & 0 & 1 \end{pmatrix}, \qquad q = \begin{pmatrix} -5.3 \\ -4.0 \\ -0.9 \end{pmatrix}.$$

M is nondegenerate strictly semimonotone, but the zero curve emanating from $(0, q-)$ disappears at $\lambda = 0.8$. Also we cannot say that the Jacobian matrix $D\Lambda(\lambda, z)$ is nonsingular along the entire zero curve γ. The ith row of the Jacobian matrix of Λ is

$$\begin{aligned}\big(D\Lambda(\lambda,z)\big)_{i\cdot} = \big(&-3|A|(A)(-I+M)_{i\cdot}z + 3(B)^2(-I+M)_{i\cdot}z,\\ &-3|A|(A)(\lambda m_{i1}) + 3(B)^2(\lambda m_{i1}), \cdots,\\ &-3|A|(A)(-\lambda+\lambda m_{ii}) + 3(B)^2(1-\lambda+\lambda m_{ii}) + 3z_i^2, \cdots,\\ &-3|A|(A)(\lambda m_{in}) + 3(B)^2(\lambda m_{in})\big),\end{aligned}$$

$$\begin{aligned}\text{where } A &= [(1-\lambda)I + \lambda M]_{i\cdot}z + q_i - z_i,\\ B &= [(1-\lambda)I + \lambda M]_{i\cdot}z + q_i.\end{aligned}$$

Observe that if $|q| > 0$, then rank $D\Lambda(0, q-) = n$, and so the starting point $z = q-$ for the zero curve is nonsingular.

PROPOSITION 4.2. *Let $M \in E^{n\times n}$ be positive definite or a P-matrix, and let $q \in E^n$. Whenever M and q are such that $(\bar{w}, \bar{z})$, the solution to (q, M), has $S(\bar{z}) \neq S(q-)$, the Jacobian matrix of Λ has singularities along the zero curve γ of Λ. There is at least one singularity for each element in the disjoint union*

$$\big(S(\bar{z}) \cup S(q-)\big) \setminus \big(S(\bar{z}) \cap S(q-)\big).$$

Proof. Let $(\bar{w}, \bar{z})$ be the solution to (q, M) and let $i \in S(\bar{z}) \setminus S(q-)$. First note that, on the (unique) zero curve γ of Λ, both z and w are continuous functions of λ. Since $z = q-$ when $\lambda = 0$, there must be a point λ_0 such that, along the zero curve, $z_i = 0$ for $0 \leqq \lambda \leqq \lambda_0$ and $z_i > 0$ for $\lambda_0 < \lambda < \lambda_0 + \epsilon$ for some ϵ. Since complementarity is maintained along the zero curve, $w_i = 0$ for $\lambda_0 < \lambda < \lambda_0 + \epsilon$. By continuity, w_i must be 0 at $\lambda = \lambda_0$. This means that both z_i and $w_i = [(1-\lambda)I + \lambda M]_{i\cdot}z + q_i$ are zero at $\lambda = \lambda_0$, and hence the Jacobian matrix $D\Lambda(\lambda_0, z(\lambda_0))$ is singular.

Similarly, let $i \in S(q-) \setminus S(\bar{z})$. There must be a point λ_1 such that, along the zero curve γ, $z_i > 0$ for $0 \leqq \lambda < \lambda_1$ and $z_i = 0$ at $\lambda = \lambda_1$. Again by complementarity and continuity, w_i must be 0 at $\lambda = \lambda_1$ and the Jacobian matrix $D\Lambda(\lambda_1, z(\lambda_1))$ is singular. □

5. Relaxation of q. We can also relax the right-hand side of the LCP keeping the matrix M fixed. This map maintains feasibility and complementarity, but uses a convex combination of the vectors q and $\|q\|_\infty e$ for the right-hand side of the equation. When $\lambda = 0$, we have the trivial problem $(\|q\|_\infty e, M)$ where the right-hand side has all components positive and, when $\lambda = 1$, we have the given problem (q, M).

Define $\Theta : [0,1) \times E^n \to E^n$ by

$$\Theta_i(\lambda, z) = -\big|M_{i\cdot}z + \lambda q_i + (1-\lambda)\|q\|_\infty - z_i\big|^3 + \big(M_{i\cdot}z + \lambda q_i + (1-\lambda)\|q\|_\infty\big)^3 + z_i^3$$

for $i = 1, \cdots, n$.

Since this is once again Mangasarian's map with a slightly different vector for q, feasibility and complementarity on the zero set of Θ is guaranteed. By Theorem 3.1

and the remark following it, we know that the LCP has a locally unique solution for any q whenever M is nondegenerate strictly semimonotone. Thus we easily have the following theorem about Θ.

THEOREM 5.1. *Let $M \in E^{n\times n}$ be positive definite or a P-matrix, and let $q \in E^n$. Then there exists a zero curve γ of Θ emanating from $(0,0)$ and reaching a point $(1,\bar{z})$, where $\bar{z}$ solves the* LCP (q,M).

Note that Theorem 5.1 does not include strictly copositive or strictly semimonotone matrices, nor any reference to the rank of the Jacobian matrix along the zero curve γ. If M is nondegenerate strictly copositive or nondegenerate strictly semimonotone, there is a solution to the LCP $(\lambda q + (1-\lambda)\|q\|_\infty e, M)$ for every $\lambda \in [0,1]$ by Theorem 3.1. However, there may be multiple solutions, and when the number of solutions changes at some $\bar{\lambda}$ some of the zero curves of Θ either "stop" or "start" at $\bar{\lambda}$. Thus there is no guarantee that a *single* zero curve of Θ will reach all the way from $\lambda = 0$ to $\lambda = 1$. For example, take

$$M = \begin{pmatrix} 1 & 5 & 10 \\ 5 & 1 & 10 \\ 1 & 1 & 1 \end{pmatrix}, \qquad q = \begin{pmatrix} -2 \\ -2 \\ -1 \end{pmatrix}.$$

M is nondegenerate strictly semimonotone, but the zero curve emanating from $(0,0)$ disappears at $\lambda = 4/5$.

Furthermore, we cannot say that the Jacobian matrix is nonsingular along the entire zero curve. The ith row of the Jacobian matrix of Θ is

$$\begin{aligned}\big(D\Theta(\lambda,z)\big)_{i\cdot} = \big(&-3|A|(A)(q_i - \|q\|_\infty) + 3(B)^2(q_i - \|q\|_\infty),\\ &-3|A|(A)(m_{i1}) + 3(B)^2 m_{i1}, \cdots,\\ &-3|A|(A)(m_{ii}-1) + 3(B)^2 m_{ii} + 3z_i^2, \cdots,\\ &-3|A|(A)(m_{in}) + 3(B)^2 m_{in}\big),\end{aligned}$$

$$\begin{aligned}\text{where } A &= M_{i\cdot}z + \lambda q_i + (1-\lambda)\|q\|_\infty - z_i,\\ B &= M_{i\cdot}z + \lambda q_i + (1-\lambda)\|q\|_\infty.\end{aligned}$$

Note that the first column and the diagonal element differ slightly in form from the rest of the entries. Also note that if z_i and $w_i = M_{i\cdot}z + \lambda q_i + (1-\lambda)\|q\|_\infty$ are both zero for some λ, then every entry in $(D\Theta)_{i\cdot}$ is 0. Hence, the Jacobian matrix is singular and we have the following proposition.

PROPOSITION 5.2. *Let $M \in E^{n\times n}$ be positive definite or a P-matrix, and let $q \in E^n$. Whenever M and q are such that $(\bar{w},\bar{z})$, the solution to (q,M), has $\bar{z} \neq 0$, the Jacobian matrix of Θ has singularities along the zero curve γ of Θ. There are at least as many singularities as there are nonzero components of $\bar{z}$.*

Proof. Let $(\bar{w},\bar{z})$ be the solution to (q,M) and let i be such that $\bar{z}_i > 0$. First note that, on the (unique) zero curve γ of Θ, both z and w are continuous functions of λ. Since $z = 0$ when $\lambda = 0$, there must be a point λ_0 such that, along the zero curve, $z_i = 0$ for $0 \leqq \lambda \leqq \lambda_0$ and $z_i > 0$ for $\lambda_0 < \lambda < \lambda_0 + \epsilon$ for some ϵ. Since complementarity is maintained along the zero curve, $w_i = 0$ for $\lambda_0 < \lambda < \lambda_0 + \epsilon$. By continuity, w_i

must be 0 at $\lambda = \lambda_0$. This means that both z_i and $w_i = M_{i\cdot}z + \lambda q_i + (1-\lambda)\|q\|_\infty$ are zero at $\lambda = \lambda_0$, and hence the Jacobian matrix $D\Theta(\lambda_0, z(\lambda_0))$ is singular. □

Geometrically, the singularity corresponds to the point at which the vector $\lambda q + (1-\lambda)\|q\|_\infty$ passes through the boundary of one complementary cone [44], [48], [56], [62] and into another. If it happens that this vector stays in such a boundary for all λ in some interval $[\lambda_0, \lambda_1]$, then z_i and w_i are simultaneously 0, and the Jacobian matrix is singular, along that entire interval. Since there are a finite number (2^n) of complementary cones, however, we can always perturb the right-hand side by adding some $(\epsilon, \epsilon^2, \cdots, \epsilon^n)$, for example, so that there are only a finite number of singularities.

6. Relaxation of complementarity. This section presents a map that uses the given matrix M and the given vector q, but does not maintain a complementary solution as we track the zero curve. Although nonnegativity of z is preserved along the curve, complementarity is enforced only at the very end of the curve, when $\lambda = 1$. Throughout this section, let $M \in E^{n\times n}$ and $q \in E^n$ be fixed.

Define $\Psi : E^n \times [0,1) \times E^n \to E^n$ by

$$\Psi_i(a,\lambda,z) = -\lambda|M_{i\cdot}z + q_i - z_i|^3 + \lambda(M_{i\cdot}z + q_i)^3 + {z_i}^3 - (1-\lambda){a_i}^3$$

for $i = 1, \cdots, n$. For fixed $a \in E^n$ let $\Psi_a(\lambda, z) = \Psi(a, \lambda, z)$. The next few lemmas show that, for suitable matrices M, there is a zero curve of Ψ that can be tracked to obtain a solution to the LCP (q, M).

LEMMA 6.1. *If $a \geqq 0$, then $z \geqq 0$ on $\Psi_a^{-1}(0)$.*

Proof. Note that if both z_k and $M_{k\cdot}z + q_k$ are negative, then the entire sum comprising $(\Psi_a(\lambda, z))_k$ is negative. If, on the other hand, $z_k < 0$ and $M_{k\cdot} + q_k \geqq 0$, then $|M_{k\cdot} + q_k| < M_{k\cdot} + q_k - z_k$, and the sum is again negative. □

LEMMA 6.2. *Let M be strictly semimonotone. Then there exists $r > 0$ such that $z \in E^n$, $z \geqq 0$, and $\|z\|_\infty = r$ implies that $z_k(Mz+q)_k > 0$ for some index k.*

Proof. First let

$$\Phi(z) = \max_{1\leqq i\leqq n} z_i(Mz)_i$$

and note that, because M is strictly semimonotone, $\Phi > 0$ for $z \geq 0$. Also note that since Φ is continuous and $\{z : z \geqq 0, \|z\|_\infty = 1\}$ is compact, Φ must assume its minimum on that set. Call that minimum $\bar{\Phi}$ and take $r > \|q\|_\infty/\bar{\Phi}$. Then for $z \geqq 0$ and $\|z\|_\infty = r$, there is some index k such that

$$\begin{aligned} z_k(Mz+q)_k &= \|z\|_\infty^2 \Phi(z/\|z\|_\infty) + z_k q_k \\ &\geqq \|z\|_\infty^2 \bar{\Phi} - \|z\|_\infty \|q\|_\infty \\ &= \|z\|_\infty(\|z\|_\infty \bar{\Phi} - \|q\|_\infty) \\ &> 0. \end{aligned}$$

□

LEMMA 6.3. *Let M be strictly semimonotone. Then there exists $r > 0$ such that $\Psi_0(\lambda, z) \neq 0$ for $0 \leqq \lambda \leqq 1$ and $\|z\|_\infty = r$.*

Proof. By Lemma 6.1, it suffices to consider $z \geqq 0$. Let r and k be as in the conclusion of Lemma 6.2 above and simply notice that, since z_k and $(Mz+q)_k$ are both positive, $\Psi_0(\lambda, z)$ cannot be 0. □

LEMMA 6.4. *Let M be strictly semimonotone. Then there exists $r > 0$ and $\delta > 0$ such that $0 \leqq \lambda \leqq 1$, $\|z\|_\infty = r$, and $\|a\|_\infty < \delta$ implies $\Psi_a(\lambda, z) \neq 0$.*

Proof. Let r be as in Lemma 6.3, and note that $\{(a, \lambda, z) \mid a = 0,\ 0 \leqq \lambda \leqq 1, \|z\|_\infty = r\}$ is disjoint from $\Psi^{-1}(0)$. Since the first of these sets is compact and the second is closed there is a positive distance $\delta > 0$ between them, measured in the max norm. This δ satisfies the conclusion of the Lemma. □

Notice that a positive definite matrix is also a P-matrix, a P-matrix is strictly semimonotone by the sign-reversal property of P-matrices [13], and a strictly copositive matrix is clearly strictly semimonotone. Hence, Lemmas 6.1–6.4 hold for any such matrix and we can state the following theorem.

THEOREM 6.5. *Let $M \in E^{n\times n}$ be positive definite, a P-matrix, strictly copositive, or strictly semimonotone, and let $q \in E^n$. Then there exists $\delta > 0$ such that for almost all $a > 0$, $\|a\|_\infty < \delta$ there is a zero curve γ of $\Psi_a(\lambda, z)$, along which the Jacobian matrix $D\Psi_a(\lambda, z)$ has full rank, emanating from $(0, a)$ and reaching a point $(1, \bar{z})$, where $\bar{z}$ solves the* LCP (q, M).

Proof. First observe that, for $a > 0$, Ψ is transversal to 0 (i.e., its Jacobian matrix has full rank on $\Psi^{-1}(0)$). To see this, note that $\partial\Psi_i/\partial a_j$ is zero if $i \neq j$, and nonzero if $i = j$. Thus, the n columns of $D\Psi$ corresponding to the partials of Ψ with respect to the a_i are linearly independent. Clearly, Ψ_a is C^2, and therefore by Lemma 2.1, for almost all $a > 0$, Ψ_a is also transversal to 0. Thus, by the implicit function theorem, Ψ_a has a zero curve γ, starting from $(0, a)$, along which the Jacobian matrix $D\Psi_a(\lambda, z)$ has full rank. All of this is true regardless of the conditions on the matrix M.

For M strictly semimonotone (positive definite, strictly copositive, or a P-matrix), Lemma 6.4 insures that there exists $\delta > 0$ such that the zero curve γ is bounded for $\|a\|_\infty < \delta$ and $0 \leqq \lambda \leqq 1$. Note that $(0, a)$ is the unique zero of Ψ_a at $\lambda = 0$, and by the implicit function theorem, γ cannot return to $(0, a)$. Since the curve can neither simply stop, nor return to $\lambda = 0$, nor go to infinity, it must reach a point $(1, \bar{z})$, where $\bar{z}$ solves the LCP (q, M). □

7. Expanded Lagrangian homotopy.

The expanded Lagrangian approach [54] may be described as an optimization/continuation approach and has in its simplest form two main steps.

Step 1. (Optimization phase).

At $r = r_0 > 0$ solve the unconstrained minimization problem

$$\min_{w,z} P(w, z, r),$$

where

$$P(w, z, r) = \frac{1}{2r}\|w - Mz - q\|_2^2 + \frac{1}{2r}\langle w, z\rangle^2 - r\sum_{i=1}^{n} \ln z_i - r\sum_{i=1}^{n} \ln w_i.$$

Step 2A. (Switch to expanded system).

A (local) solution of $\min P$ must satisfy

$$0 = \nabla_{(w,z)} P = \begin{pmatrix} I \\ -M^t \end{pmatrix} \frac{(w - Mz - q)}{r} + \begin{pmatrix} z \\ w \end{pmatrix} \frac{\langle w, z\rangle}{r} - r\left(\frac{1}{w_1}, \dots, \frac{1}{w_n}, \frac{1}{z_1}, \dots, \frac{1}{z_n}\right)^t.$$

Introduce the following variables:

$$\begin{aligned} \beta &= \frac{w - Mz - q}{r}, \\ \theta &= \frac{\langle w, z\rangle}{r}, \\ \mu_i &= \frac{r}{w_i}, \qquad i = 1, \cdots, n, \\ \eta_i &= \frac{r}{z_i}, \qquad i = 1, \cdots, n, \end{aligned}$$

which ultimately represent the Lagrange multipliers. This helps to remove the inevitable ill conditioning associated with penalty methods for small r and we thus obtain our equivalent but expanded system:

$$\begin{aligned} \begin{pmatrix} I \\ -M^t \end{pmatrix} \beta + \begin{pmatrix} z \\ w \end{pmatrix} \theta - \begin{pmatrix} \mu \\ \eta \end{pmatrix} &= 0, \\ w - Mz - q - r\beta &= 0, \\ \langle w, z\rangle - r\theta &= 0, \\ \mu_i w_i - r &= 0, \qquad i = 1, \cdots, n, \\ \eta_i z_i - r &= 0, \qquad i = 1, \cdots, n. \end{aligned}$$

(Remark. As a result of the optimization phase and the initial starting point with $r_0 > 0$, the solution $(w^{(0)}, z^{(0)})$ of $\min P(w, z, r_0)$ satisfies $z^{(0)} > 0$ and $w^{(0)} > 0$. As a consequence, $\mu^{(0)} > 0$ and $\eta^{(0)} > 0$ from the definitions of μ and η. They remain positive until $r = 0$, where we formally have

$$\begin{aligned} \begin{pmatrix} I \\ -M^t \end{pmatrix} \beta + \begin{pmatrix} z \\ w \end{pmatrix} \theta - \begin{pmatrix} \mu \\ \eta \end{pmatrix} &= 0, \\ w - Mz - q &= 0, \\ \langle w, z\rangle &= 0, \\ \mu_i w_i &= 0, \qquad i = 1, \cdots, n, \\ \eta_i z_i &= 0, \qquad i = 1, \cdots, n, \\ w, z, \theta, \mu, \eta &\geqq 0, \end{aligned}$$

which implies that we have solved the problem.)

In practice we do not solve the optimization problem $\min P$ to high accuracy since a highly accurate solution may have only a digit or two in common with the final answer. However, it is imperative that ∇P be reasonably small in magnitude, say, less than $r_0/10$. The expanded system is converted to a homotopy map by letting $r = r_0(1 - \lambda)$ and modifying the first equation to obtain:

$$\begin{aligned} \begin{pmatrix} I \\ -M^t \end{pmatrix} \beta + \begin{pmatrix} z \\ w \end{pmatrix} \theta - \begin{pmatrix} \mu \\ \eta \end{pmatrix} - \frac{r}{r_0} \nabla P(w^{(0)}, z^{(0)}, r_0) &= 0, \\ w - Mz - q - r\beta &= 0, \\ \langle w, z\rangle - r\theta &= 0, \\ \mu_i w_i - r &= 0, \qquad i = 1, \cdots, n, \\ \eta_i z_i - r &= 0, \qquad i = 1, \cdots, n. \end{aligned}$$

Write this system of $5n+1$ equations in the $5n+2$ variables λ, w, z, β, θ, μ, η as

$$\Upsilon(\lambda, w, z, \beta, \theta, \mu, \eta) = 0.$$

Step 2B. (Track the zero curve of Υ from $r = r_0$ to $r = 0$.)

Starting with arbitrary $r_0 > 0$, $w^{(0)} > 0$, and $z^{(0)} > 0$, the rest of the initial point $(0, w^{(0)}, z^{(0)}, \beta^{(0)}, \theta_0, \mu^{(0)}, \eta^{(0)})$ is given by

$$\begin{aligned}
\beta^{(0)} &= \frac{w^{(0)} - Mz^{(0)} - q}{r_0}, \\
\theta_0 &= \frac{\langle w^{(0)}, z^{(0)} \rangle}{r_0}, \\
\mu_i^{(0)} &= \frac{r_0}{w_i^{(0)}}, \qquad i = 1, \cdots, n, \\
\eta_i^{(0)} &= \frac{r_0}{z_i^{(0)}}, \qquad i = 1, \cdots, n.
\end{aligned}$$

This approach requires careful attention to implementation details. For example, the linear algebra and globalization techniques with dynamic scaling are critically important in the optimization phase. For degenerate problems the path can still be long. One possible resolution is the use of shifts and weights as developed in the method of multipliers [5], but holding $r = r_0$ fixed. (This approach is currently under investigation in the context of linear programming [53].) However, in keeping with the philosophy of the "pure" homotopy approach of the current work, we do not solve the optimization problem (Step 1), but instead use the above equations $\Upsilon(\lambda, w, z, \beta, \theta, \mu, \eta) = 0$ as a "pure" homotopy.

Logarithmic barrier potential functions are hardly new [5], and have been used recently by Kojima et al. [26], [27] and Mizuno et al. [38] to extend the ideas of Karmarkar to obtain polynomial-time algorithms for the LCP. The exact details of how the barrier parameter, step size selection, concomitant numerical linear algebra, and initial point computation are handled are crucial to the practical utility of such methods, and in practice are far more significant than theoretical polynomial complexity. It is reasonable that the pure expanded Lagrangian homotopy (without the optimization step) would behave significantly differently from other logarithmic barrier homotopies [26], [27], [38], which include a Phase 1 step equivalent to Step 1 here. These latter homotopies of Kojima et al. are certainly *not* globally convergent, since they require a nontrivial preliminary computation to get a special starting point at which to begin the homotopy.

8. Absolute Newton method. The method of this section is not a homotopy method, but is presented for the sake of comparison and as an example of what can be done with a Newton-type iterative scheme (see also [1] and [35]). Let $x = (w, z) \in E^{2n}$ and define $F : E^{2n} \to E^{2n}$ by

$$F(x) = \begin{pmatrix} w - Mz - q \\ w_1 z_1 \\ \vdots \\ w_n z_n \end{pmatrix}.$$

Then the LCP (q, M) is equivalent to $F(x) = 0$ for x nonnegative. $F(x) = 0$ is a polynomial system of equations of total degree 2^n, which in general has 2^n solutions over complex Euclidean space C^{2n}, counting multiplicities and solutions at infinity. Thus *all* solutions of the LCP (q, M) are among the zeros of $F(x)$, including degenerate solutions, which correspond to manifolds (in C^{2n}) of zeros of $F(x)$. The algebraic geometry theory of polynomial systems is rich and deep, and beyond the scope of this paper. Discussions of the pertinent aspects of algebraic and differential geometry for polynomial systems are in [39], [40], [41], and [68]. It suffices to note here that $F(x)$ is a polynomial system with a particularly simple structure.

The Jacobian matrix of F is

$$DF(x) = \begin{pmatrix} I & -M \\ \operatorname{diag}(z_1, \cdots, z_n) & \operatorname{diag}(w_1, \cdots, w_n) \end{pmatrix},$$

a $2n \times 2n$ matrix. The absolute Newton iteration is

$$x^{(k+1)} = \left| x^{(k)} - \left[DF(x^{(k)})\right]^{-1} F(x^{(k)}) \right|, \quad k = 0, 1, 2, \cdots$$

for an arbitrary starting point $x^{(0)} \in E^{2n}$. The absolute value signs mean to replace each component of the vector by its absolute value (precisely, $|x| = x+ + x-$). When this iteration is well defined is given by the following theorem:

THEOREM 8.1. *Let $M \in E^{n \times n}$ be nondegenerate and let $\bar{x} = (\bar{w}, \bar{z})$ be a zero of F. Then the Jacobian matrix $DF(\bar{x})$ is invertible if and only if $|\bar{w}| + |\bar{z}| > 0$.*

Proof. Suppose that $\bar{w}_k = \bar{z}_k = 0$. Then the $(n + k)$th row of $DF(\bar{x})$ is zero, so $DF(\bar{x})$ is not invertible.

Conversely, suppose that $|\bar{w}| + |\bar{z}| > 0$. Observe that $\bar{w}$ and $\bar{z}$ are complementary vectors, since $\bar{x} = (\bar{w}, \bar{z})$ is a zero of F. For each index k such that $\bar{z}_k \neq 0$ interchange the kth and $(n + k)$th columns of $DF(\bar{x})$. This produces a matrix of the form

$$\begin{pmatrix} A & * \\ 0 & \operatorname{diag}(\bar{w}_1 + \bar{z}_1, \cdots, \bar{w}_n + \bar{z}_n) \end{pmatrix},$$

where $A_{\cdot i} \in \{I_{\cdot i}, -M_{\cdot i}\}$ for $i = 1, \cdots, n$. $\det A$ is a principal minor of $-M$ and is thus nonzero, since M is nondegenerate by assumption. Further, since $|\bar{w}| + |\bar{z}| > 0$ and $\bar{w}$, $\bar{z}$ are complementary, $\bar{w}_i + \bar{z}_i \neq 0$ for $i = 1, \cdots, n$. Thus

$$\begin{aligned} \det DF(\bar{x}) &= \pm \det A \ \det \ \operatorname{diag}(\bar{w}_1 + \bar{z}_1, \cdots, \bar{w}_n + \bar{z}_n) \\ &= \pm \det A \prod_{i=1}^{n} (\bar{w}_i + \bar{z}_i) \\ &\neq 0, \end{aligned}$$

and $DF(\bar{x})$ is invertible. □

This absolute Newton iteration has been used for chemical equilibrium systems, which have a unique real positive solution. It has never been observed to fail for those systems with a random starting point $x^{(0)}$ [36]. The asymptotic behavior of this absolute Newton iteration is not understood, nor even the ordinary Newton iteration in complex Euclidean space C^{2n}, which is related to Julia sets and chaotic dynamical systems. Both the standard Newton iteration and the absolute Newton iteration were tried on $F(x) = 0$, where M was a P-matrix, and both completely failed for starting points distant from the solution. Why the absolute Newton method should be so successful on chemical equilibrium polynomial systems, and fail on LCP polynomial systems, is not clear.

9. Kojima–Saigal homotopy. This homotopy [25] uses the same nonlinear system as the absolute Newton method. Suppose that $w^{(0)}, z^{(0)} \in E^n$ have been obtained such that

$$w^{(0)} - Mz^{(0)} = q,$$
$$w^{(0)} > 0, \quad z^{(0)} > 0.$$

This can be done, for example, by applying Phase 1 of the simplex algorithm to the problem

$$w - Mz = q - e + Me,$$
$$w \geqq 0, \quad z \geqq 0$$

to get a feasible solution $(\bar{w}, \bar{z}) \geqq 0$. Then $w^{(0)} = \bar{w} + e > 0$ and $z^{(0)} = \bar{z} + e > 0$ will suffice. The homotopy map $K : [0,1) \times E^n \times E^n \to E^n$ is given by

$$K(\lambda, w, z) = \begin{pmatrix} w - Mz - q \\ w_1 z_1 - (1-\lambda) w_1^{(0)} z_1^{(0)} \\ \vdots \\ w_n z_n - (1-\lambda) w_n^{(0)} z_n^{(0)} \end{pmatrix}.$$

The following theorem shows that this is a reasonably good homotopy map, at least for P-matrices.

THEOREM 9.1. *Let $M \in E^{n \times n}$ be a P-matrix and let $q \in E^n$. Then there exist $w^{(0)}, z^{(0)} \in E^n$ such that*

$$w^{(0)} - M z^{(0)} = q, \quad w^{(0)} > 0, \quad z^{(0)} > 0.$$

Furthermore, there is a zero curve γ of $K(\lambda, w, z)$, along which the Jacobian matrix $DK(\lambda, w, z)$ has full rank (for $0 \leqq \lambda < 1$), emanating from $(0, w^{(0)}, z^{(0)})$ and reaching a point $(1, \bar{w}, \bar{z})$, where $\bar{z}$ solves the LCP (q, M). *λ is strictly increasing as a function of arc length s along γ $(d\lambda/ds > 0)$.*

Proof. Since M is a P-matrix, the LCP $(q - e + Me, M)$ has a solution $(\hat{w}, \hat{z})$ by Theorem 6.5. Then $w^{(0)} = \hat{w} + e > 0$ and $z^{(0)} = \hat{z} + e > 0$ have the desired properties.

The Jacobian matrix of $K(\lambda, w, z)$ is

$$DK(\lambda, w, z) = \begin{pmatrix} 0 & I & -M \\ w_1^{(0)} z_1^{(0)} & & \\ \vdots & \operatorname{diag}(z_1, \cdots, z_n) & \operatorname{diag}(w_1, \cdots, w_n) \\ w_n^{(0)} z_n^{(0)} & & \end{pmatrix}.$$

Suppose $(w, z) > 0$ and consider the last $2n$ columns $D_{(w,z)}K$ of DK:

$$\begin{aligned} \det D_{(w,z)}K &= \det \begin{pmatrix} I & -M \\ \operatorname{diag}(z_1, \cdots, z_n) & \operatorname{diag}(w_1, \cdots, w_n) \end{pmatrix} \\ &= \det \begin{pmatrix} I & -M \\ 0 & \operatorname{diag}(w_1, \cdots, w_n) + \operatorname{diag}(z_1, \cdots, z_n) M \end{pmatrix} \\ &= \det\bigl(\operatorname{diag}(w_1, \cdots, w_n) + \operatorname{diag}(z_1, \cdots, z_n) M\bigr) \\ &> 0 \end{aligned}$$

since $\operatorname{diag}(w_1, \cdots, w_n) + \operatorname{diag}(z_1, \cdots, z_n) M$ is also a P-matrix (it is easily verified that the principal minors remain positive after multiplying by and adding a positive diagonal matrix). Thus rank $DK(\lambda, w, z) = 2n$ for $0 \leqq \lambda < 1$ and $w > 0$, $z > 0$. By the Implicit Function Theorem, there is a zero curve γ of K emanating from $(0, w^{(0)}, z^{(0)})$, and the Jacobian matrix $DK(\lambda, w, z)$ has full rank along γ for $0 \leqq \lambda < 1$ since $w > 0$, $z > 0$ along γ by continuity and the definition of K.

γ can be parametrized by arc length s, giving $\lambda = \lambda(s)$, $w = w(s)$, $z = z(s)$ along γ. Furthermore, the last $2n$ columns of $DK(\lambda(s), w(s), z(s))$ being independent means that $w = w(\lambda)$, $z = z(\lambda)$, and $d\lambda/ds > 0$ along γ (this is well known, see [65], for example). Thus $\lambda = \lambda(s)$ is strictly increasing along γ.

To prove that γ reaches $\lambda = 1$, it suffices to prove that γ is bounded. Let $\alpha = \max_A\{\|A\|_\infty, \|A^{-1}\|_\infty\}$, where the maximum is taken over all matrices $A \in E^{n \times n}$ with $A_{\cdot i} \in \{I_{\cdot i}, -M_{\cdot i}\}$ for $i = 1, \cdots, n$. α is well defined since each $\det A$ is a principal minor of $-M$, which is nonzero by assumption. Fix λ_0 in $(0,1)$, and let $\epsilon = \max_i(1 - \lambda_0) w_i^{(0)} z_i^{(0)}$. Then for $\lambda_0 < \lambda(s) \leqq 1$, either $w_i(s) < \epsilon$ or $z_i(s) < \epsilon$ along γ. For $i = 1, \cdots, n$, let y_i be $w_i(s)$ or $z_i(s)$, whichever is less than ϵ, and let $\bar{y}_i$ be the complementary variable. Write $w(s) - M\, z(s) = q$ as

$$Ay + B\bar{y} = q.$$

Then

$$\|\bar{y}\|_\infty = \left\|B^{-1}(q - Ay)\right\|_\infty \leqq \left\|B^{-1}\right\|_\infty \left(\|q\|_\infty + \|A\|_\infty \|y\|_\infty\right) \leqq \alpha(\|q\|_\infty + \alpha\epsilon),$$

which says that $w(s)$ and $z(s)$ are bounded for $\lambda_0 < \lambda(s) \leqq 1$. □

Note that the theorem does not include strictly semimonotone matrices since $\operatorname{diag}(w_1, \cdots, w_n) + \operatorname{diag}(z_1, \cdots, z_n) M$ can be singular for strictly semimonotone M. Thus while K is a better homotopy than Λ, Θ, and Υ, it is not as generally applicable as ρ_a or Ψ_a.

10. Numerical experiments. The homotopy maps from the previous sections were tested on several problems, chosen to illustrate certain features of the various homotopies. A complete description of the data, tables of numerical results, and a comparative discussion of the different homotopy maps and numerical results are in [70]. The main observations from those experiments are summarized here: The probability-one homotopies ρ_a and Ψ_a work for everything that the theory predicts. The computational complexity of ρ_a and Ψ_a, measured by the number of steps along the zero curve, is relatively insensitive to n. This is in direct contrast to pivoting methods, which can exhibit exponential complexity in the number of steps [47]. The homotopies Λ and Θ frequently fail, but when they work at all, may be more efficient than the homotopies ρ_a or Ψ_a. The expanded Lagrangian homotopy Υ without the optimization phase fails for most starting points, with the zero curves of Υ either going off to infinity or returning to another solution at $r = r_0$. Υ does work very well from sufficiently close starting points, but these are not random starting points (as are used for ρ_a and Ψ_a), and the homotopy algorithm based on Υ without optimization is certainly not globally convergent. The Kojima–Saigal homotopy requires Phase 1 of the simplex algorithm just to get a starting point, which is antithetical to the homotopy philosophy of global convergence from an easily obtainable starting point. Furthermore, K and Ψ_a both essentially relax complementarity, and Ψ_a is more generally applicable.

11. Conclusion. There are many reasonable ways to construct a homotopy map for the LCP, and only a few of the possibilities have been considered here. The homotopies here fall into three different classes: *artificial*, *natural*, and *interior*. (See the discussion of the words "artificial" and "natural" in relation to homotopies in [69].) Λ and Θ are "natural" homotopies in the sense that for each $\lambda \in [0,1]$ the equation $\Lambda(\lambda, z) = 0$ or $\Theta(\lambda, z) = 0$ corresponds to an LCP. Thus, the intermediate points (λ, z) on the zero curve of the homotopy map have interpretations as solutions to a related family of LCPs. In contrast, ρ_a and Ψ_a are "artificial" homotopies in that the homotopy equations $\rho_a(\lambda, z) = 0$ and $\Psi_a(\lambda, z) = 0$ do *not* correspond to an LCP for $0 < \lambda < 1$, and the points (λ, z) on the zero curves for $0 < \lambda < 1$ have no useful interpretations as LCP solutions. Υ and K would be considered "interior" methods, since they only generate points (λ, w, z) interior to the feasible region, i.e., $(w, z) > 0$ for $0 \leqq \lambda < 1$. These class distinctions are not always clear-cut, but are useful at a high conceptual level.

The theory of globally convergent probability-one homotopy maps can be applied to the LCP in several ways; the maps ρ_a and Ψ_a are two examples. The convergence theory for the homotopy maps ρ_a and Ψ_a is very satisfactory: global convergence from an arbitrary starting point is guaranteed for a wide class of LCPs. Theorems 3.1 and 6.5 are existence results, and as such are close to the best known existence results.

Our computational experience, reported in [70], indicates that Ψ_a is the best homotopy. It never failed, is indeed globally convergent, and was frequently more efficient than Λ and Θ, even on problems where Λ and Θ did well. ρ_a takes second place, since it also never failed, but tends to be very expensive (long homotopy zero curves). This is not surprising, since Ψ_a was crafted with the benefit of ten years experience since ρ_a was created. It is quite likely that a more efficient globally convergent homotopy map than Ψ_a can yet be constructed.

Λ and Θ failed badly on problems with many singularities (corresponding to the right-hand side passing through the face of a complementary cone) along the zero curves of the homotopy maps Λ and Θ. One might hope that the curve tracking algorithms would, by chance, miss hitting the singularities exactly and thereby step past them. This does happen, to some extent, but when there are a large number of singularities close together or highly rank deficient singularities (corresponding to the right-hand side passing through a lower dimensional face of a complementary cone), the numerical linear algebra is simply overwhelmed by the ill conditioning.

Overall, the natural homotopies Λ and Θ are much worse than the artificial homotopies ρ_a and Ψ_a. For particular problems, a natural homotopy may be very efficient, but their performance is unreliable and very much data dependent. The difficulties, both theoretical (cf. Propositions 4.2 and 5.2) and practical, of natural homotopies like Λ and Θ appear to remove them from further consideration (cf. the discussions in [39]–[41] and [69]).

The numerical experiments show that the expanded Lagrangian homotopy is unacceptable as a robust homotopy without solving the optimization problem (Step 1). The zero set of Υ contains loops (in $[0, 1) \times E^{5n+1}$) starting and ending at $\lambda = 0$ as well as unbounded curves. Although the increased dimension is discouraging, we do note that $2n$ of the $5n+1$ equations result in diagonal matrices which can be exploited in the linear algebra. Furthermore, Υ does work well for fair starting points, and so

Υ may be useful for LCPs using an optimization phase to get a fairly good starting point. Although the expanded Lagrangian homotopy is an interior method based on a logarithmic barrier potential function similar in spirit to methods of Kojima et al. [26], [27], [38], it is not equivalent to any of those methods. The Kojima et al. methods converge to a solution in polynomial time from an arbitrary interior starting point (for a restricted class of LCPs), which is not true of the expanded Lagrangian homotopy method. However, generating a feasible interior starting point for K is tantamount to the optimization Step 1 for Υ, and neither K nor Υ can be considered a globally convergent homotopy for the LCP in the same sense as ρ_a and Ψ_a. Furthermore, the Kojima et al. homotopies without Phase 1 would be even less successful than the expanded Lagrangian homotopy is without Step 1.

The Kojima–Saigal homotopy is closely related to the continuous Newton homotopy of Smale. Both are theoretically interesting, but computational experience on real problems [67], [68] suggests that the globally convergent probability-one homotopies (like ρ_a and Ψ_a) are more robust and more general than the continuous Newton homotopies. Our numerical experience is that interior homotopies like Υ and K (lacking dynamic scaling) are very inefficient, but worthy of further study. At any rate, Ψ_a is more general than K (cf. Theorems 6.5 and 9.1). Similar comments apply to the polynomial-time homotopies of [26], [27], and [38], which are both less stable numerically and less generally applicable than probability-one homotopies like Ψ_a.

There are numerous fixed point iterative schemes for the LCP [2], [3], [8], [18], [35], [50], [51], [61], but they generally involve nonsmooth operators (e.g., $v+$ or $|v|$) or apply to a small class of matrices (e.g., symmetric positive definite M). Homotopy algorithms are more versatile than fixed point iteration algorithms, but whether they are competitive with fixed point iteration remains to be seen. A systematic comparison of complementary pivoting, fixed point iteration, and homotopy methods would be a worthwhile undertaking.

The LCP is a *linear* combinatorial problem. That the LCP should be reformulated as a *nonlinear* problem, which is in turn embedded in a complicated *nonlinear* homotopy, is counterintuitive. Nevertheless, a homotopy algorithm based on $\Psi_a(\lambda, z)$ is globally convergent for a wide class of LCPs, numerically robust, reasonably efficient, and (most encouraging) rather insensitive to the dimension of the problem.

12. Acknowledgment. The authors are indebted to Jong-Shi Pang, Katta Murty, and Romesh Saigal for useful comments and suggestions.

REFERENCES

[1] M. Aganagic, *Newton's method for linear complementarity problems*, Math. Programming, 28 (1984), pp. 349–362.

[2] B. H. Ahn, *Solution of nonsymmetric linear complementarity problems by iterative methods*, J. Optim. Theory Appl., 33 (1981), pp. 175–185.

[3] ———, *Iterative methods for linear complementarity problems with upper bounds on primary variables*, Math. Programming, 26 (1983), pp. 295–315.

[4] Y. Bard, *Nonlinear Parameter Estimation*, Academic Press, New York, 1974.

[5] D. P. Bertaekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[6] R. W. COTTLE AND G. B. DANZIG, *Complementary pivot theory of mathematical programming*, Linear Algebra Appl., 1 (1968), pp. 103–125.

[7] ———, *Monotone solutions of the parametric linear complementarity problem*, Math. Programming, 3 (1972), pp. 210–224.

[8] R. W. COTTLE AND J. S. PANG, *On the convergence of a block successive overrelaxation method for a class of linear complementarity problems*, Math. Programming, 17 (1982), pp. 126–138.

[9] C. CRYER, *The solution of a quadratic programming problem using systematic overrelaxation*, SIAM J. Control, 9 (1971), pp. 385–392.

[10] B. C. EAVES, *The linear complementarity problem in mathematical programming*, Management Sci., 17 (1971), pp. 612–634.

[11] ———, *Homotopies for computation of fixed points*, Math. Programming, 3 (1972), pp. 1–22.

[12] B. C. EAVES AND R. SAIGAL, *Homotopies for computation of fixed points on unbounded regions*, Math. Programming, 3 (1972), pp. 225–237.

[13] M. FIEDLER AND V. PTÁK, *On matrices with non-positive off-diagonal elements and positive principal minors*, Czech. Math. J., 12 (1962), pp. 382–400.

[14] J. T. FREDRICKSEN, L. T. WATSON, AND K. G. MURTY, *A finite characterization of K-matrices in dimensions less than four*, Math. Programming, 35 (1986), pp. 17–31.

[15] D. GALE, *The Theory of Linear Economic Models*, McGraw-Hill, New York, 1960.

[16] T. HANSEN AND H. SCARF, *On the applications of a recent combinatorial algorithm*, Cowles Commission Discussion Paper No. 272, Yale Univ., New Haven, CT, 1969.

[17] A. W. INGLETON, *A problem in linear inequalities*, Proc. London Math. Soc. 3rd Series, 16 (1966), pp. 519–536.

[18] N. W. KAPPEL AND L. T. WATSON, *Iterative algorithms for the linear complementarity problem*, Internat. J. Computer Math., 19 (1986) pp. 273–297.

[19] S. KARAMARDIAN, *The complementarity problem*, Math. Programming, 2 (1972), pp. 107–129.

[20] L. M. KELLY AND L. T. WATSON, *Q-matrices and spherical geometry*, Linear Algebra Appl., 25 (1979), pp.175–190.

[21] ———, *Erratum: Some perturbation theorems for Q-matrices*, SIAM J. Appl. Math., 34 (1978), pp. 320–321.

[22] M. KOJIMA AND R. SAIGAL, *On the number of solutions to a class of complementarity problems*, Math. Programming, 21 (1981), pp. 190–203.

[23] ———, *A study of PC^1 homeomorphisms on subdivided polyhedrons*, SIAM J. Math. Anal., 10 (1979), pp. 1299–1312.

[24] ———, *On the number of solutions for a class of linear complementarity problems*, Math. Programming, 17 (1979), pp. 136–139.

[25] ———, *Private communication*, October, 1987.

[26] M. KOJIMA, S. MIZUNO, AND A. YOHISHE, *A polynomial-time algorithm for a class of linear complementarity problems*, Math. Programming, to appear.

[27] M. KOJIMA, S. MIZUNO, AND T. NOMA, *A new continuation method for complementarity problems with uniform P-functions*, Math. Programming, to appear.

[28] M. M. KOSTREVA, *Elasto-hydrodynamic lubrication: A nonlinear complementarity problem*, Mathematics Dept., GM Research Laboratories, Warren, MI, 1982.

[29] C. E. LEMKE, *Bimatrix equilibrium points and mathematical programming*, Management Sci., 11 (1965), pp. 681–689.

[30] ———, *On complementary pivot theory*, in Mathematics of the Decision Sciences Part 1, G. B. Danzig and A. F. Veinott, Jr., eds., Amer. Math. Soc., Providence, RI, 1968, pp. 95–114.

[31] ———, *Recent results on complementarity problems*, in Nonlinear Programming, J. B. Rosen, O. L. Mangasarian and K. Ritter, eds., Academic Press, New York, 1970, pp. 349–384.

[32] C. E. LEMKE AND J. T. HOWSON, *Equilibrium points of bimatrix games*, SIAM J. Appl. Math., 12 (1964), pp. 413–423.

[33] O. L. MANGASARIAN, *Equivalence of the complementarity problem to a system of nonlinear equations*, SIAM J. Appl. Math., 31 (1976), pp. 89–92.

[34] ———, *Linear complementarity problems solvable by a single linear program*, Math. Programming, 10 (1976), pp. 263–270.

[35] ———, *Solution of symmetric linear complementarity problems by iterative methods*, J. Optimization Theory Appl., 22 (1977), pp. 465–485.

[36] K. MEINTJES AND A. P. MORGAN, *A methodology for solving chemical equilibrium systems*, Appl. Math. Comput., 22 (1987), pp. 333–361.

[37] O. H. MERRILL, *Applications and extensions of an algorithm that computes fixed points of certain non-empty, convex, upper semi-continuous point to set mappings*, Dept. of Industrial Engineering, Univ. of Michigan Tech. Report No. 71-7, 1971.

[38] S. MIZUNO, A. YOHISHE, AND T. KIKUCHI, *Practical polynomial time algorithms for linear complementarity problems*, Dept. of Management Sci. and Eng., Tokyo Institute of Technology, Tech. Report No. 13, Tokyo, Japan, April, 1988.

[39] A. P. MORGAN, *Solving polynomial systems using continuation for scientific and engineering problems*, Prentice-Hall, Englewood Cliffs, NJ, (1987).

[40] A. P. MORGAN AND A. J. SOMMESE, *A homotopy for solving general polynomial systems that respects m-homogeneous structures*, Appl. Math. Comput., 24 (1987), pp. 101–113.

[41] ———, *Computing all solutions to polynomial systems using homotopy continuation*, Appl. Math. Comput., 24 (1987), pp. 115–138.

[42] K. G. MURTY, *On the parametric complementarity problem*, Engineering Summer Conference notes, University of Michigan, Ann Arbor, MI, 1971.

[43] ———, *On a characterization of P-matrices*, SIAM J. Appl. Math., 20 (1971), pp. 378–384.

[44] ———, *On the number of solutions to the complementarity problem and spanning properties of complementary cones*, Linear Algebra Appl., 5 (1972), pp. 65–108.

[45] ———, *Note on a Bard-type scheme for solving the complementarity problem*, Opsearch, 11 (1974) pp. 123–130.

[46] ———, *Linear and Combinatorial Programming*, John Wiley, New York, 1976, pp. 481–519.

[47] ———, *Computational complexity of complementary pivot methods*, Math. Programming Study 7 (1978), pp. 61–73.

[48] ———, *On the linear complementarity problem*, Methods Oper. Res., 31 (1978), pp. 425–439.

[49] ———, *Linear Complementarity, Linear and Nonlinear Programming*, Helderman-Verlag, West Berlin, 1986.

[50] J. S. PANG, *On the convergence of a basic iterative method for the implicit complementarity problem*, J. Optim. Theory Appl., 37 (1982), pp. 139–162.

[51] ———, *Necessary and sufficient conditions for the convergence of iterative methods for the linear complementarity problem*, J. Optim. Theory Appl., 42 (1984), pp. 1–18.

[52] J. S. PANG AND R. CHANDRASEKHARAN, *Linear complementarity problems solvable by a polynomially bounded pivoting algorithm*, Math. Programming Study 25 (1985), pp. 13–27.

[53] A. B. POORE AND D. SORIA, *Continuation algorithms for linear programming*, in preparation.

[54] A. B. POORE AND Q. AL-HASSAN, *The expanded Lagrangian system for constrained optimization problems*, SIAM J. Control Optim., 26 (1988), pp. 417–427.

[55] R. SAIGAL, *A note on a special linear complementarity problem*, Opsearch, 7 (1970), pp. 175–183.

[56] ———, *On the class of complementary cones and Lemke's algorithm*, SIAM J. Appl. Math., 23 (1972), pp. 46–60.

[57] ———, *On the convergence rate of algorithms for solving equations that are based on methods of complementary pivoting*, Math. Operations Res., 22 (1977), pp. 108–124.

[58] H. SAMUELSON, R. M. THRALL AND O. WESLER, *A partitioning theorem for Euclidean n-space*, Proc. Amer. Math. Soc., 9 (1958), pp. 805–807.

[59] H. SCARF, *The Computation of Economic Equilibria*, Yale Univ. Press, New Haven, CT, 1973.

[60] A. C. STICKNEY AND L. T. WATSON, *Digraph models of Bard-type algorithms for the linear complementarity problem*, Math. Operations Res., 3 (1978), pp. 322–333.

[61] W. M. G. VAN BOKHOVEN, *A class of linear complementarity problems is solvable in polynomial time*, unpublished paper, Dept. of Electrical Eng., Univ. of Technology, The Netherlands, 1980.

[62] L. T. WATSON, *A variational approach to the linear complementarity problem*, Doctoral Dissertation, Dept. of Math., Univ. of Michigan, Ann Arbor, MI, 1974.

[63] ———, *Some perturbation theorems for Q-matrices*, SIAM J. Appl. Math., 31 (1976), pp. 379–384.

[64] ———, *An algorithm for the linear complementarity problem*, Internat. J. Computer Math., 6 (1978), pp. 319–325.

[65] ———, *A globally convergent algorithm for computing fixed points of C^2 maps*, Appl. Math. Comput., 5 (1979), pp. 297–311.

[66] ———, *Solving the nonlinear complementarity problem by a homotopy method*, SIAM J. Control Optim., 17 (1979), pp. 36–46.

[67] ———, *Computational experience with the Chow-Yorke algorithm*, Math. Programming, 19 (1980), pp. 92–101.

[68] ———, *Numerical linear algebra aspects of globally convergent homotopy methods*, SIAM Rev., 28 (1986), pp. 529–545.

[69] L. T. WATSON, S. C. BILLUPS, AND A. P. MORGAN, HOMPACK: *A suite of codes for globally convergent homotopy algorithms*, ACM Trans. Math. Software, 13 (1987), pp. 281–310.

[70] L. T. WATSON, J. P. BIXLER, AND A. B. POORE, *Continuous homotopies for the linear complementarity problem*, Tech. Report TR 87-38, Dept. of Computer Sci., VPI & SU, Blacksburg, VA, 1987.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 3, pp. 279-286, July 1989

001

CO-SQUARE ROOTS OF MATRICES AND THEIR APPLICATION TO BOUNDARY VALUE DIFFERENTIAL MATRIX PROBLEMS*

L. JÓDAR† AND E. NAVARRO†

Abstract. In this paper the concept of the co-square root of a complex matrix is introduced. Methods for obtaining co-square roots of matrices are presented. This concept is applied to solving boundary value problems related to the matrix differential equation $X^{(2)}(t) - AX(t) = 0$.

Key words. co-square root, matrix differential equation, boundary value problem, Moore–Penrose pseudoinverse, annihilating polynomial

AMS(MOS) subject classifications. 15A24, 15A30, 15A60, 65L10

1. Introduction. Second-order matrix differential equations with constant coefficient matrices appear in the theory of vibrational systems [10].

Let us consider the matrix differential equation

$$X^{(2)}(t)-AX(t)=0 \tag{1.1}$$

where A and $X(t)$ are complex $n \times n$ matrices. Boundary value problems of the following type:

$$\begin{aligned} &X^{(2)}(t)-AX(t)=0, \\ &E_1X(0)+E_2X^{(1)}(0)=0, \\ &F_1X(a)+F_2X^{(1)}(a)=0, \qquad 0 \leqq t \leqq a \end{aligned} \tag{1.2}$$

where E_i, F_i, for $i = 1, 2, A$, and $X(t)$, are $n \times n$ complex matrices, have been studied in [8] when A is nonsingular, and in [9] for the more general case where the matrix A has a pair of square roots X_0 and X_1, such that $X_1 - X_0$ is nonsingular. Under the above hypothesis, the general solution of the matrix differential equation (1.1) admits a representation of the form

$$X(t)=\exp(tX_0)C+\exp(tX_1)D \tag{1.3}$$

where C and D are arbitrary $n \times n$ complex matrices, and this fact provides existence conditions and explicit expressions for solutions of problem (1.2).

A necessary and sufficient condition for the existence of square roots of a square matrix is given in [4], and interesting methods for computing square roots of matrices may be found in [1] and [6].

In this paper we study the boundary value problem (1.2) from a more general algebraic point of view. It is shown that even for the case where the matrix A does not have square roots, as well as when the matrix A does not have a pair of square roots whose difference is a nonsingular matrix, the problem (1.2) may have nontrivial solutions. Sufficient conditions for the existence of nontrivial solutions and their closed form solutions are given.

In § 2 we introduce the concepts of co-square root a and fundamental set of co-square roots of a square complex matrix. Different methods for obtaining co-square roots and fundamental sets of co-square roots of matrices are presented.

* Received by the editors December 7, 1987; accepted for publication (in revised form) September 19, 1988. This research was partly supported by a grant from the D.G.I.C.Y.T. project PS87-0064.

† Department of Applied Mathematics, Polytechnical University of Valencia, Valencia, Spain 46022.

In § 3 we use the above concepts to obtain an expression for the general solution of (1.1) that permits us to find existence conditions and explicit closed form solutions for the boundary value problem (1.2).

Throughout this paper we denote by $\mathbb{C}_{n\times n}$ the set of all $n \times n$ complex matrices, and for a matrix B in $\mathbb{C}_{n\times n}$, the set of all eigenvalues of B, is denoted by $\sigma(B)$.

2. Co-square roots of complex matrices. If X, T are matrices in $\mathbb{C}_{n\times n}$, then it is easy to show that $Z(t) = X \exp(tT)$ is a solution of (1.1) if and only if

$$XT^2 - AX = 0. \tag{2.1}$$

This suggests the following definition.

DEFINITION 2.1. Let A be a matrix in $\mathbb{C}_{n\times n}$. We say that a pair of matrices (X, T) with X, T in $\mathbb{C}_{n\times n}$ is a co-square root of A, if $X \neq 0$ and (2.1) is satisfied.

Example 1. If $A \in \mathbb{C}_{n\times n}$ and B is a square root of A, and I denotes the identity matrix in $\mathbb{C}_{n\times n}$, then (I, B) is a co-square root of A.

The next example shows that any square matrix $A \in \mathbb{C}_{n\times n}$ has co-square roots.

Example 2. Let z be an eigenvalue of A, and let w be a complex number such that $w^2 = z$; then the kernel of the matrix $(w^2I - A)$ is nontrivial. Thus, for any nonzero matrix X in $\mathbb{C}_{n\times n}$ such that $(w^2 I - A)X = 0$, the pair (X, wI) is a co-square root of A.

Example 3. Let us suppose that (X, T) is a co-square root of A, and let H be a nonsingular matrix in $\mathbb{C}_{n\times n}$; then (XH^{-1}, HTH^{-1}) is also a co-square root of A.

Note that Example 2 shows for any square matrix $A \in \mathbb{C}_{n\times n}$, there exist co-square roots of A. Furthermore, if $z \in \sigma(A)$, and w is a complex number satisfying $w^2 = z$, and $S = w^2I - A$, then the general solution of the equation

$$SX = 0, \tag{2.2}$$

is given by the expression

$$X = (I - S^+S)Z \tag{2.3}$$

where Z is an arbitrary matrix in $\mathbb{C}_{n\times n}$, and S^+ denotes the Moore–Penrose pseudoinverse of S (see [12, Thm. 2.3.2]. Hence the following result has been proved.

THEOREM 1. *Let A be a matrix in $\mathbb{C}_{n\times n}$. Then A has co-square roots (X, T), where $T = wI$, with w a complex number satisfying $w^2 = z \in \sigma(A)$, and X the nonzero matrices defined by* (2.3), *where $S = w^2I - A$, and Z is any matrix in $\mathbb{C}_{n\times n}$.*

To compute co-square roots of a matrix $A \in \mathbb{C}_{n\times n}$, we only need an eigenvalue $z \in \sigma(A)$, and the computation of the Moore–Penrose pseudoinverse of $S = w^2I - A$. In [2, p. 12], two interesting algorithms for computing S^+ are given.

Now we are going to consider a functional method for obtaining co-square roots (X, T) of matrices, such that T is not a scalar multiple of the identity matrix I. This method is based on the reduction of the degree of the algebraic equation (2.1) and may be regarded as a continuation of § 2 of [7].

The following result provides us with a necessary condition for a pair (X, T) of matrices in $\mathbb{C}_{n\times n}$ to be a co-square root of A.

THEOREM 2. *Let $C_L = \left[\begin{smallmatrix} 0 & I \\ A & 0 \end{smallmatrix}\right]$, and let (X, T) be a co-square root of A. If $p(z)$ is a polynomial such that $p(T) = 0$ and $p(C_L)$ is the block partitioned matrix defined by*

$$p(C_L) = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \quad B_{ij} \in \mathbb{C}_{n\times n}, \quad 1 \leqq i, \quad j \leqq 2, \tag{2.4}$$

then (X, T) *satisfies the system*

$$(2.5)\qquad B_{11}X+B_{12}XT=0, \qquad B_{21}X+B_{22}XT=0.$$

Proof. Note that $X, T \in \mathbb{C}_{n\times n}$, is a co-square root of A if and only if $X \neq 0$ and

$$(2.6)\qquad \begin{bmatrix} X \\ XT \end{bmatrix} T = C_L \begin{bmatrix} X \\ XT \end{bmatrix}.$$

From (2.6) we obtain

$$(2.7)\qquad \begin{bmatrix} X \\ XT \end{bmatrix} T^p = (C_L)^p \begin{bmatrix} X \\ XT \end{bmatrix}$$

for all positive integer p. Hence, using the hypothesis $p(T) = 0$, we have that

$$(2.8)\qquad \begin{bmatrix} X \\ XT \end{bmatrix} p(T) = p(C_L) \begin{bmatrix} X \\ XT \end{bmatrix} = 0.$$

From (2.4) and (2.8), the result is established.

Note that from (2.6), which characterizes the co-square roots (X, T) of a matrix A, it follows that if $p(z)$ is an annihilating polynomial of C_L, then $Xp(T) = 0$ for any co-square root (X, T) of A. Also, from (2.6) we obtain that if (X, T) is a solution of a system of the type (2.5), where $(B_{ij}) = p(C_L)$ for some polynomial $p(z)$, then a necessary condition to be a co-square root of A is that $Xp(T) = 0$. The next result is a reciprocal one of Theorem 2.

THEOREM 3. *Let* $p(z)$ *be a polynomial, and let* $(B_{ij}) = p(C_L)$. *If* (X, T) *is a solution of the system* (2.5), *if* $X \neq 0$, *and if* B_{12} *is nonsingular, then* (X, T) *is a co-square root of* A *and* $Xp(T) = 0$.

Proof. From the equality $C_L p(C_L) = p(C_L) C_L$, it follows that

$$\begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix}.$$

Hence, we have

$$(2.9)\qquad B_{21} = B_{12}A, \qquad B_{22} = B_{11}.$$

If we multiply the first equation of (2.9) by X, and the second by XT, it follows that

$$(2.10)\qquad B_{21}X = B_{12}AX, \qquad B_{22}XT = B_{11}XT$$

so that

$$(2.11)\qquad B_{21}X + B_{22}XT = B_{12}AX + B_{11}XT.$$

As (X, T) satisfies the system (2.5), we obtain that the first member of (2.11) is equal to the matrix zero, so,

$$(2.12)\qquad B_{12}AX + B_{11}XT = 0.$$

From the first equation of system (2.5), we have $B_{11}X = -B_{12}XT$; hence, from (2.12) it follows that

$$(2.13)\qquad 0 = B_{12}AX - B_{12}XT^2 = B_{12}(AX - XT^2).$$

From (2.13) and from the invertibility of B_{12}, we obtain that (X, T) is a co-square root of A. Also, from the previous comments to the statement of Theorem 3, it follows that $Xp(T) = 0$.

To understand what class of polynomials must be considered in Theorem 3 for obtaining co-square roots of A, note that $Xp(T) = 0$, and $X \neq 0$ implies that $p(T)$ is singular. Thus, from the spectral mapping theorem [5], the polynomial $p(z)$ must annihilate a part of the spectrum $\sigma(T)$.

On the other hand, (2.6), the condition $X \neq 0$, and the Rosenblum theorem [13] imply that

$$\sigma(T) \cap \sigma(C_L) \neq \varnothing;$$

and this means that the class of polynomials that must be considered are those polynomials $p(z)$ that are multiples of a proper divisor of the minimal polynomial of C_L. This is shown in the next example.

Example 4. Let $A = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$. An easy computation yields $\sigma(A) = \{0, 2\}$, $\sigma(C_L) = \{0, 2^{1/2}, -2^{1/2}\}$, and the minimal polynomial $q(z)$ of C_L coincides with its characteristic polynomial and takes the form $q(z) = z^2(z - 2^{1/2})(z + 2^{1/2})$. Let us consider the polynomial $p(z) = z(z^2 - 2^{1/2})$, and note that $p(z)$ is a multiple of the proper divisor $r(z) = z$ of the minimal polynomial of C_L. Computing, we get that $p(C_L) = C_L(C_L^2 - 2^{1/2}I)$ takes the form

$$p(C_L) = \begin{bmatrix} 0 & A - 2^{1/2}I \\ A(A - 2^{1/2}I) & 0 \end{bmatrix}, \quad B_{11} = B_{22} = 0,$$

$$B_{12} = \begin{bmatrix} 1 - 2^{1/2} & 1 \\ 1 & 1 - 2^{1/2} \end{bmatrix}, \quad B_{21} = (2 - 2^{1/2})A.$$

Thus B_{12} is nonsingular, and the reduced system (2.5) takes the form

$$B_{12}XT = 0, \qquad B_{21}X = 0.$$

As B_{12} is nonsingular, we have $XT = 0$. Solving the system $B_{21}X = 0$, we have

$$(2 - 2^{1/2}) \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} = 0.$$

Hence X takes the form $X = \left[\begin{smallmatrix} a & b \\ -a & -b \end{smallmatrix}\right]$, for all complex values a and b, nonsimultaneously zero.

When we solve the system

$$\begin{bmatrix} a & b \\ -a & -b \end{bmatrix} \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} = 0,$$

it follows that $T = (t_{ij})$ is given by

$$T = \begin{bmatrix} \alpha b & \beta b \\ -\alpha a & -\beta a \end{bmatrix}, \qquad \beta, \alpha \in \mathbb{C}.$$

Thus an infinite set of co-square roots (X, T) of A has been obtained.

The following example shows that not every multiple of a proper divisor of the minimal polynomial of the companion matrix C_L provides co-square roots of A through Theorem 3.

Example 5. Let A be the matrix of Example 4, and let $p(z) = z^2(z - 2^{1/2})$ be a proper divisor of the minimal polynomial of C_L. Then $p(C_L) = C_L^2(C_L - 2^{1/2}I)$ takes the form

$$p(C_L) = \begin{bmatrix} -2^{1/2}A & A \\ A^2 & -2^{1/2}A \end{bmatrix},$$

and thus $(B_{ij}) = p(C_L)$ with $B_{12} = A$ singular, which means that Theorem 3 does not provide co-square roots of A.

Theorems 2 and 3 permit us to obtain co-square roots of a matrix A, but as is shown in Example 5, the class of polynomials that provides co-square roots of A is not characterized. Note that the co-square roots of A, given by Theorems 1, 2, and 3, allow us to get solutions of the matrix differential equation (1.1). However, our main interest is to obtain a pair of solutions of (1.1) by means of an appropriate pair of co-square roots of A, satisfying the property that they generate the general solution of (1.1). This desirable property is characterized by the following definition.

DEFINITION 2.2. Let $A \in \mathbb{C}_{n\times n}$, and let (X_i, T_i), for $i = 1, 2$, be co-square roots of A. We say that $\{(X_i, T_i);\ i = 1, 2\}$ is a fundamental set of co-square roots of A if the block partitioned matrix V defined by

$$V=\begin{bmatrix} X_1 & X_2 \\ X_1T_1 & X_2T_2 \end{bmatrix} \tag{2.14}$$

is invertible in $\mathbb{C}_{2n\times 2n}$.

Example 6. If $A \in \mathbb{C}_{n\times n}$ and $T_1, T_2 \in \mathbb{C}_{n\times n}$ are square roots of A, then the pair $(I, T_1)(I, T_2)$ define a fundamental set of co-square roots of A if and only if the matrix $T_2 - T_1$ is nonsingular (see [8, Lemma 1]).

The next result shows that for a very general class of matrices $A \in \mathbb{C}_{n\times n}$, there exists a fundamental system of co-square roots.

THEOREM 4. *Let $A \in \mathbb{C}_{n\times n}$, and let us suppose that $C_L = [\begin{smallmatrix} 0 & I \\ A & 0 \end{smallmatrix}]$. Then A admits a fundamental set of co-square roots if and only if C_L is similar to a block diagonal matrix*

$$J=\begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix},$$

where $J_i \in \mathbb{C}_{n\times n}$ for $i = 1, 2$. If the similarity matrix takes the form

$$P=\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} P_{ij} \in \mathbb{C}_{n\times n},$$

for $1 \leqq i, j \leqq 2$, then (P_{11}, J_1) and (P_{12}, J_2) define a fundamental set of co-square roots of A.

Proof. From the hypothesis we have that P is invertible and

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}\begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix}=\begin{bmatrix} 0 & I \\ A & 0 \end{bmatrix}\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \tag{2.15}$$

so that

$$P_{11}J_1 = P_{21}, \qquad P_{21}J_1 = AP_{11} \tag{2.16}$$

and

$$P_{12}J_2 = P_{22}, \qquad P_{22}J_2 = AP_{12}. \tag{2.17}$$

By substitution of the first equality of (2.16) and (2.17) in the corresponding second ones, we have

$$P_{11}J_1^2 = AP_{11} \quad \text{and} \quad P_{12}J_2^2 = AP_{12}.$$

Thus (P_{11}, J_1) and (P_{12}, J_2) are co-square roots of A, because $P_{11} \neq 0$ and $P_{12} \neq 0$, from (2.16), (2.17), and the invertibility of P. Also, note that $\{(P_{11}, J_1), (P_{12}, J_2)\}$ is a fundamental set of co-square roots of A because the corresponding block matrix V of (2.14) takes the form

$$V=\begin{bmatrix} P_{11} & P_{12} \\ P_{11}J_1 & P_{12}J_2 \end{bmatrix}=\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}=P.$$

Conversely, let us suppose that $\{(X_1, T_1), (X_2, T_2)\}$ is a fundamental set of co-square roots of A. Then from the definition (2.2), the matrix V defined by (2.14) is invertible, and an easy computation yields

$$V[\operatorname{diag}(T_1, T_2)] = C_L V.$$

Hence the result is proved.

Remark 1. If the elementary divisors of the matrix polynomial $\lambda I - C_L$ are denoted by $(\lambda - \lambda_{ij})^{\alpha_{ij}}, j = 1, \cdots, k_i, i = 1, 2, \cdots, r$, then as the exponents of the elementary divisors coincide with the dimensions of the Jordan blocks of C_L, the condition C_L not similar to a block diagonal matrix with diagonal blocks of dimension $n \times n$ means that for any set of possibly repeated numbers α_i, their addition is always different from n (see [11], Thm. 2, p. 270] for details).

Example 7. Let A be the matrix introduced in Example 4. From the spectral information of C_L and the fact that the minimal and the characteristic polynomials of C_L coincide and are equal to $p(z) = z^2(z - 2^{1/2})(z + 2^{1/2})$, it follows that the Jordan canonical form of C_L is given by the matrix

$$J=\begin{bmatrix} J_1 & 0 \\ 0 & J_2 \end{bmatrix}$$

where

$$J_1=\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad J_2=\begin{bmatrix} 2^{1/2} & 0 \\ 0 & 2^{-1/2} \end{bmatrix}.$$

An easy computation yields that $C_L = PJP^{-1}$, where

$$P=\begin{bmatrix} 1 & 1 & 1 & -1 \\ -1 & -1 & 1 & -1 \\ 0 & 1 & 2^{1/2} & 2^{1/2} \\ 0 & -1 & 2^{1/2} & 2^{1/2} \end{bmatrix}.$$

Thus, taking

$$P_{11}=\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, \qquad P_{12}=\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix},$$

it follows that (P_{11}, J_1) and (P_{12}, J_2) define a fundamental system of co-square roots of A because of Theorem 4.

Note that as $\sigma(A) = \{0, 2\}$, for any square root B of A it follows that $\sigma(B) = \{0, 2^{1/2}\}$, or $\sigma(B) = \{0, -2^{1/2}\}$. In the first case, the characteristic polynomial of B is $p(z) = z(z - 2^{1/2})$ and then $p(B) = B(B - 2^{1/2}I) = B^2 - 2^{1/2}B = A - 2^{1/2}B = 0$; this is $B = 2^{-1/2}A$. If $\sigma(B) = \{0, -2^{1/2}\}$, then its characteristic polynomial is $q(z) = z(z + 2^{1/2})$ and $q(B) = B(B + 2^{1/2}I) = B^2 + 2^{1/2}B = A + 2^{1/2}B = 0$. So, in this case $B = -2^{-1/2}A$. An easy computation shows that $\pm 2^{-1/2}A$ are square roots of A. As a consequence, $B_1 = 2^{-1/2}A$ and $B_2 = -2^{-1/2}A$ are the unique square roots of A and $B_1 - B_2 = 2^{1/2}A$ is singular. This fact and Example 7 shows the existence of a fundamental set of co-square roots without the existence of a pair of square roots of A, whose difference is nonsingular.

3. Applications of co-square roots to boundary value problems. We begin this section with a result that provides the general solution of the differential equation (1.1), in terms of a fundamental system of co-square roots of the matrix A as happens for the scalar case, with the obvious difference that for the scalar case the corresponding characteristic algebraic equation $z^2 - a = 0$ is always solvable. For the matrix case we are going to consider co-square roots of the matrix A.

THEOREM 5. *Let $A \in \mathbb{C}_{n\times n}$, and let $\{(X_1, T_1), (X_2, T_2)\}$ be a fundamental set of co-square roots of A.*

(i) *The general solution of* (1.1) *on the real line is given by the expression*

$$X(t) = X_1 \exp(tT_1)D_1 + X_2 \exp(tT_2)D_2, \tag{3.1}$$

where D_i, for $i = 1, 2$, are arbitrary matrices in $\mathbb{C}_{n\times n}$.

(ii) *The unique solution of* (1.1) *that satisfies the initial conditions $X(0) = C_1$, $X^{(1)}(0) = C_2$ is given by* (3.1), *where D_1 and D_2 are uniquely determined by the expression*

$$\begin{bmatrix} D_1 \\ D_2 \end{bmatrix} = V^{-1} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} \tag{3.2}$$

and V is the block matrix given by (2.14).

Proof. Considering the change $X = Y_1$, $X^{(1)} = Y_2$, it is clear that a Cauchy problem of the type (1.1) with the initial condition $X(0) = C_1$, $X^{(1)}(0) = C_2$ is equivalent to the first-order system

$$Y^{(1)}(t) = C_L Y(t), \quad Y(0) = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}, \quad Y(t) = \begin{bmatrix} Y_1(t) \\ Y_2(t) \end{bmatrix}.$$

Thus, it is clear that the above Cauchy problem has only one solution [3]. By differentiation in the expression (3.1), it follows that

$$\begin{aligned} X^{(1)}(t) &= X_1 T_1 \exp(tT_1)D_1 + X_2 T_2 \exp(tT_2)D_2, \\ X^{(2)}(t) &= X_1 T_1^2 \exp(tT_1)D_1 + X_2 T_2^2 \exp(tT_2)D_2 \end{aligned} \tag{3.3}$$

for all real numbers t, and for all matrices D_1 and D_2 in $\mathbb{C}_{n\times n}$. Thus we have that $X(t)$ given by (3.1) satisfies

$$X^{(2)}(t) - AX(t) = (X_1 T_1^2 - AX_1) \exp(tT_1)D_1 + (X_2 T_2^2 - AX_2) \exp(tT_2)D_2 = 0.$$

So, to prove (i) and (ii), it is sufficient that the solution of equation (1.1) that satisfies $X(0) = C_1$, $X^{(1)}(0) = C_2$ may be expressed by the matrix function $X(t)$ defined by (3.1) for appropriate matrices D_1 and D_2 in $\mathbb{C}_{n\times n}$. Taking into account (3.1) and (3.3), and if we impose that $X(t)$ given by (3.1) satisfies the conditions $X(0) = C_1$ and $X^{(1)}(0) = C_2$, it follows that D_1, D_2 must satisfy

$$C_1 = X_1 D_1 + X_2 D_2, \qquad C_2 = X_1 T_1 D_1 + X_2 T_2 D_2$$

or

$$\begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} X_1 & X_2 \\ X_1 T_1 & X_2 T_2 \end{bmatrix} \begin{bmatrix} D_1 \\ D_2 \end{bmatrix}. \tag{3.4}$$

From the hypothesis, the block matrix V defined by (2.14) is invertible, so solving (3.4), we have that D_1, D_2 are determined by (3.2). This concludes the proof of Theorem 5.

Now we will use the representation (3.1) for the general solution of equation (1.1) to solve the boundary value problem (1.2). For the sake of clarity in the statement of the next result, we introduce the following block matrix:

$$S=\begin{bmatrix} E_1X_1+E_2X_1T_1 & E_1X_2+E_2X_2T_2 \\ (F_1X_1+F_2X_1T_1)\exp(aT_1) & (F_1X_2+F_2X_2T_2)\exp(aT_2) \end{bmatrix}. \tag{3.5}$$

THEOREM 6. *Let us suppose that* $A\in \mathbb{C}_{n\times n}$ *has a fundamental set of co-square roots* $\{(X_1, T_1), (X_2, T_2)\}$. *Then problem* (1.2) *has nontrivial solutions if and only if the block matrix* S *defined by* (3.5) *is singular. Under this hypothesis, the general solution of problem* (1.2) *is given by the expression* (3.1), *where* D_1, D_2 *are matrices in* $\mathbb{C}_{n\times n}$ *given by*

$$\begin{bmatrix} D_1 \\ D_2 \end{bmatrix}=(I_{2n}-S^+S)Z \tag{3.6}$$

where I_{2n} *denotes the identity matrix in* $\mathbb{C}_{2n\times 2n}$, *and* Z *is an arbitrary matrix in* $\mathbb{C}_{2n\times n}$.

Proof. From Theorem 5, the general solution of the matrix differential equation (1.1) is given by (3.1), where D_1, D_2 are arbitrary matrices in $\mathbb{C}_{n\times n}$. To satisfy the boundary value conditions of problem (1.2), the matrix function $X(t)$ given by (3.1) must satisfy the boundary value conditions of (1.2). Taking into account (3.3), it follows that the matrices D_1, D_2 must satisfy the algebraic system

$$S\begin{bmatrix} D_1 \\ D_2 \end{bmatrix}=0 \tag{3.7}$$

where S is given by (3.5). Now the result is a consequence of Theorem 2.3.2 of [12, p. 24].

REFERENCES

[1] Å. BJÖRCK AND S. HAMMARLING, *A Schur method for the square root of a matrix*, Linear Algebra Appl., 52/53 (1983), pp. 127–140.

[2] S. L. CAMPBELL AND C. D. MEYER, JR., *Generalized Inverses of Linear Transformations*, Pitman, Boston, 1979.

[3] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, MacGraw-Hill, New York, 1955.

[4] G. W. CROSS AND P. LANCASTER, *Square roots of complex matrices*, Linear and Multilinear Algebra, (1974), pp. 289–293.

[5] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, Interscience, New York, 1957.

[6] N. J. HIGHAM, *Computing real square roots of a real matrix*, Linear Algebra Appl., 88/89 (1987), pp. 405–430.

[7] L. JÓDAR, *Boundary value problems for second order operator differential equations*, Linear Algebra Appl., 83 (1986), pp. 29–38.

[8] ———, *Explicit expressions for Sturm-Liouville operator problems*, Proc. Edinburgh Math. Soc. (2), 30 (1987), pp. 301–309.

[9] ———, *Explicit solutions for second order operator differential equations with two boundary value conditions*, Linear Algebra Appl., 103 (1988), pp. 73–86.

[10] P. LANCASTER, *Lambda Matrices and Vibrating Systems*, Pergamon Press, Elmsford, NY, 1966.

[11] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, 2nd ed., Academic Press, New York, 1985.

[12] C. R. RAO AND S. K. MITRA, *Generalized Inverse of Matrices and Its Applications*, John Wiley, New York, 1971.

[13] M. ROSENBLUM, *On the operator equation $BX - XA = Q$*, Duke Math. J., 23 (1956), pp. 263–270.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 3, pp. 287-293, July 1989

002

MORE ON A RAYLEIGH-RITZ REFINEMENT TECHNIQUE FOR NEARLY UNCOUPLED STOCHASTIC MATRICES*

MOSHE HAVIV†

Abstract. A Rayleigh-Ritz refinement technique is analyzed that is suitable for accelerating the convergence of iterative procedures for computing the stationary distribution of a nearly uncoupled stochastic matrix. In particular, for that case the error of the new approximation in terms of the previous error and the degree of coupling gets a special form. Cases where the refinement is promising are given as well. All the analysis requires the single assumption that the Markov chain under consideration is irreducible.

Key words. aggregation/disaggregation, stochastic matrices, near uncoupling

AMS(MOS) subject classifications. 15A09, 60S10, 65F10

1. Introduction. Let $P \in R_+^{n\times n}$ be a stochastic matrix representing an irreducible Markov chain, and let y be its stationary distribution. The irreducibility assumption implies that y is the unique row vector satisfying $yP = y$ and $yu = 1$, where u is a vector and all its entries are one. Moreover, $y_i > 0$ for $1 \leqq i \leqq n$. Let $S = \{1, 2, \cdots, n\}$ be the corresponding state space, and let $\Gamma = \{J(1), J(2), \cdots, J(p)\}$ be a partition of the state space to p subsets, $J(1), J(2), \cdots, J(p)$. Also for $K, L \subseteq S$ let P_{KL} be the submatrix of P, where rows are indexed by K and columns by L. For $L = K$ we use P_K instead of P_{KK}. A similar definition applies for x_K as a subvector of $x \in R_+^n$.

After permuting rows and corresponding columns, P can be represented with respect to the partition Γ as follows:

$$P=\begin{bmatrix} P_{J(1)} & P_{J(1)J(2)}\cdots P_{J(1)}P_{J(p)} \\ P_{J(2)J(1)} & P_{J(2)}\cdots P_{J(2)}P_{J(p)} \\ \vdots & \vdots \\ P_{J(p)J(1)} & P_{J(p)J(2)}\cdots P_{J(p)} \end{bmatrix}.$$

We say that P is nearly uncoupled, or nearly completely decomposable with respect to the partition Γ, if the row-sums of P_K for $K \in \Gamma$ are significantly greater than the row-sums of P_{KL} for $K, L \in \Gamma$ and $K \neq L$. Namely, P_K for $K \in \Gamma$ are almost stochastic. Hence, in that case $P = P^* + \varepsilon C$, where $P^* = \operatorname{diag}(P^*_{J(1)}, P^*_{J(2)}, \cdots, P^*_{J(p)})$ for some stochastic matrices $P^*_{J(1)}, P^*_{J(2)}, \cdots, P^*_{J(p)}$ close to $P_{J(1)}, P_{J(2)}, \cdots, P_{J(p)}$. Thus ε is a small number and all row-sums for $C \in R^{n\times n}$ are zero. Also the diagonal blocks of C are nonpositive, while its off-diagonal blocks are nonnegative. We will assume that ε is smaller than the modulus of the largest eigenvalue of C. Note that P^* and hence C are not uniquely defined.

Let $x \in R_+^n$ be a given probability vector that is thought to be an approximation to y. Assume that for each $J \in \Gamma$ there exists some $t \in J$ such that $x_t > 0$. Next we state the Rayleigh-Ritz refinement technique. In particular, a new approximation to y is found.

Step 0.

For each $J \in \Gamma$
Let $z_J(x) = x_J / \sum_{t\in J} x_t$

* Received by the editors May 25, 1987; accepted for publication (in revised form) September 26, 1988.

† Department of Statistics, Hebrew University, Jerusalem 91905, Israel, and Faculty of Commerce and Business Administration, University of British Columbia, Vancouver, British Columbia, Canada V6T 1W5 (USERHAVI@UBCMTSG).

```
Step 1 (aggregation step).
    (a) For all $I, J \in \Gamma$
        Let $Q_{IJ} = z_I(x)P_{IJ}u$
    (b) Find the stationary distribution of $Q$, namely, find $q \in R_+^p$ such that
        $qQ = q$ and $qu = 1$.[1]
Step 2 (disaggregation step).
    For each $J \in \Gamma$
    For each $t \in J$
    Let $\tilde{x}_t = q_J z_t(x)$
```

It is easy to see that $Q \in R_+^{p \times p}$ is a stochastic matrix and that $\tilde{x} \in R_+^n$ is a probability vector. Hence, we may consider $\tilde{x}$ as an updated approximation to y.

The above-mentioned procedure has been extensively considered in the literature, in particular in the context of nearly uncoupled stochastic matrices. The first to look at this method in that context were Simon and Ando (1961). See also Courtois (1977); Stewart (1983); Chatelin (1984); Haviv and Van der Heyden (1984); McAllister, Stewart, and Stewart (1984); Cao and Stewart (1985); and Haviv (1987).

Our main result stated as Theorem 1 and given in § 2, expresses the error of $\tilde{x}$ in terms of the error of x up to an additive term of the order $O(\varepsilon)$. Section 3 contains the proof of Theorem 1. These results will be shown in § 4 to be useful for nearly uncoupled stochastic matrices, as it leads to identifying cases where the Rayleigh–Ritz step should be applied. A numerical example is given in § 5. Finally, we refer the reader to Haviv (1987) for some numerical examples concerning this and other methods.

The method considered here was classified as a Rayleigh–Ritz method by McAllister, Stewart, and Stewart (1984). This is the case, as by solving a smaller-dimensional problem and utilizing a given approximation one finds a new approximation. However, as was pointed out to us by the referee, this terminology might be misleading, as it is usually preserved for a corresponding technique to approximate eigensystems of symmetric matrices. A detailed description on the latter can be found in Parlett (1980, pp. 213–217).

2. Preliminary notation and main results. Our main result is stated as Theorem 1 at the end of this section. First we develop the required notation.

Let x be a fixed probability vector with (at least) one $i \in J$ for all $J \in \Gamma$ such that $x_i > 0$. Then the completely uncoupled (with respect to Γ) stochastic matrix $\Pi \in R_+^{n \times n}$ is defined as follows:

$$\Pi_{st} = \begin{cases} z_t(x) & \text{if } s, t \in J \text{ for some } J \in \Gamma, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $x = x\Pi$, that $z(x) = z(x)\Pi$, that $\tilde{x} = \tilde{x}\Pi$ and that for the completely uncoupled P^*, $P^*\Pi = \Pi$.

For a stochastic matrix T representing a Markov chain (not necessarily irreducible), let $E(T)$ be its stationary matrix, namely, $E(T) = \lim_{N\to\infty} 1/N \sum_{m=0}^{N-1} T^m$. Also, let $Y(T)$ be its deviation matrix, namely, $Y(T) = [I - T + E(T)]^{-1} - E(T)$. It is well known (cf. Campbell and Meyer (1979)) that $Y(T)$ is the group inverse of $I - T$ and as such satisfies

$$(1) \qquad (I-T)Y(T) = I - E(T) = Y(T)(I-T).$$

[1] The dimension of q equals the number of subsets. The component of q corresponding to subset J will be denoted by q_J.

Returning to nearly uncoupled stochastic matrices, note that for the nearly uncoupled $P = P^* + \varepsilon C$, $P\Pi$ is also nearly uncoupled as $P\Pi = (P^* + \varepsilon C)\Pi = P^*\Pi + \varepsilon C\Pi = \Pi + \varepsilon C\Pi$. Also note that $P\Pi$ is irreducible as well. In order to emphasize the dependence on ε, let $P\Pi(\varepsilon) = \Pi + \varepsilon C\Pi$, let $E(\varepsilon) = E(P\Pi(\varepsilon))$, and let $Y(\varepsilon) = Y(P\Pi(\varepsilon))$. Also, let $Q(\varepsilon) \in R_+^{p\times p}$ be the aggregation matrix of $\Pi + \varepsilon C\Pi$, namely, for I, $J \in \Gamma$, $I \neq J$, $Q_{IJ}(\varepsilon) = \varepsilon z_I(x)(C\Pi)_{IJ}$ and for $I \in \Gamma$, $Q_{II}(\varepsilon) = 1 - \sum_{J\neq I} Q_{IJ}(\varepsilon)$. Finally, let $D \in R^{p\times p}$ be the deviation matrix of $Q(1)$.[2]

Finally, Schweitzer (1981) developed series expansions for nearly uncoupled stochastic matrices. In particular, from his results we obtain

$$Y(\varepsilon) = \frac{1}{\varepsilon} Y^{(-1)} + Y^{(0)} + O(\varepsilon) \tag{2}$$

for some matrices $Y^{(-1)}$ and $Y^{(0)}$. Moreover,[3]

$$Y_{ij}^{(-1)} = D_{IJ} z_j(x) \quad \text{for } i \in I \text{ and } j \in J. \tag{3}$$

THEOREM 1. *Let e and $\tilde{e}$ be $x - y$ and $\tilde{x} - y$, respectively. Then*

$$\begin{aligned}\tilde{e} &= e(I - \Pi)Y^{(0)} + eO(\varepsilon)\\ &= e(I + CY^{(-1)}) + eO(\varepsilon).\end{aligned}$$

3. Proofs. In this section we prove Theorem 1. Before that, we need the following Lemma 1 and Theorem 2 as prerequisites.

LEMMA 1.[4] $\tilde{x} = \tilde{x}P\Pi$.

Proof. Let $j \in J$ for some $J \in \Gamma$; then

$$\begin{aligned}[\tilde{x}P\Pi]_j &= \sum_{i\in S} \tilde{x}_i (P\Pi)_{ij}\\ &= \sum_{I\in\Gamma}\sum_{i\in I} \tilde{x}_i (P\Pi)_{ij}\\ &= \sum_{I\in\Gamma} q_I \sum_{i\in I} z_i(x) \sum_{k\in J} P_{ik} z_j(x)\\ &= z_j(x) \sum_{I\in\Gamma} q_I Q_{IJ} = z_j(x) q_J\\ &= \tilde{x}_j. \qquad \square\end{aligned}$$

Note that as P is irreducible, the same is the case with $P\Pi$ and hence $\tilde{x}$ is the unique stationary distribution of $P\Pi$. The following theorem already appears in Haviv (1987). We prove it here for completeness.

THEOREM 2. $\tilde{e} = e(I - \Pi)Y(P\Pi)$.

Proof. First note that

$$(\tilde{x} - y)(I - P\Pi) = \tilde{x}(I - P\Pi) - y(I - P)\Pi + y(\Pi - I).$$

[2] In order to ease the exposition, we assume here that $Q(1)$ is stochastic. However, if this is not the case D can be replaced with $\varepsilon D(\varepsilon)$ for any ϵ small enough where $D(\varepsilon)$ is the deviation matrix of the stochastic matrix $D(\varepsilon)$.

[3] In Schweitzer (1981) the matrix D has a different definition. We omit a detailed proof for the equivalence between the two definitions. For such a proof see Haviv and Van der Heyden (1984).

[4] This result originally appeared in Chatelin and Miranker (1984, Thm. 2). Here we give a different proof.

The first two terms on the right-hand side are zero by Lemma 1 and by the definition of y, respectively. Hence,

$$(\tilde{x}-y)(I-P\Pi)=-y(I-\Pi).$$

As noted before, $x = x\Pi$ and hence

$$(\tilde{x}-y)(I-P\Pi)=(x-y)(I-\Pi)$$

or

$$\tilde{e}(I-P\Pi)=e(I-\Pi).$$

Note that as P and Π are stochastic; $P\Pi$ is also stochastic. Postmultiplying the last equality with $Y(P\Pi)$ coupled with (1), we obtain

$$\tilde{e}[I-E(P\Pi)]=e(I-\Pi)Y(P\Pi).$$

Finally, $P\Pi$ is irreducible with stationary distribution $\tilde{x}$ (see Lemma 1), implying that $E(P\Pi) = u\tilde{x}$. Hence, $\tilde{e}E(P\Pi) = 0$, which completes the proof of the theorem. □

Proof of Theorem 1. First express $Y(\varepsilon)$, the deviation matrix of $P\Pi = \Pi + \varepsilon C\Pi$, as suggested in (2). Then by (1),

$$(4) \qquad [I-P\Pi(\varepsilon)]Y(\varepsilon)=(I-\Pi-\varepsilon C\Pi)\left(\frac{1}{\varepsilon}Y^{(-1)}+Y^{(0)}+O(\varepsilon)\right)=I-E(\varepsilon).$$

Since $E(\varepsilon)$ is stochastic, the right-hand side of the above equality is bounded. Hence, the coefficient of $1/\varepsilon$ is zero. Thus,

$$(5) \qquad (I-\Pi)Y^{(-1)}=0.$$

This, coupled with Theorem 2 and (4), implies that

$$\tilde{e}=e(I-\Pi)Y^{(0)}+eO(\varepsilon).$$

Our proof will be completed by showing next that $e(I - \Pi)Y^{(0)} = e(I + CY^{(-1)})$. From identity (4) we obtain

$$(I-\Pi)Y^{(0)}-C\Pi Y^{(-1)}=I-\lim_{\varepsilon\to 0}E(\varepsilon).$$

From (5), $\Pi Y^{(-1)} = Y^{(-1)}$. Finally, as $E(\varepsilon)$ is a matrix all its rows are identical, the same is the case with its limit.[5] Hence, $e \lim_{\varepsilon\to 0} E(\varepsilon) = 0$. This completes the proof of Theorem 1. □

4. When to apply the Rayleigh–Ritz step. Let x be a probability vector and consider it as an approximation to y. We say that x contains two types of errors. The first, "the local error," is defined by $\delta_i = [z(x) - z(y)]_i$ for $i \in S$; the second, "the global error," is defined by $\Delta_I = \sum_{i\in I}(x_i - y_i)$ for $I \in \Gamma$.[6] The names "local" and "global" are given because $z_I(y)$ is the limiting distribution conditional of being in subset I, and $\sum_{i\in I} y_i$ is the limiting probability for being in subset I. Similar interpretations, as corresponding

[5] In Schweitzer (1981) one can find the explicit expression for this limit, which is a matrix whose rows are identically the limit of the stationary probabilities of $\Pi + \varepsilon\Pi C$.

[6] See Step 0 for the definitions of $z(x)$ and $z(y)$.

approximations, exist for $z_I(x)$ and $\sum_{i\in I} x_i$. Note that for $i \in I$, $e_i = \delta_i\omega_I + \Delta_I z_i(x)$, where $\omega_I = \sum_{i\in I} y_i$. Also note that x and $\tilde{x}$ have identical local errors.

The next theorem bounds the rate of the error reduction while applying a Rayleigh–Ritz step for nearly uncoupled stochastic matrices. The bound is given in terms of the local error δ and the global error Δ.

THEOREM 3. *For any norm or seminorm* $\| \ \|$ *on* R^n,

$$\|e(I+CY^{(-1)})\| \leqq \|e\| \max_{J\in\Gamma}\max_{j\in J}\left[\frac{\delta_j\omega_j + z_j(x)f_J(e)}{\Delta_J z_j(x)+\delta_j\omega_j}\right], \tag{6}$$

where for $J \in \Gamma, f_J(e) = \sum_{I\in\Gamma}\omega_I \sum_{i\in I}\delta_i \sum_{K\in\Gamma} D_{KJ} \sum_{k\in K} C_{ik}$.

In particular,

$$\|\tilde{e}\| / \|e\| \leqq O(\max_{J\in\Gamma}\max_{j\in S} |\delta_j/\Delta_J|)+O(\varepsilon). \tag{7}$$

Proof. For $j \in J$, first consider

$$[e(CY^{(-1)})]_j = \textstyle\sum_{i\in S} e_i \sum_{k\in S} C_{ik} Y^{(-1)}_{kj} = z_j(x)\sum_{L\in\Gamma}\sum_{i\in L} e_i \sum_{K\in\Gamma}\sum_{k\in K} C_{ik}D_{KJ},$$

where the last equality follows from (3). Write e_i for $i \in L$ as $\delta_i\omega_L + \Delta_L z_i(x)$. Now partition that last summation into two parts: one for the contribution of $\delta_i\omega_L$, and one for the contribution of $\Delta_L z_i(x)$. The first summation is easily seen to equal $z_j(x)f_J(e)$. Next we show that the second part of the summation equals $-z_j(x)\Delta_J$. Indeed,

$$\begin{aligned} z_j(x)\sum_{L\in\Gamma}\Delta_L\sum_{i\in I} z_i(x)\sum_{K\in\Gamma} D_{KJ}\sum_{k\in K} C_{ik} &= z_j(x)\sum_{L\in\Gamma}\Delta_L\sum_{K\in\Gamma} D_{KJ}\sum_{i\in I} z_i(x)\sum_{k\in K} C_{ik} \\ &= z_j(x)\sum_{L\in\Gamma}\Delta_L\sum_{K\in\Gamma} D_{KJ}(Q_{LK}(1)-\Omega_{LK}) \end{aligned} \tag{8}$$

for $\Omega_{LK} = 0$ if $L \neq K$ and $\Omega_{LL} = 1$. Note that the last equality in (8) follows from the definition of the aggregation step. Then by (1) for the matrix $Q(1)$, $(Q(1) - I)D = E(Q(1)) - I$. Hence, (8) equals $z_j(x)\Delta(E(Q(1)) - I)^J$, where $(E(Q(1)) - I)^J$ is the column of the matrix $E(Q(1)) - I$ corresponding to J. But $\Delta E(Q(1)) = 0$, as all the rows of $E(Q(1))$ are identical and as the sum of the entries of Δ is zero. Thus, for $j \in J$, (8) equals $-z_j(x)\Delta_J$.

Finally, for $e_j > 0$,

$$[e(I+CY^{(-1)})]_j = e_j\left[1+\frac{(eCY^{(-1)})_j}{e_j}\right].$$

Hence from the above, we have

$$\begin{aligned} \|e(I+CY^{(-1)})\| &\leqq \|e\| \max_{J\in\Gamma}\max_{j\in J}\left|1+\frac{-z_j(x)\Delta_J + z_j(x)f_J(e)}{\Delta_J z_j(x)+\delta_j\omega_J}\right| \\ &= \|e\| \max_{J\in\Gamma}\max_{j\in J}\left|\frac{\delta_j\omega_J + z_j(x)f_J(e)}{\Delta_J z_j(x)+\delta_j\omega_J}\right|, \end{aligned}$$

completing the proof of (6). Inequality (7) is then an immediate conclusion of (6), Theorem 1, and the definition of $\delta_J(e)$. □

Inequality (7) says that the right time to apply the Rayleigh–Ritz step is when the ratio between the local and global errors is of the order $O(\varepsilon)$, making the two summands on the right-hand side of (7) the same magnitude.

Note that such a sharp bound cannot be obtained directly from Theorem 2. Theorem 2 implies that

$$\|\tilde{e}\| \leqq \|e(I-\Pi)\| \; \|Y(P\Pi)\|.$$

Indeed, in Haviv (1987) it is shown that $\|e(I-\Pi)\| = O(\max_{J\in\Gamma} \max_{j\in J} |\delta_j/\Delta_J|)$, but from (2) we obtain $\|Y(P\Pi)\| = O(1/\varepsilon)$, which is large in the nearly uncoupled case.

Finally, we note that the closest work to our own is McAllister, Stewart, and Stewart (1984). They, following Simon and Ando (1961) by using spectral decomposition and by assuming some technical and structural properties on P, considered the fast and slow phases of the convergence of the power method when applied to nearly uncoupled chains. The fast phase is intimately related to the local approximation notion: its (almost) termination implies that the approximation in hand induces a good local approximation (see their expression 4.1), but the converse is not necessarily true. Also, they show that the Rayleigh–Ritz step can ruin some of the convergence that was achieved in the fast phase (see there the discussion following the proof of Theorem 4.2).

5. A numerical example. Let $P \in R_+^{9\times 9}$ be the following stochastic matrix:

$$\left[\begin{array}{lll|lll|lll}
.85 & 0 & .149 & .0009 & 0 & 0 & .00005 & 0 & .00005 \\
.1 & .65 & .249 & 0 & .0009 & 0 & .00005 & 0 & .00005 \\
.1 & .8 & .0996 & .0003 & 0 & 0 & 0 & .0001 & 0 \\
\hline
0 & .0004 & 0 & .35 & .2995 & .35 & 0 & .0001 & 0 \\
.0005 & 0 & .0004 & .3999 & .3 & .3 & .0001 & 0 & 0 \\
0 & .0005 & .0004 & .299 & .4 & .3 & 0 & 0 & .0001 \\
\hline
0 & .00005 & 0 & 0 & .00005 & 0 & .6 & .2499 & .15 \\
.00003 & 0 & .00003 & .00004 & 0 & 0 & .1 & .8 & .0999 \\
0 & .00005 & 0 & 0 & .00005 & 0 & .1999 & .25 & .55
\end{array}\right].$$

P is nearly uncoupled with respect to the partition $\Gamma = [(1,2,3),(4,5,6),(7,8,9)]$ and $\varepsilon \cong .001$. For $x = (.1332, .1389, .0609, .1169, .1102, .1059, .0802, .1857, .0681)$,

$$\|e\| \equiv \max_{j\in S} e_j - \min_{j\in S} e_j = .1303$$

and $\max_{J\in\Gamma} \max_{j\in S} |\delta_j/\Delta_J| = .0162$. The value of $\tilde{x}$ is $(.0966, .1007, .0442, .0907, .0855, .0822, .1200, .2780, .1020)$ and then $\|\tilde{e}\| = .0006$. In particular, $\|\tilde{e}\|/\|e\| = .0046$, which is 28 percent of $\max_{J\in\Gamma} \max_{j\in S} |\delta_j/\Delta_J|$.

Acknowledgment. The author acknowledges a key comment made by Y. Ritov.

REFERENCES

S. L. Campbell and C. D. Meyer (1979), *Generalized Inverses of Linear Transformations*, Pitman, London.

W. L. Cao and W. J. Stewart (1985), *Iterative aggregation/disaggregation techniques for nearly uncoupled Markov chains*, J. Assoc. Comput. Mach., 32, pp. 702–719.

F. Chatelin (1984), *Iterative aggregation/disaggregation methods*, in Mathematical Computer Performance and Reliability, G. Iazeolla, P. J. Courtois, and A. Hordijk, eds., Elsevier, North Holland.

F. CHATELIN AND W. L. MIRANKER (1984), *Aggregation/disaggregation for eigenvalue problems*, SIAM J. Numer. Anal., 21, pp. 567–582.

P. J. COURTOIS (1977), *Decomposability*, Academic Press, New York.

M. HAVIV AND L. VAN DER HEYDEN (1984), *Perturbation bounds for the stationary probabilities of a finite Markov chain*, Adv. in Appl. Probab., 16, pp. 804–818.

M. HAVIV (1987), *Aggregation/disaggregation methods for computing the stationary distribution of a Markov chain*, SIAM J. Numer. Anal., 24, pp. 952–966.

D. F. MCALLISTER, G. W. STEWART, AND W. J. STEWART (1984), *On a Rayleigh-Ritz refinement technique for nearly uncoupled stochastic matrices*, Linear Algebra Appl., 60, pp. 1–25.

B. PARLETT (1980), *The Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs, NJ.

P. J. SCHWEITZER (1981), *Perturbation series expansions of nearly completely decomposable Markov chains*, Working Paper Series 8122, Graduate School of Management, University of Rochester, Rochester, NY.

H. SIMON AND A. ANDO (1961), *Aggregation of variables in dynamics systems*, Econometrica, 29, pp. 111–138.

G. M. STEWART (1983), *Computable error bounds for aggregate Markov chains*, J. Assoc. Comput. Mach., 30, pp. 271–285.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 3, pp. 294–315, July 1989

003

ANALYSIS AND PROPERTIES OF THE GENERALIZED TOTAL LEAST SQUARES PROBLEM $AX \approx B$ WHEN SOME OR ALL COLUMNS IN A ARE SUBJECT TO ERROR*

SABINE VAN HUFFEL† AND JOOS VANDEWALLE†

Abstract. The Total Least Squares (TLS) method has been devised as a more global fitting technique than the ordinary least squares technique for solving overdetermined sets of linear equations $AX \approx B$ when errors occur in all data. This method, introduced into numerical analysis by Golub and Van Loan, is strongly based on the Singular Value Decomposition (SVD). If the errors in the measurements A and B are uncorrelated with zero mean and equal variance, TLS is able to compute a strongly consistent estimate of the true solution of the corresponding unperturbed set $A_0X = B_0$. In the statistical literature, these coefficients are called the parameters of a classical errors-in-variables model.

In this paper, the TLS problem, as well as the TLS computations, are generalized in order to maintain consistency of the parameter estimates in a general errors-in-variables model; i.e., some of the columns of A may be known exactly and the covariance matrix of the errors in the rows of the remaining data matrix may be arbitrary but positive semidefinite and known up to a factor of proportionality. Here, a computationally efficient and numerically reliable Generalized TLS algorithm GTLS, based on the Generalized SVD (GSVD), is developed. Additionally, the equivalence between the GTLS solution and alternative expressions of consistent estimators, described in the literature, is proven. These relations allow the main statistical properties of the GTLS solution to be deduced. In particular, the connections between the GTLS method and commonly used methods in linear regression analysis and system identification are pointed out. It is concluded that under mild conditions the GTLS solution is a consistent estimate of the true parameters of any general multivariate errors-in-variables model in which all or some subsets of variables are observed with errors.

Key words. total least squares, generalized singular value decomposition, errors in variables, 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 $AX \approx B$. Whenever *both* the data matrix A and observation matrix B are *inaccurate*, the Total Least Squares (TLS) technique is appropriate for solving this set. The problem of *linear parameter estimation* arises in a broad class of scientific disciplines such as signal processing, automatic control, system theory, general engineering, statistics, physics, economics, biology, and medicine. It can be described by a linear equation:

$$a_1x_1 + \cdots + a_nx_n = b \tag{1}$$

where $a_1, \cdots, a_n$ and b denote the variables and $x = [x_1, \cdots, x_n]^T \in \mathscr{R}^n$ plays the role of a parameter vector that characterizes the special system ($\mathscr{R}$ denotes the set of real numbers). A basic problem of applied mathematics is to determine an estimate of the true but unknown parameters from certain measurements of the variables. This gives rise to an overdetermined set of m linear equations ($m \geqq n$):

$$Ax \approx b \tag{2}$$

* Received by the editors March 29, 1988; accepted for publication (in revised form) November 4, 1988. The first author is a Senior Research Assistant of the Belgian National Fund of Scientific Research.

† Electronica-Systemen-Automatisatie-Technologie Laboratory, Department of Electrical Engineering, Katholieke Universiteit Leuven, Kardinaal Mercierlaan 94, B-3030 Heverlee, Belgium (vanhuffel@kulesat.uucp and vanhuffel%kulesat.uucp@BLEKUL60).

where the ith row of the data matrix $A \in \mathcal{R}^{m\times n}$ and the vector of observations $b \in \mathcal{R}^m$ contain the measurements of the variables $a_1, \cdots, a_n$ and b, respectively.

In the classical least squares (LS) approach the measurements A of the variables a_i (the left-hand side of (2)) are assumed to be free of error and, hence, all errors are confined to the observation vector b (the right-hand side of (2)). However, this assumption is frequently unrealistic: sampling errors, human errors, modelling errors, and instrument errors may imply inaccuracies of the data matrix A. For those cases, TLS has been devised and amounts to fitting a "best" subspace to the measurement points (A_i^T, b_i), $i = 1, \cdots, m$, where A_i is the ith row of A.

Much of the literature concerns the classical TLS problem in which all variables are observed with errors (see, e.g., [31], [14], [11], [34]) with particular emphasis on the univariate linefitting problem, i.e., $n = 1$ in (1) (see, e.g., [1], [29]). If the errors on the measurements A and b are uncorrelated with zero mean and equal variance, then under mild conditions the TLS solution $\hat{x}$ of (2) is a strongly consistent estimate of the true but unknown parameters x in (1), i.e., $\hat{x}$ converges to x with probability one as the number of equations m tends to infinity. However in many linear parameter estimation problems, some of the variables a_i in (1) may be observed without error. This implies that *some of the columns* of A in (2) are assumed to be *known exactly*. For instance, every intercept model

$$\alpha + a_1x_1 + \cdots + a_nx_n = b \tag{3}$$

gives rise to an overdetermined set of equations

$$[1_m; A]\begin{bmatrix}\alpha\\ x\end{bmatrix} = b \tag{4}$$

with $1_m = [1, \cdots, 1]^T$ in which the first column of the left-hand side matrix is known exactly [11], [13]. To maximize the accuracy of the estimated parameters x it is natural to require that the corresponding columns of A be unperturbed since they are known exactly. 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 classical TLS formulation can be *generalized* as follows (E denotes the expected value operator).

GENERALIZED TLS FORMULATION. Given a set of m linear equations in $n \times d$ unknowns X:

$$AX \approx B, \qquad A \in \mathcal{R}^{m\times n}, B \in \mathcal{R}^{m\times d}, \text{ and } X \in \mathcal{R}^{n\times d}, \tag{5}$$

$$\text{Partition } A = [A_1; A_2], \qquad A_1 \in \mathcal{R}^{m\times n_1}, \quad A_2 \in \mathcal{R}^{m\times n_2}, \quad n = n_1 + n_2,$$

$$X = [X_1^T; X_2^T]^T, \qquad X_1 \in \mathcal{R}^{n_1\times d}, \quad X_2 \in \mathcal{R}^{n_2\times d}. \tag{6}$$

Assume that the columns of A_1 are known exactly and that the covariance matrix $E(\Delta^T\Delta)$ of the errors $\Delta \in \mathcal{R}^{m\times(n_2+d)}$ in the perturbed data matrix $[A_2; B]$ is given by $C \in \mathcal{R}^{(n_2+d)\times(n_2+d)}$, up to a factor of proportionality. Let $C = R_C^TR_C$ be nonsingular. Then a generalized TLS (GTLS) solution of (5)–(6) is any solution of the set

$$\hat{A}X = A_1X_1 + \hat{A}_2X_2 = \hat{B} \tag{7}$$

where $\hat{A} = [A_1; \hat{A}_2]$ and $\hat{B}$ are determined such that

$$\text{Range}(\hat{B}) \subseteq \text{Range}(\hat{A}), \tag{8}$$

$$\|[\Delta\hat{A}_2; \Delta\hat{B}]R_C^{-1}\|_F = \|[A_2 - \hat{A}_2; B - \hat{B}]R_C^{-1}\|_F \text{ is minimal.} \tag{9}$$

The *problem of finding* $[\Delta\hat{A}_2;\ \Delta\hat{B}]$ such that (8)-(9) are satisfied, is referred to as the GTLS *problem*. Whenever the solution is *not unique*, GTLS singles out the *minimum norm* solution, denoted by $\hat{X} = [\hat{X}_1^T; \hat{X}_2^T]^T$.

An even more general GTLS formulation, that allows for correlations between the errors in each column of $[A_2;\ B]$, is given in [35]. It is worth noting that when all columns of A are known exactly and when $C \sim I$, the GTLS solution reduces to the ordinary Least Squares (LS) estimate. By varying n_1 from zero to n, this formulation can handle the ordinary LS problem, as well as every TLS ($C \sim I$) and GTLS problem.

Although the name "total least squares" has appeared only recently in the literature [14], this method of fitting is certainly not new and has a long history in the statistical literature where the method is known as *orthogonal regression* or *errors-in-variables* regression. Indeed, the univariate linefitting problem ($n = 1$) was already scrutinized in the previous century [1]. About 20 years ago, the technique was extended to multivariate problems ($n > 1$) and later on to multidimensional problems that deal with more than one observation vector ($d > 1$ in (5)), e.g., [31], [11]. More recently, the TLS approach to fitting has also attracted interest outside of statistics. In the field of *numerical analysis*, this problem was first studied by Golub and Van Loan [14]. Their analysis, as well as their algorithm, is strongly based on the *Singular Value Decomposition* (SVD). Geometrical insight into the properties of the SVD has brought us independently to the same concept. We have *generalized* the algorithm of Golub and Van Loan [14] to all cases in which their algorithm fails to produce a solution, described the properties of these so-called nongeneric TLS problems and proved that the proposed generalization still satisfies the TLS criteria (8)-(9) if additional constraints are imposed on the solution space [39]-[40].

Although this linear algebraic approach is quite different, it is easy to see that the multivariate errors-in-variables regression estimate, given by Gleser [11], coincides with the TLS solution given by Golub and Van Loan [14] whenever the TLS problem has a unique minimizer. The only difference between both methods is the algorithm used: Gleser's method is based on an eigenvalue-eigenvector analysis, while the TLS Algorithm uses the SVD, which is numerically more robust. Furthermore, the TLS algorithm computes the minimum norm solution whenever the TLS problem lacks a unique solution. These extensions are not considered by Gleser. Also in the field of *experimental modal analysis*, the TLS technique (more commonly known as the H_v technique), has recently been studied [25]. And finally in the field of *system identification*, Levin [26] first studied the same problem. His method, called the *eigenvector method* or the *Koopmans-Levin method* [6], computes the same estimate as the TLS Algorithm whenever the TLS problem has a unique solution.

Much less considered is the case in which *some* of the columns of A in (5) are *known exactly*. It is quite easy to generalize the classical TLS Algorithms, given in [14], [34], and [39], in order to compute the more general TLS estimate $\hat{X} = [\hat{X}_1^T;\ \hat{X}_2^T]^T$ satisfying the TLS criteria (7)-(8)-(9) with $R_C \sim I$. The technique involves computing a QR *factorization of the "known" columns* A_1 and then solving a TLS *problem of reduced dimension* [12], [34, § 1.7]. Using a generalization of the Eckart–Young-Mirsky matrix approximation theorem [13], Golub, Hoffman, and Stewart have proved that this procedure indeed finds the best rank $r\ (\leqq n)$ approximation $[A_1; \hat{A}_2; \hat{B}]$ to $[A; B]$ that leaves A_1 fixed such that

$$(10)\qquad \|[A_1;A_2;B]-[A_1;\hat{A}_2;\hat{B}]\|_F = \min_{\text{rank}([A_1;\tilde{A}_2;\tilde{B}])\leqq r} \|[A_1;A_2;B]-[A_1;\tilde{A}_2;\tilde{B}]\|_F.$$

In particular, this algorithm is able to compute the *Compensated Least Squares* (CLS) estimate as derived by Guidorzi [17] and Stoica and Söderström [33]. When the only disturbance of the input-output sequences is given by zero mean white noise of equal variance, the CLS, GTLS and eigenvector methods all give the same estimate. Observe that our TLS Algorithm, that allows for exactly known columns in A and coincides with our GTLS Algorithm in § 2 for the case that $C \sim I$, is computationally more efficient than the computation procedure presented in [33].

The *generalization* of the TLS problem, presented in this paper, that allows for *correlations* between the measurement errors in the data A and B, is inspired by a generalization of the classical errors-in-variables model discussed in [9]. As said before, the TLS solution is not very meaningful unless the errors in the measurements A and B in (5) are independently derived and equilibrated. In statistical terms, this means that the errors must be uncorrelated with zero mean and all have the same variance, i.e., the associated error covariance matrix C in (6) is $\sim I$. The best statistical approach, directly related to the classical TLS concept, is the "*errors-in-variables*" model [11] that considers all observations as coming from some unknown true values plus measurement errors. The true values are assumed to follow a linear relation (1). The estimation of the parameters in this model is a problem with a long history in the statistical literature [1], yet one with a considerable recent emphasis. Much less considered is the following *general "errors-in -variables" model*, directly related to our generalized TLS formulation given above, in which *some subset* of variables is observed *with errors*.

General Errors-In-Variables Model Formulation.

$$\begin{aligned} \underset{m\times d}{B_0} &= A_0X = \underset{m\times n_1}{A_1}\ \underset{n_1\times d}{X_1} + \underset{m\times n_2}{(A_2)_0}\ \underset{n_2\times d}{X_2} \\ A_2 &= (A_2)_0 + \Delta A_2 \\ B &= B_0 + \Delta B. \end{aligned} \tag{11}$$

X_1 and X_2 are the true but unknown parameters to be estimated; A_1 and $(A_2)_0$ are of full column rank. They consist of constants as well as B_0. A_1 is known but $(A_2)_0$ and B_0 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 C_Δ, up to a factor of proportionality c^2, i.e.,

$$C_\Delta = c^2 C = c^2 \begin{bmatrix} C_a & C_{ab} \\ C_{ab}^T & C_b \end{bmatrix} \text{ with } C_{(n_2+d)\times(n_2+d)} \text{ known.} \tag{12}$$

Observe that this model requires that the rows of $[\Delta A_2;\ \Delta B]$ are independently derived. If this assumption is not satisfied, we can premultiply the data $[A;\ B]$ in advance with an appropriate $m \times m$ matrix D such that the preprocessed data $D[A; B] = [DA_1; DA_2; DB]$ satisfy the assumptions of model (11). If D is ill-conditioned, this premultiplication must be performed implicitly, as outlined in [35]. This preprocessing operation does not affect the true solution X of model (11).

Now to compute strongly consistent estimates of the true but unknown parameters X of model (11)–(12), the classical TLS Algorithm, as given in [14], [34, § 1.8], and [39] can be used whenever $C_\Delta \sim I$. However, in case that the covariance matrix C_Δ is more *general*, the classical TLS algorithm may *not* be used *straightforwardly*. To maintain consistency, the data can be *pretreated* appropriately such that the covariance matrix of the transformed data is diagonal with equal error variances (i.e., $C_\Delta = c^2 I$). The classical TLS Algorithm can then be used to solve this transformed set of equations and finally the solution of the transformed system must be converted to a solution of the original set of equations. Such transformation procedures are described in [10], [11], and [34,

§ 4.5] for the case that $n_1 = 0$ and C in (12) is positive definite. This approach, however, is not recommended in general since computing $[A;\ B]R_C^{-1}$ (with $C = R_C^T R_C$) usually leads to unnecessarily large numerical errors if R_C is ill-conditioned with respect to the solution of equations.

The *objective* of this paper is to *solve the generalized* TLS *problem*, defined above, by making use of the *Generalized* SVD (GSVD). Hereto, a computationally efficient and numerically reliable *Generalized Total Least Squares* (GTLS) *Algorithm* is developed. As shown in § 3, this algorithm is able to compute consistent estimates of the parameters in any errors-in-variables model (11) directly *without transforming* the data *explicitly*. The great advantage of the GSVD is that it replaces these transformation procedures by *one*, which is numerically reliable and can more easily handle (nearly) *singular* covariance matrices C in (12). Moreover, the GSVD reveals the *structure* of the general errors-in-variables model (11) more clearly than the usual transformation procedures. Additionally, statistical properties of the GTLS solution are deduced by proving the equivalence between the GTLS solution and alternative expressions of consistent estimators described in the statistical literature [9], [33]. The GSVD of a matrix pair (A, B) is defined as follows [15], [41].

THEOREM 1. GSVD of (A, B). *If* $A \in \mathcal{R}^{m\times n}(m \geqq n)$ *and* $B \in \mathcal{R}^{p\times n}$, *then there exist orthogonal* $T \in \mathcal{R}^{m\times m}$ *and* $W \in \mathcal{R}^{p\times p}$ *and a nonsingular* $Z \in \mathcal{R}^{n\times n}$ *such that*

$$T^T A Z = D_A \quad \textit{and} \quad W^T B Z = D_B \tag{13}$$

with

$$D_A = \operatorname{diag}(\alpha_1, \cdots, \alpha_n) \in \mathcal{R}^{m\times n}, \qquad \alpha_i \geqq 0$$

and

$$D_B = \operatorname{diag}(\beta_1, \cdots, \beta_q) \in \mathcal{R}^{p\times n}, \qquad \beta_i \geqq 0 \quad q = \min\{p, n\}$$

$$\beta_1 \geqq \cdots \geqq \beta_r > \beta_{r+1} = \cdots = \beta_q = 0, \qquad r = \operatorname{rank}(B).$$

The assumption $m \geqq n$ is not restrictive from the applications point of view. The elements of the set $\sigma(A, B) = \{\alpha_i/\beta_i,\ i = 1, \cdots, r\}$ are referred to as the ordinary *generalized singular values* of A and B. The remaining generalized singular values α_i/β_i in which α_i is nonzero (respectively, zero) and β_i is zero, are called *infinite* (respectively, *indefinite*) [3]. It is worth emphasizing that infinite generalized singular values are not necessarily badly behaved. In fact, the infinite generalized singular values of (A, B) are the zero generalized singular values of (B, A) since the roles of A and B are interchangeable. Theorem 1 is a *generalization of the ordinary* SVD in that if $B = I_n$ then $\sigma(A, B)$ equals the singular value spectrum $\sigma(A)$ of matrix A. Note that there exists an intimate theoretical link between the GSVD of the matrix pair (A, B) and the *generalized symmetric eigenvalue problem* [15]. Indeed

$$\sigma_i \in \sigma(A, B) \Leftrightarrow \sigma_i^2 \in \lambda(A^T A, B^T B) \tag{14}$$

where $\lambda(A^T A, B^T B)$ is the set of generalized eigenvalues of the matrix pair $(A^T A, B^T B)$ and correspondingly

$$A^T A z_i = \sigma_i^2 B^T B z_i \tag{15}$$

where the generalized eigenvector z_i is given by the ith column of the matrix Z in Theorem 1. This matrix diagonalizes $A^T A$ and $B^T B$ simultaneously. The value of the GSVD is that these diagonalizations can be achieved without forming $A^T A$ and $B^T B$. These connections to the generalized symmetric eigenvalue problem allow us to prove the interesting statistical properties of the GTLS solution (see § 3).

In this paper, the main emphasis is put on the *linear algebraic approach* of the GTLS problem. The statistical analysis of the GTLS solution is limited and relies on showing the correspondence of particular cases of our GTLS problem with that of others in the literature. For a more detailed statistical appraisal and consistency analysis of the GTLS solution in different applications, the reader is referred to the list of references, e.g. [2], [9], [11], [22], [33]. This paper is organized into four sections. In § 2, the GTLS Algorithm is presented. Its main difference with respect to the classical TLS Algorithm is the fact that the GSVD is used instead of the ordinary SVD. Section 3 describes the *properties* of the GTLS solution. Alternative expressions of the GTLS solution are deduced that allow us to derive the main *statistical* properties of the GTLS solution. Finally, § 4 gives the conclusions.

2. The generalized TLS Algorithm GTLS. The original motivation is to compute *consistent estimates* of the parameters in a *general errors-in-variables model* of the form (11), thereby improving the computational efficiency and numerical robustness of the methods currently used in statistics, linear regression [9], and identification [6], [30], [33]. Therefore, the classical TLS problem has been generalized in the previous section. In this section, the GTLS Algorithm which solves the GTLS problem is described in detail. As shown below, the GTLS Algorithm mainly performs *orthogonal transformations* and does not need to square or invert matrices (as done in [11]), which guarantees its *numerical reliability*. Moreover, by first performing a QR factorization [15] we need only to compute the GSVD *of a smaller submatrix*, which makes the GTLS Algorithm *computationally more efficient* than methods described in [6], [30], and [33]. The GTLS Algorithm is outlined below.

ALGORITHM. GTLS

Given

- An $m \times d$ matrix B and an $m \times n$ matrix $A = [A_1; A_2]$ whose first n_1 columns A_1 have full column rank and are known exactly, $n = n_1 + n_2$ and $m \geqq n_1$.
- An $(n_2 + d) \times (n_2 + d)$ matrix C, proportional with the covariance matrix $E(\Delta^T\Delta)$ of the errors $\Delta_{m \times (n_2+d)}$ in the matrix $[A_2; B]$, or any square root R_C of C (such that $C = R_C^T R_C$).

Step 1: QR factorization and Cholesky decomposition.

```
1.a. If $n_1 > 0$   then   begin
                           compute the QR factorization of $[A_1; A_2; B]$:
```

$$Q^T[A_1; A_2; B] = R = \begin{array}{cc} \left[\begin{array}{cc} R_{11} & R_{12} \\ 0 & R_{22} \end{array}\right] & \begin{array}{l} n_1 \\ m-n_1 \end{array} \\ \begin{array}{cc} n_1 & \quad n_2+d \end{array} & \end{array} \tag{16}$$

```
                           with $Q$ being orthogonal and $R_{11}$ upper triangular.
                           if $n_1 = n$ then begin
                                          solve $R_{11}\hat{X} = R_{12}$ by back substitution
                                          stop
                                          end
                           end
                    else   $R_{22} \leftarrow [A; B]$
```

1.b. If C is given and $\not\propto I_{n_2+d}$, compute the Cholesky decomposition of C:

$$C = R_C^T R_C \tag{17}$$

Step 2: GSVD.

2.a. Compute the GSVD (or SVD if $C \sim I$) of the matrix pair (R_{22}, R_C) as in (13):

$$\begin{aligned} &T^T R_{22} Z = \operatorname{diag}(\alpha_1, \cdots, \alpha_s), \alpha_{s+1} = \cdots = \alpha_{n_2+d} = 0, \qquad s = \min\{m - n_1, n_2 + d\} \\ &W^T R_C Z = \operatorname{diag}(\beta_1, \cdots, \beta_{n_2+d}) \end{aligned} \tag{18}$$

where the generalized singular values $\sigma_i = \alpha_i/\beta_i$, $i = 1, \cdots, n_2 + d$ are organized in decreasing order of magnitude (i.e., $\sigma_i \geqq \sigma_{i+1}$) and the corresponding columns, z_i, of Z are normalized to unit norm.

2.b. If not user determined, compute the rank $r(\leqq n_2)$ of the matrix pair (R_{22}, R_C) by

$$\sigma_1 \geqq \cdots \geqq \sigma_r > R_0 \geqq \sigma_{r+1} \geqq \cdots \geqq \sigma_{n_2+d} \tag{19}$$

with R_0 a user-defined rank determinator.

2.c. Solve by back substitution:

$$Z_2 \leftarrow [z_{r+1}, \cdots, z_{n_2+d}]; \; R_{11} Z_1 = -R_{12} Z_2 \tag{20}$$

Step 3: GTLS solution $\hat{X} = [\hat{X}_1^T; \hat{X}_2^T]^T$.

3.a. If $C \nsim I_{n+d}$, $d > 1$ and $r < n_2$, orthonormalize $\left[\begin{smallmatrix} Z_1 \\ Z_2 \end{smallmatrix}\right]$ using a QR factorization:

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = Q_z R_z; \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \leftarrow Q_z \qquad Q_z^T Q_z = I_{n_2 - r + d} \text{ and } R_z \text{ upper triangular}$$

3.b. Perform Householder transformations such that

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} Q = \begin{array}{c c} \begin{bmatrix} W & Y \\ 0 & \Gamma \end{bmatrix} & \begin{matrix} n \\ d \end{matrix} \\ \begin{matrix} n_2 - r & d \end{matrix} & \end{array} \tag{21}$$

with Q being orthogonal and Γ d by d upper triangular.

```
      If Γ is nonsingular  then  {GTLS problem is generic}
(22)                             solve $\hat{X}\Gamma = -Y$
                           else  {GTLS problem is nongeneric}
                                 begin
                                 $r \leftarrow r - \rho$   where $\rho$ is the multiplicity of $\sigma_r$
                                 go back to Step 2.c.
                                 end
```

END

The following comments are in order:

- The GSVD can readily be applied to (5) to yield the solution of the generalized TLS problem. Just as the SVD is a valuable tool for the solution and analysis of the classical TLS problem, so the GSVD plays the same role for the generalized problem. Stable numerical methods have emerged for computing the GSVD [3], [27], [32], [42]. The methods proposed in [3] and [27], being based on an implicit *Kogbetliantz approach*, have potential for systolic implementation [4]. For background information on the GSVD the reader is strongly recommended to consult [28] and [41].

- To compute the decomposition (17) of any positive semidefinite covariance matrix C, a *Cholesky decomposition with complete pivoting* can be used. This method is proven to be one of the most numerically stable methods [5, p. 3.16]. Software for computing this decomposition is readily available, notably in the LINPACK library [5, Chap. 8]. An error analysis of this Cholesky decomposition is given in [19]. Whenever C is singular, R_C is not of full row rank. Hence, indefinite generalized singular values (i.e., $\alpha_i = \beta_i = 0$ in (18)) may occur. These values and corresponding columns in Z are to be considered in the GTLS computation.
- The QR *factorization* of $[A_1; A_2; B]$ can be computed with the LINPACK routine SQRDC [5, Chap. 9]. *Pivoting* may be done within three groups of columns: the first n_1 columns A_1, the next n_2 columns A_2 and the last d columns B. Columns may not be pivoted with columns from another group. Even for full rank problems, column pivoting seems to produce more accurate solutions [21]. If pivoting is done, the columns in R_C must be permuted correspondingly if $C \not\sim I_{n_2+d}$, and the inverse permutations must be performed in the last step of the GTLS Algorithm.
- To orthonormalize the columns of $[{Z_1 \atop Z_2}]$ in Step 3.a., a QR factorization is performed. This can again be computed with the LINPACK routine SQRDC [5, Chap. 9].
- If *no columns* of A *are known exactly* ($n_2 = n$) and $C \sim I_{n+d}$, the GTLS Algorithm reduces to the classical TLS Algorithm, given in [14], [34, § 1.8.1], and [39].
- If a subset A_1 of A is known exactly and $C \sim I_{n_2+d}$, the GTLS Algorithm reduces to the TLS *Algorithm with exactly known columns*, as described in [34, § 1.8.2]. Observe also that the GTLS algorithm solves the *ordinary* LS *problem*, using a QR factorization, if all columns of A are known exactly ($n_1 = n$).
- Mostly, the matrix pair (R_{22}, R_C) has maximal rank $r = n_2$. If $r < n_2$ (e.g., when the set of equations $AX \approx B$ is *underdetermined*), the GTLS solution is *no longer unique*. In this case, GTLS singles out the *minimum norm* solution. Indeed if the solution $[{\hat{X} \atop -I}]$ is deduced from an orthonormal basis $[{Y \atop \Gamma}]\,{n \atop d}$, then

$$\|\hat{X}\|_F^2 = \|\Gamma^{-1}\|_F^2 - d \quad \text{and} \quad \|\hat{X}\|_2^2 = (1-\sigma_{\min}^2(\Gamma))/\sigma_{\min}^2(\Gamma)$$

as proved in [15, p. 422] and [34, p. 69], based on the CS decomposition [15, Thm. 2.4-1]. Hence we have to select d base vectors $[{Y \atop \Gamma}]\,{n \atop d}$ within the Range $([{Z_1 \atop Z_2}])$ such that $\|\Gamma^{-1}\|_F$ is minimized and the minimal singular value $\sigma_{\min}(\Gamma)$ of Γ is maximized. This is done by computing an orthonormal matrix Q (e.g., by using Householder transformations) such that

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} Q = \begin{bmatrix} Z_1 Q \\ Z_2 Q \end{bmatrix} \begin{matrix} n \\ d \end{matrix} = \begin{array}{c} \begin{bmatrix} W & Y \\ 0 & \Gamma \end{bmatrix} \begin{matrix} n \\ d \end{matrix} \\ \begin{matrix} n_2 - r & \quad d \quad \end{matrix} \end{array}$$

Z_2 and Z_2Q have the same singular values. Denote by $\tilde{\gamma}_i$, $i = 1, \cdots, d$, the singular values of a submatrix $\tilde{\Gamma}$ of Z_2 or Z_2Q (obtained by deleting $n_2 - r$ columns) and by σ_i, $i = 1, \cdots, d$, the singular values of Z_2 or Z_2Q, in decreasing order of magnitude. Then, the interlacing property for singular values [15, p. 286] yields

$$\tilde{\gamma}_i \leqq \sigma_i \text{ or equivalently } \tilde{\gamma}_i^{-1} \geqq \sigma_i^{-1}\, i = 1, \cdots, d.$$

Hence, $\|\tilde{\Gamma}^{-1}\|_F^2 = \sum_{i=1}^d \tilde{\gamma}_i^{-2}$ and $\tilde{\gamma}_d^{-2}$ are minimized if $\tilde{\gamma}_i = \sigma_i$, for all i. Since the d by d submatrix Γ, defined above, has the same singular values as Z_2, it follows directly that the TLS solution $\hat{X} = -Y\Gamma^{-1}$, computed from (21)–(22), has minimal norm $\|\hat{X}\|_2$ and $\|\hat{X}\|_F$.

Observe that the expressions of $\|\hat{X}\|_2$ and $\|\hat{X}\|_F$ are deduced from the orthonormality of $[\begin{smallmatrix} Y \\ \Gamma \end{smallmatrix}]$. Therefore the base vectors in $[\begin{smallmatrix} Z_1 \\ Z_2 \end{smallmatrix}]$ must first be orthonormalized in Step 3.a.

If $C \sim I_{n+d}$, the columns in $[\begin{smallmatrix} Z_1 \\ Z_2 \end{smallmatrix}]$ are already orthonormal since they are the right singular vectors of $R_{22} = [A; B]$, obtained from its SVD in Step 2.a.

If $d = 1$, Γ is a scalar. To minimize $\|\hat{X}\|_F$, this scalar must be maximized. This can be accomplished by (21) such that the last column $[\begin{smallmatrix} Y \\ \Gamma \end{smallmatrix}]\,{}^{n}_{1}$ has the largest $(n + 1)$th component of all unit vectors within the Range $([\begin{smallmatrix} Z_1 \\ Z_2 \end{smallmatrix}])$. Therefore, the columns of $[\begin{smallmatrix} Z_1 \\ Z_2 \end{smallmatrix}]$ need not be orthogonal.

If $r = n$, $\hat{X}$ is unique. The columns of $[\begin{smallmatrix} Z_1 \\ Z_2 \end{smallmatrix}]$ need not be orthonormalized since the GTLS solution $\hat{X}$ is invariant with respect to any base transformation P in its solution space. Indeed, with $[\begin{smallmatrix} Y \\ \Gamma \end{smallmatrix}]$ in (21) a basis (not necessarily orthogonal) of the TLS solution space, Range $([\begin{smallmatrix} \hat{X} \\ -I \end{smallmatrix}])$, we have that $[\begin{smallmatrix} Y \\ \Gamma \end{smallmatrix}]P = [\begin{smallmatrix} YP \\ \Gamma P \end{smallmatrix}]$. Hence the GTLS solution $\hat{X} = -(YP)(\Gamma P)^{-1} = -YPP^{-1}\Gamma^{-1} = -Y\Gamma^{-1}$ remains invariant.

- *If* Γ *in* (21) is nonsingular (respectively, *singular*), the GTLS solution is called generic (respectively, *nongeneric*). As shown in § 3, Γ can only be singular when A is (nearly) rank-deficient or when the set of equations $AX \approx B$ is highly incompatible. In this case, the generic GTLS solution does not exist but the GTLS computations are generalized in order to solve these *nongeneric* GTLS *problems* in the same way as the nongeneric TLS problem [39], [40].
- If A_1 does *not* have *full column rank*, i.e., rank $(A_1) = r_A < n_1$, we can always replace A_1 with a matrix having r_A independent columns selected from A_1, apply the GTLS Algorithm and set the coefficients of $\hat{X}_1$ corresponding to this missing column to zero without changing either the rank of the result or the norm of the difference (9). Note, however, that the GTLS solution $\hat{X}$ no longer has minimal norm in this case.
- Finally, observe that we only need to compute *a few vectors* z_i associated with the *smallest* generalized singular values of (R_{22}, R_C) in order to obtain the GTLS solution $\hat{X}$. Moreover, we only need to compute a *basis* of the solution space Range $([\begin{smallmatrix} Z_1 \\ Z_2 \end{smallmatrix}])$. Indeed, as proven before the GTLS solution $\hat{X}$ is invariant with respect to any base transformation P in its solution space.

Based on these properties, we were able to improve the efficiency of the TLS computations by computing the SVD only "*partially*" [36]. This results in the development of an improved algorithm *Partial Total Least Squares* (PTLS) [38]. PTLS is about two times faster than the classical TLS algorithm [14], [39], while the same accuracy can be maintained. The same modifications could be applied to the generalized SVD and GTLS Algorithms insofar as they are based on the QR Algorithm [15, § 8.2].

3. Properties of the generalized TLS solution. In this section a number of theorems are proven that *link the* GTLS *solution* with alternative *expressions of consistent estimators* given in literature. These links allow us to deduce the *main statistical properties* of the GTLS solution as shown below.

Throughout this section we will make use of the following notation and assumptions:

(23a) Consider the *set of equations* given in (5)–(6), and assume that A_1 has full column rank n_1 and is known exactly.

(23b) The n by n *identity matrix* is denoted by I_n.

(23c) Let the *covariance matrix* $E(\Delta^{*T}\Delta^*)$ of the errors Δ^* in $[A; B] = [A_1; A_2; B]$ be given by C^*, up to a factor of proportionality:

$$C^* = \begin{array}{cc} \left[\begin{array}{cc} C_a^* & C_{ab}^* \\ C_{ab}^{*T} & C_b^* \end{array}\right] & \begin{array}{c} n \\ d \end{array} \\ \begin{array}{cc} n & \quad d \end{array} & \end{array} = \begin{array}{cc} \left[\begin{array}{cc} 0 & 0 \\ 0 & C \end{array}\right] & \begin{array}{c} n_1 \\ n_2+d \end{array} \\ \begin{array}{cc} n_1 & n_2+d \end{array} & \end{array}$$

where

$$C = \begin{array}{cc} \left[\begin{array}{cc} C_a & C_{ab} \\ C_{ab}^T & C_b \end{array}\right] & \begin{array}{c} n_2 \\ d \end{array} \\ \begin{array}{cc} n_2 & \quad d \end{array} & \end{array}$$

is the positive definite covariance matrix $E(\Delta^T\Delta)$ of the errors Δ in $[A_2; B]$, up to a factor of proportionality.

(23d) Let R_C^* be any *square root* of C^*, defined by $C^* = R_C^{*T}R_C^*$ and partitioned as follows:

$$R_C^* = \begin{array}{cc} \left[\begin{array}{cc} R_{Ca}^* & R_{Cab}^* \\ R_{Cba}^* & R_{Cb}^* \end{array}\right] & \begin{array}{c} n \\ d \end{array} \\ \begin{array}{cc} n & \quad d \end{array} & \end{array} = \begin{array}{cc} \left[\begin{array}{cc} 0 & 0 \\ 0 & R_C \end{array}\right] & \begin{array}{c} n_1 \\ n_2+d \end{array} \\ \begin{array}{cc} n_1 & n_2+d \end{array} & \end{array}$$

where R_C is any square root of C, i.e., $C = R_C^T R_C$.

(23e) Denote by

$$\hat{X}_{n\times d} = \begin{array}{cc} \left[\begin{array}{c} \hat{X}_1 \\ \hat{X}_2 \end{array}\right] & \begin{array}{c} n_1 \\ n_2 \end{array} \end{array}$$

the GTLS solution, as computed by the GTLS algorithm (given in § 2).

(23f) Let $\sigma(A, B)$ (respectively, $\lambda(A, B)$) be the set of *generalized singular values* σ_i (respectively, *generalized eigenvalues* λ_i) of the matrix pair (A, B), organized in decreasing order of magnitude, i.e., $\sigma_i \geqq \sigma_{i+1}$ (respectively, $\lambda_i \geqq \lambda_{i+1}$). Analogously, if $B \sim I$, then $\sigma(A)$ (respectively, $\lambda(A)$) denotes the ordinary singular values (respectively, ordinary eigenvalues) of A.

Based on the properties of the (generalized) eigenvalue decomposition and the (generalized) SVD [15, § 8.6], the following links between the different problems can be established (assume $C = C^*$ and $m \geqq n$ for simplicity):

The generalized eigenvalue problem: $[A;B]^T[A;B]z_i = \lambda_i C z_i$

$$\Updownarrow$$

The ordinary eigenvalue problem: $C^{-1}[A;B]^T[A;B]z_i = \lambda_i z_i$

$$\Updownarrow$$

The generalized SVD of $([A;B], R_C)$:

$$T^T[A;B]Z = \operatorname{diag}(\alpha_1, \cdots, \alpha_{n+d}), \qquad \sigma_i^2 = \lambda_i = \alpha_i^2/\beta_i^2$$

$$W^T R_C Z = \operatorname{diag}(\beta_1, \cdots, \beta_{n+d}),\ z_i = i\text{th column of } Z$$

$$\Updownarrow$$

The symmetric eigenvalue problem: $R_C^{-T}[A;B]^T[A;B]R_C^{-1}v_i=\lambda_i v_i$ and $z_i=R_C^{-1}v_i$

$$\Updownarrow$$

$$\text{The ordinary SVD of } [A;B]R_C^{-1}=\sum_{i=1}^{n+d}\sigma_i u_i v_i^T \quad \sigma_i^2=\lambda_i \quad z_i=R_C^{-1}v_i.$$

These links are used to prove the main theorems in this section.

THEOREM 2. *Consider the notation and assumptions* (23). *Let* $r^*=n_1+r\leqq n$ *be the rank of* $([A_1;\ A_2;\ B],\ R_C^*)$ *as computed by the* GTLS *Algorithm from* (19), *and assume that* Γ *in* (21) *is nonsingular*; *then*

$$\text{Range}\left(\begin{bmatrix}\hat{X}\\-I_d\end{bmatrix}\right)\subseteq \text{Range}\,(Z^*) \textit{ such that } \|\hat{X}\|_F \textit{ and } \|\hat{X}\|_2 \textit{ are minimal}$$

where $Z^*_{(n+d)\times(n+d-r^*)}$ *contains the vectors* z_i *associated with the* $(n+d-r^*)$ *smallest generalized singular values, obtained from the* GSVD (13) *of the matrix pair* $([A;B],\ R_C^*)$.

Proof. Set $s=n+d-r^*$. Let Σ_2 be a diagonal matrix containing on its diagonal the s smallest generalized singular values of the GSVD (13) of the matrix pair $([A;B],\ R_C^*)$ and let

$$Z^*=\begin{bmatrix}Z_1\\Z_2\end{bmatrix}\begin{matrix}n_1\\n_2+d\end{matrix}$$

be the s corresponding columns of the nonsingular matrix $Z_{(n+d)\times(n+d)}$. Now use the link (15) between the GSVD of $([A;B],\ R_C^*)$ and the generalized symmetric eigenvalue problem $([A;B]^T[A;B],\ C^*)$:

$$[A;B]^T[A;B]\begin{bmatrix}Z_1\\Z_2\end{bmatrix}=C^*\begin{bmatrix}Z_1\\Z_2\end{bmatrix}\Sigma_2^2. \tag{24}$$

Compute the QR factorization (16) of $[A;B]$.

Since Q is orthogonal, the generalized eigenvalues Σ_2^2 and corresponding eigenvectors $\left[\begin{smallmatrix}Z_1\\Z_2\end{smallmatrix}\right]$ of $([A;B]^T[A;B],\ C^*)$ and $(R^TR,\ C^*)$ coincide. Hence, (24) yields

$$R^TR\begin{bmatrix}Z_1\\Z_2\end{bmatrix}=\begin{bmatrix}R_{11}^TR_{11} & R_{11}^TR_{12}\\ R_{12}^TR_{11} & R_{12}^TR_{12}+R_{22}^TR_{22}\end{bmatrix}\begin{bmatrix}Z_1\\Z_2\end{bmatrix}=\begin{bmatrix}0 & 0\\0 & C\end{bmatrix}\begin{bmatrix}Z_1\\Z_2\end{bmatrix}\Sigma_2^2$$

$$\text{or } R_{11}^T(R_{11}Z_1+R_{12}Z_2)=0 \tag{25}$$

$$R_{12}^T(R_{11}Z_1+R_{12}Z_2)+R_{22}^TR_{22}Z_2=CZ_2\Sigma_2^2. \tag{26}$$

Since A_1 has full column rank, R_{11} is nonsingular. Hence, the columns of R_{11} span the whole n_1-dimensional space $\mathscr{R}^{n_1}$ and thus, (25) is only satisfied if

$$R_{11}Z_1+R_{12}Z_2=0 \tag{27}$$

Substituting (27) into (26), we obtain

$$R_{22}^TR_{22}Z_2=CZ_2\Sigma_2^2. \tag{28}$$

Equation (28) implies that the generalized eigenvalues Σ_2^2 and corresponding eigenvectors Z_2 are obtained from the symmetric eigenvalue decomposition of $(R_{22}^TR_{22},\ C)$ or equivalently, from the GSVD of $(R_{22},\ R_C)$. It is precisely this matrix Z_2 which is computed in Step 2 of the GTLS Algorithm. Once Z_2 is computed, Z_1 is obtained from (27), as also done in Step 2.c of our GTLS algorithm

$$(29) \qquad Z_1 = -R_{11}^{-1} R_{12} Z_2.$$

Hence, $\left[\begin{smallmatrix} Z_1 \\ Z_2 \end{smallmatrix}\right]$, as computed in Step 2 of our GTLS algorithm, equals precisely the vectors z_i in Z^*, associated with the $(n + d - r^*)$ smallest generalized singular values obtained from the GSVD (13) of the matrix pair $([A; B], R_C^*)$.

Since the orthonormalization of $\left[\begin{smallmatrix} Z_1 \\ Z_2 \end{smallmatrix}\right]$ in Step 3.a of our GTLS algorithm (if needed) does not change its range and since Range $(\left[\begin{smallmatrix} Y \\ \Gamma \end{smallmatrix}\right]) \subseteq$ Range $(\left[\begin{smallmatrix} Z_1 \\ Z_2 \end{smallmatrix}\right])$ from (21)–(22), it follows that

$$\text{Range}\left(\begin{bmatrix} \hat{X} \\ -I_d \end{bmatrix}\right) = \text{Range}\left(\begin{bmatrix} Y \\ \Gamma \end{bmatrix}(-\Gamma^{-1})\right) \subseteq \text{Range}\,(Z^*) = \text{Range}\left(\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}\right)$$

As proven in § 2, $\hat{X} = -Y\Gamma^{-1}$ has minimal norm $\|\hat{X}\|_F$ and $\|\hat{X}\|_2$. □

This theorem allows us to deduce the following *relationships* between the GSVD of the matrix pairs $([A; B], R_C^*)$, (R, R_C^*) and (R_{22}, R_C) (with R, R_{22} defined in (16)).

COROLLARY 1. *Consider the notation and assumptions* (23) *and let* (16) *be the* QR *factorization of* $[A_1; A_2; B]$; *then*

$$\forall i = 1, \cdots, n_2 + d\colon \sigma_i \in \sigma(R_{22}, R_C) \Leftrightarrow \sigma_{n_1+i} \in \sigma([A; B], R_C^*) \Leftrightarrow \sigma_{n_1+i} \in \sigma(R, R_C^*)$$

and the $(n_2 + d)$ *corresponding vectors* $Z_{R_{22}}^*$ *of* (R_{22}, R_C), Z_{AB}^* *of* $([A; B], R_C^*)$ *and* Z_R^* *of* (R, R_C^*) *obtained from their respective* GSVD (13) *are, up to a normalization factor, related by*

$$Z_{AB}^* = Z_R^* = \begin{bmatrix} -R_{11}^{-1} R_{12} Z_{R_{22}}^* \\ Z_{R_{22}}^* \end{bmatrix}$$

Proof. The proof follows straightforwardly from the proof of Theorem 2. □

Theorem 2 and Corollary 1 imply that the GTLS solution can also be computed from the GSVD of $([A; B], R_C^*)$, namely, from the vectors z_i corresponding to its smallest generalized singular values, as follows:

$$\text{Range}\left(\begin{bmatrix} \hat{X} \\ -I \end{bmatrix}\right) = \text{Range}\,(Z^*) \quad \text{and} \quad \hat{X} = -Z_1^* Z_2^{*-1}$$

where

$$Z^* = \begin{array}{c} \begin{bmatrix} Z_1^* \\ Z_2^* \end{bmatrix} \end{array} \begin{array}{l} n \\ d \end{array}$$

are the vectors z_i associated with the d smallest generalized singular values obtained from the GSVD (13) of $([A; B], R_C^*)$.

For the *one-dimensional* GTLS problem (i.e., $d = 1$ in (5)) with A of *full column rank* and Γ in (21) nonsingular (this is the problem mostly considered in literature), this means that the GTLS solution can also be computed from the vector z_{n+1} corresponding to the smallest singular value σ_{n+1} of the GSVD (13) of $([A; B], R_C^*)$ or equivalently, from the generalized eigenvector z_{n+1} corresponding to the smallest generalized eigenvalue σ_{n+1}^2 of the matrix pair $([A; B]^T[A; B], C^*)$. However computing the GTLS solution in this way requires the GSVD of a *larger* matrix pair than the matrix pair (R_{22}, R_C) used in the GTLS Algorithm. It is evident that our GTLS algorithm is *computationally more efficient* than the Koopmans–Levin method described in [6], and the compensated LS estimation method described in [33]. These methods are based on computing the GSVD of the matrix pair $([A; B], R_C^*)$ even if R_C^* is given by

$$\begin{bmatrix} 0 & 0 \\ 0 & I_{n_2+d} \end{bmatrix}.$$

Since $R_C \sim I$ in the latter case, the solution can be computed by the TLS algorithm with exactly known columns, given in [34, § 1.8.2], which first performs a QR factorization of the known columns and then proceeds with an ordinary SVD of the submatrix R_{22}.

In Theorem 1 we assumed that Γ in (21) is *nonsingular*, i.e., the GTLS solution is *generic*. Analogously to the classical TLS problem (see [34], [39]), we can deduce conditions that guarantee the *existence and uniqueness* of the generic GTLS solution. Hereto, we apply the results of Theorem 2.

THEOREM 3. Existence and uniqueness of the generic GTLS solution. *Consider the notation and assumptions* (23). *Let* $m \geqq n$ *and denote by* σ' (*respectively*, σ) *the smallest* (*respectively*, $(n+1)th$) *generalized singular value of* (A, R_{Ca}^*) (*respectively*, $([A; B], R_C^*)$), *then*

$$\sigma' > \sigma \Rightarrow \textit{the}\ \text{GTLS}\ \textit{solution}\ \hat{X} = \begin{bmatrix} \hat{X}_1 \\ \hat{X}_2 \end{bmatrix}\ \textit{is unique and generic}.$$

Proof. Compute the QR factorization (16) of $[A_1; A_2; B]$ and let R be partitioned as follows:

$$(30) \qquad R = \begin{array}{cc|c} \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} & \begin{array}{c} n_1 \\ m-n_1 \end{array} \\ \begin{array}{cc} n_1 & n_2+d \end{array} & \end{array} = \begin{array}{c} \begin{bmatrix} R_{11} & R_{1a} & R_{1b} \\ 0 & R_{2a} & R_{2b} \end{bmatrix} \\ \begin{array}{ccc} n_1 & n_2 & d \end{array} \end{array}.$$

Denote

$$R_{Ca}^* = \begin{array}{c} \begin{bmatrix} 0 & 0 \\ 0 & R_{Ca} \end{bmatrix} \begin{array}{c} n_1 \\ n_2 \end{array} \\ \begin{array}{cc} n_1 & n_2 \end{array} \end{array}.$$

Since

$$A^TA = \begin{bmatrix} R_{11} & 0 \\ R_{1a} & R_{2a} \end{bmatrix} \begin{bmatrix} R_{11} & R_{1a} \\ 0 & R_{2a} \end{bmatrix},$$

we can prove analogous relationships between the GSVD of

$$(A, R_{Ca}^*), \quad \left(\begin{bmatrix} R_{11} & R_{1a} \\ 0 & R_{2a} \end{bmatrix}, R_{Ca}^* \right),$$

and (R_{2a}, R_{Ca}) as given in Corollary 1. This and Corollary 1 imply that

$$(31) \qquad \sigma' = \sigma'_{n_2} = \min \{\sigma(R_{2a}, R_{Ca})\} \quad \text{and} \quad \sigma = \sigma_{n_2+1} \in \sigma(R_{22}, R_C).$$

Since R_{Ca} and R_C are nonsingular, we can use the link with the corresponding ordinary SVD problems (as given in the beginning of this section). Hence, (31) yields

$$(32) \quad \sigma' = \sigma'_{n_2} = \min \{\sigma(R_{2a}R_{Ca}^{-1})\}, \quad \sigma = \sigma_{n_2+1} \in \sigma(R_{22}R_C^{-1}), \quad \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = R_C^{-1} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix},$$

where

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}\begin{matrix} n_2 \\ d \end{matrix} \qquad \left(\text{respectively, } \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}\begin{matrix} n_2 \\ d \end{matrix}\right)$$

are the d vectors z_i (respectively, right singular vectors v_i) associated with the d smallest (generalized) singular values σ_i, obtained from the GSVD (18) of (R_{22}, R_C) (respectively, SVD of $R_{22}R_C^{-1}$). The assumption $\sigma' > \sigma$ guarantees that the TLS solution $\tilde{X}_2 = -V_1 V_2^{-1}$ of the classical TLS problem $R_{22}R_C^{-1}[\begin{smallmatrix} X_2 \\ -I \end{smallmatrix}] \approx 0$, is unique and generic (i.e., V_2 nonsingular) according to the existence and uniqueness theorems [34, Thms. 1-1, and 1-2], [39] of the classical generic TLS solution. Since the GTLS solution $\hat{X}_2$ of the GTLS problem, $R_{22}[\begin{smallmatrix} X_2 \\ -I \end{smallmatrix}] \approx 0$ with corresponding error covariance matrix $C = R_C^T R_C$, is given by $\hat{X}_2 = -Z_1 Z_2^{-1}$ (Z_1, Z_2 defined by (32)), $\hat{X}_2$ and $\tilde{X}_2$ are related by

$$R_C^{-1}\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = R_C^{-1}\begin{bmatrix} \tilde{X}_2 \\ -I \end{bmatrix}(-V_2) = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} \hat{X}_2 \\ -I \end{bmatrix}(-Z_2).$$

Since R_C is nonsingular, the existence and uniqueness of the generic TLS solution $\tilde{X}_2$ imply that the generic GTLS solution $\hat{X}_2$ is unique and generic (i.e., Z_2 nonsingular). Since A_1 has full column rank, R_{11} is nonsingular. This implies that $\hat{X}_1 = R_{11}^{-1}R_{12}[\begin{smallmatrix} Z_1 \\ Z_2 \end{smallmatrix}]Z_2^{-1}$ (using (20)–(21)–(22)) exists and is unique. Thus, the GTLS solution

$$\hat{X} = \begin{bmatrix} \hat{X}_1 \\ \hat{X}_2 \end{bmatrix}$$

of the GTLS problem is unique and generic. □

Whenever Γ in (21) is *singular*, the GTLS problem is called *nongeneric*. Using Theorem 3 this happens when $\sigma' \leqq \sigma$, i.e., when A *is* (*nearly*) *rank-deficient* ($\sigma' \approx 0$) or when *the set of equations* $AX \approx B$ (or at least one subset $AX_i \approx B_i$) is *highly incompatible* ($\sigma' \approx \sigma$) (see also [34, § 1.6], [39], and [40]).

Gallo [9] used a statistical approach to prove under which condition his estimate (which equals our GTLS solution as proven by Theorem 4 below) is generic. These statistical results agree with our algebraic approach.

Now using the previously proven theorems, the correspondence between particular cases of our GTLS solution and alternative expressions of consistent estimators described in the literature can be proved, as done in the following theorem. These equivalences allow us to derive the main statistical properties of the GTLS solution in different statistical situations.

THEOREM 4. *Consider the notation and assumptions* (23). *Let* $\sigma' = \min\{\sigma(A, R_{Ca}^*)\}$ *and assume further that* $\sigma = \sigma_{n+1} = \cdots = \sigma_{n+d} \in \sigma([A;B], R_C^*)$ *has multiplicity* d.

If $\sigma' > \sigma$, *the* GTLS *solution* $\hat{X}$ *is given by*

$$\hat{X} = (A^T A - \sigma^2 C_a^*)^{-1}(A^T B - \sigma^2 C_{ab}^*). \tag{33}$$

Proof. Using the link (15) between the GSVD of $([A;B], R_C^*)$ and the symmetric generalized eigenvalue problem $([A;B]^T[A;B], C^*)$, we obtain

$$([A;B]^T[A;B] - \sigma^2 C^*)\begin{bmatrix} Y \\ \Gamma \end{bmatrix} = 0 \tag{34}$$

where $[{Y \atop \Gamma}]^n_d$ are d base vectors of the d-dimensional eigensubspace associated with the smallest generalized eigenvalue σ^2 of $([A;B]^T[A;B], C^*)$ of multiplicity d. Partitioning C^* in (34) yields

$$\begin{bmatrix} A^TA-\sigma^2C_a^* & A^TB-\sigma^2C_{ab}^* \\ B^TA-\sigma^2C_{ab}^{*T} & B^TB-\sigma^2C_b^* \end{bmatrix}\begin{bmatrix} Y \\ \Gamma \end{bmatrix}=0. \tag{35}$$

Since σ^2 has multiplicity d, the left-hand-side matrix of (35) has rank n and solutions to (35) will be determined by equations corresponding to any n linearly independent rows of that matrix. Consider the first n equations of (35):

$$(A^TA-\sigma^2C_a^*)Y+(A^TB-\sigma^2C_{ab}^*)\Gamma=0; \tag{36}$$

then, the assumption $\sigma'^2 > \sigma^2$ guarantees that $A^TA - \sigma^2C_a^*$ is invertible and also guarantees that Γ^{-1} exists. Hence, (36) yields

$$-Y\Gamma^{-1}=(A^TA-\sigma^2C_a^*)^{-1}(A^TB-\sigma^2C_{ab}^*). \tag{37}$$

Theorem 1 and the assumptions above imply that the GTLS solution space

$$\text{Range}\left(\begin{bmatrix} \hat{X} \\ -I \end{bmatrix}\right)=\text{Range}\left(\begin{bmatrix} Y \\ \Gamma \end{bmatrix}\right)$$

and thus $\hat{X} = -Y\Gamma^{-1}$ since the GTLS solution is invariant with respect to any base transformation in its solution space (see § 2). Hence, (37) yields (33). □

Theorem 4 allows us to derive the main *statistical properties* of the GTLS solution.

First, assume that *none of the columns of* A are *known exactly* ($n_1 = 0$ and $C^* = C$).

- If $C \sim I_{n+d}$, the expression (33) reduces to

$$\hat{X}=(A^TA-\sigma^2I_n)^{-1}A^TB, \tag{38}$$

which is a well-known expression of the *classical* TLS *solution* as proved in [14] for $d = 1$ and in [34, § 2.2] for $d \geqq 1$. The consistency, distributional, and asymptotic properties of the classical TLS estimate have been proved by Gleser for any $d \geqq 1$ [11]. Assuming that the rows of the error matrix $[\Delta A; \Delta B]$ in (11) are independently and identically distributed (i.i.d.) zero mean vectors with common covariance matrix $C \sim I_{n+d}$ and that $\lim_{m\to\infty}(1/m)A_0^TA_0$ exists and is positive definite, Gleser has proved that the TLS solution $\hat{X}$ is a *strongly consistent* estimate of the true but unknown parameters X of the corresponding unperturbed model $A_0X = B_0$. This result holds, regardless of the common distribution of the errors. When this common error distribution has finite fourth moments, $\hat{X}$ is shown to be *asymptotically normally distributed*. Expressions for the covariance matrix of this distribution are given in [11], as well as large-sample approximate confidence regions.

While Gleser assumes that the elements of A_0 and B_0 in model (11) are fixed (i.e., the functional equations model [23]), Kelly [22] considers the case in which these elements are *random* (called the structural equations model [23]). More specifically, Kelly assumes that the rows of $[A_0; B_0]$ are i.i.d. with common mean vector and common covariance matrix. By calculating the influence function of Gleser's errors-in-variables estimator (which equals the classical TLS solution), Kelly is able to derive an explicit expression for the asymptotic covariance matrix of this estimator in the *structural* equations model.

Finally, Aoki and Yue [2] have studied the statistical properties of the TLS solution (called the solution of the eigenvector method or the Koopmans–Levin method) of Toeplitz-like sets of equations arising in *autoregressive moving average* (ARMA) *modelling* and *system identification*. These models are given by

$$y(t)+a_1y(t-1)+\cdots+a_{n_a}y(t-n_a)=b_1u(t-1)+\cdots+b_{n_b}u(t-n_b) \tag{39}$$

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. The *observations* at the input and output are assumed to be perturbed by mutually independent *white noise* sequences (i.e., i.i.d. random variables) with *zero mean* and *equal variances*. If sufficient observations are taken, (39) gives rise to an overdetermined set of equations of the form (2) where the corresponding error covariance matrix on the data $\sim I_{n+d}$. Assuming that the given system is stable (i.e., the polynomial $1+\sum_{i=1}^{n_a} a_iz^i$ has all zeros outside the unit circle) and the input sequence $\{u(t)\}$ is uniformly bounded, the TLS solution of this set is *strongly consistent* if and only if the matrix

$$\lim_{m\to\infty}(1/m)\sum_{t=1}^{m} s_ts_t^T,$$

where $s_t=[-y(t+n_a-1),\cdots,-y(t),u(t+n_b-1),\cdots,u(t)]^T$, is positive definite [2] (observe that these conditions are analogous to the consistency conditions imposed by Gleser). This is the case if the input sequence is persistently exciting of order n_a+n_b (i.e., $\lim_{m\to\infty}(1/m)\sum_{t=1}^{m} q_tq_t^T$, where

$$q_t=[u(t+n_a+n_b-1),\cdots,u(t)]^T,$$

is positive definite) and if the polynomials $1+\sum_{i=1}^{n_a} a_iz^i$ and $\sum_{i=1}^{n_b} b_iz^i$ are relatively prime. The first condition is always satisfied if the input is white noise, whereas the second condition means that (39) is a minimal realization of the input-output sequences. Additionally, Aoki and Yue have given explicit expressions for the mean square error of the TLS estimates as a function of the observation noise variances and the number of observations. As demonstrated in [6], the accuracy of these TLS estimates is comparable with that of the joint output method [30] and superior to all other methods described by Söderström [30]. Moreover the TLS method based on the SVD is numerically much more robust and plays an important role in ARMA modelling. See [2], [6], [30] and [34] for more details.

- Now if *the error covariance matrix* C_Δ has the *more general* form c^2C—where C is known and positive definite—Gleser proposed [10], [11] to *transform the original data* $[A;B]$ to new data $[A;B]R_C^{-1}$ (where $C=R_C^TR_C$, R_C upper triangular) such that the error covariance matrix corresponding to the transformed system $\sim I_{n+d}$. Computing the classical TLS solution of this transformed system and converting this result back to the original set, *preserves consistency* of the result. It is easy to prove that this estimate equals our GTLS solution that can be obtained straightforwardly *without pretransforming* the original data. Indeed, in the transformed system the TLS estimate is obtained from the eigenvectors V corresponding to the smallest eigenvalues Λ of the data matrix $[A;B]R_C^{-1}$:

$$R_C^{-1}[A;B]^T[A;B]R_C^{-1}V=V\Lambda. \tag{40}$$

Substituting R_C^{-1} by Z, we obtain the corresponding generalized eigenvector equations:

$$[A;B]^T[A;B]Z = CZ\Lambda. \tag{41}$$

The GTLS solution is computed straightforwardly from the eigenvectors Z, using a GSVD. The transformation formulas given by Gleser [10], [11] just transform the solution obtained from (40) to the GTLS solution obtained from (41). Observe that Gleser needs to compute the inverse of the square root of the covariance matrix that may cause numerical problems especially when R_C is ill-conditioned. This inversion is avoided in our GTLS algorithm.

Even if the true covariance matrix C_Δ is not known but an *estimate* S of it, up to an unknown factor of proportionality, the GTLS solution $\hat{X}$ can still be consistent [7]. An experiment wherein several observations for each row of $[A_0;\ B_0]$ in (11) are available is an example of this case. Assume that the estimator S is distributed as a multiple of a Wishart matrix with $\eta^{-1}m$ degrees of freedom independently of A and B, where η is a fixed positive number. Further assume that $d = 1$ in (11) and that the rows of the error matrix $[\Delta A;\ \Delta b]$ are i.i.d. as a multivariate normal random variable with zero mean and covariance matrix C_Δ so that $E(S) \sim C_\Delta$. Under these assumptions, Fuller [7] has proved that $\sqrt{m}(\hat{X} - X)$ converges in distribution to a normal random variable with zero mean and computed an explicit expression for the covariance matrix of this variable.

In ARMA *modelling* and system identification, the situation of known noise covariance functions is treated in [24] and [8]. In [8] Furuta and Paquet discuss the case of *correlated noise* when all the correlation functions are known, up to a factor of proportionality. The suggested procedures, based on solving a generalized eigenvalue problem, are extensions of the eigenvector method described by Levin [26]. It is easy to see that these solutions coincide with our GTLS solution whenever the GTLS problem has a unique minimizer. Hence, the results in [24] and [8] can be applied straightforwardly to the GTLS solution. Conditions for strong consistency of the GTLS estimates in *multi-input multi-output* system models of the form (39) have been derived in [24] and are similar to those for strong consistency of the TLS estimates in the single-input single-output case [2] (see above).

Consider now the case that n_1 *columns* A_1 of A are *known exactly*.

- If $\mathbf{C} \sim \mathbf{I}_{n_2+d}$, (33) reduces to

$$\hat{X} = \left(A^TA - \sigma^2\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}\right)^{-1} A^TB. \tag{42}$$

Equation (42) equals the expression of the *compensated least squares estimate*, derived by Stoica and Söderström [33]. Assume that the given system (39) is stable and that the polynomials $1 + \sum_{i=1}^{n_a} a_i z^i$ and $\sum_{i=1}^{n_b} b_i z^i$ are relatively prime. If the observations at the *input* are *noise-free*, persistently exciting of order $n_a + n_b$ and independent of the observation noise, and if the observations at the *output* are perturbed by *zero mean white noise of equal variance*, (42) is a *consistent* estimate of the parameters a_i and b_i in the ARMA model (39). Additionally, Stoica and Söderström [33] have proved that this estimate is *asymptotically Gaussian distributed* and an expression of the covariance matrix is explicitly given. As shown before, our GTLS Algorithm is, however, computationally more efficient than the computation procedure presented in [33].

- If the *error covariance matrix* has the *general* form c^2C—where C is known and positive definite—the statistical properties of the GTLS solution for the one-dimensional problem ($d = 1$) can be derived from Gallo [9]. Indeed, Theorem 4 proves the correspondence between our GTLS solution and the consistent estimate derived by Gallo in [9, Thm. 1]. This link allows us to investigate the properties of our GTLS solution as an estimator of the parameters in the general one-dimensional errors-in-variables model, given by (11) for $d = 1$. More specifically, when the joint distribution of the errors possesses finite fourth moment and when

$$\frac{1}{\sqrt{m}} \min\{\lambda(A_0^TA_0)\} \to \infty \quad \text{and} \quad \frac{(\min\{\lambda(A_0^TA_0)\})^2}{\max\{\lambda(A_0^TA_0)\}} \to \infty \quad \text{as } m \to \infty, \tag{43}$$

Gallo has proved that the GTLS *solution* $\hat{X}$ is a *weakly consistent* estimate of the parameters X in model (11). This property holds, regardless of the joint distribution of the errors. Observe that the conditions (43) are less restrictive than those assumed by Gleser (i.e., $\lim_{m\to\infty}(1/m)A_0^TA_0$ exists and is positive definite).

Finally, in ARMA *modelling* and system identification, systems of the form (39) whose *inputs* are observed exactly and whose observed *outputs* are disturbed by zero mean *correlated noise*, have been treated in [20] and [16] for the case that $n_a = n_b$. James, Souter, and Dixon have used the same basic principle of *bias correction* as Stoica and Söderström in [33] and have derived an expression of the form (33). Grosjean and Foulard have *extended* the eigenvector method of Levin [26] to the identification of *multi-input multi-output* systems whose *outputs* are disturbed by *correlated noise*. Assuming that the order $n_a = n_b$ of the system and the correlation functions of the output noise are known, these estimates coincide with the GTLS solution. Conditions for *strong consistency* of the GTLS solution in these cases have been described and are similar to those given by Stoica and Söderström in [33]. For more details, see [24] and [16]. Observe that multi-input multi-output systems in [16] and [24] are treated as s multi-input one output problems where s is the number of outputs, i.e., as s one-dimensional GTLS problems ($d = 1$).

Since Theorem 4 proves the link between our GTLS solution and Gallo's estimate, the following *alternative expressions for our* GTLS *solution* can be deduced straightforwardly from Theorem 1 of [9].

THEOREM 5. *Consider the notation and assumptions* (23). *Let* $\sigma' = \min\{\sigma(A, R_{Ca}^*)\}$ *and assume further that* $\sigma = \sigma_{n+1} = \cdots = \sigma_{n+d} \in \sigma([A; B], R_C^*)$ *has multiplicity d. If* $\sigma' > \sigma$, *the* GTLS *solution* $\hat{X}$ *is given by*

$$\hat{X}_2 = (A_2^TPA_2 - \sigma^2C_a)^{-1}(A_2^TPB - \sigma^2C_{ab}), \tag{44}$$

$$\hat{X}_1 = (A_1^TA_1)^{-1}A_1^T(B - A_2\hat{X}_2) \tag{45}$$

with

$$P = I_m - A_1(A_1^TA_1)^{-1}A_1^T$$

and

$$\hat{X}_2 = (R_{2a}^TR_{2a} - \sigma^2C_a)^{-1}(R_{2a}^TR_{2b} - \sigma^2C_{ab}), \tag{46}$$

$$\hat{X}_1 = R_{11}^{-1}(R_{1b} - R_{1a}\hat{X}_2) \tag{47}$$

where the matrices R_{ij} are defined by the QR *factorization of* $[A_1; A_2; B]$:

$$[A_1;A_2;B]=\mathrm{QR}=\begin{array}{cc}[Q_1 & Q_2]\\ n_1 & m-n_1\end{array}\begin{array}{c}\left[\begin{array}{ccc}R_{11}&R_{1a}&R_{1b}\\0&R_{2a}&R_{2b}\end{array}\right]\begin{array}{l}n_1\\m-n_1\end{array}\\ \begin{array}{ccc}n_1&n_2&d\end{array}\end{array} \tag{48}$$

Proof. Equations (44) and (45) follow straightforwardly from Theorem 1 of [9]. Substituting A_1 in the definition of P by Q_1R_{11} from (48) yields

$$P=I_m-Q_1Q_1^T. \tag{49}$$

Now substitute (49) in (44) and replace A_1, A_2, and B by their equivalents Q_1R_{11}, $Q_1R_{1a}+Q_2R_{2a}$ and $Q_1R_{1b}+Q_2R_{2b}$ obtained from the QR factorization. Making use of the orthonormality of Q, we obtain

$$A_2^TPA_2=A_2^TA_2-R_{1a}^TR_{1a}=R_{2a}^TR_{2a}, \tag{50}$$

$$A_2^TPB=A_2^TB-R_{1a}^TR_{1b}=R_{2a}^TR_{2b}, \tag{51}$$

$$(A_1^TA_1)^{-1}A_1^TB=R_{11}^{-1}Q_1^TB=R_{11}^{-1}R_{1b}, \tag{52}$$

$$(A_1^TA_1)^{-1}A_1^TA_2=R_{11}^{-1}Q_1^TA_2=R_{11}^{-1}R_{1a}, \tag{53}$$

which prove (46) and (47). □

In case that the GTLS *solution* is either *nonunique* or *nongeneric* or when the *subset* A_1 of A in the GTLS problem (5)–(6) is *rank-deficient*, the GTLS solution is *no longer consistent* but is in fact a *biased* estimator of the parameters in (11). Our approach in these situations can be justified as follows. All these cases mentioned above refer to the presence of *multicollinearities*, i.e., there is a (nearly) exact linear relation among the columns of A in the model (11). The consequences are well known; in particular, coefficient estimates obtained by ordinary LS or TLS (without rank reduction) tend to be inflated and can have extremely large variances. One way of handling the multicollinearity problem and *stabilizing* the coefficients is *reducing the rank* of the data matrix A and amounts to filtering out the smallest (generalized) singular values from the estimator. This approach has been adopted in our GTLS Algorithm and is *similar* to the biased estimation techniques: *principal component* regression [18] and *latent root* regression [43], used in linear regression (see also [37] and [40]).

Observe however that in these cases of nonuniqueness or nongenericity the GTLS solution *no longer equals* the solution obtained by applying the *usual transformation procedures* (see, e.g., [11] or [34, § 4.5]). This is evident from (40) and (41). Indeed, the minimum norm or nongeneric TLS solution computed in the transformed system of equations, i.e.,

$$[A;B]R_C^{-1}\begin{bmatrix}\hat{X}\\-I\end{bmatrix}\approx 0,$$

and converted back to a solution of the original set, does not coincide with the minimum norm or nongeneric GTLS solution of the original set of equations. The properties of the GTLS solution in these cases are not yet fully analyzed.

Summarizing, *consistency results of the* GTLS *solution* have been derived in this section:

- For *any multidimensional* ($d \geqq 1$) *errors-in-variables model* given by (11), in which *none of the columns* of A are *known exactly* ($n_1 = 0$) (based on [11]); and

- For any one-dimensional ($d = 1$) errors-in-variables model given by (11) in which some columns of A are known exactly ($n_1 \geqq 0$) (based on [9]). However the authors strongly believe that the consistency results of Gallo and Gleser can be generalized in order to prove consistency of the GTLS solution for any multidimensional errors-in-variables model given by (11) in which some columns of A are known exactly ($n_1 \geqq 1$).

4. Conclusions. Every *linear parameter estimation* problem arising in signal processing, system identification, automatic control, or in general engineering, statistics, and medicine, gives rise to an *overdetermined* set of linear equations $AX \approx B$ that are usually solved with the ordinary least squares method. Very often, errors occur in *both* A and B. For those cases, the *Total Least Squares* (TLS) technique was devised as a better method of fitting. This method introduced into numerical analysis by Golub and Van Loan, is strongly based on the *Singular Value Decomposition* (SVD). If the errors on the measurements A and B are *uncorrelated with zero mean and equal variance*, TLS is able to compute a *strongly consistent* estimate of the true solution of the corresponding unperturbed set $A_0X = B_0$. In this paper the TLS problem is *generalized* to maintain consistency of the solution in the following cases: first of all, some columns of A may be error-free and second, the errors on the remaining data may be *correlated* and not equally sized provided the covariance matrix of the errors on the rows of the remaining data matrix is known, up to a factor of proportionality. Here, a numerically reliable *Generalized* TLS *algorithm* GTLS, based on the Generalized Singular Value Decomposition (GSVD), is developed. This GSVD avoids transforming the data A, B explicitly and is *numerically more robust* with respect to ill-conditioned covariance matrices. This explains the better numerical performance of the GTLS Algorithm with respect to the explicit transformation procedures. Moreover, by first performing a QR factorization, the GTLS Algorithm only needs to compute the GSVD of a smaller submatrix. This makes the GTLS Algorithm *computationally more efficient* than other methods described in literature.

Additionally, the correspondence between the GTLS solution and alternative expressions of consistent estimators, described in the literature, is proven. From these relations, the main statistical properties of the GTLS solution are deduced. In particular, the equivalence between the GTLS method and the errors-in-variables regression estimator, well known in statistics, is shown. It is concluded that under mild conditions the GTLS solution is a *consistent* estimate of the true parameters of any *general multivariate errors-in-variables model* in which *all or some subset* of variables are observed *with errors*. Furthermore, it is shown that the GTLS algorithm computes the same estimate as the eigenvector method, also called the Koopmans–Levin method, and the Compensated Least Squares (CLS) method. These methods, commonly used in system identification, were developed in order to provide consistent parameter estimates in ARMA *modelling* using noise-corrupted data. If the only disturbances in the observed outputs and the inputs (if they cannot be measured exactly) are given by mutually independent *zero mean white noise sequences of equal variance*, the CLS, TLS, and eigenvector methods all compute strongly consistent estimates. Using the GTLS Algorithm, consistency can be maintained in cases where the disturbances are *not* necessarily white provided the covariance matrix of the correlated noise on the input-output data is known, up to a factor of proportionality.

Acknowledgment. The author would like to thank Hongyuan Zha for his helpful comments.

REFERENCES

[1] R. J. Adcock, *A problem in least squares*, The Analyst, 5 (1878), pp. 53–54.

[2] M. Aoki and P. C. Yue, *On a priori error estimates of some identification methods*. IEEE Trans. Automat. Contr. AC-15 (1970), pp. 541-548.

[3] Z. Bai, *The direct* GSVD *algorithm and its parallel implementation*, Tech. Report CS-TR-1901, Dept. Computer Science, Univ. of Maryland, College Park, MD, August 1987.

[4] R. P. Brent, F. T. Luk, and C. F. Van Loan, *Computation of the generalized singular value decomposition using mesh-connected processors*, TR-CS-83-563, Dept. Computer Science, Cornell Univ. Ithaca, New York, 1983.

[5] J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart, LINPACK *User's Guide*, SIAM Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.

[6] K. V. Fernando and H. Nicholson, *Identification of linear systems with input and output noise: The Koopmans-Levin method*, IEE Proc. D, 132 (1985), pp. 30–36.

[7] W. A. Fuller, *Properties of some estimators for the errors-in-variables model*, Ann. Statist., 8 (1980), pp. 407–422.

[8] K. Furuta and J. G. Paquet, *On the identification of time-invariant discrete processes*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 153–155.

[9] P. P. Gallo, *Consistency of regression estimates when some variables are subject to error*, Comm. Statist.–Theory Methods, 11 (1982), pp. 973–983.

[10] L. J. Gleser, *Calculation and simulation in errors-in-variables regression problems*, Mimeograph Series 78-5, Dept. of Statistics, Purdue Univ., West Lafayette, IN, 1978.

[11] ———, *Estimation in a multivariate "errors in variables" regression model: Large sample results*, Ann. Statist., 9 (1981), pp. 24–44.

[12] G. H. Golub, personal communication, November 1983.

[13] 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. 322-327.

[14] 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.

[15] ———, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[16] 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 IFAC Symposium on Identification and System Parameter Estimation, Tbilisi, USSR, 1976, pp. 2003–2010.

[17] R. P. Guidorzi, *Canonical structures in the identification of multivariable systems*, Automatica, 11 (1975), pp. 361–374.

[18] D. M. Hawkins, *On the investigation of alternative regressions by principal component analysis*, Appl. Statist., 22 (1973), pp. 275–286.

[19] N. J. Higham, *Analysis of the Cholesky decomposition of a semi-definite matrix*, Numerical Analysis Report no. 128, Univ. of Manchester, January 1987; in Reliable, Numerical Computation, M. Cox and S. Hammarling, eds., Oxford University Press.

[20] P. N. James, P. Souter, and D. C. Dixon, *Suboptimal estimation of the parameters of discrete systems in the presence of correlated noise*, Electron. Lett. 8 (1972), pp. 411–412.

[21] L. S. Jennings and M. R. Osborne, *A direct error analysis for least squares*, Numer. Math., 22 (1975), pp. 322-332.

[22] G. Kelly, *The influence function in the errors in variables problem*, Ann. Statist., 12 (1984), pp. 87–100.

[23] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Fourth edition, Griffin, London, 1979.

[24] Ü. Kotta, *Structure and parameter estimation of multivariable systems using the eigenvector method*, in Proc. 5th IFAC Symposium on Identification and System Parameter Estimation, Darmstadt, FRG, 1979, pp. 453–458.

[25] J. Leuridan, D. De Vis, H. Van der Auweraer, and F. Lembregts, *A comparison of some frequency response function measurement techniques*, in Proc. 4th Internat. Modal Analysis Conference, Los Angeles, CA, February 3-6, 1986, pp. 908–918.

[26] M. J. Levin, *Estimation of a system pulse transfer function in the presence of noise*, IEEE Trans. Automat. Control, AC-9 (1964), pp. 229–235.

[27] C. C. Paige, *Computing the generalized singular value decomposition*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1126–1146.

[28] C. C. Paige and M. A. Saunders, *Towards a generalized singular value decomposition*, SIAM J. Numer. Anal., 18 (1981), pp. 398–405.

[29] K. PEARSON, *On lines and planes of closest fit to points in space*, Philos. Mag., 2 (1901), pp. 559–572.

[30] T. SÖDERSTRÖM, *Identification of stochastic linear systems in presence of input noise*, Automatica, 17 (1981), pp. 713–725.

[31] P. SPRENT, *Models in Regression and Related Topics*, Methuen, London, 1969.

[32] G. W. STEWART, *A method for computing the generalized singular value decomposition*, in Matrix Pencils, B. Kagstrom and A. Ruhe, eds., Springer-Verlag, Berlin, New York, 1983, pp. 207–220.

[33] P. STOICA AND T. SÖDERSTRÖM, *Bias correction in least squares identification*, Internat. J. Control, 35 (1982), pp. 449–457.

[34] S. VAN HUFFEL, *Analysis of the total least squares problem and its use in parameter estimation*, Ph.D. dissertation, Dept. of Electrical Engineering, Katholieke Universiteit Leuven, Belgium, June, 1987.

[35] ———, *The generalized total least squares problem: Formulation, algorithm and properties*, in Proc. NATO Advanced Study Institute on Numerical Linear Algebra, Digital Signal Processing and Parallel Algorithms, Leuven, Belgium, August 1–12, 1988.

[36] S. VAN HUFFEL, J. VANDEWALLE, AND A. HAEGEMANS, *An efficient and reliable algorithm for computing the singular subspace of a matrix, associated with its smallest singular values*, J. Comput. Appl. Math., 19 (1987), pp. 313–330.

[37] S. VAN HUFFEL AND J. VANDEWALLE, *Algebraic relationships between classical regression and total least-squares estimation*, Linear Algebra Appl., 93 (1987), pp. 149–162.

[38] ———, *The partial total least squares algorithm*, J. Comput. and Appl. Math., 21 (1988), pp. 333–341.

[39] ———, *Analysis and solution of the nongeneric total least squares problem*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 360–372.

[40] ———, *Comments on the solution of the nongeneric total least squares problem*, Internal Report ESAT-KUL-88/3, ESAT Lab., Dept. of Electrical Engineering, Katholieke Universiteit, Leuven, February 1988.

[41] C. F. VAN LOAN, *Generalizing the singular value decomposition*, SIAM J. Numer. Anal., 13 (1976), pp. 76–83.

[42] ———, *Computing the CS and generalized singular value decomposition*, Tech. Report CS-604, Department of Computer Science, Cornell University, Ithaca, NY, 1984.

[43] J. T. WEBSTER, R. F. GUNST, AND R. L. MASON, *Latent root regression analysis*, Technometrics, 16 (1974), pp. 513–522.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 3, pp. 316-325, July 1989

004

PARTIAL POLE AND ZERO DISPLACEMENT BY CASCADE CONNECTION*

I. GOHBERG†, M. A. KAASHOEK‡, AND A. C. M. RAN‡

Abstract. Recent interpolation results for rational matrix functions with incomplete data are applied to solve a problem of shifting a part of the poles and the zeros of a given rational matrix function, keeping the other poles and zeros unchanged and the McMillan degree as small as possible.

Key words. pole placement, interpolation, rational matrix functions

AMS(MOS) subject classifications. 93B55, 15A54, 15A18

Introduction. This paper concerns a problem of partial pole and zero shifting. Let W be an $m \times m$ rational matrix function that, say, is regular at infinity, has poles in the points $p_1, \cdots, p_i, p_{i+1}, \cdots, p_k$, and zeros in the points $z_1, \cdots, z_j, z_{j+1}, \cdots, z_l$. Furthermore, a nonempty set Ω is specified such that the points $p_1, \cdots, p_i, z_1, \cdots, z_j$ are not in Ω. The problem we deal with is the following. Construct a $m \times m$ rational matrix function R such that we have the following:

(1) R has no poles and zeros in the set $\Lambda = \{p_1, \cdots, p_i, z_1, \cdots, z_j\}$.

(2) R is regular at infinity.

(3) WR has all its poles and zeros in the set $\Omega \cup \{z_1, \cdots, z_j, p_1, \cdots, p_i\}$.

(4) The McMillan degree $\delta(WR)$ of WR is as small as +possible.

In other words, we want to shift the poles at $p_{i+1}, \cdots, p_k$ and the zeros at $z_{j+1}, \cdots, z_l$ to Ω by a cascade connection of W with R, keeping the pole and zero structure at $p_1, \cdots, p_i$ and $z_1, \cdots, z_j$, respectively, and such that $\delta(WR)$ is as small as possible.

In this paper we shall describe explicitly all functions R that solve the problem just stated. Our solution is based on our recent study [GKR] of interpolation problems for rational matrix functions with incomplete data (see also [GK]).

Another version of this problem, in which we only desire to shift either the poles $p_{i+1}, \cdots, p_k$ or the zeros $z_{j+1}, \cdots, z_l$, has been solved recently with a different method by Van Dooren [VD].

In § 1 of this paper we state our main result, which is proved in § 4. In §§ 2 and 3 we explain the necessary results from [GKR] and [GK]. In § 5 we shall consider as a special case the problem stated in [VD], and give a complete solution for the case when all functions considered are square and regular at infinity.

1. The main result. In this section we state the main theorem that will be proved later. Let $W(\lambda) = I_m + C(\lambda I_n - A)^{-1}B$ be a given rational matrix function, analytic and invertible for $\lambda \in \Gamma$, where Γ is a given contour in the complex plane. By γ_+ we shall denote the region inside Γ, by γ_- the region outside Γ. Let P, respectively $P^\times$, be the spectral projection of A, respectively, $A^\times$, corresponding to the eigenvalues in γ_+. We define the *γ_+-spectral triple* of W by

$$\tau_+ = \{(C|_M, A|_M), (A^\times|_{\operatorname{Im} P^\times}, P^\times B), P^\times|_M : M \rightarrow \operatorname{Im} P^\times\}$$

where $M = \operatorname{Im} P$.

* Received by the editors March 30, 1988; accepted for publication (in revised form) August 25, 1988.

† School of Mathematical Sciences, Tel Aviv University, Ramat Aviv, Israel.

‡ Faculteit Wiskunde en Informatica, Vrije Universiteit, Amsterdam, the Netherlands (ran@cs.vu.nl (UUCP network)).

The problem we consider is the following. Given a nonempty set Ω, $\Omega \cap \bar{\gamma}_+ = \varnothing$, find all rational matrix functions R such that

$$W_1(\lambda) := W(\lambda)R(\lambda)$$

has the following properties:

(1.1) W_1 has all its poles and zeros in $\Omega \cup \gamma_+$.

(1.2) The γ_+-spectral triple of W_1 is τ_+.

(1.3) The McMillan degree of W_1, $\delta(W_1)$, is as small as possible.

(1.4) $W_1(\infty) = I_m$.

This problem is exactly the problem stated in the Introduction. Indeed, condition (1.2) is equivalent to saying that R has no poles and zeros in γ_+. Note also that instead of γ_+ we could just consider the poles and zeros of W inside Γ, as is done in the Introduction.

Before stating the solution of this problem we must introduce several spaces and operators. Let $M^\times = \operatorname{Ker} P^\times$. Note that $\operatorname{Ker} P^\times|_M = M \cap M^\times$. Choose a complement N of $M \cap M^\times$ in M, and a complement K of $\operatorname{Im} P^\times|_M$ in $\operatorname{Im} P^\times$. Introduce the projections ρ_p in M onto $M \cap M^\times$ along N and ρ_z in $\operatorname{Im} P^\times$ onto K along $\operatorname{Im} P^\times|_M$. Also, let η_p be the imbedding of $M \cap M^\times$ into M and η_z the imbedding of K in $\operatorname{Im} P^\times$.

Since $(\rho_p C|_{M\cap M^\times}, \rho_p A|_{M\cap M^\times})$ is observable, there exist operators $G : \mathbb{C}^m \to M \cap M^\times$ and $S : M \cap M^\times \to M \cap M^\times$ such that

$$\rho_p(A - GC)|_{M\cap M^\times} = S, \tag{1.5}$$

$$\sigma(S) \subset \Omega. \tag{1.6}$$

Likewise, $(\rho_z A^\times|_K, \rho_z P^\times B)$ is controllable and there exist operators $F : K \to \mathbb{C}^m$ and $T : K \to K$ such that

$$\rho_z(A^\times - P^\times BF)|_K = T, \tag{1.7}$$

$$\sigma(T) \subset \Omega. \tag{1.8}$$

(Compare Rosenbrock's theorem [R]; see also [GKR], § 1.) Choose any such pair (S, G) and any such pair (F, T).

Let $X : K \to M$ and $Y : \operatorname{Im} P^\times \to M \cap M^\times$ be the unique solutions of the Lyapunov equations

$$A|_M X - XT = A_{12}, \tag{1.9}$$

$$YA^\times|_{\operatorname{Im} P^\times} - SY = A_{21}, \tag{1.10}$$

where $A_{12} : K \to M$ and $A_{21} : \operatorname{Im} P^\times \to M \cap M^\times$ are defined via the following equations:

$$A_{21}(I - \rho_z) = \rho_p(A|_M - S\rho_p - GC|_M)\Gamma^\dagger, \tag{1.11}$$

$$(I - \rho_p)A_{12} = \Gamma^\dagger(A^\times|_M - \eta_z T - P^\times BF)\eta_z, \tag{1.12}$$

$$A_{21}\rho_z = (\rho_p A_{12} - GF + \Gamma_1 T - S\Gamma_1)\rho_z. \tag{1.13}$$

Here $\Gamma_1 : K \to M \cap M^\times$ is arbitrary, and $\Gamma^\dagger$ is a generalized inverse of $P^\times|_M : M \to \operatorname{Im} P^\times$, such that $\Gamma^\dagger P^\times|_M = (I - \rho_p)$ and $P^\times \Gamma^\dagger = (I - \rho_z)$. The operator $\rho_p A_{12} : K \to M \cap M^\times$ is chosen arbitrarily. Then (1.13) defines $A_{21}\rho_z$, or, in other words, it defines the action of A_{21} on K. Since $\operatorname{Ker} \Gamma^\dagger = \operatorname{Ker}(I - \rho_z)$ the right-hand side of (1.11) defines $A_{21}(I - \rho_z)$. Next, since $\operatorname{Im} \Gamma^\dagger = \operatorname{Im}(I - \rho_p)$, $(I - \rho_p)A_{12}$ is well defined by (1.12).

Finally, let $\Gamma_0 = YP^\times X - Y\eta_z - \rho_p X + \Gamma_1$. Note that

$$\begin{bmatrix} P^\times|_M & -P^\times X+\eta_z \\ -YP^\times|_M+\rho_p & \Gamma_0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ -Y & I \end{bmatrix} \begin{bmatrix} P^\times|_M & \eta_z \\ \rho_p & \Gamma_1 \end{bmatrix} \begin{bmatrix} I & -X \\ 0 & I \end{bmatrix}.$$

Take L: Im $P^\times \to M$ such that

$$\rho_p L = -\Gamma_1 \rho_z, \qquad P^\times|_M (I-\rho_p)L + \rho_z = I_M.$$

(This way L is well defined in a unique manner.) Then the inverse of the middle operator is easily seen to be $[\begin{smallmatrix} L & \eta_p \\ \rho_z & 0 \end{smallmatrix}]$. It follows that

$$\text{(1.14)} \qquad \begin{bmatrix} P^\times|_M & -P^\times X+\eta_z \\ -YP^\times|_M+\rho_p & \Gamma_0 \end{bmatrix}^{-1} = \begin{bmatrix} \Gamma_{11}^\times & \eta_p \\ \rho_z & 0 \end{bmatrix},$$

where $\Gamma_{11}^\times = L + \eta_p Y + X\rho_z$: Im $P^\times \to M$.

THEOREM 1.1. *Let the pair* (S, G) *satisfy* (1.5), (1.6) *and let* (F, T) *satisfy* (1.7) *and* (1.8). *Further, let X and Y be given by* (1.9) *and* (1.10), *and* $\Gamma_{11}^\times$ *by* (1.14). *Then all functions* $R(\lambda)$ *such that* $W_1(\lambda) = W(\lambda)R(\lambda)$ *satisfies* (1.1)–(1.4) *are given by*

$$\text{(1.15)} \qquad \begin{aligned} R(\lambda) = I - (C, C(P-P^\times)X + F + C\eta_z)\left(\lambda - \begin{bmatrix} A^\times|_{M^\times} & V_+^\times \\ 0 & T \end{bmatrix}\right)^{-1} \\ \cdot \begin{bmatrix} B - (\Gamma_{11}^\times - \eta_p Y)P^\times B - \eta_p G \\ \rho_z P^\times B \end{bmatrix}, \end{aligned}$$

$$\text{(1.16)} \qquad \begin{aligned} R(\lambda)^{-1} = I + (C + \{C(PX\rho_z - \Gamma_{11}^\times) - F\rho_z\}P^\times, C\eta_p)\left(\lambda - \begin{bmatrix} A|_{\operatorname{Ker} P} & 0 \\ V_+ & S \end{bmatrix}\right)^{-1} \\ \cdot \begin{bmatrix} (I-P)B \\ \{\rho_p P + YP^\times(I-P)\}B - G \end{bmatrix}, \end{aligned}$$

where

$$V_+ = -YP^\times BC + GC, \qquad V_+^\times = -BCPX - BF.$$

The corresponding functions $W_1(\lambda)$ *are given by*

$$\text{(1.17)} \qquad \begin{aligned} W_1(\lambda) = I + C(\lambda - A|_M)^{-1}\{(\Gamma_{11}^\times - \eta_p Y)P^\times B + \eta_P G\} \\ + (-CPX - F)(\lambda - T)^{-1}\rho_z P^\times B, \end{aligned}$$

$$\text{(1.18)} \qquad \begin{aligned} W_1(\lambda)^{-1} = I - \{C(\Gamma_{11}^\times - PX\rho_z) + F\rho_z\}(\lambda - A^\times)^{-1}P^\times B \\ - C\eta_p(\lambda - S)^{-1}(-YP^\times B + G). \end{aligned}$$

The McMillan degree of $W_1(\lambda)$ *is given by*

$$\delta(W_1) = \dim M + \dim K = \dim \operatorname{Im} P^\times + \dim M \cap M^\times.$$

Theorem 1.1 describes all functions $R(\lambda)$ such that $W_1 = WR$ satisfies (1.1)–(1.4). This description does not provide an explicit parametrization of the set of all such divisors $R(\lambda)$. However, there is considerable freedom in the choice of R. In particular, there is freedom in the choice of the subspaces K and N, in the choices of the pairs of operators (G, S) and (F, T), as described in Rosenbrock's theorem [R], and finally in the choice of the operators Γ_1 and $\rho_p A_{12}$.

2. Admissible triples and factorization. In this section and the next, which have an auxiliary character, we recall several concepts and results from [GK] (see also [GKR]).

Let $W(\lambda) = I_m + C(\lambda I_n - A)^{-1}B$ be a minimal realization of W, and let σ be a set in the complex plane. First we introduce the following triple:

$$\tau = \{(C|_M, A|_M), (A^\times|_{\operatorname{Im} P^\times}, P^\times B), P^\times|_M : M \to \operatorname{Im} P^\times\} \tag{2.1}$$

where P, $P^\times$ are the spectral projections of A and $A^\times$, respectively, corresponding to the eigenvalues in σ, and $M = \operatorname{Im} P$. Any triple

$$\tau_+ = \{(C_p, A_p), (A_z, B_z), \Gamma\} \tag{2.2}$$

obtained from τ by

$$C_p = C|_M E_1, \qquad E_1^{-1} A|_M E_1 = A_p, \tag{2.3}$$

$$E_2^{-1} A^\times|_{\operatorname{Im} P^\times} E_2 = A_z, \qquad B_z = E_2^{-1} P^\times B, \tag{2.4}$$

$$\Gamma = E_2^{-1} P^\times|_M E_1, \tag{2.5}$$

is called a *σ-spectral triple for W*. In fact τ_+ contains a right pole pair for W: (C_p, A_p), a left zero pair for W: (A_z, B_z), and the matrix Γ coupling these two pairs via

$$\Gamma A_p - A_z \Gamma = B_z C_p. \tag{2.6}$$

(We easily check that (2.6) holds, by checking it for the special case of τ and using (2.3)-(2.5).)

Note that the pair (C_p, A_p) in a σ-spectral triple for a rational matrix function W is observable, and the pair (A_z, B_z) is controllable. From now on we shall call any triple $\tau = \{(C_p, A_p), (A_z, B_z), \Gamma\}$ with (C_p, A_p) observable, (A_z, B_z) controllable, $\sigma(A_p)$ and $\sigma(A_z)$ inside a set σ, and such that Γ satisfies (2.6) a *σ-admissible triple*.

If we take for σ a set containing all eigenvalues of A and $A^\times$, then M and $\operatorname{Im} P^\times$ just become the whole state space, and $P^\times|_M : M \to \operatorname{Im} P^\times$ is the identity operator. In that case the operator Γ in any σ-spectral triple for W will be invertible, and we have the following minimal realizations for W and W^{-1}:

$$W(\lambda) = I + C_p(\lambda - A_p)^{-1}\Gamma^{-1}B_z, \tag{2.7}$$

$$W(\lambda)^{-1} = I - C_p\Gamma^{-1}(\lambda - A_z)^{-1}B_z \tag{2.8}$$

(see [GKLR], [BR]). An admissible triple $\tau = \{(C_p, A_p), (A_z, B_z), \Gamma\}$ with Γ invertible will be called a *global spectral triple*. The function given by (2.7) will be called the function corresponding to τ in this case.

The following theorem describes the situation when two functions have a common σ-spectral triple.

THEOREM 2.1. *Let $W(\lambda) = I_m + C(\lambda I - A)^{-1}B$ and $W_1(\lambda) = I_m + C_1(\lambda - A_1)^{-1}B_1$ have a common σ-spectral triple. Then $R(\lambda) := W(\lambda)^{-1}W_1(\lambda)$ has no poles and zeros on σ. Let P, respectively, P_1, be the spectral projection of A, respectively A_1, corresponding to eigenvalues in σ, and let $P^\times$, respectively, $P_1^\times$, denote the same spectral projection for $A^\times$, respectively $A_1^\times$. Then the function $R(\lambda)$ has the following realization*:

$$\begin{aligned} R(\lambda) = {} & (I - C(I - P^\times)(\lambda - A^\times)^{-1}B)(I + C_1(I - P_1)(\lambda - A_1)^{-1}B) \\ & + C(\lambda - A^\times)^{-1}(I - P^\times)E_1P_1B_1 - CE_2P_1^\times(I - P_1)(\lambda - A_1)^{-1}B_1, \end{aligned} \tag{2.9}$$

$$\begin{aligned} R(\lambda)^{-1} = {} & (I - C_1(I - P_1^\times)(\lambda - A_1^\times)^{-1}B_1)(I + C(I - P)(\lambda - A)^{-1}B) \\ & + C_1(\lambda - A_1^\times)^{-1}(I - P_1^\times)E_1^{-1}PB - C_1E_2^{-1}P^\times(I - P)(\lambda - A)^{-1}B, \end{aligned} \tag{2.10}$$

where $E_1 : \mathrm{Im}\,P_1 \to \mathrm{Im}\,P$, $E_2 : \mathrm{Im}\,P_1^\times \to \mathrm{Im}\,P^\times$ *are invertible operators such that*

$$A|_{\mathrm{Im}\,P}E_1 = E_1A_1|_{\mathrm{Im}\,P_1}, \qquad C|_{\mathrm{Im}\,P}E_1 = C_1|_{\mathrm{Im}\,P_1}, \tag{2.11}$$

$$A^\times|_{\mathrm{Im}\,P^\times}E_2 = E_2A_1^\times|_{\mathrm{Im}\,P_1^\times}, \qquad P^\times B = E_2P_1^\times B_1, \tag{2.12}$$

$$P^\times|_{\mathrm{Im}\,P}E_1 = E_2P_1^\times|_{\mathrm{Im}\,P_1}. \tag{2.13}$$

Proof. Consider the σ-spectral triple for W given by

$$\tau = \{(C|_{\mathrm{Im}\,P}, A|_{\mathrm{Im}\,P}), (A^\times|_{\mathrm{Im}\,P^\times}, P^\times B), P^\times|_{\mathrm{Im}\,P} : \mathrm{Im}\,P \to \mathrm{Im}\,P^\times\}$$

and the σ-spectral triple for W_1 given by

$$\tau_1 = \{(C_1|_{\mathrm{Im}\,P_1}, A_1|_{\mathrm{Im}\,P_1}), (A_1^\times|_{\mathrm{Im}\,P_1^\times}, P_1^\times B_1), P_1^\times|_{\mathrm{Im}\,P_1} : \mathrm{Im}\,P_1 \to \mathrm{Im}\,P_1^\times\}.$$

Since W and W_1 have a common σ-spectral triple, there exist invertible operators E_1 and E_2 such that (2.11)–(2.13) hold. Using these identities and the fact that $BC = (\lambda - A^\times) - (\lambda - A)$, $B_1C_1 = (\lambda - A_1^\times) - (\lambda - A_1)$, it is straightforward to check (2.9) and (2.10). From these formulas we see that $R(\lambda)$ has no poles and zeros. □

The theorem has been proved in Theorem 5.1 of [GK] and also in Lemma 1.3 of [BGR]. In fact, in [GK] and [BGR] more has been proved. It turns out that the converse to Theorem 2.1 also holds, i.e., if $W(\lambda)^{-1}W_1(\lambda)$ has no poles and zeros on σ, then W_1 and W have a common σ-spectral triple. Since we do not use the converse in the sequel, we choose not to prove it.

3. Minimal complements. In this section $\tau = \{(C_p, A_p), (A_z, B_z), \Gamma\}$ is an admissible triple, where A_p and A_z act on X_p and X_z, respectively. We shall assume that Γ is not invertible, so that τ is not a global spectral triple for some function. Let $\sigma(A_p) \cup \sigma(A_z)$ be denoted by σ. In this section we shall give a description of all rational matrix functions $W(\lambda)$ with $W(\infty) = I$ such that τ is a σ-spectral triple for W, and with McMillan degree as small as possible.

Recall from [GK] that if $\tau_0 = \{(C_{p0}, A_{p0}), (A_{z0}, B_{z0}), \Gamma_0\}$ is an ε-admissible triple, where $\varepsilon \cap \sigma = \varnothing$, then $\tau \oplus \tau_0$ is defined by

$$\tau \oplus \tau_0 = \left\{([C_p, C_{p0}], A_p \oplus A_{p0}), \left(A_z \oplus A_{z0}, \begin{bmatrix} B_z \\ B_{z0} \end{bmatrix}\right), \begin{bmatrix} \Gamma & \Gamma_{12} \\ \Gamma_{21} & \Gamma_0 \end{bmatrix}\right\},$$

where Γ_{12} and Γ_{21} are the unique solutions of

$$\Gamma_{12}A_{p0} - A_z\Gamma_{12} = B_zC_{p0}, \qquad \Gamma_{21}A_p - A_{z0}\Gamma_{21} = B_{z0}C_p.$$

An ε-admissible triple τ_0 will be called a *complement* of τ if the coupling operator $\Gamma_{\tau \oplus \tau_0}$ of $\tau \oplus \tau_0$ is invertible. We say that τ_0 is a *minimal complement* of τ if it is a complement and for any other complement τ_0' we have

$$\operatorname{rank} \Gamma_{\tau \oplus \tau_0} \leqq \operatorname{rank} \Gamma_{\tau \oplus \tau_0'}.$$

If τ_0 is a minimal complement of τ then the function $W(\lambda)$ corresponding to $\tau \oplus \tau_0$ will be a solution to the problem stated in the first paragraph of this section, and any solution is obtained this way. So our problem can be rephrased as follows. Describe all minimal complements of τ.

To give this description we first introduce some spaces and projections. Let N be a complement in X_p to $\operatorname{Ker} \Gamma$, and let K be a complement in X_z to $\operatorname{Im} \Gamma$. So

$$X_p = N \dotplus \operatorname{Ker} \Gamma, \qquad X_z = \operatorname{Im} \Gamma \dotplus K.$$

Let ρ_p be the projection along N onto Ker Γ and ρ_z the projection along Im Γ onto K. Further, let η_p be the imbedding of Ker Γ into X_p, η_z the imbedding of K into X_z.

We easily verify that Ker Γ is (C_p, A_p)-invariant and that Im Γ is (A_z, B_z)-invariant. By Rosenbrock's theorem [R], there exist $S : \text{Ker}\,\Gamma \to \text{Ker}\,\Gamma$ and $G : \mathbb{C}^m \to \text{Ker}\,\Gamma$ such that

$$(3.1) \qquad \rho_p(A_p - GC_p)|_{\text{Ker}\,\Gamma} = S.$$

Likewise there exist $T : K \to K$ and $F : K \to \mathbb{C}^m$ such that

$$(3.2) \qquad \rho_z(A_z - B_zF)|_K = T.$$

In fact this can be done in such a way that $\sigma(S) \subset \varepsilon$, $\sigma(T) \subset \varepsilon$, where $\varepsilon \subset \mathbb{C}$ is an arbitrary set.

We also select a generalized inverse $\Gamma^\dagger$ of Γ such that $\Gamma\Gamma^\dagger = I - \rho_z$, $\Gamma^\dagger\Gamma = I - \rho_p$.

The next theorem describes all minimal complements of τ.

THEOREM 3.1. *Let $\tau = \{(C_p, A_p), (A_z, B_z), \Gamma\}$ be a σ-admissible triple. Choose a set $\varepsilon \subset \mathbb{C}$ such that $\sigma \cap \varepsilon = \emptyset$, and choose pairs (S, G) and (F, T) such that* (3.1) *and* (3.2) *hold, respectively, and $\sigma(S) \subset \varepsilon$, $\sigma(T) \subset \varepsilon$. Then*

$$(3.3) \qquad \tau_0 = \{(-C_pX - F, T), (S, -YB_z + G), \Gamma_0\}$$

is a minimal complement for τ, where $X : K \to X_p$ and $Y : X_z \to$ Ker Γ are the unique solutions of the Lyapunov equations

$$(3.4) \qquad A_pX - XT = A_{12},$$

$$(3.5) \qquad YA_z - SY = A_{21}.$$

Here $A_{12} : K \to X_p$ and $A_{21} : X_z \to$ Ker Γ are defined as follows. Choose $\Gamma_1 : K \to$ Ker Γ and ρ_pA_{12} arbitrarily, and let

$$(3.6) \qquad A_{21}(I - \rho_z) = \rho_p(A_p - S\rho_p - GC_p)\Gamma^\dagger,$$

$$(3.7) \qquad (I - \rho_p)A_{12} = \Gamma^\dagger(A_z - \eta_zT - B_zF)\eta_z,$$

$$(3.8) \qquad A_{21}\rho_z = \rho_pA_{12} + GF + \Gamma_1T - S\Gamma_1.$$

Finally, $\Gamma_0 = Y\Gamma X - Y\eta_z - \rho_pX + \Gamma_1$.

Conversely, every minimal complement

$$\tau_0 = \{(C_{p0}, A_{p0}), (A_{z0}, B_{z0}), \Gamma_0\}$$

with $A_{p0} : K \to K$ and A_{z0} : Ker $\Gamma \to$ Ker Γ is obtained as in the first part of the theorem up to a change of basis in K and Ker Γ. *Consequently, every minimal complement is similar to a minimal complement obtained as in the first part of the theorem.*

The theorem has been proved in [GKR], and for a special case also in [GK].

Here we will sketch only part of the proof, namely, that τ_0 is a minimal complement of τ. For full details and the proof of the converse part we refer the reader to [GKR].

First, it is an easy check that (3.1) and (3.6) are not contradictory, and that (3.2) and (3.7) do not contradict each other. Next, we show that $(C_pX + F, T)$ is observable and $(S, -YB_z + G)$ is controllable. Indeed, suppose $0 \neq x \in K$ is an unobservable eigenvector of the pair $(C_pX + F, T)$ corresponding to $\lambda \in \varepsilon$. Then $Tx = \lambda x$, $-C_pXx = Fx$. By (3.7) we have

$$B_zFx = -\Gamma A_{12}x - Tx + A_zx = -\Gamma A_{12}x - (\lambda - A_z)x.$$

On the other hand, by (2.6) and (3.4) we have

$$-B_zC_pXx=A_z\Gamma Xx-\Gamma A_pXx=A_z\Gamma Xx-\Gamma A_{12}x-\lambda\Gamma Xx.$$

So, by equating these two, we obtain

$$(\lambda-A_z)(x-\Gamma Xx)=0.$$

Since $\lambda\in\varepsilon$, $\lambda\notin\sigma(A_z)$, so $x=+\Gamma Xx$. But then $x\in K\cap\operatorname{Im}\Gamma=(0)$, which contradicts $x\neq 0$. So $(C_pX+F,\ T)$ is observable. The controllability of $(S,\ -YB_z+G)$ is proved likewise.

A straightforward computation using (2.6) and (3.4)–(3.8) shows

$$\Gamma_0T-S\Gamma_0=(-YB_z+G)(-C_pX-F).$$

So τ_0 is indeed an ε-admissible triple.

One next computes that the coupling operator of $\tau\oplus\tau_0$ is given by

$$\Gamma_{\tau\oplus\tau_0}=\begin{bmatrix}\Gamma & -\Gamma X+\eta_z\\ -Y\Gamma+\rho_p & \Gamma_0\end{bmatrix}:X_p\oplus K\rightarrow X_z\oplus\operatorname{Ker}\Gamma. \tag{3.9}$$

It is easily seen that this operator is invertible. Hence τ_0 is a complement to τ. Note that if τ_0' is a minimal complement of τ then

$$\operatorname{rank}\Gamma_{\tau\oplus\tau_0'}=\dim X_p\oplus K=\dim X_z\oplus\operatorname{Ker}\Gamma;$$

it then follows that τ_0 is a minimal complement.

4. Proof of the main theorem. In this section we give a proof of Theorem 1.1, using the concepts developed in the previous sections.

Let τ_+ be the γ_+-spectral triple of W given by

$$\tau_+=\{(C|_M,A|_M),(A^\times|_{\operatorname{Im}P^\times},P^\times B),P^\times|_M\},$$

and let $(S,\ G)$ satisfy (1.5), (1.6), and let $(F,\ T)$ satisfy (1.7), (1.8). Let X and Y be given by (1.9) and (1.10) and $\Gamma_{11}^\times$ by (1.14). Note that according to the result of § 3

$$\tau_0:=\{(-CPX-F,T),(S,-YP^\times B+G),\Gamma_0\}$$

is a Ω-minimal complement of τ_+, and that any Ω-minimal complement of τ_+ is of this form. Also, the function $W_1(\lambda)$ given by (1.17) is the function corresponding to $\tau_+\oplus\tau_0$ as we easily verify using (2.7). Define

$$R(\lambda)=W(\lambda)^{-1}W_1(\lambda).$$

Clearly, by construction W_1 satisfies (1.1)–(1.4). As (1.17), (1.18) are minimal realizations, the formulas for $\delta(W_1)$ hold. To obtain the formulas for $R(\lambda)$ and its inverse, it remains to use the formulas in Theorem 2.1, using the realizations (1.17) and (1.18) for $W_1(\lambda)$ and $W_1(\lambda)^{-1}$, and $W(\lambda)=I+C(\lambda-A)^{-1}B$, $W(\lambda)^{-1}=I-C(\lambda-A^\times)^{-1}B$.

Indeed, using the realization (1.18) for $W_1(\lambda)^{-1}$ we have

$$A_1^\times=\begin{bmatrix}A^\times|_{\operatorname{Im}P^\times} & 0\\ 0 & S\end{bmatrix},\qquad B_1=\begin{bmatrix}P^\times B\\ -YP^\times B+G\end{bmatrix},$$

$$C_1=[C(\Gamma_{11}^\times-PX\rho_z)+F\rho_z,C\eta_p],$$

$$A_1=\Gamma_{\tau_+\oplus\tau_0}\begin{bmatrix}A|_M & 0\\ 0 & T\end{bmatrix}\Gamma_{\tau_+\oplus\tau_0}^{-1}.$$

Using (3.9), we check that

$$E_1^{-1}=\begin{bmatrix} P^{\times}|_M \\ -YP^{\times}|_M+\rho_p \end{bmatrix}, \qquad E_2^{-1}=\begin{bmatrix} I|_{\operatorname{Im} P^{\times}} \\ 0 \end{bmatrix}.$$

Inserting in (2.9), we obtain

$$\begin{aligned} R(\lambda)^{-1}=&(I-C\eta_p(\lambda-S)^{-1}(-YP^{\times}B+G))(I+C(\lambda-A|_{\operatorname{Ker} P})^{-1}(I-P)B)\\ &+C\eta_p(\lambda-S)^{-1}(-YP^{\times}+\rho_p)PB\\ &-(C(\Gamma_{11}^{\times}-PX\rho_z)+F\rho_z)(\lambda-A|_{\operatorname{Ker} P})^{-1}(I-P)B, \end{aligned}$$

which is (1.16) after a little rewriting.

Likewise we prove (1.15) from (2.10) using the realization (1.17) of $W_1(\lambda)$. □

5. Displacement of either poles or zeros by cascade connection. In this section we consider the following problems that were considered before in [VD]. Given a rational $m \times m$ matrix function $W(\lambda) = I_m + C(\lambda I_n - A)^{-1}B$ and a nonempty set γ_+ we wish to factorize W as $W(\lambda) = W_1(\lambda)W_2(\lambda)$, where W_1, W_2 are square, regular, $W_1(\infty) = W_2(\infty) = I_m$ and

(5.1) The poles of W_2 and the zeros of W_1 are in γ_+,

(5.2) $\delta(W_1)$ is as small as possible,

or, where (5.1) is replaced by

(5.1′) the poles of W_1 and the zeros of W_2 are in γ_+.

We shall solve two problems, imposing seemingly stronger conditions on W_1 and W_2. Later we shall see that these conditions are actually equivalent to (5.1), (5.2) and (5.1′), (5.2), respectively.

Let W be as above, and let P be the spectral projection corresponding to the eigenvalues of A in γ_+. We introduce the admissible triple

$$\tau_-=\{(C|_{\operatorname{Ker} P},A|_{\operatorname{Ker} P}),(0,0),\Gamma\},$$

where $\Gamma : \operatorname{Ker} P \to (0)$ is the zero operator. The problem we consider is that of finding a function $R(\lambda)$ such that for

$$W_1(\lambda):=W(\lambda)R(\lambda)$$

we have the following:

(5.3) W_1 has all its poles and zeros in $\mathbb{C}$,

(5.4) The γ_--spectral triple of W_1 is τ_-,

(5.5) $\delta(W_1)$ is as small as possible,

(5.6) $W_1(\infty) = I_m$,

which is basically problem (1.1)–(1.4) for a special case. Clearly, if W_1 is a solution to this problem, then W_1 has all its zeros in γ_+. We shall construct a solution $R(\lambda)$ and show that with $W_2(\lambda)$ defined by

$$W_2(\lambda)=R(\lambda)^{-1}$$

we have a solution to problem (5.1)–(5.2), i.e., all the poles of W_2 are in γ_+.

Conversely, if W_1, W_2 solve (5.1), (5.2) it is clear that the zero pair for W_1 corresponding to zeros in γ_- is (0, 0). Also since W_2 has no poles in γ_-, W_1 has a pole pair corresponding to poles in γ_- that is as follows:

$$\left((C|_{\operatorname{Ker} P}, C_0), \begin{bmatrix} A|_{\operatorname{Ker} P} & * \\ 0 & A_0 \end{bmatrix} \right),$$

where P is the spectral projection of A corresponding to γ_+. However, since we want $\delta(W_1)$ to be as small as possible, it follows that the pole pair for W_1 corresponding to γ_- must in fact be $(C|_{\operatorname{Ker} P}, A|_{\operatorname{Ker} P})$. So the γ_--spectral triple for W_1 is

$$\tau_- = \{(C|_{\operatorname{Ker} P}, A|_{\operatorname{Ker} P}), (0,0), \Gamma\}$$

where $\Gamma : \operatorname{Ker} P \to (0)$ is the zero operator. So, with $R(\lambda) = W_2(\lambda)^{-1}$ we have a solution to (5.3)–(5.6). Hence the two problems are equivalent.

The construction of a solution to (5.3)–(5.6) is straightforward. Let $S : \operatorname{Ker} P \to \operatorname{Ker} P$ and $G : \mathbb{C}^m \to \operatorname{Ker} P$ be such that

$$(5.7) \qquad (A - GC)|_{\operatorname{Ker} P} = S, \quad \sigma(S) \subset \gamma_+.$$

Then a minimal complement to τ_- is given by

$$\tau_+ = \{(0,0), (S, G), I_{\operatorname{Ker} P}\}.$$

Let $W_1(\lambda)$ be the function corresponding to $\tau_- \oplus \tau_+$; then W_1 satisfies (5.3)–(5.6).

THEOREM 5.1. *Let* (S, G) *satisfy* (5.7). *Then a solution to* (5.1), (5.2) *is given by*

$$(5.8) \qquad W_1(\lambda) = I + C(\lambda - A|_{\operatorname{Ker} P})^{-1} G,$$

$$(5.9) \qquad W_1(\lambda)^{-1} = I - C(\lambda - S)^{-1} G,$$

$$(5.10) \qquad W_2(\lambda) = I + [C|_{\operatorname{Ker} P}, C|_M] \left(\lambda - \begin{bmatrix} S & -GC|_M \\ 0 & A|_M \end{bmatrix} \right)^{-1} \begin{bmatrix} (I-P)B - G \\ PB \end{bmatrix}$$

$$(5.11) \qquad W_2(\lambda)^{-1} = I - C(\lambda - A^\times)^{-1}(B - \eta G)$$

where $\eta : \operatorname{Ker} P \to \mathbb{C}^n$ *is the natural imbedding. Here* $M = \operatorname{Im} P$ *and* $A^\times = A - BC$.

Proof. Clearly, the function W_1 corresponding to $\tau_- \oplus \tau_+$ is given by (5.8), (5.9). Further,

$$\begin{aligned} W_2(\lambda) = W_1(\lambda)^{-1} W(\lambda) &= (I - C(\lambda - S)^{-1} G)(I + C(\lambda - A)^{-1} B) \\ &= I - C(\lambda - S)^{-1} G + C(\lambda - A)^{-1} B - C(\lambda - S)^{-1} GC(\lambda - A|_{\operatorname{Ker} P})^{-1}(I-P)B \\ &\quad - C(\lambda - S)^{-1} GC(\lambda - A|_M)^{-1} PB. \end{aligned}$$

Using (5.7), we obtain the formula for W_2. Note that indeed W_2 has all its poles in γ_+. Finally,

$$\begin{aligned} W_2(\lambda)^{-1} = W(\lambda)^{-1} W_1(\lambda) &= (I - C(\lambda - A^\times)^{-1} B)(I + C(\lambda - A)^{-1} \eta G) \\ &= I - C(\lambda - A^\times)^{-1} B + C(\lambda - A)^{-1} \eta G - C(\lambda - A^\times)^{-1} BC(\lambda - A)^{-1} G. \end{aligned}$$

Using $BC = (\lambda - A^\times) - (\lambda - A)$, we get the formula in the theorem. □

Comparing it with the Algorithm 3.1 in Van Dooren's paper [VD], we see that our solution is precisely the same as his solution. Note that the minimal McMillan degree of W_1, $\delta(W_1)$ turns out to be dim Ker P, i.e., the number of poles of R in γ_- (counting multiplicities).

Now consider problem (5.1′), (5.2). A reasoning analogous to the one used in the previous case shows that the γ_--spectral triple for W_1 is now given by

$$\tau_- = \{(0,0),(A^\times|_{M^\times},(I-P^\times)B),0\},$$

where $P^\times$ is the spectral projection of $A^\times$ corresponding to γ_+ and $M^\times = \operatorname{Ker} P^\times$.

THEOREM 5.2. *Let $T: M^\times \to M^\times$ and $F: M^\times \to \mathbb{C}^m$ satisfy*

$$(A^\times - (I-P^\times)BF)|_{M^\times} = T, \quad \sigma(T) \subset \gamma_+.$$

Then a solution to problem (5.1′), (5.2) *is given by*

$$W_1(\lambda) = I - F(\lambda - T)^{-1}(I-P^\times)B,$$

$$W_1(\lambda)^{-1} = I + F(\lambda - A^\times|_{M^\times})^{-1}(I-P^\times)B,$$

$$W_2(\lambda) = I + (C - F(I-P^\times))(\lambda - A)^{-1}B,$$

$$W_2(\lambda)^{-1} = I - [C, F-C]\left(\lambda - \begin{bmatrix} A^\times|_{\operatorname{Im} P^\times} & -P^\times BF \\ 0 & T \end{bmatrix}\right)^{-1} \begin{bmatrix} P^\times B \\ (I-P^\times)B \end{bmatrix}.$$

Proof. Any γ_+-minimal complement of τ_- is of the form $\{(F, T), (0, 0), -I\}$ where F and T are as in the theorem. The rest of the proof is analogous to the proof of Theorem 5.1. □

REFERENCES

[BGR] J. A. BALL, I. GOHBERG, AND L. RODMAN, *Minimal factorization of meromorphic matrix functions in terms of local data*, Integral Equations Operator Theory, 10 (1987), pp. 309–348.

[BR] J. A. BALL AND A. C. M. RAN, *Global inverse spectral problems for rational matrix functions*, Linear Algebra Appl., 86 (1987), pp. 237–282.

[VD] P. VAN DOOREN, *Rational and polynomial matrix factorizations via recursive pole-zero cancellation*, Philips Research Laboratory, 1986, preprint.

[GK] I. GOHBERG AND M. A. KAASHOEK, *An inverse spectral problem for rational matrix functions and minimal divisibility*, Integral Equations Operator Theory, 10 (1987), pp. 437–465.

[GKLR] I. GOHBERG, M. A. KAASHOEK, L. LERER, AND L. RODMAN, *Minimal divisors of rational matrix functions with prescribed zero and pole structure*, in Topics in Operator Theory, Systems and Networks, H. Dym and I. Gohberg, eds., Operator Theory 12, Birkhäuser, Basel, 1984, pp. 241–275.

[GKR] I. GOHBERG, M. A. KAASHOEK, AND A. C. M. RAN, *Interpolation problems for rational matrix functions with incomplete data and Wiener-Hopf factorization*, in Topics in Interpolation Theory of Rational Matrix Functions, I. Gohberg, ed., Operator Theory 33, Birkhäuser, Basel, 1988, pp. 73–108.

[R] H. H. ROSENBROCK, *State Space and Multivariate Theory,* Nelson, London, 1970.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 3, pp. 326-346, July 1989

005

ON THE CONVERGENCE OF THE CYCLIC JACOBI METHOD FOR PARALLEL BLOCK ORDERINGS*

GAUTAM SHROFF† AND ROBERT SCHREIBER†

Abstract. Convergence of the cyclic Jacobi method for diagonalizing a symmetric matrix has never been conclusively settled. Forsythe and Henrici [*Trans. Amer. Math. Soc.*, 94 (1960), pp. 1-23] proved convergence for a cyclic by rows ordering. Here orderings are investigated that can be obtained from the cyclic by rows ordering through convergence preserving combinatorial transformations. First the class of "cyclic wavefront" orderings is introduced and it is shown that the class consists of exactly those orderings that are "equivalent" to the cyclic by rows ordering. It is also shown that certain block Jacobi methods are cyclic wavefront orderings when viewed as cyclic Jacobi methods. While discussing convergence proofs for parallel implementations of Jacobi methods and block Jacobi methods, the notions of "weak equivalence" and "P-equivalence" of Jacobi orderings is developed. Next the class of "P-wavefront" orderings is introduced that includes all orderings related to the cyclic by rows ordering through any known convergence preserving transformations. Finally, it is shown that the "P-wavefront" orderings are characterized by simple properties that can be verified efficiently (in polynomial time).

Key words. Jacobi method, parallel, singular value decomposition, eigenvalues, orderings

AMS(MOS) subject classification. 65F15

1. Cyclic Jacobi methods.

A Jacobi method for diagonalizing a symmetric $n \times n$ matrix A performs a sequence of similarity transformations

$$A_{k+1} = U_k A_k U_k^T, \qquad k = 0, 1, 2 \cdots \tag{1}$$

where $A_0 = A$ and U_k, $k = 0, 1, \cdots$ is an orthogonal plane rotation. Let $A_k = [a_{ij}^{(k)}]$. The elements of $U_k = [u_{ij}]$ are defined as follows. For every $k = 0, 1, \cdots$ there is a pair $(i, j) = (i_k, j_k)$ with $1 \leqq i < j \leqq n$, such that

$$u_{pq} = \begin{cases} 1 & \text{if } p = q \text{ and } p \neq i, j, \\ \cos \phi_k & \text{if } p = i \text{ and } q = i, \\ \cos \phi_k & \text{if } p = j \text{ and } q = j, \\ \sin \phi_k & \text{if } p = i \text{ and } q = j, \\ -\sin \phi_k & \text{if } p = j \text{ and } q = i, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

and

$$\tan 2\phi_k = \frac{2a_{ij}^{(k)}}{a_{ii}^{(k)} - a_{jj}^{(k)}}. \tag{3}$$

The choice of ϕ_k according to (3) assures that

$$a_{ji}^{(k+1)} = a_{ij}^{(k+1)} = 0.$$

(Note that there are four values of ϕ_k that satisfy (3), and one of them always lies in the interval $(-\pi/4, \pi/4)$.) We say that U_k is a Jacobi rotation that annihilates (or rotates)

* Received by the editors April 25, 1988; accepted for publication (in revised form) November 7, 1988. This research was supported by the Office of Naval Research under contract N00014-86-k-0610.

† Computer Science Department, Rensselaer Polytechnic Institute, Troy, New York 12180 (Shroffg@turing.cs.rpi.edu).

$a_{ij}^{(k)}$. U_k will be written as $U_k^{(ij)}$ or $U^{(ij)}$ to indicate this. We want A_k to converge to a diagonal matrix Σ that contains the eigenvalues of A on the diagonal. The closeness of A_k to Σ is measured by the quantity

$$S_k = \sum_{\substack{p \neq q \\ 1 \leqq p \leqq n \\ 1 \leqq q \leqq n}} |a_{pq}^{(k)}|^2. \tag{4}$$

If $S_k \to 0$ as $k \to \infty$ for any A, we say that the Jacobi method converges.

The Jacobi method for computing the singular value decomposition of a matrix was introduced by Kogbetliantz [9]. Here we will focus on the eigenvalue problem although the discussion can be easily translated to the SVD setting.

There are various strategies for choosing the order in which the off-diagonal elements are annihilated in a Jacobi method (i.e., the sequence of pairs (i_k, j_k), $k = 0, 1, \cdots$). In the classical Jacobi method, we choose the element largest in magnitude as the next one to be rotated. This guarantees convergence [16], but incurs the cost of searching the matrix for the largest element before every rotation. Moreover, it is a highly sequential algorithm.

In a cyclic Jacobi method, the $N = (n(n-1)/2)$ off-diagonal elements are rotated in some predetermined order, each element being rotated exactly once in any "sweep" of N rotations. Convergence of any cyclic Jacobi method can be guaranteed by omitting the annihilation of elements that are smaller than some threshold [16]. Although this ensures convergence, the rate of convergence may be slow if the threshold is not decreased repeatedly from an initially large value [13]. Further, choosing the thresholds efficiently requires information about the entire matrix and is therefore to be avoided in a parallel setting. Proving convergence for any cyclic ordering without using thresholds is more difficult because it is possible for an element to be pushed around ahead of the sequence of rotations and never be annihilated [6].

If we assume that each sweep starts with the (1, 2) element, there are

$$[n(n-1)/2-1]!$$

different cyclic orderings. However, of the possible orderings, some are equivalent, as discussed below.

Let $T_{ij}(A)$ be the transformation of A that occurs when the (i, j) element is rotated, i.e.,

$$T_{ij}(A) = U^{(ij)} A U^{(ij)^T}.$$

Consider the transformations T_{ij} and T_{rs}, where i, j, r, s are distinct. It is easily verified that because of the special structure of the matrices $U^{(ij)}$ and $U^{(rs)}$ (defined in (2)), these two matrices commute, i.e.,

$$U^{(ij)} U^{(rs)} = U^{(rs)} U^{(ij)}.$$

Therefore, the transformations T_{ij} and T_{rs} also commute, i.e.,

$$T_{ij} T_{rs}(A) = T_{rs} T_{ij}(A).$$

Commuting Rotations. In any cyclic Jacobi ordering O, the rotations of elements (i, j) and (r, s) are said to commute if i, j, r, s are distinct and the rotation (i, j) immediately precedes the rotation (p, q) in the ordering O.

Equivalent Orderings. Let T_1 and T_2 each be a product of N transformations of the form T_{ij} representing a single sweep of the Jacobi method with two different cyclic orderings O_1 and O_2. We say that O_1 is equivalent to O_2 if the transformations T_1 can be changed into T_2 through a sequence of transpositions of commuting rotations. Note that for equivalent orderings,

$$T_1(A)=T_2(A) \quad \text{for any } A,$$

i.e., equivalent orderings give the same matrices at the end of the sweep. Therefore if the Jacobi method converges using ordering O_1, it also converges using O_2.

Cyclic By Rows Ordering. The rotations are chosen according to the following rule. The first rotation in the sweep is $(1, 2)$. A rotation (p, q) is followed by

$$\begin{array}{lll} (p,q+1) & \text{if } p<n-1, & q<n, \\ (p+1,p+2) & \text{if } p<n-1, & q=n, \\ (1,2) & \text{if } p=n-1, & q=n. \end{array} \tag{5}$$

THEOREM 1 (Forsythe and Henrici [5]). *Let a sequence of Jacobi transformations be applied to a symmetric matrix A. Further, let the angle ϕ_k be restricted as follows:*

$$\phi_k\in[a,b] \quad \textit{and} \quad -\frac{\pi}{2}<a<b<\frac{\pi}{2}. \tag{6}$$

If the off-diagonal elements are annihilated using a cyclic by rows ordering, then this Jacobi method converges.

That (6) is always realizable is clear from (3), since tan $2\phi_k$ takes all values in any open interval of length $\pi/2$.

In this paper, we first identify the class of orderings that can be obtained from the cyclic by rows ordering through a sequence of transpositions of commuting rotations. The orderings so obtained will be equivalent to the row ordering. Therefore, they will converge.

Cyclic Wavefront Orderings. In a cyclic ordering O of the pairs

$$\{(i,j), 1\leqq i<j\leqq n\},$$

let $I(i, j)$ be the index at which the pair (i, j) occurs. If

$$I(i,j-1)<I(i,j)<I(i+1,j) \tag{7}$$

for all $1 \leqq i < j \leqq n$, then O is called a cyclic wavefront ordering.

In other words, in a cyclic wavefront ordering the element immediately to the left of (i, j) in the same row and that above (i, j) in the same column is rotated before it. Also the element immediately to the right of (i, j) in the same row and that below (i, j) in the same column is rotated after it. Since this holds for all pairs (i, j) as stated in (7), we can easily show the following lemma.

LEMMA 1.1. *In a cyclic wavefront ordering, for all $1 \leqq i < j \leqq n$, and $1 \leqq p < q \leqq n$,*

$$\begin{array}{ll} \text{(i)} & I(i,j)\geqq I(p,q) \quad \textit{if } p\leqq i \textit{ and } q\leqq j, \\ \text{(ii)} & I(i,j)\leqq I(p,q) \quad \textit{if } p\geqq i \textit{ and } q\geqq j. \end{array} \tag{8}$$

This property of cyclic wavefront orderings is illustrated in Fig. 1. We see that in a cyclic wavefront ordering, every element to the right and below (i, j) is rotated after it, and every element to the left and above (i, j) is rotated before it.

In § 5 we will prove the following theorem regarding the class of cyclic wavefront orderings.

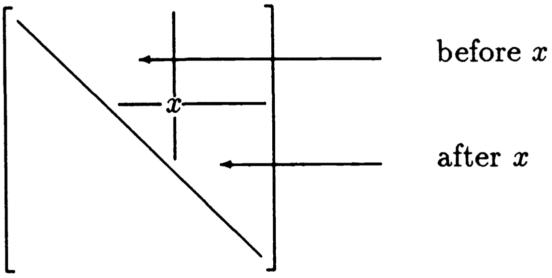


FIG. 1. *Lemma* 1.1.

$$\begin{pmatrix} x & 1 & 2 & 3 & 4 \\ & x & 5 & 6 & 7 \\ & & x & 8 & 9 \\ & & & x & 10 \\ & & & & x \end{pmatrix} \quad \begin{pmatrix} x & 1 & 2 & 4 & 7 \\ & x & 3 & 5 & 8 \\ & & x & 6 & 9 \\ & & & x & 10 \\ & & & & x \end{pmatrix} \quad \begin{pmatrix} x & 1 & 2 & 3 & 5 \\ & x & 4 & 6 & 7 \\ & & x & 8 & 9 \\ & & & x & 10 \\ & & & & x \end{pmatrix}$$

Cyclic by rows Cyclic by columns Antidiagonals

FIG. 2. *Examples of cyclic orderings.*

THEOREM 2. *A cyclic Jacobi ordering is equivalent to the cyclic by rows ordering if and only if it is a cyclic wavefront ordering.*

In Fig. 2 we illustrate some well-known cyclic orderings that fall in the class of cyclic wavefront orderings. The cyclic by rows ordering was defined in (5). The cyclic by columns ordering was proved equivalent to the row ordering by Hansen [6]. The antidiagonals ordering was used by Luk and Park [11] in their discussion of parallel Jacobi methods. In § 7 we will show that membership of a cyclic ordering in the class of wavefront orderings can easily be established. In the next section we will introduce some more cyclic wavefront orderings in the context of block Jacobi methods.

2. Block Jacobi methods. The motivation for developing the block algorithms is that in almost every modern computer, computation is significantly cheaper than input/output. Block algorithms for matrix problems typically work with blocks of data having $O(d^2)$ elements, performing $O(d^3)$ work on the block before requiring another memory access. The $O(d)$ ratio of work to storage means that processors with an $O(d)$ ratio of computing speed to input/output bandwidth can be tolerated.

In the Jacobi methods discussed above, we can view the rotation of an off-diagonal element (i, j) as solving a 2×2 eigenvalue problem:

$$\begin{pmatrix} u_{ii} & u_{ij} \\ u_{ji} & u_{jj} \end{pmatrix} \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{pmatrix} \begin{pmatrix} u_{ii} & u_{ij} \\ u_{ji} & u_{jj} \end{pmatrix}^T = \begin{pmatrix} a'_{ii} & 0 \\ 0 & a'_{jj} \end{pmatrix}.$$

In a block Jacobi method we work with larger subproblems. Consider the matrix A as a $b \times b$ block matrix with b^2 blocks, each of $d \times d$ size. If (i, j) identifies an off-diagonal block A_{ij}, we associate the submatrix

$$P = \begin{pmatrix} A_{ii} & A_{ij} \\ A_{ji} & A_{jj} \end{pmatrix}$$

with this block element. There are different ways in which we may "solve" this subproblem, not all of which will assure convergence of the matrix as a whole to diagonal form. Here we will consider solving the subproblem by one sweep of Jacobi rotations on some or all elements of the matrix P. By accumulating the rotations in this sweep, we get a "block rotation" that "solves" the subproblem:

$$\begin{pmatrix} U_{ii} & U_{ij} \\ U_{ji} & U_{jj} \end{pmatrix} \begin{pmatrix} A_{ii} & A_{ij} \\ A_{ji} & A_{jj} \end{pmatrix} \begin{pmatrix} U_{ii} & U_{ij} \\ U_{ji} & U_{jj} \end{pmatrix}^T = \begin{pmatrix} A'_{ii} & A'_{ij} \\ A'_{ji} & A'_{jj} \end{pmatrix}.$$

This block rotation is then applied to the ith and jth block rows and columns, just as in the scalar Jacobi methods.

The subproblems themselves may be chosen using various cyclic orderings. A sequence of subproblems that includes every block index $\{(i, j);\ 1 \leqq i < j \leqq b\}$ will be called a block sweep. We now describe a class of block Jacobi methods that are also cyclic Jacobi methods. We will then show that these methods are cyclic wavefront methods.

Block Wavefront Jacobi. Let each block sweep be generated as follows.

(1) Let a subproblem (i, j) be solved by applying a sweep of any cyclic wavefront Jacobi process to the subproblem, but including only certain elements as described below:

(9)
1) If $j > i + 1$ and $i < b - 1$, then the elements in A_{ij} are included;
2) If $j = i + 1$ and $i < b - 1$, then the elements in A_{ii} and A_{ij} are included;
3) If $(i, j) = (b - 1, b)$, then the elements in the entire subproblem are included.

(2) Let the subproblems be chosen using a cyclic wavefront ordering.

When the block ordering is a cyclic by rows ordering, with the subproblems themselves being solved using the cyclic by rows ordering, the subproblems chosen and the elements included in their solution are shown in Fig. 3 (for the case $b = 3$ and $d = 3$).

THEOREM 3. *The block wavefront Jacobi method converges to a diagonal matrix.*

Proof. Note that if one block sweep of this method is viewed as a sequence of scalar Jacobi rotations applied to the matrix as a whole, each element is rotated only once, so block wavefront methods are a particular class of cyclic Jacobi methods. To prove convergence, all that needs be shown is that this is indeed a cyclic wavefront Jacobi ordering. Then convergence is assured by Theorem 2.

Consider three adjacent off-diagonal elements in the matrix as shown below. Let X, Y, and Z be the elements $(i, j - 1)$, (i, j), and $(i + 1, j)$, respectively. Let Y lie in block

$$\left[\begin{array}{ccc|ccc|ccc}
x & 1 & 2 & 3 & 4 & 5 & 13 & 14 & 15 \\
 & x & 6 & 7 & 8 & 9 & 16 & 17 & 18 \\
 & & x & 10 & 11 & 12 & 19 & 20 & 21 \\
\hline
 & & & x & 22 & 23 & 24 & 25 & 26 \\
 & & & & x & 27 & 28 & 29 & 30 \\
 & & & & & x & 31 & 32 & 33 \\
\hline
 & & & & & & x & 34 & 35 \\
 & & & & & & & x & 36 \\
 & & & & & & & & x
\end{array}\right]$$

FIG. 3. *Block cyclic by rows.*

(l, m). Let $h = I(i, j-1)$, $r = I(i, j)$, and $s = I(i, j+1)$ be times at which these elements are annihilated in a particular block sweep. We have to show that $h < r < s$:

$$\begin{array}{cc} X^{(h)} & Y^{(r)} \\ & Z^{(s)}. \end{array}$$

Case 1. X, Y, Z all lie in the same block. Since a wavefront ordering is used on the subproblem, $h < r < s$ is guaranteed by (7).

Case 2. If X lies in an adjacent block that is an off-diagonal block $(l, m-1)$, then since the subproblems are solved in a cyclic wavefront order, the block $(l, m-1)$ is solved before (l, m), so $h < r$ holds.

Case 3. If Z lies in an adjacent, off-diagonal block $(l+1, m)$, then that block is solved after (l, m), and $s > r$ follows.

Case 4. If X lies in an adjacent diagonal block, then by (9)(1)2), the elements in that block will be included in the wavefront process used on this subproblem $(m-1, m)$, and so $h < r$.

Case 5. Z lies in an adjacent diagonal block (m, m). If $m \neq b$, then by (9)(1)2), the elements in that block will be rotated during the solution of subproblem $(m, m+1)$. Since the subproblems are chosen in a wavefront order, $(m, m+1)$ is solved after $(m-1, m)$, therefore $s > r$. If $m = b$, then by (9)(1)3), Z will be rotated during the solution of this block $(b-1, b)$, and since a wavefront ordering is used on the subproblem, $s > r$ is guaranteed.

So we have shown that block wavefront Jacobi methods are cyclic wavefront orderings and therefore converge. □

3. Parallel Jacobi methods. Jacobi methods may be ideally suited for parallel implementation. In any cyclic Jacobi ordering O, the order in which commuting rotations are performed does not affect the transformation T that represents one sweep of the method. Commuting rotations can therefore be performed in parallel. Several parallel Jacobi orderings have been proposed [10]. Luk and Park [11] demonstrate the convergence of some of these methods. Here we will briefly review their arguments and formalize some of the concepts involved. Consider the following example.

Antidiagonals Ordering. A rotation (p, q), $1 \leqq p < q \leqq n$, is followed by

$$\begin{array}{ll} (p+1, q-1) & \text{if } q-p>2, \\ (1, p+q) & \text{if } q-p \leqq 2, p+q \leqq n, \\ (p+q+1-n, n) & \text{if } q-p \leqq 2, n<p+q<2n-1, \\ (1,2) & \text{if } q=n \text{ and } p=n-1. \end{array}$$

The antidiagonals ordering is shown in Fig. 4. It is easily verified that this is a cyclic wavefront ordering. We can obtain a parallel antidiagonals ordering by performing the commuting rotations that occur on each antidiagonal in parallel.

$$\begin{pmatrix} x & 1 & 2 & 3 & 5 \\ & x & 4 & 6 & 7 \\ & & x & 8 & 9 \\ & & & x & 10 \\ & & & & x \end{pmatrix} \qquad \begin{pmatrix} x & 1 & 2 & 3 & 4 \\ & x & 3 & 4 & 5 \\ & & x & 5 & 6 \\ & & & x & 7 \\ & & & & x \end{pmatrix}$$

Antidiagonals Parallel antidiagonals

FIG. 4.

In the parallel antidiagonals ordering for the $n \times n$ case, all off-diagonal elements are rotated in $2n-3$ stages. The number of rotations performed in parallel at stage j is $\lfloor (j+1)/2 \rfloor$ for $1 \leqq j \leqq n-1$, and $\lfloor n-(j+1)/2 \rfloor$ for $n-1 < j \leqq 2n-3$. To achieve more parallelism at every stage, we need to extend the idea of equivalent orderings to that of weakly equivalent orderings.

Shift-Equivalent Orderings. Let $I(p, q)$ and $I'(p, q)$ denote the times at which (p, q) is rotated in orderings O and O', respectively. O and O' are shift-equivalent if

$$I(p,q)=(I'(p,q)-1+c) \bmod N+1 \quad \text{for } 1 \leqq p<q \leqq n \tag{10}$$

where $N = n(n-1)/2$ and $0 \leqq c < N$. We will say that O is obtained from O' by a shift of c. It is easily seen that this also means that O' is obtained from O by a shift of $-c$.

Weakly Equivalent Orderings. Two orderings O and O' are weakly equivalent if either

(1) The ordering O is shift-equivalent to O'' and O'' is equivalent to O', or

(2) If O and O'' are weakly equivalent and O'' and O' are weakly equivalent.

We can express this in another way as follows. Let $S(c)$ denote a shift by an amount c. Let t_i denote a transposition of two commuting rotations. Then if two orderings are weakly equivalent, one can be obtained from the other by a transformation of the form

$$S(c_1)t_1t_2\cdots S(c_2)t_kt_{k+1}\cdots . \tag{11}$$

For example in Fig. 5, O' is obtained from O by a shift of 2, and O'' from O' by the transposition of a pair of commuting rotations. So O'' is weakly equivalent to O.

LEMMA 3.1. *If O and O' are two weakly equivalent orderings, and the cyclic Jacobi method using O' converges, then the cyclic Jacobi method using O converges.*

Proof. We need only consider shift equivalent orderings. Let ordering O be obtained from O' by a shift of c. Let A be a given symmetric matrix. Let A_c be the matrix obtained by applying the first c rotations from the sequence O to A. If the cyclic Jacobi method using O' converges then it converges for A_c. But applying Jacobi rotations to A_c using ordering O' generates the same sequence of matrices as applying Jacobi rotations to A using ordering O. Since A is arbitrary, the cyclic Jacobi method using O converges. □

Weakly Wavefront Orderings. An ordering O is a weakly wavefront ordering if it is weakly equivalent to a wavefront ordering.

THEOREM 4. *The class of weakly wavefront orderings converges.*

Proof. The proof is obtained by directly applying Theorem 2 and Lemma 3.1. □

In § 6 we will characterize weakly wavefront orderings by a simple property that can be easily tested. We now show how to use this notion to develop convergent parallel orderings.

The basic idea is due to Luk and Park [11]. Let O be the antidiagonals ordering. For simplicity we will illustrate the argument for the case $n = 5$. Consider the ordering O' that is obtained from O by a shift of 2, as shown in Fig. 5.

$$\begin{array}{ccc}
\begin{pmatrix} x & 1 & 2 & 3 & 5 \\ & x & 4 & 6 & 7 \\ & & x & 8 & 9 \\ & & & x & 10 \\ & & & & x \end{pmatrix} &
\begin{pmatrix} x & \mathbf{3} & 4 & 5 & 7 \\ & x & 6 & 8 & 9 \\ & & x & 10 & 1 \\ & & & x & \mathbf{2} \\ & & & & x \end{pmatrix} &
\begin{pmatrix} x & \mathbf{2} & 4 & 5 & 7 \\ & x & 6 & 8 & 9 \\ & & x & 10 & 1 \\ & & & x & \mathbf{3} \\ & & & & x \end{pmatrix} \\
O & O' & O''
\end{array}$$

FIG. 5.

$$\begin{pmatrix} x & 1 & 2 & 3 & 4 \\ & x & 3 & 4 & 5 \\ & & x & 5 & 1 \\ & & & x & 2 \\ & & & & x \end{pmatrix}$$

FIG. 6. *Modulus ordering.*

The ordering O'' is obtained from O' by transposing the commuting rotations (1, 2) and (4, 5). Therefore O'' is weakly equivalent to O by definition.

Now note that in the ordering O'', in addition to the rotations on the antidiagonals, the pairs of rotations (1, 2), (3, 5) and (1, 3), (4, 5) also commute. By performing commuting rotations in parallel we obtain a parallel ordering in which two rotations are performed in parallel at every stage. This ordering is shown in Fig. 6 and is called the modulus ordering [11].

Modulus Ordering. If $I(p, q)$ denotes the time at which (p, q) is rotated in the modulus ordering, then

$$I(p,q)=((p+q-3) \bmod n)+1.$$

In the modulus ordering, $\lfloor n/2 \rfloor$ rotations are performed in parallel at each stage. The above arguments can be easily generalized in the $n \times n$ case to show that the modulus ordering is weakly equivalent to the antidiagonals ordering. Thus it is a weakly wavefront ordering and converges due to Theorem 4.

4. Parallel block Jacobi methods. In this section we develop a class of provably convergent parallel block Jacobi methods in which multiple subproblems can be solved in parallel. Consider the following class of block Jacobi methods.

Block Weakly Wavefront Jacobi. Let each block sweep be generated in the following manner:

(1) The subproblems are chosen using a weakly wavefront ordering.

(2) Each subproblem is solved using one sweep of any cyclic wavefront Jacobi process on particular elements of the subproblem, as in the definition of block wavefront methods (9).

Remark. In the case when the subproblems are chosen using a modulus ordering, we get a class of parallel block Jacobi methods.

THEOREM 5. *Block weakly wavefront Jacobi methods converge to a diagonal matrix.*

Proof. We will prove the theorem by showing the following lemma.

LEMMA 4.1. *Let M_1 and M_2 be two block Jacobi methods that use orderings O_1 and O_2 to choose the subproblems. The ordering used within the subproblems as well as the way the matrix is divided into blocks is the same for both methods. Let M_2 converge:*

(1) *If O_1 is shift-equivalent to O_2, then M_1 converges;*

(2) *If O_1 is equivalent to O_2, then M_1 converges.*

Proof. The proof for part (1) is similar to the proof of Lemma 3.1. Let O_1 be obtained from O_2 by a shift of c. Let A_c be obtained from A by solving the first c subproblems of the sequence O_1. Then if M_2 converges, it converges for A_c. But M_1 applied to A generates the same iterates as M_2 applied to A_c. So M_1 converges.

To show (2), we first note the following easily established fact. Let $T = t_1 t_2 \cdots t_n$ and $R = r_1 r_2 \cdots r_n$ be defined by two sequences of Jacobi transformations. If t_i and r_j commute for all i, j, then $TR = RT$.

Define T_{ij} to be the sequence of Jacobi rotations used in the solution of the subproblem (i, j). If p, q, r, s are distinct integers, then the rotations in the solution of (p, q) lie in the (p, q), (p, p) and (q, q) blocks. The rotations used in the solution of (r, s) lie in the (r, s), (r, r) and (s, s) blocks of the matrix. By the above fact, the sequences $T_{pq}T_{rs}$ and $T_{rs}T_{pq}$ are equivalent. So, the subproblems (p, q) and (r, s) are "independent," since the order in which they are solved does not change the effect of one block sweep of the block Jacobi method.

Therefore, independent subproblems (p, q) and (r, s) can be transposed in the ordering O_2 without affecting convergence. □

Since a weakly wavefront ordering is weakly equivalent to a wavefront ordering by definition, the theorem follows. □

5. Convergence of cyclic wavefront Jacobi. In this section we will prove Theorem 2. Let C_R be the cyclic by rows ordering. Let O be a cyclic wavefront ordering. We will prove the "if" part of the theorem by induction on the length of the leading common subsequence of O' and C_R, where O' is equivalent to C_R. We will prove the "only if" part by showing that (a) C_R is wavefront, and (b) transpositions of commuting rotations preserve the wavefront property.

It easily follows from Lemma 1.1 that every wavefront ordering starts with the $(1, 2)$ rotation. Suppose $O = (1, 2), (1, 3), \cdots, (p, q), (x, y) \cdots$, where $(1, 2)$, $(1, 3) \cdots (p, q)$ is a leading subsequence of C_R. We first consider the case when $(p, q + 1)$ follows (p, q) in C_R.

LEMMA 5.1. *If* $O = (1,2), (1,3), \cdots, (p,q), (x,y), \cdots, (x',y'), (p,q+1), \cdots$ *is a cyclic wavefront ordering, then neither* p *nor* $q + 1$ *occurs as any of the indices in the sequence* $(x, y), \cdots, (x', y')$.

Proof. Let $h = I(p, q)$, $s = I(p, q + 1)$ in O. Suppose (m, n) is an element in the sequence $(x, y), \cdots, (x', y')$, i.e., $r = I(m, n)$, and $h < r < s$. Referring to Fig. 7, only the following situations can occur.

Case 1. (m, n) is (p, j) with $j < q$.

Case 2. (m, n) is $(i, q + 1)$ with $i > p$.

Case 3. (m, n) is (p, j) with $j > q + 1$.

Case 4. (m, n) is (i, p) with $i < p$.

Case 5. (m, n) is $(q + 1, j)$ with $j > q + 1$.

Case 6. (m, n) is $(i, q + 1)$ with $i < p$.

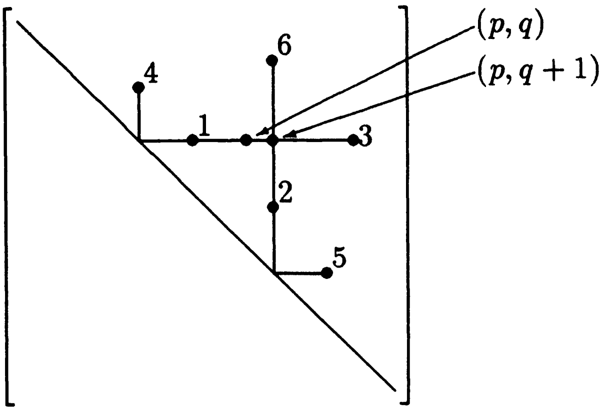


FIG. 7.

We know that O is a cyclic wavefront ordering, so by Lemma 1.1 none of Cases 1–5 can occur. Case 6 cannot occur since $(i, q+1)$ precedes (p, q) in C_R and by assumption (p, q) is the end of a leading common subsequence of O and C_R. □

We must now consider the case when (p, q) is followed by $(p+1, p+2)$ in C_R, i.e., when $q = n$.

LEMMA 5.2. *If* $O = (1, 2), (1, 3), \cdots, (p, q), (x, y), \cdots, (x', y'), (p+1, p+2), \cdots$ *is a cyclic wavefront ordering, then neither* $p+1$ *nor* $p+2$ *occurs as any of the indices in the sequence* $(x, y), \cdots, (x', y')$.

Proof. Let $h = I(p, q)$ and $s = I(p+1, p+2)$ in O. Let (m, n) be in the sequence $(x, y), \cdots, (x', y')$, i.e., if $r = I(m, n)$ then $h < r < s$. Referring to Fig. 8, we see that there are the following cases.

Case 1. $(m, n) = (p+1, j)$ with $j > p+2$.

Case 2. $(m, n) = (p+2, j)$ with $j > p+2$.

Case 3. $(m, n) = (i, p+2)$ with $i \leqq p$.

Case 4. $(m, n) = (i, p+1)$ with $i \leqq p$.

Since O is a cyclic wavefront ordering, none of the above cases can occur by Lemma 1.1. □

Thus by transposing $(p, q+1)$ or $(p+1, p+2)$ with the elements $(x, y), \cdots, (x', y')$ in turn we arrive at an equivalent ordering $O' = (1, 2), (1, 3), \cdots, (p, q), (p, q+1), (x, y), \cdots, (n-1, n)$ or $O' = (1, 2) \cdots (p, q), (p+1, p+2), (x, y) \cdots (n-1, n)$, which has a longer leading subsequence in common with C_R. Furthermore, the trailing sequence of O', $(x, y), \cdots, (n-1, n)$ is a proper subsequence of O, so it also satisfies the wavefront property (7). This allows us to continue with this process of extending the leading common subsequence of O' and C_R until $O' = C_R$. We have therefore shown that if O is a cyclic wavefront ordering, it can be obtained from the cyclic by rows ordering through a sequence of transpositions of commuting rotations.

To show the converse, let $O = (1, 2), \cdots, (i, j), (r, s), (m, n), (p, q), \cdots, (n-1, n)$ be a wavefront ordering, and suppose that (r, s) and (m, n) commute. Then $O' = (1, 2), \cdots, (i, j), (m, n), (r, s), (p, q), \cdots, (n-1, n)$ is also a cyclic wavefront ordering. This is because the sequences $(1, 2), \cdots, (i, j)$ and $(p, q), \cdots, (n-1, n)$ are subsequences of O and therefore satisfy the wavefront property. Further, the positions of (r, s) and (m, n) relative to these sequences do not change when we go from O to O'. Thus any ordering O' obtained from a cyclic wavefront ordering O through a sequence

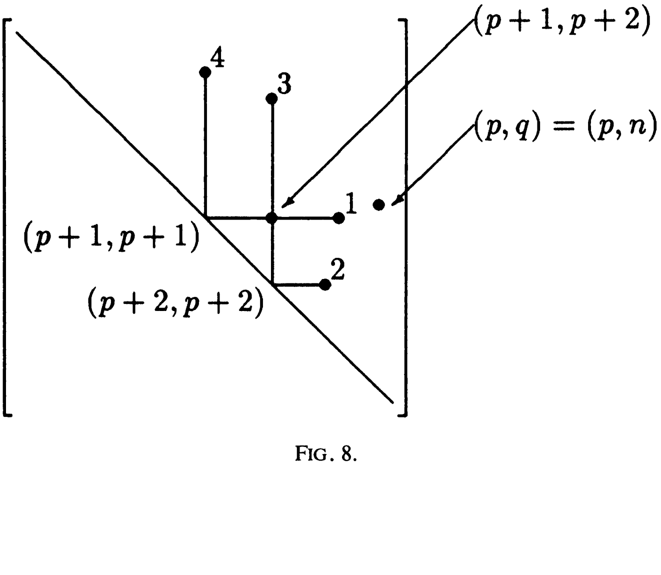


FIG. 8.

of transpositions of commuting rotations is also a cyclic wavefront ordering. From the definition (7) and Lemma 1.1, C_R is a wavefront ordering. Therefore the result is proved. □

6. Weakly wavefront orderings. The cyclic wavefront orderings are exactly the orderings equivalent to the cyclic by rows ordering, and they are characterized by a simple property (7).

The weakly wavefront orderings have been introduced in § 3 as those orderings that are weakly equivalent to a wavefront ordering. In this section we will show that these orderings are also easy to characterize.

Splitting. Consider two disjoint subsets X, Y of $P = \{(i, j);\ 1 \leqq i < j \leqq n\}$ such that $X \cup Y = P$. If the subsets satisfy the following:

(12) (1) If $(p, q) \in X$, then all (i, j), $i \leqq p$ and $j \leqq q$ are also in X.
(2) If $(p, q) \in Y$, then all (i, j), $i \geqq p$ and $j \geqq q$ are also in Y.
Then the pair (X, Y) is called a splitting.

Restricting O to X or Y gives a partial ordering of the pairs $\{(i, j);\ 1 \leqq i < j \leqq n\}$. When we refer to the *ordering* X or Y under the ordering O, we mean the restriction mentioned above. The *set* X, or Y, however, is just a set of pairs, independent of the ordering.

Good Splitting. A splitting (X, Y) is called a good splitting of the ordering O if the following conditions are satisfied:

(13) (1) The orderings X and Y under O each satisfy the wavefront property (7).
(2) For every $(r, s) \in Y$ and $(i, j) \in X$ such that $I(i, j) < I(r, s)$, i, j, r, s are distinct integers.
(3) The pair $(1, 2) \in X$.

A splitting divides the ordering O into two subsets. The pairs to the left and above any element of X are also in subset X, and those to the right and below any element of Y are also in subset Y. A splitting is good if, first, the pairs in X satisfy the wavefront property (7) among themselves and the same is true of the pairs in Y. Second, every pair in Y has different indices than all those pairs in X that occur before it in the ordering O.

Figure 9 shows two orderings O and O' and a splitting (X, Y) for each ordering. The pairs in the set Y are shown enclosed in boxes. The splitting of O' is not a good splitting because the pair $(3, 5) \in Y$ does not commute with $(2, 5) \in X$, which occurs before it in O', thus violating property (13)(2), although (13)(1) is satisfied. It is easily verified that the splitting of O is a good splitting.

Note that any wavefront ordering has a trivial good splitting with $X = P$ and $Y = \varnothing$.

$$\underset{O}{\left(\begin{array}{cccccc}
 & 1 & 3 & 5 & 8 & 10 \\
 & & 7 & 9 & 11 & 13 \\
 & & & 12 & 14 & 15 \\ \cline{5-6}
 & & & & \multicolumn{1}{|c}{2} & \multicolumn{1}{c|}{4} \\
 & & & & \multicolumn{1}{|c}{} & \multicolumn{1}{c|}{6} \\ \cline{5-6}
 & & & & &
\end{array}\right)}
\quad
\underset{O'}{\left(\begin{array}{cccccc}
 & 1 & 3 & 5 & 8 & 9 \\
 & & 6 & 7 & 10 & 11 \\ \cline{4-6}
 & & & \multicolumn{1}{|c}{2} & 4 & \multicolumn{1}{c|}{12} \\
 & & & \multicolumn{1}{|c}{} & 13 & \multicolumn{1}{c|}{14} \\
 & & & \multicolumn{1}{|c}{} & & \multicolumn{1}{c|}{15} \\ \cline{4-6}
 & & & & &
\end{array}\right)}$$

FIG. 9.

Canonical Ordering. An ordering is canonical if $I(1, 2) = 1$.

It is obvious that any ordering is shift equivalent, and therefore weakly equivalent, to a canonical ordering. Also, each ordering is shift equivalent to a unique canonical ordering, and so it makes sense to talk of the canonical form of an ordering. Further, two orderings are weakly equivalent if and only if their canonical forms are weakly equivalent. So, without loss of generality, we can restrict our attention to canonical orderings. Note that every wavefront ordering is in canonical form.

We now show that the notion of a good splitting allows us to characterize the weakly wavefront orderings. In the next section we will give an efficient algorithm that checks if an ordering has a good splitting.

THEOREM 6. *A canonical ordering is a weakly wavefront ordering if and only if it has a good splitting.*

Before we prove the theorem, we show that the definition of weak equivalence can be reformulated in terms of transpositions of rotations, by also allowing certain transpositions other than those of commuting rotations.

The following notation will be used in the sequel. $I(i, j)$ will refer to the position of pair (i, j) in the ordering O, $I'(i, j)$ to its position in ordering O', etc.

Admissible Transpositions. The transposition of two rotations (i, j) and (r, s) will be called admissible if i, j, r, s are distinct and $|I(i, j) - I(r, s)| \equiv 1 \bmod N$.

Admissible transpositions allow the transposition of rotations involving distinct indices if the rotations are either consecutive or are the first and last rotations in the ordering.

LEMMA 6.1. *Let O_1, O_2, O_3, O_4 be orderings such that we have the following:*

(1) O_2 is obtained from O_1 through a shift of c.

(2) O_3 is obtained from O_2 through transposition of two commuting rotations (p, q) and (r, s).

(3) O_4 is obtained from O_3 by a shift of $-c$.

Then O_4 can be obtained from O_1 through an admissible transposition of the rotations (p, q) and (r, s).

Proof. Suppose $I_2(p, q) = I_2(r, s) - 1$ and $I_3(r, s) = I_3(p, q) - 1$, since these are commuting rotations in O_2. Consider the case $c > 0$. The case $c < 0$ is similar.

If $I_2(p, q) \neq c$, then $I_1(p, q) = I_2(p, q) - c$ and $I_1(r, s) = I_2(r, s) - c$. So $I_1(p, q) = I_1(r, s) - 1$. Similarly, $I_4(p, q) = I_3(p, q) - c$ and $I_4(r, s) = I_3(r, s) - c$, so $I_4(r, s) = I_4(p, q) - 1$. Therefore (p, q) and (r, s) are commuting rotations in O_1, and O_4 is obtained when they are transposed.

If $I_2(p, q) = c$, then $I_1(p, q) = N$ and $I_1(r, s) = 1$. Also, $I_4(p, q) = 1$ and $I_4(r, s) = N$. So (p, q) and (r, s) are the first and last rotations in O_1 and their transposition yields the ordering O_4.

If $c < 0$, the argument is similar. □

We can now reformulate the definition of weak equivalence in term of admissible transpositions through the following lemma.

LEMMA 6.2. *If two orderings are weakly equivalent, one can be obtained from the other through a sequence of admissible transpositions followed by a shift, i.e., through a transformation of the form*

$$a_1 a_2 a_3 \cdots a_k S(c), \tag{14}$$

where a_i denotes an admissible transposition.

Proof. A shift of c followed by a shift of $-c$ is the identity transformation, so $S(c_1)t$ is equivalent to $S(c_1)tS(-c_1)S(c_1)$ (where t denotes a transposition of commuting rotations). But by Lemma 6.1, this is equivalent to $aS(c_1)$, (where a is an admissible transposition). Using this, (11) can be shown to be equivalent to (14). □

Proof of Theorem 6. We will first prove the "only if" part of the argument. We have to show that every canonical weakly wavefront ordering has a good splitting. From Lemma 6.2 it follows that a weakly wavefront ordering can be obtained from a wavefront ordering through a sequence of admissible transpositions followed by a shift. We have already noted that a wavefront ordering has a trivial good splitting. Since shift equivalent orderings have the same canonical form, all we need to show is that admissible transpositions preserve the property of possessing a good splitting.

Consider an ordering O' obtained from a canonical ordering O by a single admissible transposition. Let (X, Y) be a good splitting of O.

First consider the case when $(1, 2)$ is not involved in the transposition. Then, the transposition must be a transposition of commuting rotations (p, q) and (r, s) with $I(p, q) = I(r, s) - 1$. Since transposing two adjacent rotations does not change the relative position of any other pair in O with respect to the pairs (p, q) and (r, s), it is easily verified that all the properties of a good splitting still hold, and (X, Y) is a good splitting of O' as well.

Now suppose $(1, 2)$ is involved in the transposition. $I(1, 2) = 1$ since O is canonical. There are only two possible admissible transpositions, $(1, 2)$ with (p, q), where $I(p, q) = N$ or $I(p, q) = 2$. Consider $I(p, q) = N$.

CLAIM 1. $(p, q) \in X$.

Proof. Suppose $(p, q) \in Y$. Then $(p, q) = (n - 1, n)$ since Y satisfies the wavefront property. Consider the pair $(1, n)$. It cannot be in X since $(1, n)$ and $(n - 1, n)$ do not have disjoint indices, and (X, Y) was assumed to be a good splitting of O. It cannot be in Y since $(1, 2)$ and $(1, n)$ do not have disjoint indices. Therefore $(p, q) \in X$. □

In the ordering O', $I'(1, 2) = N$ and $I'(p, q) = 1$. Let O'' be the canonical form of O', obtained after a shift by $+1$. $I''(1, 2) = 1$ and $I''(p, q) = 2$.

An example of each of the orderings O, O', and O'' is shown in Fig. 10.

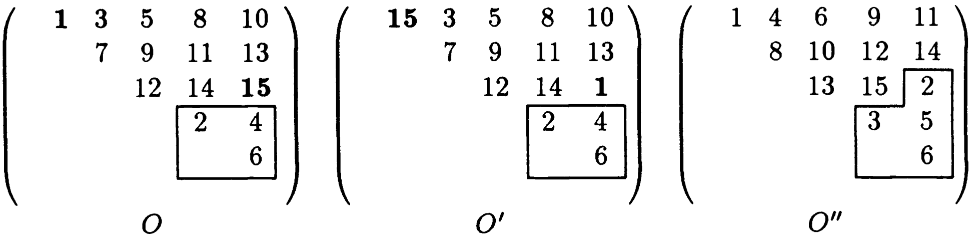
CLAIM 2. *For every* (i, j) *such that* $i \geqq p, j \geqq q$ *and* $(i, j) \neq (p, q)$, $(i, j) \in Y$.

Proof. Since $(p, q) \in X$ and $I(p, q) = N$, if $(i, j) \in X$, it would violate the assumption that ordering X satisfies the wavefront property. □

Now consider the splitting (X'', Y'') of O'', where $X'' = X - \{(p, q)\}$ and $Y'' = Y \cup \{(p, q)\}$ (as shown in Fig. 10). Any pair $(i, j) \neq (1, 2)$ or (p, q) was unaffected in going from O to O'. Also, O'' is just a shift of 1 applied to O'. So,

$$I''(i,j) = I(i,j) + 1 \quad \text{for } (i,j) \neq (1,2) \text{ or } (p,q). \tag{15}$$

The orderings X and Y satisfy the wavefront property since (X, Y) is a good splitting of O. Therefore, by (15), all the pairs in X'' besides $(1, 2)$ satisfy the wavefront property and all the pairs in Y'' besides (p, q) satisfy the wavefront property. But $I''(1, 2) = 1$, so X'' satisfies the wavefront property. Also $I''(p, q) = 2$ and Claim 2 show that Y'' satisfies the wavefront property.



$$\begin{pmatrix} \mathbf{1} & 3 & 5 & 8 & 10 \\ & 7 & 9 & 11 & 13 \\ & & 12 & 14 & \mathbf{15} \\ & & & 2 & 4 \\ & & & & 6 \end{pmatrix} \quad \begin{pmatrix} \mathbf{15} & 3 & 5 & 8 & 10 \\ & 7 & 9 & 11 & 13 \\ & & 12 & 14 & \mathbf{1} \\ & & & 2 & 4 \\ & & & & 6 \end{pmatrix} \quad \begin{pmatrix} 1 & 4 & 6 & 9 & 11 \\ & 8 & 10 & 12 & 14 \\ & & 13 & 15 & 2 \\ & & & 3 & 5 \\ & & & & 6 \end{pmatrix}$$

O $\qquad$ O' $\qquad$ O''

FIG. 10. *Transposing* $(1, 2)$ *and* $(3, 6)$, $I(3, 6) = N = 15$.

The remaining property of a good splitting (13)(2) is true for all pairs in Y. The set X'' is smaller than X, therefore (15) implies that (13)(2) is true for all pairs in Y'' except (p, q). The only pair in X'' that precedes (p, q) in O'' is $(1, 2)$. But (p, q) and $(1, 2)$ have distinct indices since they have been transposed. Therefore the ordering O'' satisfies (13). So (X'', Y'') is a good splitting of O''.

Now consider the case when $(1, 2)$ and (p, q), such that $I(p, q) = 2$, are transposed.

CLAIM 3. $(p, q) \in Y$.

Proof. If $(p, q) \in X$, then because X satisfies the wavefront property, $(p, q) = (1, 3)$. (If $(r, s) = (1, 3) \neq (p, q)$, then $I(r, s) > I(p, q) = 2$ would violate the wavefront property.) But $1, 2, p, q$ must be distinct if $(1, 2)$ and (p, q) are to be transposed. Hence $(p, q) \in Y$. □

The ordering O' is obtained by transposing $(1, 2)$ and (p, q). In O', $I'(1, 2) = 2$ and $I'(p, q) = 1$. To get the canonical form O'', we now have to shift by -1. In O'', $I''(1, 2) = 1$ and $I''(p, q) = N$.

An example of each of the orderings O, O', and O'' is shown in Fig. 11.

CLAIM 4. *For every (i, j) such that $i \leqq p, j \leqq q$ and $(i, j) \neq (p, q)$, $(i, j) \in X$.*

Proof. Since $(p, q) \in Y$ and $I(p, q) = 2$, if $(i, j) \in Y$, it would violate the assumption that ordering Y satisfies the wavefront property. □

Now consider the splitting (X'', Y''), where $X'' = X \cup \{(p, q)\}$ and $Y'' = Y - \{(p, q)\}$. By the same argument as used to obtain (15), we have

$$I''(i,j) = I(i,j) - 1 \quad \text{for } (i,j) \neq (1,2) \text{ or } (p,q). \tag{16}$$

The orderings X and Y satisfy the wavefront property. So, by (16), all the pairs in Y'' satisfy the wavefront property and all the pairs in X'' besides (p, q) and $(1, 2)$ satisfy the wavefront property. But $I''(1, 2) = 1$, $I''(p, q) = N$ and Claim 4 show that X'' satisfies the wavefront property.

The remaining property of a good splitting (13)(2) also follows for Y'' since it was true for Y, and the only pair added to X to get X'' is the last one in the ordering O'' and so it cannot precede any pair in Y''.

Thus we have shown that given a canonical ordering O with a good splitting, the ordering O' obtained from O by an admissible transposition has a canonical form O'' that also has a good splitting. So every canonical weakly wavefront ordering has a good splitting.

Now to prove the "if" part of the theorem, we have to show that if a canonical ordering has a good splitting, it is weakly equivalent to a wavefront ordering.

Consider a canonical ordering O and (X, Y) a good splitting of O. Let

$$X = (1,2), \cdots, (p,q), \qquad Y = (r,s), \cdots, (n-1,n),$$

$$\underset{O}{\left(\begin{array}{ccccc} \mathbf{1} & 3 & 5 & 8 & 10 \\ & 7 & 9 & 11 & 13 \\ & & 12 & 14 & 15 \\ \cline{4-5} & & & \multicolumn{1}{|c}{\mathbf{2}} & \multicolumn{1}{c|}{4} \\ & & & \multicolumn{1}{|c}{} & \multicolumn{1}{c|}{6} \\ \cline{4-5} \end{array}\right)} \quad \underset{O'}{\left(\begin{array}{ccccc} \mathbf{2} & 3 & 5 & 8 & 10 \\ & 7 & 9 & 11 & 13 \\ & & 12 & 14 & 15 \\ \cline{4-5} & & & \multicolumn{1}{|c}{\mathbf{1}} & \multicolumn{1}{c|}{4} \\ & & & \multicolumn{1}{|c}{} & \multicolumn{1}{c|}{6} \\ \cline{4-5} \end{array}\right)} \quad \underset{O''}{\left(\begin{array}{ccccc} 1 & 2 & 4 & 7 & 9 \\ & 6 & 9 & 10 & 12 \\ & & 11 & 13 & 14 \\ \cline{5-5} & & & 15 & \multicolumn{1}{|c|}{3} \\ & & & & \multicolumn{1}{|c|}{5} \\ \cline{5-5} \end{array}\right)}$$

FIG. 11. *Transposing* $(1, 2)$ *and* $(4, 5)$, $I(4, 5) = 2$.

$$\underset{\text{Has good splitting}}{\left(\begin{array}{ccccc} 1 & 3 & 5 & 8 & 10 \\ & 7 & 9 & 11 & 13 \\ & & 12 & 14 & 15 \\ & & & \multicolumn{2}{c}{\boxed{\begin{matrix} 2 & 4 \\ & 6 \end{matrix}}} \end{array}\right)} \quad \underset{\text{Transpositions}}{\begin{array}{c} (1,2) \leftrightarrow (4,5) \\ (1,3) \leftrightarrow (4,6) \\ (1,2) \leftrightarrow (4,6) \\ (1,4) \leftrightarrow (5,6) \\ (1,3) \leftrightarrow (5,6) \\ (1,2) \leftrightarrow (5,6) \end{array}} \quad \underset{\text{Has property } P}{\left(\begin{array}{ccccc} 4 & 5 & 6 & 8 & 10 \\ & 7 & 9 & 11 & 13 \\ & & 12 & 14 & 15 \\ & & & \multicolumn{2}{c}{\boxed{\begin{matrix} 1 & 2 \\ & 3 \end{matrix}}} \end{array}\right)}$$

FIG. 12.

where the pairs are arranged in order of their position in O. We will show how to transform the ordering O by a sequence of transpositions of commuting rotations so that Y satisfies the property.

PROPERTY P. For every $(i, j) \in Y$ and $(l, m) \in X$, the positions of these pairs in O satisfy $I(i, j) < I(l, m)$.

Suppose we have a maximal leading subsequence S of Y, $(r, s), \cdots, (f, g)$, that satisfies P. Then if $Y = (r, s) \cdots (f, g)(s, t) \cdots$, we know that there exists a pair $(a, b) \in X$, $I(a, b) < I(s, t)$. Also $I(a, b) = I(s, t) - 1$, since otherwise property P would not hold for the sequence S. But now we can transpose (a, b) and (s, t) in O, because of the property (13)(2). Eventually, we will have exhausted all $(i, j) \in X$ such that $I(i, j) < I(s, t)$. So the pair (s, t) can be added to S and S still satisfies P. Proceeding in this fashion, Y will eventually satisfy P. The situation is explained in Fig. 12. Note that (X, Y) is still a good splitting of the ordering obtained, since transpositions of commuting rotations preserve this property.

Now that Y satisfies P, let $I(1, 2) = d$. Consider a shift of ordering O by $-d$, to obtain ordering O'.

CLAIM 5. *O' is a wavefront ordering.*

Proof. Before the shift, for $(i, j) \in Y$, $1 \leqq I(i, j) \leqq d - 1$ follows from P, so after the shift $N - d + 1 \leqq I'(i, j) \leqq N$. Also for $(i, j) \in X$, before the shift, $d \leqq I(i, j) \leqq N$. After the shift $1 \leqq I'(i, j) \leqq N - d$. Therefore in O', all the pairs in X occur before the pairs in Y. Also note that since (X, Y) was a good splitting of O, the orderings X and Y under O' are still wavefront orderings.

Consider (i, j) in Y. If any neighboring element is in Y, then the wavefront property (7) is satisfied. (By neighboring element we mean one of $(i-1, j)$, $(i+1, j)$, $(i, j-1)$, and $(i, j+1)$.) If it is in X, it must occur before (i, j) in O' and it must be either $(i-1, j)$ or $(i, j-1)$ by the definition of a splitting. So the wavefront property is satisfied for all pairs in O' that are in Y.

Consider (i, j) in X. If any neighboring element is in X, then the wavefront property (7) is satisfied. If it is in Y, it must occur after (i, j) in O' and it must be either $(i+1, j)$ or $(i, j+1)$ by the definition of a splitting. So the wavefront property is satisfied for all pairs in O'. □

We have shown that O' obtained from O by transpositions of commuting rotations and shifts is a wavefront ordering if O has a good splitting. Hence the theorem is proved. □

7. Verifying the weakly wavefront property.

In this section we describe an algorithm that checks if a given canonical ordering has a good splitting. We will also be able to identify if the good splitting is the trivial one associated with a wavefront ordering, and thus check for the wavefront property as well.

The algorithm will go through the pairs in the order specified by the given ordering, marking some pairs as it scans them, according to the following rule. The first pair is marked. For other pairs, a pair is marked if its north and west neighbors have already been marked. At the end of the pass, the marked and unmarked pairs form a splitting of the ordering. A second pass is required to check if the splitting is a good splitting.

Let I be a matrix such that $I(i, j)$ is the position of pair (i, j) in the ordering O. We are also given a function next (i, j) that returns the pair following (i, j) in the ordering O.

```
ALGORITHM GOOD SPLITTING.
For $i, j \notin [1, n]$ or $i = j$, $I(i, j) \leftarrow 0$.
$S \leftarrow$ empty list
$(i, j) \leftarrow (1, 2)$
$I(i, j) \leftarrow 0$
for $c = 1$ to $N - 1$
   $(i, j) = \text{next}(i, j)$
   if $I(i-1, j) = 0$ and $I(i, j-1) = 0$
      $I(i, j) \leftarrow 0$
   else
      $I(i, j) \leftarrow -1$
   endif
endfor
```

This completes the first pass of the algorithm. At this stage, we have identified a splitting (X, Y) with X = all those pairs marked with a 0 and Y = all the pairs marked with a -1. Further, we have constructed X by proceeding through the pairs in the order they appear in O. So the partial ordering X satisfies the wavefront property (7). Now we do another pass through the pairs to check the remaining properties of a good splitting.

```
$(i, j) \leftarrow (1, 2)$
for $c = 1$ to $N$
   if $I(i, j) = 0$
      append the indices (not the pair) $i, j$ to the list $S$.
   endif
   if $I(i, j) = -1$
      { Here we check for the remaining properties that $Y$ must satisfy }
      if any of the following are true, return a NO:
        $I(i-1, j) = -1$ or $I(i, j-1) = -1$   {violates(13)(1)}
        $S \cap \{i, j\} \neq \emptyset$              {violates(13)(2)}
   endif
      $I(i, j) \leftarrow -2$
      endif
   $I(i, j) \leftarrow \text{next}(i, j)$
endfor
return YES
end.
```

LEMMA 7.1. *If* (X, Y) *is the splitting of* O *identified by Algorithm Good Splitting, then we have the following*:

(1) X *satisfies the wavefront property* (7).

(2) *If* (X_1, Y_1) *is any other splitting of* O *in which* X_1 *satisfies* (7), *then* $X_1 \subseteq X$.

$$\underset{\text{Before}}{\begin{pmatrix} 0 & 1 & 3 & 5 & 8 & 10 \\ & 0 & 7 & 9 & 11 & 13 \\ & & 0 & 12 & 14 & 15 \\ & & & 0 & 2 & 4 \\ & & & & 0 & 6 \\ & & & & & 0 \end{pmatrix}} \quad \underset{\text{After pass 1}}{\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 & 0 & 0 \\ & & & 0 & 0 & 0 \\ & & & & -1 & -1 \\ & & & & & -1 \\ & & & & & \end{pmatrix}} \quad \underset{\text{After pass 2}}{\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 & 0 & 0 \\ & & & 0 & 0 & 0 \\ & & & & -2 & -2 \\ & & & & & -2 \\ & & & & & \end{pmatrix}}$$

FIG. 13.

Proof. The ordering X satisfies (7) by construction since a pair is placed in X only if its north and west neighbors occur before it in O. Let $(i, j) \in X_1$ but not in X. Therefore when (i, j) is encountered in the first pass of the algorithm, either $(i, j-1)$ or $(i-1, j)$ has not been marked, i.e., it occurs after (i, j) in O. This contradicts the assumption that X_1 satisfies the wavefront property. So $X_1 \subseteq X$. □

LEMMA 7.2. *A good splitting* (X, Y) *of an ordering* O *is unique and satisfies property* (2) *in Lemma* 7.1.

Proof. Let (X_1, Y_1) be another splitting in which X_1 satisfies the wavefront property. Let $S = X_1 \cap Y$. Since X_1 satisfies the wavefront property, the ordering S does too. Let $(p, q) \in S$ be the first according to ordering O. Clearly, $(p, q) \neq (1, 2)$ since it is in Y. Let $(r, s) = (p-1, q)$, or $(p, q-1)$ (either will do, and at least one of them is always defined). Since (p, q) is in X_1, by (12) the pair (r, s) is also in X_1. Since X_1 satisfies the wavefront property, $I(r, s) < I(p, q)$, so $(r, s) \notin S$. Therefore $(r, s) \in X \cap X_1$.

Now, by (13)(2), (p, q) cannot share any indices with a pair in X that precedes it in O. But $(r, s) \in X$ contradicts this. Therefore S must be the empty set, so $X_1 \subseteq X$.

To show that a good splitting is unique, we now suppose that (X_1, Y_1) is a good splitting. Then since X satisfies the wavefront property, we can reverse the above arguments to show that $X \subseteq X_1$. So $(X, Y) = (X_1, Y_1)$. □

We have shown above that in good splitting (X, Y), X is the largest set that satisfies the wavefront property. Further, Algorithm Good Splitting returns this unique splitting. So, if the algorithm returns a YES, the splitting identified is a good splitting. If the algorithm returns a NO, then we know that there is no good splitting. If in addition the set X in the splitting is the entire set O, we know we have a wavefront ordering. An example is shown in Fig. 13. The matrix I is shown initially, after the first pass and after the second pass.

8. P-wavefront orderings. The weakly wavefront orderings identify a class of provably convergent Jacobi orderings. But it is easy to construct convergent orderings from weakly wavefront orderings by permuting the indices of the pairs [11].

Permutation Equivalent Orderings. Let O be any cyclic ordering and let Π be a permutation matrix. If we permute the row and column indices in O according to Π, we get another ordering O'. (Alternatively, the matrix I' that contains the positions of the pairs in O' is given by $I' = \Pi^T I \Pi$.) We will say that the ordering O' is a permutation of the ordering O, or that O is permutation equivalent to O'.

Performing a cyclic Jacobi method on a matrix A using ordering O' is exactly the same as using ordering O on $\Pi^T A \Pi$. Therefore, if O converges, so does O'. Also, since every ordering weakly equivalent to a convergent ordering converges, we have the following enlarged class of provably convergent orderings.

P-*wavefront orderings*. If O is weakly equivalent to O_1, and O_1 is a permutation of a weakly wavefront ordering O_2, then O is a P-wavefront ordering.

THEOREM 7. *All* P-*wavefront orderings converge.*

Proof. A P-wavefront ordering is obtained from a weakly wavefront ordering through convergence preserving transformations. So convergence follows from Theorem 4. □

An example of permutation equivalence is shown in Fig. 14. The ordering O' is obtained from O by the permutation $(12345) \rightarrow (12543)$.

P-wavefront orderings are important in practice. The Brent–Luk ordering [2] is used in implementing the Jacobi method on a systolic array. Schreiber [14] describes an implementation of a parallel block Jacobi method using this ordering to choose the subproblems. Luk and Park [11] show that the Brent–Luk ordering is a permutation of the weakly wavefront modulus ordering for odd values of n, and is therefore a P-wavefront ordering.

One way to test if an ordering O is a P-wavefront ordering is to run Algorithm Good Splitting on every permutation of every ordering O_1 that is weakly equivalent to O, i.e., an exponential number of orderings consider. However, we can improve this to a polynomial time algorithm by reducing the number of permutations that need to be considered to $O(n^2)$.

LEMMA 8.1. *If* (X, Y) *is a good splitting of an ordering* O *then the set*

$$\{(1,j); 1<j\leqq n\} \in X.$$

Proof. Suppose $(1, k) \in Y$ for some $k > 2$. Now, $I(1, 2) = 1$ and $(1, 2) \in X$ by (13)(3) and (12). But this contradicts (13)(2). Hence $(1, k) \in X$. □

We will associate an edge-weighted graph G with an ordering O as follows. G is the complete graph on n vertices with edge (i, j) having $I(i, j)$ (i.e., index of the pair (i, j) in the ordering O) as weight. Permuting the vertices of the graph G gives a permutation of the corresponding ordering O. We will label the vertices of G so that the resulting ordering can have a good splitting.

Since $I(1, 2) = 1$ is required by (13), we must assign the labels 1 and 2 to the ends of the edge with weight 1. After having done this in one of the two possible ways, we assign labels to the other vertices as follows. By Lemma 8.1, the pairs (edges) $(1, k)$ are all in the set X of any possible good splitting. By (13)(1) they must satisfy the wavefront property. This means that we have only one way to label the remaining vertices, i.e., in the order of the weights of the edges joining them to vertex 1. Thus, only the above two permutations of the ordering O can be candidates for the good splitting property.

Now consider transposing two commuting rotations in ordering O. The corresponding operation in the graph G is to exchange the weights of two disjoint edges, with the

$$\left(\begin{array}{ccccc} 1 & 3 & 5 & 8 & 10 \\ & 7 & 9 & 11 & 13 \\ & & 12 & 14 & 15 \\ & & & 2 & 4 \\ & & & & 6 \end{array}\right) \quad \underset{\longrightarrow}{\Pi : \begin{pmatrix} 12345 \\ 12543 \end{pmatrix}} \quad \left(\begin{array}{ccccc} 1 & 8 & 5 & 3 & 10 \\ & 11 & 9 & 7 & 13 \\ & & 2 & 14 & 6 \\ & & & 12 & 4 \\ & & & & 15 \end{array}\right)$$

FIG. 14. *Permutation of an ordering.*

weights being consecutive integers. Note that this will not affect the relative order of the edges incident to vertex 1 in the procedure described above. So the permutations produced will be the same. Further, the good-splitting property is preserved under transpositions of commuting rotations. We have shown that to verify if any permutation of an ordering equivalent to O has a good splitting, we need to check only two permutations of O for the good splitting property.

However, if we consider an ordering shift equivalent to O, the permutations generated in our procedure will change. So we need to check every shift of the ordering O as well. But this is only $N = n(n - 1)/2$ orderings. Therefore, to check if an ordering O is a P-wavefront ordering, we only need to check $n(n - 1)$ orderings for the good splitting property using Algorithm Good Splitting. Since Algorithm Good Splitting takes at most $O(n^2)$ time, we have shown that membership in the provably convergent class of P-wavefront orderings can be checked in $O(n^4)$ time.

9. Other convergent orderings. The question arises whether the large class of orderings introduced here completely covers all known convergent cyclic Jacobi orderings. Interestingly, the answer is no. Nazareth [12] proves the convergence of a class of orderings. Some of the orderings presented here, such as the block wavefront offerings, do not fall in Nazareth's class. The converse is also true. In Fig. 15 O is an ordering in Nazareth's class that is not a P-wavefront ordering. So this is a provably convergent ordering that is not a permutation of a weakly wavefront ordering.

Finally, we mention again that there are orderings for which no proof of convergence exists, and there is evidence to suggest that a counterexample to convergence may exist [6]. The Brent–Luk ordering for even values of n is the best known example (refer to Fig. 16). It is also easily seen that a whole class of orderings can be obtained that are not provably convergent by considering orderings weakly equivalent to the Brent–Luk ordering (even n).

$$\begin{pmatrix} & 1 & 2 & 4 & 5 \\ & & 3 & 8 & 9 \\ & & & 6 & 7 \\ & & & & 10 \\ & & & & \end{pmatrix}$$

FIG. 15. *An ordering in Nazareth's class.*

$$\begin{pmatrix} & 1 & 5 & 4 & 2 \\ & & 3 & 2 & 5 \\ & & & 1 & 4 \\ & & & & 3 \\ & & & & \end{pmatrix} \qquad \begin{pmatrix} & 1 & 5 & 4 & 2 & 3 \\ & & 3 & 2 & 5 & 4 \\ & & & 1 & 4 & 2 \\ & & & & 3 & 5 \\ & & & & & 1 \\ & & & & & \end{pmatrix}$$

$n = 5$, Provably convergent $n = 6$, No convergence proof

FIG. 16. *Brent-Luk ordering.*

10. Summary and conclusions. In this paper we have introduced a class of P-wavefront orderings. The cyclic Jacobi methods for the symmetric eigenvalue and SVD problems using orderings from this class converge without the need of thresholds. This class is characterized by properties that are easy to test (polynomial time). We have also shown that there are some parallel block Jacobi methods that fall in this class. However, there are provably convergent orderings that are not P-wavefront orderings.

Can block methods that solve subproblems in other ways be proved convergent? One important possibility is a parallel ordering of Jacobi rotations within the subproblem. Alternatively, some method other than Jacobi such as QR, or a totally new method could be used. In [1] and [15], parallel block Jacobi methods are described that involve completely diagonalizing the subproblem. However, no convergence proofs are available for these methods. The methods described here perform only one Jacobi sweep on some elements of the subproblem, which is cheaper than complete diagonalization. How the method used to solve the subproblems affects the overall rate of convergence, in terms of the number of block sweeps required to diagonalize the whole matrix, is not well understood.

Regarding the SVD problem, it has recently been conjectured that preserving triangular form of the matrix A during the Jacobi iterations can improve the rate of convergence in certain situations [3], [8]. It has also been shown that the well-known rows ordering preserves triangular form [8]. Since wavefront orderings are equivalent to the cyclic by rows ordering, it follows that all wavefront orderings, including some parallel orderings (for example, the parallel antidiagonals ordering), also preserve triangular form and have the same good convergence properties.

A variant of the Jacobi method for the unsymmetric eigenvalue problem has been proposed by Eberlein [4]. Hari [7] has proved convergence for a similar algorithm that uses a cyclic by rows ordering. It is easy to verify that the results proved here can be used to show convergence of the Eberlein algorithm using P-wavefront orderings.

REFERENCES

[1] C. Bischof, *Computing the singular value decomposition on a distributed system of vector processors*, Tech. Report 87-869, Department of Computer Science, Cornell University, Ithaca, NY, 1987.

[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] J. P. Charlier and P. Van Dooren, *On Kogbetliantz's* SVD *algorithm in the presence of clusters*, Linear Algebra Appl., 95 (1987), pp. 135–160.

[4] P. J. Eberlein, *A Jacobi method for the automatic computation of eigenvalues and eigenvectors of an arbitrary matrix*, J. Soc. Indust. Appl. Math., 10 (1962), pp. 74–88.

[5] 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.

[6] E. R. Hansen, *On cyclic Jacobi methods*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 448–459.

[7] V. Hari, *On the global convergence of the Eberlein method for real matrices*, Numer. Math., 39 (1982), pp. 361–369.

[8] V. Hari and K. Veselic, *On Jacobi methods for singular value decompositions*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 741–754.

[9] E. Kogbetliantz, *Diagonalization of general complex matrices as a new method for solution of linear systems*, in International Congress on Mathematics, Amsterdam, the Netherlands, 1954, pp. 356–357.

[10] F. T. Luk and H. T. Park, *On parallel Jacobi orderings*, Tech. Report EE-CEG-86-5, School of Electrical Engineering, Cornell University, Ithaca, NY, 1986.

[11] ———, *A proof of convergence for two parallel Jacobi* SVD *algorithms*, Tech. Report EE-CEG-86-12, School of Electrical Engineering, Cornell University, Ithaca, NY, 1986.

[12] L. NAZARETH, *On the convergence of the cyclic Jacobi method*, Linear Algebra Appl., 12 (1975), pp. 151–164.

[13] D. A. POPE AND C. TOMKINS, *Maximizing functions of rotations—experiments concerning speed of diagonalization of symmetric matrices using Jacobi's method*, J. Assoc. Comput. Mach., 4 (1957), pp. 459–466.

[14] R. SCHREIBER, *Solving eigenvalue and singular value problems on an undersized systolic array*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 441–451.

[15] C. VAN LOAN, *The block Jacobi method for computing the singular-value decomposition*, Technical Report 85-680, Department of Computer Science, Cornell University, Ithaca, NY, 1985.

[16] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 3, pp. 347-382, July 1989

006

ON MAXIMIZING THE MINIMUM EIGENVALUE OF A LINEAR COMBINATION OF SYMMETRIC MATRICES*

J. C. ALLWRIGHT†

Abstract. The problem considered is that of maximizing, with respect to the weights, the minimum eigenvalue of a weighted sum of symmetric matrices when the Euclidean norm of the vector of weights is constrained to be unity. A procedure is given for determining the sign of the maximum of the minimum eigenvalue and for approximating the optimal weights arbitrarily accurately when that sign is positive or zero. Linear algebra, a conical hull representation of the set of $n \times n$ symmetric positive semidefinite matrices and convex programming are employed.

Key words. positive semidefinite matrices, eigenvalue optimization

AMS(MOS) subject classifications. 52A40, 49D45, 15A18

1. Introduction.

Consider given $n \times n$ symmetric real matrices $H_1, \cdots, H_l$, not all zero, and the following problem:

(1.1) Find $\hat{\lambda} = \max_{\mu \in \mathbb{B}^l} \lambda(\mu)$ and an associated maximizer $\hat{\mu}$

where

$$\lambda(\mu) = \lambda_{\min}\left(\sum_{i=1}^{l} \mu_i H_i\right). \tag{1.2}$$

Here $\lambda_{\min}(A)$ denotes the minimal (i.e., most negative) eigenvalue of symmetric A and $\mathbb{B}^l = \{x \in \mathbb{R}^l : \|x\| = 1\}$. Throughout, $\| \quad \|$ denotes $\| \quad \|_2$ and $n \geqq 2$.

The initial motivation for this problem came from optimal output feedback [1]. In that context, the operating cost for a system with an initial condition $x_0 \in \mathbb{R}^n$ and a parameter vector $f \in \mathbb{R}^l$ can be written as $x_0' K(f) x_0$, where $K(f)' = K(f) \in \mathbb{R}^{n \times n}$. Given a parameter vector f, it would be desirable to try to find a search direction $\mu \in \mathbb{R}^l$ such that the cost is reduced for all initial conditions $x_0 \in \mathbb{R}^n$ by taking a sufficiently small positive step w along μ from f, i.e., such that $x_0' K(f) x_0 - x_0' K(f + w\mu) x_0 \geqq 0$ for all $x_0 \in \mathbb{R}^n$, i.e., such that $K(f) - K(f + w\mu) \geqq 0$. Whether such a μ exists depends on the solution of (1.1) when $H_i = -\partial K(f)/\partial f_i$, because then, in terms of a first-order expansion of K about f, a sufficient condition for there to be a μ such that $K(f) - K(f + w\mu) \geqq 0$ is that there exists a μ such that $\sum_{i=1}^{l} \mu_i H_i > 0$, i.e., is that there exists a $\mu \in \mathbb{B}^l$ such that $\lambda(\mu) > 0$, for λ of (1.2), i.e., is that $\hat{\lambda} > 0$ for $\hat{\lambda}$ of (1.1). If $\hat{\lambda} > 0$ then a suitable search direction μ is a $\hat{\mu}$ for (1.1). Similarly, there is no μ such that $K(f) - K(f + w\mu) \geqq 0$ for some small positive w if $\hat{\lambda} < 0$. Hence it is not necessary to compute a $\hat{\mu}$ if $\hat{\lambda} < 0$. The situation when $\hat{\lambda} = 0$ depends on second-order effects. This has outlined the significance of problem (1.1) in this control context.

It seems that solution of (1.1) is, in general, difficult. However it is clear that it is sufficient, in the control context mentioned above, to be able to solve the following subproblem:

(1.3) Find sign $(\hat{\lambda})$ and, if $\hat{\lambda} \geqq 0$, find an associated maximizer $\hat{\mu}$ where
sign $(\hat{\lambda})$ is 1 if $\hat{\lambda} > 0$, is 0 if $\hat{\lambda} = 0$ and is -1 if $\hat{\lambda} < 0$.

* Received by the editors November 9, 1987; accepted for publication (in revised form) November 23, 1988.

† Department of Electrical Engineering, Imperial College of Science and Technology, London SW7 2BT, United Kingdom (JANET:m.davis@uk.ac.ic.cc.vaxa).

This paper is mostly about the solution of problem (1.3).

The maximum in (1.1) exists because the constraint set $\mathbb{B}^l$ is compact and because the function λ of (1.2) is continuous. In fact λ is also concave and is not necessarily differentiable everywhere. The nondifferentiability could be approached using Clarke's theory of nonsmooth optimization [2], Fletcher's results on positive semidefinite matrices [3], and ideas from Overton [4] and Overton and Womersley [5]. However even if λ were differentiable everywhere, the global optimality required for problems (1.1) and (1.3) could not easily be achieved using standard techniques owing to the nonconvexity of $\mathbb{B}^l$. Overton [4] has given a second-order method for solving (1.1) without the constraint $\mu \in \mathbb{B}^l$ but that does not provide information that is decisive in this context because $\sup \{ \lambda(\mu) : \mu \in \mathbb{R}^l \}$ is always nonnegative, since $\lambda(0) = 0$. An approach is used here that is designed specifically for the constraint $\mu \in \mathbb{B}^l$ and that turns out to avoid all issues associated with nondifferentiability.

Problem (1.3) is a fundamental problem of linear algebra and is of interest in its own right. A related problem that has been considered before is that of deciding whether there is a linear combination of given symmetric matrices that is positive definite [6], but a general solution is not given in [6]. In the above context, the related problem reduces to that of deciding whether sign $(\hat{\lambda}) = 1$. The results of [1] or [4] could be applied to solve that problem but cannot be used to decide whether $\hat{\lambda} = 0$ or $\hat{\lambda} < 0$.

Actually, in [1] the problem considered was that of minimizing the maximal eigenvalue of $\sum_{i=1}^{l} \mu_i H_i$ with respect to $\mu \in \mathbb{B}^l$ instead of maximizing the minimal eigenvalue (as in (1.1)). It is more convenient to consider here the formulation of (1.1)—the only consequence is that some results from [1] need obvious sign changes that are included here in quotations from [1].

In Appendix C of [1] a method was given for approximating $\hat{\mu}$ as accurately as desired if it turns out that $\hat{\lambda} > 0$, but in general that method does not determine sign $(\hat{\lambda})$ and certainly does not yield a $\hat{\mu}$ if $\hat{\lambda} = 0$. Hence the part of problem (1.3) that does not appear to have been solved before is:

(1.4) Find sign $(\hat{\lambda})$ and if sign $(\hat{\lambda}) = 0$, then find a corresponding $\hat{\mu}$.

The theory developed in § 5 enables problem (1.4) to be solved without using any iterations in some cases, but in general it seems that an iterative algorithm is needed. A consequence is that it is not generally possible to determine whether sign $(\hat{\lambda})$ is exactly zero using a finite amount of computational effort. In practice it would usually be adequate to solve the following approximation problem:

(1.5) For a prespecified $\delta > 0$, iterate until it is possible to decide whether $\hat{\lambda} > 0$ or to decide whether $\hat{\lambda} < 0$ or to find a $\bar{\mu} \in \mathbb{B}^l$ that approximates a $\hat{\mu}$ in that $\lambda(\bar{\mu}) \in [\hat{\lambda} - \delta, \hat{\lambda}]$.

The main purpose of this paper is to study problems (1.3)–(1.4) and to develop a procedure (Procedure 5.1) that solves the following problem, using only algebra and convex programming:

(1.6) Solve problem (1.4) when that can be done using finite work, else solve approximation problem (1.5).

In § 2, relations are obtained between $\hat{\lambda}$, $\hat{\mu}$, the origin, and a convex set Π derived from the H_i. Those relations extend the results of [1] and contribute to the understanding of problem (1.1) and of problems (1.3)–(1.4). The study in § 3 of a conical-hull description of the set $S^n_{\geqq}$ of real, symmetric, positive semidefinite, $n \times n$ matrices yields results that play a vital role in § 5. In § 4, a new convex programming algorithm (Algorithm 4.1) is presented for approximating, with prespecified precision, the minimal value of α from $\mathbb{R}$ such that, for a specified vector h, αh belongs to a given convex compact set. That algorithm is needed in § 5, where a procedure is developed for solving problem (1.6). Section 6 contains numerical results for some examples.

The proofs for results in §§ 2–5 that are nontrivial or are not presented in those sections are given in Appendices A–D, respectively.

The set of nonnegative real numbers is denoted by $\mathbb{R}_{\geqq}$ and the set of strictly positive real numbers by $\mathbb{R}_{>}$. The sets $\mathbb{R}_{\leqq}$ and $\mathbb{R}_{<}$ are defined similarly.

The line $\{\alpha x + (1-\alpha)y \colon \alpha \in [0, 1]\}$ between points x and y in $\mathbb{R}^n$ will sometimes be written as $[x, y]$.

For a nonzero vector x from $\mathbb{R}^n$, the normalized vector $x\|x\|^{-1}$ is denoted by $\langle\!\langle x \rangle\!\rangle$.

The interior of a set S is written as int (S), the boundary as ∂S, the convex hull as conv (S), and the conical hull (i.e., $\{\alpha s \colon \alpha \in \mathbb{R}_{\geqq}, s \in S\}$) as cone (S). The set MS denotes $\{Ms \colon s \in S\}$, where M is a matrix of order compatible with S, and $x + S$ denotes the set $\{x + s \colon s \in S\}$. The set of points in S that are closest, in the Euclidean sense, to a point x is denoted minpoints $[x, S]$ when there could be several closest points and minpoint $[x, S]$ when the closest point is definitely unique. The corresponding minimal distance is often called mindist $[x, S]$. When S has the form MF for a set F, the set of points in F that minimize $\|x - Mf\|$ with respect to f from F is written minpoints$_F$ $[x, MF]$, and $\arg_F \min_{z \in MF} v(z)$ denotes the set of points in F that minimize $v(Mf)$ with respect to f from F. The above notation regarding closest points will only be used when closest points exist.

The hyperplane H with normal η that supports a set $S \subset \mathbb{R}^n$ at a point y refers to the set $H = \{x \in \mathbb{R}^n \colon \eta' x = \eta' y\}$ where $y \in \arg\max \{\eta' z \colon z \in S\}$. That point y is sometimes called a contact point for the hyperplane and the set.

The range of a matrix M is written as $R[M]$, its null space as $N[M]$, its Frobenius norm as $\|M\|_{\mathcal{F}}$, and $M^{\dagger}$ denotes the pseudoinverse of M.

The orthogonal complement of, for example, $R[M]$ is written as ${}^{\perp}R[M]$.

2. Some relationships between eigenvalue maximization and a convex set specified by the given symmetric matrices. For the matrices H_i of problem (1.1), consider the function $p \colon \mathbb{R}^n \to \mathbb{R}^l$ defined by

$$p(x) = [x'H_1x \; x'H_2x \cdots x'H_lx]' \tag{2.1}$$

and the associated convex compact set

$$\Pi = \text{conv}\,[p(\mathbb{B}^n)] \subset \mathbb{R}^l. \tag{2.2}$$

Now

$$\begin{aligned}\lambda(\mu) = \lambda_{\min}\left[\sum_{i=1}^{l} \mu_i H_i\right] &= \min_{x \in \mathbb{B}^n} x'\left[\sum_{i=1}^{l} \mu_i H_i\right]x \\ = \min_{x \in \mathbb{B}^n} \mu' p(x) \quad &= \min_{\pi \in p(\mathbb{B}^n)} \mu'\pi = \min_{\pi \in \Pi} \mu'\pi\end{aligned}$$

so there is the following connection between Π and $\lambda(\mu)$.

LEMMA 2.1. $\lambda(\mu) = \min\{\mu'\pi \colon \pi \in \Pi\}$. □

That connection enables the following relationships between $\hat{\mu}$, $\hat{\lambda}$, the origin, and Π to be established.

THEOREM 2.1. (i) $\hat{\lambda} < 0$ *if and only if* $0 \in \operatorname{int}(\Pi)$. *If* $0 \in \operatorname{int}(\Pi)$, *then a* $\hat{\mu}$ *that maximizes* λ *on* $\mathbb{B}^l$ *is* $-\langle\!\langle \tilde{\pi} \rangle\!\rangle$, *for any* $\tilde{\pi} \in$ minpoints $[0, \partial\Pi]$, *and* $\hat{\lambda} = -\|\tilde{\pi}\|$.

(ii) $\hat{\lambda} = 0$ *if and only if* $0 \in \partial\Pi$. *If* $0 \in \partial\Pi$, *then a* $\hat{\mu}$ *that maximizes* λ *on* $\mathbb{B}^l$ *is* $\hat{\mu} = -\langle\!\langle \mu \rangle\!\rangle$ *for any* μ *that is the normal to a hyperplane that supports* Π *at* 0.

(iii) $\hat{\lambda} > 0$ *if and only if* $0 \notin \Pi$. *If* $0 \notin \Pi$, *then there is a unique* $\hat{\mu}$ *that maximizes* λ *on* $\mathbb{B}^l$, *given by* $\hat{\mu} = \langle\!\langle \hat{\pi} \rangle\!\rangle$, *where* $\hat{\pi} =$ minpoint $[0, \Pi]$, *and* $\hat{\lambda} = \|\hat{\pi}\|$. □

Theorem 2.1 reveals the close connection between sign $(\hat{\lambda})$ and whether $0 \in \Pi$, $0 \in \partial\Pi$, or $0 \in \operatorname{int}(\Pi)$. Part (iii) was stated and proved in Appendix C of [1] but will be proved here in Appendix A, together with the rest of Theorem 2.1, for completeness.

Problem (1.3) is concerned with the determination of whether sign $(\hat{\lambda}) \geqq 0$ and with the computation of a maximizing μ if sign $(\hat{\lambda}) \geqq 0$. By Theorem 2.1, sign $(\hat{\lambda}) \geqq 0$ if and only if $0 \notin \operatorname{int}(\Pi)$. Owing to the way Π is defined it does not seem to be a simple matter to determine whether $0 \notin \operatorname{int}(\Pi)$. At least conceptually, it is possible to compute $\hat{\pi} =$ minpoint $[0, \Pi]$, which, since $\|\hat{\pi}\| = 0$ if and only if $0 \in \Pi$, reveals some useful information about the location of the origin with respect to Π. If $\hat{\pi} \neq 0$ then it is clear that $0 \notin \Pi$ and, by Theorem 2.1(iii), $\hat{\mu} = \langle\!\langle \hat{\pi} \rangle\!\rangle$. However, if $\hat{\pi} = 0$ then that could occur either because $0 \in \operatorname{int}(\Pi)$ or because $0 \in \partial\Pi$ and there is no obvious way to decide which is true. Furthermore, even if $0 \in \partial\Pi$ so that $\hat{\lambda} = 0$, a $\hat{\mu}$ is required to complete the solution of problem (1.3). Theorem 2.1(ii) reveals that then $\hat{\mu}$ can be obtained from the normal to any hyperplane that supports Π at 0, but it is not clear how such a normal can be computed.

In practice, usually it is not possible to compute $\hat{\pi}$ exactly. However, it is shown in Appendix C of [1] that $\hat{\pi}$ can be approximated arbitrarily accurately using an implementable iterative algorithm for minimizing $\|\pi\|$ on Π. The algorithm generates a sequence $\pi_j \in \Pi$ convergent to $\hat{\pi}$ and operates by computing supporting hyperplanes for Π that have particular normals. That turns out to be quite easy to do even though the definition of Π in (2.2) does not seem to render computation with Π attractive. Unfortunately, numerical evaluation of a sequence $\pi_j \rightarrow \hat{\pi}$ generally does not enable us to decide, using finite work, whether $\hat{\pi}$ is nonzero or zero. Consequently, in general it is not practicable to decide whether $0 \notin \Pi$ so that it is not possible to evaluate sign $(\hat{\lambda})$.

Therefore, in connection with problem (1.3), there is a requirement for a general method for determining sign $(\hat{\lambda})$. If sign $(\hat{\lambda}) > 0$, then the algorithm in Appendix C of [1] can be used to find $\hat{\pi}$, and hence (by Theorem 2.1(iii)) $\hat{\mu}$, to any prespecified accuracy. If sign $(\hat{\lambda}) = 0$, then some method for finding a corresponding $\hat{\mu}$ is needed. This has explained in more detail the motivation for problem (1.4) given in § 1. The determination of sign $(\hat{\lambda})$ and of a global maximizer $\hat{\mu}$ if sign $(\hat{\lambda}) = 0$ are considered in § 5.

It is fortunate that the motivation for this work did not require the determination of a global minimizer $\hat{\mu}$ when $0 \in \operatorname{int}(\Pi)$ because Theorem 2.1(i) suggests that the determination of $\hat{\mu}$ and $\hat{\lambda}$ are difficult problems in that case since then minimization of $\|\pi\|$ on $\partial\Pi$ is a nonconvex optimization problem with, possibly, many local minima that are not global minima. It can actually be shown that if π^* is a local minimizer of $\|\pi\|$ on $\partial\Pi$ then $-\langle\!\langle \pi^* \rangle\!\rangle$ is a local maximizer of λ on $\mathbb{B}^l$. The existence of at least one global minimizer $\tilde{\pi}$ follows from the fact that $\partial\Pi$ is compact because it is the boundary of a compact set.

Theorem 2.1 exploits the convexity of Π, which is defined to be the convex hull of the set $p(\mathbb{B}^n)$. Brickman [7] has shown that $p(\mathbb{B}^n)$ is convex for the case with $l = 2$ and $n > 2$ and is not necessarily convex for $n = 2$ or for $l \geqq 3$. Taussky [6, § 1] has stated that in the case $l = 2$ and $n > 2$, if $[x'H_1x = x'H_2x = 0] \Rightarrow [x = 0]$, then there are μ_1 and μ_2 such that $\mu_1 H_1 + \mu_2 H_2 > 0$ but the relationships between $\hat{\lambda}$ and Π of Theorem 2.1 seem to be new.

3. A conical hull characterization of the set of symmetric positive semidefinite matrices. Let

$$S^n = \{A \in \mathbb{R}^{n \times n} \colon A' = A\},$$

$$S^n_{\geqq} = \{A \in \mathbb{R}^{n \times n} \colon A' = A \geqq 0\}.$$

The characterization given here of the convex cone $S^n_{\geqq}$ of $n \times n$ symmetric positive semidefinite matrices is one of those mentioned in [8]. It is based on the fact that any symmetric positive semidefinite matrix $A \in \mathbb{R}^{n \times n}$ may be represented as $A = \alpha B^2$ for some $\alpha \in \mathbb{R}_{\geqq}$ and some symmetric $B \in \mathbb{R}^{n \times n}$ with $\|B\|_{\mathcal{F}} = 1$. So

$$S^n_{\geqq} = \text{cone}\,(\{B^2 \colon B \in U\}) \tag{3.1}$$

where

$$U = \{B \in S^n \colon \|B\|_{\mathcal{F}} = 1\}. \tag{3.2}$$

It will often turn out to be more convenient to work with vectors characterizing positive semidefinite matrices than with the matrices themselves. The machinery for doing that is introduced next.

For $C \in \mathbb{R}^{m \times n}$, let vec $[C]$ be the following vector containing all the entries of C:

$$\text{vec}\,[C] = [c'_{1*} c'_{2*} \cdots c'_{m*}]' \in \mathbb{R}^{mn}$$

where c_{i*} denotes row i of C.

Consider the linear subspace vec $[S^n] = \{\text{vec}\,[A] : A \in S^n\} \subset \mathbb{R}^{n^2}$ of all vectors vec $[A]$ associated with symmetric A. It has dimension $r = n(n+1)/2$. Suppose $w_1, w_2, \cdots, w_r$ is an orthonormal basis-set for vec $[S^n]$. For example, suitable w_i for $n = 2$ might be $w_1 = [1\ 0\ 0\ 0]'$, $w_2 = [0\ 2^{-1/2}\ 2^{-1/2}\ 0]'$, $w^3 = [0\ 0\ 0\ 1]'$. Consequently,

$$W = [w_1 w_2 \cdots w_r] \in \mathbb{R}^{n^2 \times r} \tag{3.3}$$

is a basis-matrix for vec $[S^n]$ and

$$W'W = I_r, \quad R[W'] = \mathbb{R}^r, \quad R[W] = \text{vec}\,[S^n], \tag{3.4}$$

$$WW' \text{ projects } \mathbb{R}^{n^2} \text{ orthogonally onto vec}\,[S^n]. \tag{3.5}$$

For symmetric A, let $\overline{\text{vec}}\,[A]$ denote the vector of coordinates of vec $[A]$ with respect to the basis-set $w_1, w_2, \cdots, w_r$, i.e., with respect to the columns of W. Then, in view of (3.3)–(3.4)

$$\text{vec}\,[A] = W\, \overline{\text{vec}}\,[A] \in \mathbb{R}^{n^2}, \qquad \overline{\text{vec}}\,[A] = W' \,\text{vec}\,[A] \in \mathbb{R}^r. \tag{3.6}$$

The function $\overline{\text{vec}}^{-1} \colon \mathbb{R}^r \to S^n$ will be useful later.

Many of the calculations in this paper will be carried out using the vector $\overline{\text{vec}}\,[A]$ instead of A itself. Clearly,

$$\{A \in S^n_{\geqq}\} \Leftrightarrow \{\overline{\text{vec}}\,[A] \in \overline{\text{vec}}\,[S^n_{\geqq}]\}$$

where

$$\overline{\text{vec}}\,[S^n_{\geqq}] = \{\overline{\text{vec}}\,[A] \colon A \in S^n_{\geqq}\}.$$

So, in view of (3.1) and (3.6)

$$\overline{\text{vec}}\,[S^n_{\geqq}] = \text{cone}\,(\{\overline{\text{vec}}\,[B^2] \colon B \in U\}) = \text{cone}\,(\{W' \,\text{vec}\,[B^2] \colon B \in U\}).$$

Therefore it is not surprising that $\overline{\text{vec}}\,[S^n_{\geqq}]$ may be represented as the conical hull of the convex set $\Gamma = W' \operatorname{conv}(\{\text{vec}\,[B^2]\colon B \in U\})$. That characterization is summarized below in Theorem 3.1, where some properties of Γ are given that will be useful later.

THEOREM 3.1. *Let*

$$\Gamma = W'\Omega \subset \mathbb{R}^r \textit{ for } \Omega = \operatorname{conv}(\{\text{vec}\,[B^2]\colon B \in U\}). \tag{3.7}$$

Then

(i) $\overline{\text{vec}}\,[S^n_{\geqq}] = \text{cone}\,[\Gamma]$;

(ii) Γ *is a convex compact subset of* $\mathbb{R}^r$;

(iii) *For* $g \in \mathbb{R}^q$ *and* $M \in \mathbb{R}^{q \times r}$: $\min_{x \in M\Gamma} g'x = \lambda_{\min}[Z]$; $W'\,\text{vec}\,[vv'] \in \arg_\Gamma \min_{x \in M\Gamma} g'x$ *and* $MW'\,\text{vec}\,[vv'] \in \arg\min_{x \in M\Gamma} g'x$. *Here*

$$Z = [\text{vec}^{-1}[\bar{g}] + \text{vec}^{-1}[\bar{g}]']/2 \in \mathbb{R}^{n \times n} \textit{ for } \bar{g} = WM'g \tag{3.8}$$

and v is a normalized eigenvector of Z corresponding to the minimal eigenvalue, $\lambda_{\min}[Z]$, *of* Z. □

Furthermore, it will turn out to be very useful that there are sets Γ_1 and Γ_2 that have a simple structure and are super- and subsets of Γ, respectively.

THEOREM 3.2. *For* Γ *of* (3.7)

$$\Gamma_2 \subset \Gamma \subset \Gamma_1, \tag{3.9}$$

$$n^{-1/2} \leqq \|\gamma\| \leqq 1, \quad \forall \gamma \in \Gamma. \tag{3.10}$$

Here

$$\Gamma_1 = n^{-1}\iota + M\Xi_1, \qquad \Gamma_2 = n^{-1}\iota + M\Xi_2 \tag{3.11}$$

where

(3.12) $\iota = \overline{\text{vec}}\,[I_n] = W'\,\text{vec}\,[I_n]$ *and I_n is the $n \times n$ identity matrix,*

(3.13) $M \in \mathbb{R}^{r \times (r-1)}$ *and has orthonormal columns that span the orthogonal complement, with respect to* $\mathbb{R}^r$, *of the linear subspace spanned by the vector* ι *of* (3.12),

$$\Xi_1 = \{x \in \mathbb{R}^{r-1}\colon \|x\| \leqq (1 - n^{-1})^{1/2}\}, \tag{3.14}$$

$$\Xi_2 = \{x \in \mathbb{R}^{r-1}\colon \|x\| \leqq n^{-1}\}. \tag{3.15}$$

4. Finding the first point at which a line in a given direction intersects a given convex compact set.

The basic problem here is:

(4.1) Determine both $\hat{\alpha} = \min\{\alpha \in \mathbb{R}\colon \alpha h \in MF\}$ and an $\hat{f} \in F$ such that $\hat{\alpha}h = M\hat{f}$, assuming $\hat{\alpha}$ exists

where

(4.2) F is a convex compact subset of $\mathbb{R}^r$; $\zeta\phi \in F$ for some given $\zeta \in \mathbb{R}$ and some given $\phi \in \mathbb{R}^r$; $M \in \mathbb{R}^{q \times r}$; $h = M\phi \neq 0$ and q, r, F, M are all given.

Hence the line $L = \{\alpha h\colon \alpha \in \mathbb{R}\}$ intersects MF (since $\zeta h = M[\zeta\phi] \in MF$) and the problem is to find $\hat{\alpha}$, the most negative α such that the point αh in the line L belongs to the set MF, and to find a point $\hat{f}$ in F such that $M\hat{f} = \hat{\alpha}h$. Clearly if $\hat{\alpha}$ exists then $\hat{\alpha} \leqq \zeta$, since $\zeta h \in MF$.

Algorithm 4.1, stated later, solves the following approximation problem associated with problem (4.1):

(4.3) For any given $\varepsilon_{11}, \varepsilon_{12}, \varepsilon_2 > 0$, determine the following:
(i) An approximation $\bar{\alpha} \in \mathbb{R}$ to $\hat{\alpha}$ such that $|\hat{\alpha} - \bar{\alpha}| < \varepsilon_{11}|\hat{\alpha}| + \varepsilon_{12}$;
(ii) An $\bar{f} \in F$ that satisfies both $\|M\bar{f} - \bar{\alpha}h\| \leqq (1 + \varepsilon_2)$ mindist $[\bar{\alpha}h, MF]$ and $\|M\bar{f} - \bar{\alpha}h\| \leqq \|\zeta h - \bar{\alpha}h\|$;

when it is assumed that

(4.4) $\hat{\alpha}$ exists and there is a hyperplane with normalized normal $\eta \in \mathbb{R}^q$ that satisfies $\eta' h < 0$ and supports MF at $\hat{\alpha}h$;

(4.5) $\arg_F \max \{g'y \colon y \in MF\}$ can be computed exactly, for all $g \in \mathbb{R}^q$;

(4.6) MF contains a subset $\Xi = \zeta h + \{\theta \in \mathbb{R}^q \colon \|\theta\| \leqq k\}$ for some known $k \in \mathbb{R}_>$.

Assumption (4.4) actually guarantees that Algorithm 4.1, with its stopping condition omitted, generates a sequence α_i convergent to $\hat{\alpha}$ of (4.1). Assumption (4.5) ensures that key calculations in the algorithm can be carried out. Assumption (4.6) is required only to enable the construction of a stopping condition for Algorithm 4.1 that guarantees that the required approximations $\bar{\alpha}$ and $\bar{f}$, of (4.3), are obtained at termination.

Algorithm 4.1 requires the calculation of an approximation x_i to $\hat{x}_{\alpha_i} :=$ minpoint $[\alpha_i h, MF]$ for each member of the sequence $\alpha_i \in \mathbb{R}$ computed by the algorithm. For $\varepsilon_2 \in (0, 1)$, the approximation x_i is required to be in the set ε_2-minpoint $[\alpha_i h, MF]$ of approximations to minpoint $[\alpha_i h, MF]$. That set of approximations is specified in Definition 4.1 using the function $\tau_\alpha \colon \mathbb{R}^q \to \mathbb{R}$ defined in the following lemma, which concerns it. In that lemma and in Definition 4.1, given later, an important consequence of the hypothesis $\alpha < \hat{\alpha}$ is that (in view of (4.1)) $\alpha h \notin MF$. A graphical interpretation of $\tau_{\alpha_i}(x_i)$ is given in Fig. 4.1.

LEMMA 4.1. *Suppose $\alpha < \hat{\alpha}$ and $x \in MF$. Consider the hyperplane with normal $\alpha h - x$ that supports MF. Suppose $\hat{y}_\alpha(x) \in$ arg max $\{(\alpha h - x)'y \colon y \in MF\}$. Let $\tau_\alpha(x)$ be the value of a corresponding to the point ah in the line $L = \{ah \colon a \in \mathbb{R}\}$ at which the supporting hyperplane intersects L. Then*

(4.7) *If* $(\alpha h - x)'h < 0$ *then* $\tau_\alpha(x) = \alpha + ((\alpha h - x)'(\hat{y}_\alpha(x) - \alpha h)/(\alpha h - x)'h) \leqq \hat{\alpha}$,

(4.8) $(\alpha h - \hat{x}_\alpha)'h < 0$ *and* $\alpha < \tau_\alpha(\hat{x}_\alpha) = \alpha + (\|\hat{x}_\alpha - \alpha h\|^2/h'(\hat{x}_\alpha - \alpha h)) \leqq \hat{\alpha}$

where $\hat{x}_\alpha =$ minpoint $[\alpha h, MF]$. □

The set ε-minpoint $[\alpha h, MF]$ of approximations to minpoint $[\alpha h, MF]$ is defined next. Points x in it approximate $\hat{x}_\alpha$ both in terms of their distance from MF and in terms of $\tau_\alpha(x)$ being near $\tau_\alpha(\hat{x}_\alpha)$, as follows.

DEFINITION 4.1. Suppose $\alpha < \hat{\alpha}$. Then

$$x \in \varepsilon\text{-minpoint}\,[\alpha h, MF] \quad \text{iff}$$

(i) $x \in MF$;
(ii) $\|x - \alpha h\| \leqq (1 + \varepsilon)$ mindist $[\alpha h, MF]$ and $\|x - \alpha h\| \leqq \|\zeta h - \alpha h\|$;
(iii) $\tau_\alpha(x)$ exists, $\tau_\alpha(x) \leqq \hat{\alpha}$ and $[\tau_\alpha(\hat{x}_\alpha) - \tau_\alpha(x)] \leqq \varepsilon[\tau_\alpha(\hat{x}_\alpha) - \alpha]$. □

An algorithm for finding a point in ε-minpoint $[\alpha h, MF]$, Algorithm 4.2, is given later. The algorithm for approximating $\hat{\alpha}$ by solving problem (4.3) follows. In it, the notation $\varepsilon :\in (0, 1)$ means "choose ε from the interval (0, 1)," etc.

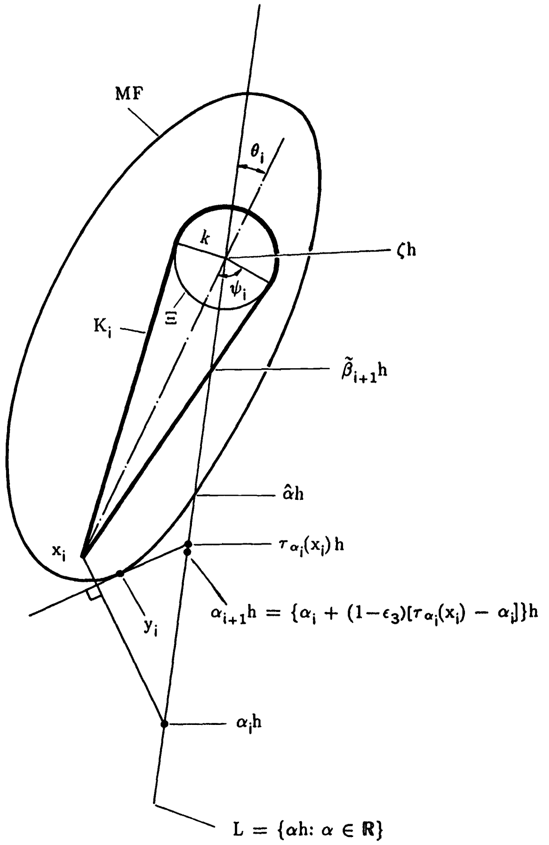


FIG. 4.1

ALGORITHM 4.1.

```
0  Choose tolerance parameters (see Remark 4.1)
(4.9)    $\varepsilon_{11} :\in (0,\infty)$;  $\varepsilon_{12} :\in (0,\infty)$;  $\varepsilon_2 :\in (0,1)$;  $\varepsilon_3 :\in (0,1)$;
I  Initialization
         $i := 0$; (the initial value of the iteration index)
         (determine an initial approximation $\alpha_0$ to $\hat{\alpha}$ with $\alpha_0 < \hat{\alpha}$)
(4.10)   $y_0 :\in \arg\max \{(-h)'y: y \in MF\}$;
(4.11)   $\alpha_0 := y_0'h\|h\|^{-2} - \varepsilon_3$;
         (select an initial upper-bound (associated with stopping condition (4.21)) for $\hat{\alpha}$)
(4.12)   $\beta_0 := \zeta - \frac{k}{\|h\|}$;
II  Determine $\alpha_{i+1}$ from $\alpha_i$
(4.13)   $x_i :\in \varepsilon_2$-minpoint $[\alpha_i h, MF]$;
(4.14)   $y_{i+1} :\in \arg\max \{(\alpha_i h - x_i)'y: y \in MF\}$;
(4.15)   $\alpha_{i+1} := \alpha_i + (1-\varepsilon_3)\frac{(y_{i+1} - \alpha_i h)'(\alpha_i h - x_i)}{h'(\alpha_i h - x_i)}$;
III  Decide whether to stop iterating
         (compute an upper-bound $\tilde{\beta}_{i+1}$ for $\hat{\alpha}$)
         if  $\{\|x_i - \zeta h\| \leqq k\}$
```

```
               then
(4.16)             $\tilde{\beta}_{i+1} := \zeta - \frac{k}{\|h\|}$
               else
(4.17)             $\psi_i := \cos^{-1}\left(\frac{k}{\|x_i - \zeta h\|}\right)$ (for $k$ of (4.6));     $\theta_i := \cos^{-1}\left(\frac{(\zeta h - x_j)'h)}{\|\zeta h - x_i\| \; \|h\|}\right)$;
                   if $\cos(\theta_i) \leqq \cos(\psi_i)$
                     then
(4.18)                 $\tilde{\beta}_{i+1} := \zeta - \frac{k}{\|h\|}$
                   else
(4.19)                 $\tilde{\beta}_{i+1} := \zeta - \frac{k}{\cos(\psi_i - \theta_i)\|h\|}$
                   endif
               endif;
             (compute the least upper-bound for $\hat{\alpha}$ that has been found so far)
(4.20)       $\beta_{i+1} := \min\{\beta_i, \tilde{\beta}_{i+1}\}$;
             (decide whether or not to stop iterating)
(4.21)       if $|\beta_{i+1} - \alpha_{i+1}| < \varepsilon_{11}\max\{|\alpha_{i+1}|, |\beta_{i+1}|\} + \varepsilon_{12}$
               then $\bar{\alpha} := \alpha_{i+1}; \bar{f} :\in \varepsilon_2$-minpoints$_F$ $[\bar{\alpha}h, MF]$; stop;
             $i := i + 1$; go to II.     □
```

Remark 4.1 (*Tolerance parameters in Algorithm* 4.1). In (4.9), ε_{11} and ε_{12} are the relative and absolute tolerances associated with the determination of $\hat{\alpha}$ (recall problem (4.3)), and ε_2 is the tolerance allowed in the approximation x_i to $\hat{x}_{\alpha_i} =$ minpoint $[\alpha_i h, MF]$ (recall Definition 4.1).

The purpose of ε_3 is to ensure that, for α_{i+1} of (4.15), $\alpha_{i+1}h$ never quite belongs to MF (i.e., to ensure that $\alpha_{i+1} < \hat{\alpha}$) because $\alpha_{i+1}h \notin MF$ is a precondition for Algorithm 4.2 for the determination of an $x_{i+1} \in \varepsilon_2$-minpoint $[\alpha_{i+1}h, MF]$, that is needed in (4.13) during the next iteration. The value of ε_3 used for the numerical examples reported in § 6 was 0.05. Algorithm 4.2 is stated later. □

Some important properties of Algorithm 4.1 are given next.

THEOREM 4.1. *For Algorithm* 4.1 *with stopping condition* (4.21) *omitted, so that it iterates indefinitely,*

$$\alpha_i \uparrow \hat{\alpha} \text{ with } \alpha_i < \alpha_{i+1} < \hat{\alpha} \quad \forall i \geqq 0, \qquad \beta_i \downarrow \hat{\alpha} \text{ with } \hat{\alpha} \leqq \beta_i \quad \forall i \geqq 0, \tag{4.22}$$

$$(\hat{\alpha} - \alpha_i) \leqq [1 + (1 - \varepsilon_2)(1 - \varepsilon_3)\eta' h \|h\|^{-1}]^i (\hat{\alpha} - \alpha_0). \tag{4.23}$$

With stopping condition (4.21) *included, the algorithm stops after a finite number of iterations with*

$$\bar{\alpha} < \hat{\alpha} \quad \text{and} \quad |\hat{\alpha} - \bar{\alpha}| < \varepsilon_{11}|\hat{\alpha}| + \varepsilon_{12}, \tag{4.24}$$

$$\bar{f} :\in \varepsilon_2\text{-minpoints}_F\,[\bar{\alpha}h, MF] \quad (\textit{in that } M\bar{f} \in \varepsilon_2\text{-minpoints}\,[\bar{\alpha}h, MF]). \tag{4.25}$$

□

Remark 4.2 (*Geometric motivation for Algorithm* 4.1). The discussion here makes the results of Theorem 4.1 seem plausible—the full proof is given in Appendix C. The basic mechanism by which Algorithm 4.1 determines α_{i+1} from α_i with $\alpha_i < \alpha_{i+1} < \hat{\alpha}$ can be understood geometrically from Fig. 4.1. That property of the sequence α_i suggests strongly that α_i actually converges to $\hat{\alpha}$, as claimed in (4.22). The value of $\tilde{\beta}_{i+1}$ computed in Algorithm 4.1 will be shown, in the proof of Theorem 4.1, to be the value of α

corresponding to the point in the line $\{\alpha h \colon \alpha \in \mathbb{R}\}$ at which that line first enters the set $K_i = \operatorname{conv}(x_i \cup \Xi)$, for Ξ of (4.6). Hence, in Fig. 4.1, the value of $\tilde{\beta}_{i+1}$ corresponds to the point shown. Figure 4.1 suggests that $\tilde{\beta}_{i+1} \geqq \hat{\alpha}$, for all i. Since $\alpha_i \to \hat{\alpha}$, it seems, from Fig. 4.1, that $\tilde{\beta}_{i+1} \to \hat{\alpha}$. Then, in view of (4.20), $\beta_i \downarrow \hat{\alpha}$, as claimed in (4.22). The upper- and lower-bounds on $\hat{\alpha}$ provided by α_i and β_i, and the fact that they both converge to $\hat{\alpha}$, enable the stopping condition of (4.21) to terminate the algorithm at an i for which α_{i+1} is within a prespecified distance of $\hat{\alpha}$. Since the final value α_{i+1} is called $\bar{\alpha}$ in Algorithm 4.1, this makes post-condition (4.24) seem reasonable. The definition of $\bar{f}$ in (4.21) leads immediately to (4.25). □

Remark 4.3 (*Algorithm* 4.1 *and problem* (4.3)). In view of Theorem 4.1, Algorithm 4.1 solves approximation problem (4.3).

From (4.24), it is clear that, by suitable choice of ε_{11} and ε_{12}, an approximation $\bar{\alpha}$ to $\hat{\alpha}$ of any required accuracy can be obtained. In (4.23), η is the normalized normal of assumption (4.4) so $\eta' h < 0$. Consequently, it is clear from (4.23) that the smaller is ε_2 (i.e., the more accurately each x_i is required to approximate $\hat{x}_{\alpha_i}$), the more rapidly will α_i be guaranteed to approach $\hat{\alpha}$ as i increases. □

The subproblems of determining an $x_i \in \varepsilon_2$-minpoint $[\alpha_i h, MF]$ (which is needed in step (4.13) of Algorithm 4.1) and of finding an $\bar{f} \in \varepsilon_2$-minpoints$_F$ $[\bar{\alpha} h, MF]$ (needed in step (4.21)), can be solved by adapting an existing proximal point algorithm for approximating the point in a closed convex set that is nearest to the origin. In order to be able to guarantee termination with an $\bar{x} \in \varepsilon_2$-minpoint $[\alpha h, MF]$ and an $\bar{f} \in \varepsilon_2$-minpoints$_F$ $[\alpha h, MF]$, it is necessary to include a suitable stopping condition in the proximal point algorithm. Such a stopping condition is given in the following proximal point algorithm.

ALGORITHM 4.2.

```
I Initialization
          (choose an initial approximation $f_0$ to a point in $\varepsilon_2$-minpoints$_F$ $[\alpha h, MF]$)
(4.26)    $f_0 :\in F$   (e.g., $f_0 := \zeta\phi$ for $\phi$ of (4.2));
          (compute the associated approximation $x_0$ to a point in $\varepsilon_2$-minpoint
          $[\alpha h, MF]$)
(4.27)    $x_0 := Mf_0$;
          $i := 0$;

II Compute terms necessary for updating $f_i$ and $x_i$
(4.28)    $t_i :\in \arg_F \min_{y \in MF} (x_i - \alpha h)' y$;      $y_i := Mt_i$;

III Decide when to stop iterating
(4.29)    $z_i := \alpha h + \left\{ \frac{(y_i - \alpha h)'(x_i - \alpha h)}{\|x_i - \alpha h\|^2} \right\} (x_i - \alpha h)$;
(4.30)    if $\|z_i - \alpha h\| = 0$ or $\|z_i - \alpha h\| > \|x_i - \alpha h\|$ then go to IV;
(4.31)    $\chi_i := \cos^{-1} \left\{ \frac{\|z_i - \alpha h\|}{\|x_i - \alpha h\|} \right\}$;    $\nu_i := \cos^{-1} \left\{ \frac{h'(z_i - \alpha h)}{\|h\| \, \|z_i - \alpha h\|} \right\}$;
(4.32)    If
(4.33)        $\|x_i - \alpha h\| \leqq (1 + \varepsilon_2) \|z_i - \alpha h\|$   and   $\|x_i - \alpha h\| \leqq \|\zeta h - \alpha h\|$
(4.34)        $(\alpha h - x_i)' h < 0$   and   $\cos(\chi_i - \nu_i) > 0$
(4.35)        $(1 - \varepsilon_2) \|x_i - \alpha h\| \leqq \|h\| \left[ \frac{(\alpha h - x_i)'(y_i - \alpha h)}{(\alpha h - x_i)' h} \right] \cos(\chi_i + \nu_i)$
          then $\bar{f} := f_i$; $\bar{x} := x_i$; stop;
```

IV *Update f_i and x_i*

(4.36) $f_{i+1} := \text{minpoint}_{[f_i, t_i]}\,[\alpha h, [f_i, t_i]]$; $x_{i+1} := Mf_{i+1}$; $i := i + 1$; go to II. □

Remark 4.4 (Algorithm 4.2 and existing proximal point algorithms). The above algorithm is essentially the Gilbert proximal point algorithm [9] with a suitable stopping condition, which is the only novel feature. In practice, it would probably be better to use in step IV the update for f_i corresponding to that of Algorithm 2.4.8 in [10], which according to the evidence in [10], should yield much faster convergence. For the sake of brevity, that update has not been presented here, although it was used in the program which generated the numerical results of § 6. □

THEOREM 4.2. *Suppose $\alpha h \notin MF$ and $\varepsilon_2 \in (0, 1)$.*

Consider Algorithm 4.2 *with any update for f_i in step* IV (*such as that shown*) *that, if stopping condition* (4.32) *were omitted from the algorithm, would give $f_i \in F$, for all i, and would give $x_i \to$* minpoint $[\alpha h, MF]$ *with $\|x_i - \alpha h\| \downarrow \|\hat{x}_\alpha - \alpha h\|$.*

Then Algorithm 4.2 *terminates in a finite number of iterations with $\bar{x} \in \varepsilon_2$*-minpoint $[\alpha h, MF]$ *and $\bar{f} \in \varepsilon_2$*-minpoints$_F$ $[\alpha h, MF]$. □

5. On the eigenvalue maximization problem. The problem studied first here is that of (1.4): the determination of sign $(\hat{\lambda})$ and of a $\hat{\mu}$ if $\hat{\lambda} = 0$.

Recall that the results of Theorem 2.1 reveal that sign $(\hat{\lambda}) = 1$ if $0 \notin \Pi$, $= 0$ if $0 \in \partial\Pi$, $= -1$ if $0 \in \text{int}(\Pi)$. One way to attempt to find sign $(\hat{\lambda})$ would be to take any nonzero point in Π, e.g., $p(z)$ for any z such that not all the scalars $z'H_iz$ are zero, and then to consider the line $\{\alpha p(z) : \alpha \in \mathbb{R}\}$. Then the value of $\hat{\alpha} = \min\{\alpha \in \mathbb{R} : \alpha p(z) \in \Pi\}$ would reveal that $0 \notin \Pi$ if it turned out that $\hat{\alpha} \geqq 0$ and would reveal that $0 \in \partial\Pi$ if $\hat{\alpha} = 0$. Provided a hyperplane supporting Π at $\hat{\alpha}p(z)$ has normal η satisfying $\eta' p(z) < 0$, Algorithm 4.1 (with $q = r$, $M = I_q$, $F = \Pi$, $h = p(z)$) could be used to generate a sequence $\alpha_i \to \hat{\alpha}$. However, a $\zeta \in \mathbb{R}$ and a $k > 0$ such that $\zeta h + \{\theta \in \mathbb{R}^q : \|\theta\| \leqq k\} \subset \Pi$ do not seem to be available, so it is not clear that Assumption (4.6) regarding Algorithm 4.1 can be satisfied. Consequently the stopping condition for that algorithm cannot be implemented and it does not seem obvious how to decide when to stop iterating in such a way as to guarantee that an approximation to $\hat{\alpha}$ of prespecified accuracy will be obtained. Consequently, a different approach to the determination of sign $(\hat{\lambda})$ will be developed here.

In the context of problem (1.4), the following theorem and remark reveal that there is no loss of generality in assuming that the symmetric matrices $H_1, \cdots, H_l$ are linearly independent.

THEOREM 5.1. *Suppose $H_1, H_2, \cdots, H_l$ are linearly dependent and span the subspace $\mathscr{H}$. Suppose that $\bar{H}_1, \bar{H}_2, \cdots, \bar{H}_{\bar{l}}$ are linearly independent symmetric matrices that also span $\mathscr{H}$.*

Let $\bar{\lambda}(\bar{\mu}) = \lambda_{\min}(\sum_{i=1}^{\bar{l}} \bar{\mu}_i \bar{H}_i)$, $\hat{\bar{\lambda}} = \max\{\bar{\lambda}(\bar{\mu}) : \bar{\mu} \in \mathbb{B}^{\bar{l}}\}$ and let $\hat{\bar{\mu}}$ maximize $\bar{\lambda}(\bar{\mu})$ on $\mathbb{B}^{\bar{l}}$. Then

(i) $(\hat{\bar{\lambda}} > 0) \Rightarrow (\hat{\lambda} > 0)$;

(ii) $(\hat{\bar{\lambda}} = 0) \Rightarrow (\hat{\lambda} = 0$, *a $\hat{\mu}$ that maximizes λ on $\mathbb{B}^l$ is $\hat{\mu} = \langle\!\langle \mu \rangle\!\rangle$ for any $\mu \in \mathbb{R}^l$ such that $\sum_{i=1}^{l} \mu_i H_i = \sum_{i=1}^{\bar{l}} \hat{\bar{\mu}}_i \bar{H}_i$, and $\sum_{i=1}^{l} \hat{\mu}_i H_i \neq 0$ for all such $\hat{\mu}$*);

(iii) $(\hat{\bar{\lambda}} < 0) \Rightarrow (\hat{\lambda} = 0$, *a $\hat{\mu}$ that maximizes λ on $\mathbb{B}^l$ is $\hat{\mu} = \langle\!\langle \mu \rangle\!\rangle$ for any $\mu \in \mathbb{R}^l$ such that $\sum_{i=1}^{l} \mu_i H_i = 0$, and $\sum_{i=1}^{l} \hat{\mu}_i H_i = 0$ for all such $\hat{\mu}$*). □

Hence if $\hat{\bar{\lambda}} > 0$ then $\hat{\lambda} > 0$ and if $\hat{\bar{\lambda}} \leqq 0$ then $\hat{\lambda} = 0$ and $\hat{\mu}$ may be found in the way specified in the appropriate part of Theorem 5.1. Therefore the solution of problem (1.4) for linearly dependent H_i can be obtained by finding sign $(\hat{\bar{\lambda}})$ for appropriate linearly independent $\bar{H}_i$ and by finding the associated $\hat{\hat{\mu}}$ if $\hat{\hat{\lambda}} = 0$. Hence in the context of problem (1.4), there is no loss of generality in studying the case when the H_i are linearly independent.

The determination of $\hat{\lambda}$ and $\hat{\mu}$ when the H_i are linearly independent will be considered next.

Clearly, $\hat{\lambda} > 0$ if there is a $\mu \in \mathbb{R}^l$ such that $\sum_{i=1}^l \mu_i H_i = I_n$, i.e., after taking the $\overline{\text{vec}}$ (recall § 3) of each side of the above equation, if $\sum_{i=1}^l \mu_i \overline{\text{vec}}\,[H_i] = \overline{\text{vec}}\,[I_n]$ for some $\mu_i \in \mathbb{R}$, i.e., if $H\mu = \iota$ for some $\mu \in \mathbb{R}^l$, where H and ι are those of (5.1)–(5.2) below, i.e., if $d = 0$, for d of (5.3), where

$$H = [\overline{\text{vec}}\,[H_1]\,\overline{\text{vec}}\,[H_2]\cdots\overline{\text{vec}}\,[H_l]] \in \mathbb{R}^{r\times l}, \tag{5.1}$$

$$\iota = \overline{\text{vec}}\,[I_n] \in \mathbb{R}^r, \tag{5.2}$$

$$d = D\iota \in \mathbb{R}^r, \tag{5.3}$$

$$D = I_r - HH^\dagger \in \mathbb{R}^{r\times r}. \tag{5.4}$$

Summarizing: $\hat{\lambda} > 0$ if $d = 0$. This result can be generalized somewhat.

THEOREM 5.2. *If* $\|d\| < (\leqq)\ 1$, *then* $\hat{\lambda} > (\geqq)\ 0$ *and* $H^\dagger\iota \neq 0$. *Furthermore*, $\bar{\mu} = \langle\langle H^\dagger \iota\rangle\rangle$ *is an approximation to* $\hat{\mu}$ *in that*

$$\hat{\lambda} \geqq \lambda(\bar{\mu}) \geqq \frac{1-\|d\|}{\|H^\dagger \iota\|} > (\geqq) 0. \qquad \square$$

Example 2 (Example 3) in § 6 shows that $\|d\| < (\leqq)\ 1$ is not a necessary condition for $\hat{\lambda} > (\geqq)\ 0$, so the sufficient condition of Theorem 5.2 for $\hat{\lambda} > (\geqq)\ 0$ is not necessary.

Theorem 5.2 reveals that sign $(\hat{\lambda}) = 1$ if $\|d\| < 1$ so problem (1.4) has been solved in that case. The following analysis will enable problem (1.4), or approximation problem (1.5), to be solved when $\|d\| \geqq 1$.

Recall from Theorem 3.1(i) that $\overline{\text{vec}}\,[S^n_{\geqq}] = \text{cone}\,(\Gamma)$. Now

$$\lambda_{\min}\left(\sum_{i=1}^{l} \mu_i H_i\right) = \delta$$

if and only if $-\delta$ is the most negative real number α such that $\sum_{i=1}^l \mu_i H_i + \alpha I_n \in S^n_{\geqq}$, i.e., such that $\overline{\text{vec}}\,[\sum_{i=1}^l \mu_i H_i + \alpha I_n] \in \overline{\text{vec}}\,[S^n_{\geqq}] = \text{cone}\,(\Gamma)$, i.e., (after making use of (5.1)–(5.2)) such that $H\mu + \alpha\iota \in \text{cone}\,(\Gamma)$. The reason for stating the condition in terms of $-\delta$ being the most negative number α such that $H\mu + \alpha\iota \in \text{cone}\,(\Gamma)$ rather than saying that δ is the most positive number α such that $H\mu - \alpha\iota \in \text{cone}\,(\Gamma)$ is that the former choice renders consequential optimization problems easier to interpret geometrically. Hence we might expect there to be a connection, which will be given in Theorem 5.3 later, between $\hat{\lambda}$ and the number $\hat{\alpha}$ defined below.

DEFINITION 5.1. Let $\hat{\alpha}$ be the most negative real number α for which there is a $\mu \in \mathbb{R}^l$ such that $H\mu + \alpha\iota \in \Gamma$. Let $\mu_{\hat{\alpha}}$ be a $\mu \in \mathbb{R}^l$ associated with $\hat{\alpha}$ in that $H\mu_{\hat{\alpha}} + \hat{\alpha}\iota \in \Gamma$. $\square$

In view of Definition 5.1, if $\hat{\alpha}$ exists it is the most negative α such that

$$\min_{\gamma\in\Gamma}\min_{\mu\in\mathbb{R}^l} \|\gamma - [H\mu + \alpha\iota]\| = 0. \tag{5.5}$$

For each $\gamma \in \Gamma$, the minimizing μ is clearly $H^\dagger[\gamma - \alpha\iota]$. Hence, for d of (5.3) and D of (5.4), condition (5.5) is equivalent to the condition

$$\min_{\gamma\in\Gamma} \|D\gamma - \alpha d\| = 0, \tag{5.6}$$

i.e., to

$$\min_{x\in D\Gamma}\|x-\alpha d\|=0. \tag{5.7}$$

Consequently, if $\hat{\alpha}$ exists then

$$\hat{\alpha}=\min\{\alpha\in\mathbb{R}:\alpha d\in D\Gamma\} \tag{5.8}$$

and a $\mu_{\hat{\alpha}}$ is given by

$$\mu_{\hat{\alpha}}=H^{\dagger}[\gamma_{\hat{\alpha}}-\hat{\alpha}\iota] \tag{5.9}$$

where

$$\gamma_{\hat{\alpha}}\in\arg_{\Gamma}\text{minpoints}\,[\hat{\alpha}d,D\Gamma]. \tag{5.10}$$

The useful connection between $\hat{\lambda}$ and $\hat{\alpha}$ mentioned earlier is among the following properties concerning d and $\hat{\alpha}$.

THEOREM 5.3. *Suppose that $H_1,\cdots,H_l$ are linearly independent and that $\iota\notin R[H]$, i.e., that $d\neq 0$. Then*

(i) *$D\Gamma$ is compact; $\hat{\alpha}$ exists; $|\hat{\alpha}|\leqq\|d\|^{-1}$; if $H'\iota=0$ then $\hat{\alpha}=n^{-1}$ else $\hat{\alpha}\leqq n^{-1}$;*

(ii) *If $H'\iota\neq 0$, then $d\in R[DM]$ (for M of Theorem 3.2);*

(iii) $\operatorname{sign}(\hat{\lambda})=-\operatorname{sign}(\hat{\alpha})$;

(iv) $(\hat{\alpha}<0)\Rightarrow(\mu_{\hat{\alpha}}\neq 0$ and $-n^{1/2}\hat{\alpha}\|H\|\geqq\hat{\lambda}\geqq\lambda(\langle\!\langle\mu_{\hat{\alpha}}\rangle\!\rangle)>0)$;

(v) $(\hat{\alpha}=0)\Rightarrow(\hat{\lambda}=0,\ \mu_{\hat{\alpha}}\neq 0$ *and a $\hat{\mu}$ is* $\hat{\mu}=\langle\!\langle\mu_{\hat{\alpha}}\rangle\!\rangle)$;

(vi) $(\hat{\alpha}>0)\Rightarrow(\hat{\lambda}<0)$. □

From part (iii), $\operatorname{sign}(\hat{\lambda})$ can be determined from $\operatorname{sign}(\hat{\alpha})$. Hence the solution of problem (1.4) can be determined by evaluating $\operatorname{sign}(\hat{\alpha})$ and, from part (v), by evaluating $\mu_{\hat{\alpha}}$ if $\operatorname{sign}(\hat{\alpha})=0$.

By Theorem 5.3(i), $\hat{\alpha}=n^{-1}$ if $H'\iota=0$. So, from Theorem 5.3(iii), $\operatorname{sign}(\hat{\lambda})=-1$ in that case. Of course, in general $H'\iota\neq 0$. Then the determination of $\operatorname{sign}(\hat{\lambda})$ is more complicated. It turns out that, when $d\neq 0$ and $H'\iota\neq 0$, specification (5.8) for $\hat{\alpha}$ can easily be transformed into a specification that is more convenient computationally, as follows.

Suppose $d\neq 0$ and $H'\iota\neq 0$.

Theorem 3.2 reveals that

$$n^{-1}\iota+M\Xi_2\subset\Gamma\subset n^{-1}\iota+M\Xi_1\subset n^{-1}\iota+R[M].$$

Premultiplying by D and using the result of Theorem 5.3(ii) that $d\in R[DM]$ when $H'\iota\neq 0$, we obtain

$$n^{-1}d+DM\Xi_2\subset D\Gamma\subset n^{-1}d+DM\Xi_1\subset n^{-1}d+R[DM]\subset R[DM]. \tag{5.11}$$

Now $DM\neq 0$ when $d\neq 0$ and $H'\iota\neq 0$ since then, by Theorem 5.3(ii), $0\neq d\in R[DM]$. Therefore the singular value decomposition of DM can be written as

$$DM=P\,\text{blockdiag}\,\{\Lambda,0\}\,Q'\in\mathbb{R}^{r\times(r-1)}\quad\text{for } 0<\Lambda\in\mathbb{R}^{q\times q} \tag{5.12}$$

for some $q>0$, with $q\leqq r$, and where $P\in\mathbb{R}^{r\times r}$ and $Q\in\mathbb{R}^{(r-1)\times(r-1)}$ are orthogonal matrices.

Let

$$T=\text{blockdiag}\,\{\Lambda^{-1},I_{r-q}\}P'\in\mathbb{R}^{r\times r},\qquad S=[I_q\quad 0]\in\mathbb{R}^{q\times r}. \tag{5.13}$$

Premultiply (5.11) by T and make use of (5.12) to give

$$\begin{aligned} & n^{-1}Td+\begin{bmatrix} I_q & 0 \\ 0 & 0 \end{bmatrix} Q'\{x\in\mathbb{R}^{r-1}: \|x\| \leqq n^{-1}\} \\ (5.14) \qquad & \subset TD\Gamma \\ & \subset n^{-1}Td+\begin{bmatrix} I_q & 0 \\ 0 & 0 \end{bmatrix} Q'\{x\in\mathbb{R}^{r-1}: \|x\| \leqq \sqrt{1-n^{-1}}\}\subset\mathbb{R}^q \end{aligned}$$

where, since $0 \neq d \in R[DM]$ and T is nonsingular,

$$(5.15) \qquad Td=\begin{bmatrix} \tilde{d} \\ 0 \end{bmatrix}, \quad 0\neq \tilde{d}\in\mathbb{R}^q.$$

Premultiply (5.15) and (5.14) by S of (5.13) to give, respectively,

$$(5.16) \quad \tilde{d}=STd=STD\iota,$$

$$(5.17) \quad n^{-1}\tilde{d}+\{y\in\mathbb{R}^q: \|y\|\leqq n^{-1}\}\subset STD\Gamma\subset n^{-1}\tilde{d}+\{y\in\mathbb{R}^q: \|y\|\leqq\sqrt{1-n^{-1}}\}\subset\mathbb{R}^q.$$

Since T of (5.13) is nonsingular, $\hat{\alpha}$ of (5.8) is also given by

$$(5.18) \qquad \hat{\alpha}=\min\{\alpha\in\mathbb{R}: \alpha Td\in TD\Gamma\}=\min\{\alpha\in\mathbb{R}: \alpha\tilde{d}\in STD\Gamma\}$$

where the second equality is a consequence of the fact that, by (5.14)-(5.15),

(5.19) the last $r-q$ entries of Td and of every vector in $TD\Gamma$ are all zero.

This has proved the first two parts of the following theorem. Part (iii) is a straightforward consequence of (5.17) and (5.18).

THEOREM 5.4. *For linearly independent H_i when $d \neq 0$ and $H'\iota \neq 0$*

(i) $\tilde{d} \neq 0$ *and* $\hat{\alpha} = \min\{\alpha: \alpha\tilde{d} \in STD\Gamma\}$ *for $\tilde{d}$ of* (5.15) *and T, S of* (5.13);

(ii) $n^{-1}\tilde{d} + \{y \in \mathbb{R}^q: \|y\| \leqq n^{-1}\} \subset STD\Gamma \subset n^{-1}\tilde{d} + \{y \in \mathbb{R}^q: \|y\| \leqq \sqrt{1-n^{-1}}\}$;

(iii) $\hat{\alpha} \in [\alpha^-, \alpha^+]$ *where* $\alpha^- = n^{-1} - \sqrt{1-n^{-1}}/\|\tilde{d}\|$, $\alpha^+ = n^{-1} - n^{-1}/\|\tilde{d}\|$;

(iv) *$STD\Gamma$ is a compact convex subset of $\mathbb{R}^q$ and there exists a hyperplane that supports $STD\Gamma$ at $\hat{\alpha}\tilde{d}$ and has a normalized normal η that satisfies $\eta'\tilde{d} < 0$.* □

Theorem 5.4(iii) might enable sign ($\hat{\alpha}$) to be determined for it reveals that if $\alpha^+ < 0$ then sign $(\hat{\alpha}) = -1$ and if $\alpha^- > 0$ then sign $(\hat{\alpha}) = 1$. However in general the values of α^- and α^+ will not be sufficient to specify $\hat{\alpha}$. Then more information can be obtained by considering the specification of $\hat{\alpha}$ given by Theorem 5.4(i). Algorithm 4.1 can be applied to approximate $\hat{\alpha}$ to any required accuracy by approximating the solution of $\min\{\alpha \in \mathbb{R}: \alpha\tilde{d} \in STD\Gamma\}$. The identifications of the terms h, F, M, and k in Algorithm 4.1 necessary to be able to do that are given in (5.20) below. The reasons the assumptions for Algorithm 4.1 are then all satisfied will be explained next.

Assumption 4.2 is satisfied for the F, ζ, ϕ, M, h, k of (5.20) because

(i) $F = \Gamma$ is a compact convex set (by Theorem 3.1(ii));

(ii) $\zeta\phi \in F$ since $\zeta\phi = n^{-1}\iota$ and, by Theorem 3.2, $n^{-1}\iota \in \Gamma$;

(iii) $h = M\phi = STD\iota = STd = \tilde{d} \neq 0$ (by (5.16) and Theorem 5.4(i)).

Assumption (4.4) is valid because $\hat{\alpha}$ exists (by Theorem 5.3(i)) and because there is a hyperplane with normalized normal $\eta \in \mathbb{R}^q$ that satisfies $\eta'\tilde{d} < 0$ and supports $STD\Gamma$ at $\hat{\alpha}\tilde{d}$ (by Theorem 5.4(iv)). Assumption (4.5) is valid, after making the identifications of (5.20), since $\arg_F \max\{g'y: y \in MF\} = \arg_\Gamma \min\{(-g)'y: y \in M\Gamma\}$ and a point in $\arg_\Gamma \min\{(-g)'y: y \in M\Gamma\}$ can be found using Theorem 3.1(iii). Finally, Theorem 5.4(ii) reveals that Assumption (4.6) is satisfied for $\zeta = k = n^{-1}$.

Stopping condition (4.21) of Algorithm 4.1 can easily be modified to (4.21′) below to enable that algorithm to solve approximation problem (1.5). The key results are stated in the following theorem, where $\|d\| \geqq 1$ is assumed because the situation when $\|d\| < 1$ is covered adequately by Theorem 5.2.

THEOREM 5.5. *Suppose the H_i are linearly independent, $\|d\| \geqq 1$ and $H'\iota \neq 0$. Consider employing Algorithm* 4.1 *for approximating $\tilde{\alpha} = \min\{\alpha \in \mathbb{R}\colon \alpha\tilde{d} \in STD\Gamma\}$ when its data F, ζ, ϕ, M, h, k are specified by*

$$(5.20) \qquad F=\Gamma, \quad \zeta=n^{-1}, \quad \phi=\iota, \quad M=STD\in\mathbb{R}^{q\times r}, \quad h=\tilde{d}\in\mathbb{R}^{q}, \quad k=n^{-1}$$

where q, T, and S are from (5.12), (5.13), $r = \frac{1}{2}n(n+1)$ (*from* § 3) *and where $\tilde{d}$ is from* (5.16).

Choose $\delta \in \mathbb{R}_>$ and let ε_2 be as specified in (4.9) *of Algorithm* 4.1.

Suppose Algorithm 4.1′ *is defined to be Algorithm* 4.1 *with termination condition* (4.21) *replaced by*

$$\text{if}\,\alpha_{i+1}\geqq 0 \textit{ or } \beta_{i+1}<0 \textit{ then stop};$$

$$(4.21') \qquad \delta_{i+1}:=-\alpha_{i+1}n^{1/2}\|H\|+\frac{|\alpha_{i+1}+(1+\epsilon_2)(\beta_{i+1}-\alpha_{i+1})\|d\|\,|\,\|\iota\|\;\|H\|}{(n^{-1}-\alpha_{i+1})\|HH^{\dagger}\iota\|^{2}};$$

$$\textit{if}\,\alpha_{i+1}<0 \textit{ and } \beta_{i+1}>0 \textit{ and } \delta_{i+1}\leqq\delta \textit{ then } \bar{\gamma}:\in\varepsilon_2\text{-minpoints}_{\Gamma}\,[\alpha_{i+1}d, D\Gamma];$$

$$\bar{\mu}:=\langle\!\langle H^{\dagger}[\bar{\gamma}-\alpha_{i+1}\iota]\rangle\!\rangle;\textit{stop};.$$

Then Algorithm 4.1′ *will stop after a finite number of iterations and, at termination, approximation problem* (1.5) *will have been solved because*

if $\alpha_{i+1} \geqq 0$ then $\hat{\lambda} < 0$; if $\beta_{i+1} < 0$ then $\hat{\lambda} > 0$;
if $\alpha_{i+1} < 0$ and $\beta_{i+1} \geqq 0$ and $\delta_{i+1} \leqq \delta$ then $\lambda(\bar{\mu}) \in [\hat{\lambda} - \delta, \hat{\lambda}]$. □

Remark 5.1 (Determination of $\bar{\gamma}$ in Theorem 5.5 and of x_i in Algorithm 4.1). The vector $\bar{\gamma}$ of (4.21′) in Theorem 5.5 can be determined by applying Algorithm 4.2 with the following identifications: $F = \Gamma$, $\zeta = n^{-1}$, $\phi = \iota$, $M = D$, $h = d$, $k = n^{-1}$. Then it turns out from the proof of Theorem 5.5 that to obtain $\lambda(\bar{\mu}) \in [\hat{\lambda} - \delta, \hat{\lambda}]$, it is only necessary to find a γ such that both $\|D\gamma - \alpha_{i+1}d\| \leqq (1 + \varepsilon_2)$ mindist $[\alpha_{i+1}d, D\Gamma]$ and $\|D\gamma - \alpha_{i+1}d\| \leqq \|\zeta d - \alpha_{i+1}d\|$ for a suitable α_{i+1}. Consequently, stopping conditions (4.34) and (4.35) could be omitted from Algorithm 4.2 when it is applied to compute $\bar{\gamma} \in \varepsilon_2$-minpoints$_\Gamma$ $[\alpha_{i+1}d, D\Gamma]$ in the modified stopping condition (4.21′) that is stated in Theorem 5.5, possibly leading to termination after a smaller number of iterations than otherwise. Furthermore, in the operation of the main body of Algorithm 4.1 itself, using the identifications of (5.20), the condition $\|x_i - \alpha h\| \leqq \|\zeta h - \alpha h\|$ in Definition 4.1 of the set ε-minpoint $[\alpha h, MF]$ is not necessary, so that condition could be removed from stopping condition (4.33) of Algorithm 4.2 when Algorithm 4.2 is used to implement step (4.13) of Algorithm 4.1. The requirements of Definition 4.1 were chosen to yield the least complicated exposition. □

Clearly the results of this section can be used together to solve problem (1.6), namely, to solve problem (1.4) when that seems possible using finite work and to solve approximation problem (1.5) otherwise. The procedure for doing that is summarized next.

PROCEDURE 5.1.

Suppose $H_1, H_2, \cdots, H_l$ are linearly independent.

Compute d of (5.3). If $\|d\| < 1$, then (by Theorem 5.2), sign $(\hat{\lambda}) = 1$ and problem (1.4) has been solved, so stop.

If $\|d\| \geqq 1$, then compute $H'\iota$.

If $H'\iota = 0$ then $\hat{\alpha} = n^{-1}$ (by Theorem 5.3(i)) and sign$(\hat{\lambda}) = -1$ (by Theorem 5.3(iii)) so stop since problem (1.4) has been solved.

If this point is reached then $H'\iota \neq 0$ so next compute α^- and α^+ of Theorem 5.4(iii).

If $\alpha^- > 0$ then sign$(\hat{\alpha}) = +1$ (by Theorem 5.4(iii)) and consequently sign$(\hat{\lambda}) = -1$ (by Theorem 5.3(iii)) so stop since problem (1.4) has been solved. Similarly, if $\alpha^+ < 0$ then sign$(\hat{\lambda}) = 1$ so stop since problem (1.4) has been solved.

However, if $\alpha^- < 0$ and $\alpha^+ > 0$ then sign$(\hat{\alpha})$ cannot be determined from consideration of α^- and α^+ so apply Algorithm 4.1 to solve approximation problem (1.5), using Theorem 5.5. □

6. Computed examples. When Procedure 5.1 was applied here, the value of ε_2 for Algorithm 4.1 was taken to be 0.5 as that seemed to yield a good compromise between convergence rate, with respect to iterations, for Algorithm 4.1 and the number of iterations required by Algorithm 4.2 when called by Algorithm 4.1. The parameter ε_3 was taken to be 0.05. The value $\delta = 10^{-3}$ was found to be adequate for the examples considered because stopping condition (4.21′) of Theorem 5.5 then gave very nearly optimal $\bar{\mu}$s (e.g., see Example 3 below).

Example 1. Here

$$H_1 = \begin{bmatrix} -1 & -2 \\ -2 & 0 \end{bmatrix}, \qquad H_2 = \begin{bmatrix} 0 & 4 \\ 4 & 1 \end{bmatrix}.$$

In this case, with $n = 2$, it was possible to plot the set Π of § 2 and it turned out that $0 \in \operatorname{int}(\Pi)$ and that a $\tilde{\pi}$ (minimizing $\|\pi\|$ on $\partial\Pi$) is approximately $[0.54\ 0.84]'$ so that $\hat{\mu}$ is approximately $[-0.49\ -0.87]'$ and gives $\hat{\lambda}$ as approximately -0.43. Of course, this technique would not be practical when $n > 3$ since such plotting would not be feasible.

Application of Procedure 5.1 gave $\|d\| \approx 1.33$ (so Theorem 5.2 provides no information regarding sign$(\hat{\lambda})$), and gave $\alpha^- \approx 0.314$ (consequently for this problem $\hat{\alpha} > 0$), which reveals that $\hat{\lambda} < 0$, which is consistent with the above results obtained from consideration of Π. □

Example 2. Here

$$H_1 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad H_2 = \begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 1 \end{bmatrix}.$$

Application of Procedure 5.1 gave $\|d\| \approx 1.04$ (so Theorem 5.2 provides no information regarding sign$(\hat{\lambda})$) and $\alpha^- \approx -0.291$, $\alpha^+ \approx 0.008$ (which do not specify sign$(\hat{\lambda})$) so then Algorithm 4.1′ was applied and gave the following results:

i	α_i	β_i	δ_i
0	-0.274	7.84×10^{-2}	-
1	-5.31×10^{-2}	-1.203×10^{-2}	0.355

Iteration stopped because β_1 was negative indicating that $\hat{\alpha} < 0$ and hence that $\hat{\lambda} > 0$. Therefore the method of Appendix C of [1] was actually applied to approximate $\hat{\mu}$ and $\hat{\lambda}$, giving $\hat{\mu} \approx [0.9747\ 0.2237]'$ and $\hat{\lambda} \approx 0.1146$. This has demonstrated the utility of Procedure 5.1 for a case when $\hat{\lambda} > 0$. Note that Algorithm 4.1′ decided that sign$(\hat{\lambda}) = 1$ after only one iteration.

Example 3. In this case $n = 7$ and $l = 4$ with

$$H_1 = \text{blockdiag}\left\{\begin{bmatrix} 2 & -1 & -1 & 0 & 0 \\ -1 & 0 & 0 & -2 & 0 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 & -3 \\ 0 & 0 & 0 & -3 & 0 \end{bmatrix}, 0_{2\times 2}\right\},$$

$$H_2 = \text{diag}\,\{0, 10, 5, 4, 3, 0, 0\},$$

$$H_3 = \text{blockdiag}\left\{0_{5\times 5}, \begin{bmatrix} 4 & -2 \\ -2 & 0 \end{bmatrix}\right\},$$

$$H_4 = \text{blockdiag}\left\{0_{5\times 5}, \begin{bmatrix} 10 & 0 \\ 0 & -10 \end{bmatrix}\right\}$$

where $0_{i\times j}$ denotes the null matrix from $\mathbb{R}^{i\times j}$.

This example has $\hat{\lambda} = 0$ and is therefore suitable for testing Procedure 5.1 for that case. The reason $\hat{\lambda} = 0$ is explained next.

It can be checked easily that $\lambda_{\min}[\mu_3 H_3 + \mu_4 H_4] \leqq 0$ with equality holding only when $\mu_3 = -5\mu_4$ and $\mu_4 \leqq 0$. Consequently, owing to the block-diagonal structure of $\sum_{i=1}^{5} \mu_i H_i$, $\hat{\lambda} \leqq 0$. By running Procedure 5.1 on H_1 and H_2 alone, it was found that $\max\{\lambda_{\min}(\mu_1 H_1 + \mu_2 H_2) \colon \mu \in \mathbb{B}^2\} = 0.958$. Hence, again owing to the block-diagonal structure of $\sum_{i=1}^{5} \mu_i H_i$, for the above problem with all five matrices H_i, $\hat{\lambda} = 0$.

Application of Procedure 5.1 gave $\|d\| \approx 1.63$ (so Theorem 5.2 provides no information regarding sign$(\hat{\lambda})$) and $\alpha \approx -0.305 \times 10^{-3}$, $\alpha^+ \approx 0.0738$ (which do not specify sign$(\hat{\lambda})$) so then Algorithm 4.1′ was applied and gave the following results:

i	α_i	β_i	δ_i
0	-2.33×10^{-1}	7.38×10^{-2}	–
2	-9.93×10^{-3}	3.41×10^{-3}	1.84
4	-6.07×10^{-5}	5.62×10^{-5}	0.214
6	-1.52×10^{-7}	2.40×10^{-7}	5.54×10^{-5}

at which point the algorithm stopped because $\alpha_6 < 0$, $\beta_6 \geqq 0$ and $\delta_i \leqq \delta = 0.001$. The value of $\bar{\mu}$ obtained was $\bar{\mu} \approx (5.81 \times 10^{-1}, 8.14 \times 10^{-1}, 7.23 \times 10^{-7}, -1.45 \times 10^{-7})$. Since $\hat{\lambda} = 0$ in this case, Theorem 5.1 claims that $\lambda(\bar{\mu})$ should belong to $[-\delta, 0] = [-.001, 0]$. It is clear that a much better approximation to $\hat{\lambda} = 0$ than that was achieved since in fact $\lambda(\bar{\mu}) \approx -1.24 \times 10^{-17}$. This has illustrated the usefulness of Procedure 5.1 for a case with $\hat{\lambda} = 0$. □

These examples have demonstrated the value of the results and algorithms that have been presented in this paper.

Appendix A. Proofs for § 2. The proof of Theorem 2.1 is facilitated by the next result.

LEMMA A1. *If* $0 \in \text{int}(\Pi)$ *and* $\tilde{\pi}$ *minimizes* $\|\pi\|$ *on* $\partial\Pi$, *then*

$$\max\{\tilde{\pi}'\pi \colon \pi \in \Pi\} = \|\tilde{\pi}\|^2. \tag{A1}$$ □

Proof. Since $0 \in \text{int}(\Pi)$, $\|\tilde{\pi}\| > 0$. For proof by contradiction, suppose

(A2) There exists a $\bar{\pi} \in \Pi$ with $\tilde{\pi}'\bar{\pi} > \|\tilde{\pi}\|^2$.

Consider

$$\pi^* := \tilde{\pi} - a(\bar{\pi} - \tilde{\pi}) \tag{A3}$$

for some $a \in R_>$ that is chosen so that

$$\tilde{\pi} \in \Pi \quad \text{and} \quad 0 < \|\pi^*\| < \|\tilde{\pi}\|, \tag{A4}$$

which, in view of (A2) and the convexity of Π, can be done.

Let

$$\pi^0 = \pi^* m[\pi^*]^{-1} \in \partial\Pi \tag{A5}$$

where $m\colon \mathbb{R}^l \to \mathbb{R}$ is the Minkowski functional for Π, defined by

$$m[z] = \inf\{r \in R_>: r^{-1}z \in \Pi\}. \tag{A6}$$

Since $0 \in \operatorname{int}(\Pi)$ and Π is compact, it follows [11, Lemma 5.12.1] that

$$\begin{aligned} &0 < m[z] < \infty \quad \forall z \in R^n - \{0\}, \\ &\operatorname{int}(\Pi) = \{z \in \mathbb{R}^l: m(z) < 1\}, \quad \Pi = \{z \in \mathbb{R}^l: m(z) \leqq 1\}. \end{aligned} \tag{A7}$$

Now $\tilde{\pi}$ minimizes $\|\pi\|$ on $\partial\pi$ so, by (A5)

$$\|\pi^* m[\pi^*]^{-1}\| \geqq \|\tilde{\pi}\|.$$

Hence, in view of (A4) and (A7),

$$0 < m[\pi^*] < 1. \tag{A8}$$

Consider $\pi^\# = b\bar{\pi} + (1 - b)\pi^0$, where $b = a(m[\pi^*] + a)^{-1}$. Then $\pi^\# \in \Pi$ because $b \in (0, 1)$ and $\bar{\pi}$, $\pi^0 \in \Pi$. Direct computation using (A3), (A5) reveals that $\pi^\# = c\tilde{\pi}$ where $c = (1 + a)(m[\pi^*] + a)^{-1}$. Hence, since $\pi^\# \in \Pi$, $c\tilde{\pi} \in \Pi$ where, in view of (A8) and the choice of a from $R_>$, $c > 1$. Since $\tilde{\pi} \in \partial\Pi$ and $\pi^\# \in \Pi$, it follows from (A7) that $m(\tilde{\pi}) = 1$ and $m(\pi^\#) \leqq 1$. Therefore

$$1 \geqq m(\pi^\#) = m(c\tilde{\pi}) = cm(\tilde{\pi}) = c > 1,$$

which is a contradiction. Hence (A2) is false that, since $\tilde{\pi} \in \Pi$, establishes (A1) and thereby completes the proof. □

Proof of Theorem 2.1. From Theorem 13.1 of [12]

$$0 \in \operatorname{int}(\Pi) \text{ iff } 0 < \max\{\mu'\pi: \pi \in \Pi\} \quad \forall \mu \in \mathbb{B}^l,$$

$$0 \in \Pi \text{ iff } 0 \leqq \max\{\mu'\pi: \pi \in \Pi\}, \quad \forall \mu \in \mathbb{B}^l.$$

By Lemma 2.1, $\max\{\mu'\pi: \pi \in \Pi\} = -\lambda(-\mu)$. Hence

$$0 \in \operatorname{int}(\Pi) \text{ iff } \hat{\lambda} < 0, \qquad 0 \in \Pi \text{ iff } \hat{\lambda} \leqq 0. \tag{A9}$$

This has proved the if and only if results of parts (i) and (iii). The rest of the proof follows, below.

(i) Suppose that $0 \in \operatorname{int}(\Pi)$ and $\tilde{\pi} \in$ minpoints $[0, \partial\Pi]$. Then $\|\tilde{\pi}\| > 0$ and, by Lemma 2.1 and Lemma A1

$$\lambda(-\langle\!\langle\tilde{\pi}\rangle\!\rangle) = \min\{-\langle\!\langle\tilde{\pi}\rangle\!\rangle'\pi: \pi \in \Pi\} = -\max\{\langle\!\langle\tilde{\pi}\rangle\!\rangle'\pi: \pi \in \Pi\} = -\|\langle\!\langle\tilde{\pi}\rangle\!\rangle\| < 0. \tag{A10}$$

It will be shown next that $-\langle\!\langle\tilde{\pi}\rangle\!\rangle$ maximizes $\lambda(\mu)$ on $\mathbb{B}^l$.

For any $\mu \in \mathbb{B}^l$, $\mu m[\mu]^{-1} \in \partial\Pi$, where m is the Minkowski functional for Π, of (A6). Therefore, since $\tilde{\pi}$ minimizes $\|\pi\|$ on $\partial\Pi$,

$$\|\mu m[\mu]^{-1}\| \geqq \|\tilde{\pi}\| \quad \forall \mu \in \mathbb{B}^l.$$

Hence, since it follows from (A7) that $m[\mu] > 0$ for all $\mu \in \mathbb{B}^l$, and since $\|\mu\| = 1$ for $\mu \in \mathbb{B}^l$,

$$\text{(A11)} \qquad m[\mu]^{-1} \geqq \|\hat{\pi}\| \quad \forall \mu \in \mathbb{B}^l.$$

Consequently, since $\partial\Pi = \{\bar{\mu} m[\bar{\mu}]^{-1} \colon \bar{\mu} \in \mathbb{B}^l\}$,

$$\begin{aligned}\lambda(\mu) &= \min\{\mu'\pi \colon \pi \in \Pi\} = \min\{\mu'\pi \colon \pi \in \partial\Pi\} = \min\{\mu'\bar{\mu} m[\bar{\mu}]^{-1} \colon \bar{\mu} \in \mathbb{B}^l\} \\ &\leqq \mu'(-\mu) m[-\mu]^{-1} = -m[-\mu]^{-1} \leqq -\|\hat{\pi}\| = \lambda(-\langle\!\langle \hat{\pi} \rangle\!\rangle) \quad \forall \mu \in \mathbb{B}^l\end{aligned}$$

where the last inequality is from (A11) and the last equality is from (A10). This reveals that $-\langle\!\langle \hat{\pi} \rangle\!\rangle$ maximizes $\lambda(\mu)$ on $\mathbb{B}^l$, which completes the proof of part (i).

(ii) From (A9), $[0 \in \partial\Pi]$ if and only if $[\hat{\lambda} = 0]$. Suppose $0 \in \partial\Pi$. Then $\hat{\lambda} = 0$ and there is at least one hyperplane, with some nonzero normal μ, which supports Π at 0 [12, Cor. 11.6.1]. Consequently, $\max\{\mu'\pi \colon \pi \in \Pi\} = \mu' 0$ so that, by Lemma 2.1, $\lambda(-\langle\!\langle \mu \rangle\!\rangle) = 0$. Since $\hat{\lambda} = 0$ in this case, $-\langle\!\langle \mu \rangle\!\rangle$ maximizes $\lambda(\mu)$ on $\mathbb{B}^l$; as claimed in part (ii).

(iii) Since $\hat{\pi} = \text{minpoint}\,[0, \Pi]$ and Π is convex, $(\pi - \hat{\pi})'\hat{\pi} \geqq 0$, for all $\pi \in \Pi$. Consequently, by Lemma 2.1,

$$\text{(A12)} \qquad \lambda(\langle\!\langle \hat{\pi} \rangle\!\rangle) = \min\{\langle\!\langle \hat{\pi} \rangle\!\rangle'\pi \colon \pi \in \Pi\} = \|\hat{\pi}\|,$$

where, since in this case $0 \notin \Pi$, $\|\hat{\pi}\| > 0$. Furthermore, since $\hat{\pi}$ minimizes $\|\pi\|$ on Π,

$$\text{(A13)} \qquad \lambda(\mu) = \min\{\mu'\pi \colon \pi \in \Pi\} \leqq \mu'\hat{\pi} \leqq \|\mu\|\,\|\hat{\pi}\| = \|\hat{\pi}\| = \lambda(\langle\!\langle \hat{\pi} \rangle\!\rangle) \quad \forall \mu \in \mathbb{B}^l.$$

Hence $\hat{\lambda} = \|\hat{\pi}\|$ and $\lambda(\langle\!\langle \hat{\pi} \rangle\!\rangle) = \hat{\lambda}$. Furthermore, since the second inequality in (A13) is strict when μ and $\hat{\pi}$ are not collinear, $\hat{\pi}$ is the unique global maximizer of $\lambda(\mu)$ with respect to $\mu \in \mathscr{B}^l$, which completes the proof of Theorem 2.1. □

Appendix B. Proofs for § 3.

Proof of Theorem 3.1. Apart from part (ii), the results of Theorem 3.1 are an immediate consequence of Theorem 2.1 of [8]. The proof of part (ii) follows.

Theorem 2.1 [8] asserts that Ω is a convex compact set. Therefore, since $\Gamma = W'\Omega$, Γ is convex and bounded and is closed if $N[W'] \cap 0^+\Omega = \{0\}$ where $0^+\Omega$ denotes the recession cone of Ω [12, Thm. 9.1]. Since Ω is compact, its recession cone is just $\{0\}$ [12, Thm. 8.4] so $N[W'] \cap 0^+\Omega = \{0\}$. Consequently, Γ is closed and is therefore a convex compact set, as claimed. □

Some well-known or easily proved facts are summarized first.

LEMMA B1. (i) *For any* $C \in \mathbb{R}^{m \times n}$ *and any* $A \in S^n$

$$\begin{aligned}&\|C\| \leqq \|C\|_{\mathcal{F}} = \|\text{vec}\,[C]\|, \quad \|\overline{\text{vec}}\,[A]\| = \|\text{vec}\,[A]\| = \|A\|_{\mathcal{F}}, \\ &\|A\|_{\mathcal{F}}^2 = \text{trace}\,[A^2], \quad |\lambda_{\min}(A)| \leqq \|A\|.\end{aligned}$$

(ii) *For any* $A, B \in \mathbb{R}^{n \times n}$

$$\text{trace}\,[AB] = \text{vec}\,[A]'\,\text{vec}\,[B']. \qquad □$$

Proof of Theorem 3.2. (a) Proof that $\Gamma_2 \subset \Gamma$. For $x \in \Gamma_2 \subset \mathbb{R}^r$, let $X(x)$ denote the associated matrix, i.e., let

$$\text{(B1)} \qquad X(x) = \overline{\text{vec}}^{-1}\,[x] \in S^n.$$

Then

$$\begin{aligned}
&[x \in \Gamma_2] \\
&\Rightarrow [x - n^{-1}\iota = M\Xi_2 (\text{by}\,(3.11)] \\
&\Rightarrow [\,\|x - n^{-1}\iota\| \leqq n^{-1} (\text{from } (3.15), \text{ since } M \text{ has orthonormal columns})] \\
\text{(B2)}\quad &\Rightarrow [\,\|X(x) - n^{-1}I_n\|_{\mathcal{F}} \leqq n^{-1} (\text{by Lemma B1 (i), B1, and } (3.12))] \\
&\Rightarrow [X(x) = n^{-1}I_n + \Delta \text{ for some } \Delta \in S^n \text{ with } \|\Delta\|_{\mathcal{F}} \leqq n^{-1}] \\
&\Rightarrow [X(x)' = X(x) \geqq 0\ (\text{since } \lambda_{\min}[\Delta] \geqq -\|\Delta\|_{\mathcal{F}})].
\end{aligned}$$

By Lemma B1 (ii), the fact that $X(x)' = X(x)$, (B1), and by (3.6)

$$\begin{aligned}
\text{(B3)}\quad \operatorname{trace}[X(x)] &= \operatorname{trace}[I_n X(x)'] = \operatorname{vec}[I_n]' \operatorname{vec}[X(x)] = \operatorname{vec}[I_n]' W\, \overline{\operatorname{vec}}\,[X(x)] \\
&= \operatorname{vec}[I_n]' W x.
\end{aligned}$$

Clearly, $\operatorname{vec}[I_n] \in \operatorname{vec}[S^n]$ so, by (3.5),

$$\text{(B4)}\qquad W W' \operatorname{vec}[I_n] = \operatorname{vec}[I_n].$$

So, from (B3)-(B4) and because [by (3.4)] $W'W = I_r$

$$\text{(B5)}\quad \operatorname{trace}[X(x)] = (W W' \operatorname{vec}[I_n])' W x = (W' \operatorname{vec}[I_n])' x = \iota' x = \|\iota\|^2 n^{-1} = 1$$

where the fourth equality is a result of x belonging to Γ_2 of (3.11).

Hence, from (B2) and (B5)

$$\begin{aligned}
&[x \in \Gamma_2] \\
&\Rightarrow [X(x)' = X(x) \geqq 0 \text{ and } \operatorname{trace}[X(x)] = 1] \\
&\Rightarrow [X(x) = B^2 \text{ for some } B \in U \text{ of } (3.2)\ (\text{by Lemma B1 (i)})] \\
&\Rightarrow [\operatorname{vec}[X(x)] \subset \Omega, \text{ for } \Omega \text{ of } (3.7)] \\
&\Rightarrow [x = \overline{\operatorname{vec}}\,[X(x)] = W' \operatorname{vec}[X(x)] \in W'\Omega = \Gamma\ (\text{by (B1), (3.6), and (3.7)})]
\end{aligned}$$

so $\Gamma_2 \subset \Gamma$, as claimed.

(b) Proof that $\Gamma \subset \Gamma_1$. It will be shown first that $\Gamma \subset n^{-1}\iota + R[M]$, for M of (3.13). Now, since B^2 is symmetric when $B \in U$

$$\operatorname{vec}[I_n]' \operatorname{vec}[B^2] = \operatorname{trace}[I_n B^2] = 1 \quad \forall B \in U$$

where the first equality is from Lemma B1 (ii) and the second is from both Lemma B1 (i) and definition (3.2) of U. So, since $\|\operatorname{vec}[I_n]\|^2 = n$,

$$\begin{aligned}
0 &= \operatorname{vec}[I_n]' \{\operatorname{vec}[B^2] - n^{-1} \operatorname{vec}[I_n]\} \\
&= (W' \operatorname{vec}[I_n])' W' (\operatorname{vec}[B^2] - n^{-1} \operatorname{vec}[I_n]) \\
&= \iota' (W' \operatorname{vec}[B^2] - n^{-1}\iota) \quad \forall B \in U
\end{aligned}$$

where the second equality is from (3.5) and the third is from (3.12). Hence, for all $B \in U$, $W' \operatorname{vec}[B^2] - n^{-1}\iota$ is orthogonal to ι and therefore belongs to $R[M]$ for M of Theorem 3.2. Consequently, $W' \operatorname{vec}[B^2] \in n^{-1}\iota + R[M]$, for all $B \in U$. So, as Ω, of (3.7), is the convex hull of $\{\operatorname{vec}[B^2]: B \in U\}$, it follows that $W'\Omega \subset n^{-1}\iota + R[M]$. Since $\Gamma = W'\Omega$, this reveals that

$$\Gamma \subset n^{-1}\iota + R[M], \tag{B6}$$

as claimed earlier.

It will be shown next that actually $\Gamma \subset n^{-1}\iota + M\Xi_1$, for Ξ_1 of (3.14).

Since Ω is a convex subset of R^{n^2}, by the Caratheodory Representation Theorem [12, Thm. 17.1], every $\omega \in \Omega$ can be represented as

$$\omega = \overline{\sum} \alpha_i \operatorname{vec}[B_i^2] \quad \text{where } \overline{\sum} \text{ denotes } \sum_{i=1}^{n^2+1}, \tag{B7}$$

for some $\alpha_i \in \mathbb{R}_{\geqq}$ with $\overline{\sum}\, \alpha_i = 1$ and for some $B_i \in U$.

Put B_i^2 in the spectral form $V\Lambda_i V'$ and let $\lambda = [\lambda_1\lambda_2 \cdots \lambda_n]'$ where the λ_j are the eigenvalues of B_i^2. Then, since $B_i \in U$ of (3.2), trace $[B_i^2] = 1$ so $\|\lambda\|_1 = 1$ and consequently

$$\operatorname{trace}[B_i^4] = \operatorname{trace}[\Lambda_i^2] = \sum \lambda_i^2 = \|\lambda\|_2^2 \leqq \|\lambda\|_1^2 = 1 \tag{B8}$$

where the penultimate inequality occurs since $\|x\|_2 \leqq \|x\|_1$ for all x [13, § 2.1].

Suppose $\gamma \in \Gamma = W'\Omega$. Then $\gamma = W'\omega$ for some $\omega \in \Omega$ which can be represented as in (B7) and

$$\begin{aligned}
\|\gamma - n^{-1}\iota\| &= \|W'\omega - n^{-1}W' \operatorname{vec}[I_n]\| && \text{(from (3.12))}\\
&= \|\omega - n^{-1} \operatorname{vec}[I_n]\| && \text{(from Lemma B1(i))}\\
&= \|\overline{\sum} \alpha_i\{\operatorname{vec}[B_i^2] - n^{-1}\operatorname{vec}[I_n]\}\| && \text{(since } \overline{\sum}\, \alpha_i = 1)\\
&\leqq \overline{\sum} \alpha_i \|\operatorname{vec}[B_i^2] - n^{-1}\operatorname{vec}[I_n]\| && \text{(since } \alpha_i \geqq 0)\\
&= \overline{\sum} \alpha_i \|B_i^2 - n^{-1}I_n\|_{\mathcal{F}} && \text{(by Lemma B1(i))}\\
&= \overline{\sum} \alpha_i \sqrt{\operatorname{trace}\{(B_i^2 - n^{-1}I_n)^2\}} && \text{(by Lemma B1(i))}\\
&= \overline{\sum} \alpha_i \sqrt{\operatorname{trace}\{B_i^4\} - n^{-1}} && \text{(since } B_i \in U)\\
&\leqq \overline{\sum} \alpha_i \sqrt{1 - n^{-1}} && \text{(from (B8))}\\
&= \sqrt{1 - n^{-1}} && \text{(since } \overline{\sum}\, \alpha_i = 1).
\end{aligned} \tag{B9}$$

Now (B6) reveals that $\gamma - n^{-1}\iota = Mx$ for some $x \in \mathbb{R}^{r-1}$ and, from (B9) and the orthonormality of the columns of M,

$$\|x\| = \|Mx\| = \|\gamma - n^{-1}\iota\| \leqq \sqrt{1 - n^{-1}}.$$

Consequently, $\gamma - n^{-1}\iota \in M\Xi_1$ for $\Xi_1 = \{x \in \mathbb{R}^{r-1}\colon \|x\| \leqq \sqrt{1 - n^{-1}}\}$, for all $\gamma \in \Gamma$, i.e.,

$$\Gamma \subset n^{-1}\iota + M\Xi_1,$$

as required.

(c) Proof of (3.10). In view of (3.9), (3.11) and the definitions of ι and M in (3.12) and (3.13)

$$\begin{aligned}
[\gamma \in \Gamma] &\Rightarrow [\gamma = n^{-1}\iota + Mx \text{ for some } x \text{ with } \|x\| \leqq \sqrt{1 - n^{-1}}]\\
&\Rightarrow [\|\gamma\| = \sqrt{\|n^{-1}\iota\|^2 + \|Mx\|^2}, \text{ where } \|x\| \leqq \sqrt{1 - n^{-1}}]\\
&\Rightarrow [n^{-1/2} = \|n^{-1}\iota\| \leqq \|\gamma\| \leqq 1],
\end{aligned}$$

which completes the proof of Theorem 3.2. □

Appendix C. Proofs for § 4.

Proof of Lemma 4.1. The hyperplane with normal $\alpha h - x$ that supports MF is $\{z \in \mathbb{R}^q\colon (\alpha h - x)'z = (\alpha h - x)'\hat{y}_\alpha(x)\}$. Consequently, the point $\tau_\alpha(x)h$ in the line L is also in the hyperplane if and only if

$$\text{(C1)} \qquad \tau_\alpha(x)(\alpha h - x)'h = (\alpha h - x)'\hat{y}_\alpha(x),$$

from which the formula for τ_α of (4.7) follows. Since

$$\hat{y}_\alpha(x) \in \arg\max\{(\alpha h - x)'y\colon y \in MF\}$$

and $\hat{\alpha}h \in MF$, $(\alpha h - x)'\hat{y}_\alpha(x) \geqq (\alpha h - x)'h\hat{\alpha}$. Hence, since it is assumed in Lemma 4.1 that $(\alpha h - x)'h < 0$, it follows from (C1) that

$$\text{(C2)} \qquad \tau_\alpha(x) = \frac{(\alpha h - x)'\hat{y}_\alpha(x)}{(\alpha h - x)'h} \leqq \hat{\alpha} \quad \forall x \in MF,$$

which completes the proof of (4.7).

Since $\hat{x}_\alpha$ minimizes $\|x - \alpha h\|$ with respect to $x \in MF$,

$$\text{(C3)} \qquad (z - \hat{x}_\alpha)'(\hat{x}_\alpha - \alpha h) \geqq 0 \quad \forall z \in MF.$$

So, because $\hat{\alpha}h \in MF$, it follows that $\hat{\alpha}h'(\hat{x}_\alpha - \alpha h) \geqq \hat{x}'_\alpha(\hat{x}_\alpha - \alpha h)$, i.e.,

$$\text{(C4)} \qquad h'(\hat{x}_\alpha - \alpha h) \geqq \|\hat{x}_\alpha - \alpha h\|^2/(\hat{\alpha} - \alpha).$$

Now

$$\text{(C5)} \qquad \hat{x}_\alpha - \alpha h \neq 0$$

since $\hat{x}_\alpha \in MF$ and since $\alpha h \notin MF$ (because it is assumed in Lemma 4.1 that $\alpha < \hat{\alpha}$ and because $\hat{\alpha} = \min\{\alpha \in \mathbb{R}\colon \alpha h \in MF\}$). Subsequently, from (C4) and the assumption that $\alpha < \hat{\alpha}$,

$$\text{(C6)} \qquad h'(\hat{x}_\alpha - \alpha h) > 0,$$

which establishes the first result of (4.8).

From (C2), $\tau_\alpha(\hat{x}_\alpha) \leqq \hat{\alpha}$. In view of (C3), $(\hat{y}_\alpha(\hat{x}_\alpha) - \hat{x}_\alpha)'(\hat{x}_\alpha - \alpha h) \geqq 0$. A consequence of $\hat{y}_\alpha(\hat{x}_\alpha) \in \arg\max\{(\alpha h - \hat{x}_\alpha)'y\colon y \in MF\}$ is that

$$(\alpha h - \hat{x}_\alpha)'\hat{y}_\alpha(\hat{x}_\alpha) \geqq (\alpha h - \hat{x}_\alpha)'\hat{x}_\alpha.$$

Hence $(\hat{y}_\alpha(\hat{x}_\alpha) - \hat{x}_\alpha)'(\hat{x}_\alpha - \alpha h) = 0$. Therefore, from (4.7), (C5), and (C6)

$$\tau_\alpha(\hat{x}_\alpha) = \alpha + \frac{\|\hat{x}_\alpha - \alpha h\|^2}{h'(\hat{x}_\alpha - \alpha h)} > \alpha,$$

which completes the proof of Lemma 4.1. □

Proof of Theorem 4.1. It will be shown first that

$$\text{(C7)} \qquad \alpha_0 < \alpha_1 < \cdots < \alpha_i < \alpha_{i+1} < \cdots < \hat{\alpha}.$$

Since $\hat{\alpha}h \in MF$, it follows from the definition of y_0 in (4.10) that $h'y_0 \leqq \hat{\alpha}h'h$. Consequently, and since ε_3 of (4.9) is in (0, 1), it follows from (4.11) that

$$\text{(C8)} \qquad \alpha_0 < \hat{\alpha}$$

(which, incidentally, verifies the assertion made about α_0 in the statement of Algorithm 4.1).

Suppose $\alpha_i < \hat{\alpha}$. A consequence of (4.7) in Lemma 4.1 and of formula (4.15) for α_{i+1} is that

$$\text{(C9)} \qquad \alpha_{i+1} - \alpha_i = (1 - \varepsilon_3)[\tau_{\alpha_i}(x_i) - \alpha_i].$$

Since $x_i \in \varepsilon_2$-minpoint $[\alpha_i h, MF]$ and $(1 - \varepsilon_2) \in (0, 1)$, it follows from Definition 4.1(iii), (4.8), and from (4.7) that

$$\tau_{\alpha_i}(x_i) - \alpha_i \geqq (1 - \varepsilon_2)[\tau_{\alpha_i}(\hat{x}_{\alpha_i}) - \alpha_i] > 0, \qquad \tau_{\alpha_i}(x_i) \leqq \hat{\alpha}.$$

So, from (C9), Lemma 4.1 and the fact that $1 - \varepsilon_3 \in (0, 1)$,

$$\text{(C10)} \qquad \alpha_{i+1} - \alpha_i \geqq (1 - \varepsilon_2)(1 - \varepsilon_3)[\tau_{\alpha_i}(\hat{x}_{\alpha_i}) - \alpha_i] > 0,$$

and, from (C9) and (4.7) of Lemma 4.1,

$$\text{(C11)} \qquad \alpha_{i+1} = \alpha_i + (1 - \varepsilon_3)[\tau_{\alpha_i}(x_i) - \alpha_i] < \tau_{\alpha_i}(x_i) \leqq \hat{\alpha}.$$

Then (C7) follows from (C8), (C10), and (C11).

Convergence rate result (4.23) will be established next. From (4.8)

$$\text{(C12)} \qquad \tau_{\alpha_i}(\hat{x}_{\alpha_i}) - \alpha_i = \frac{\|\hat{x}_{\alpha_i} - \alpha_i h\| \; \|h\|}{h'(\hat{x}_{\alpha_i} - \alpha_i h)} \frac{\|\hat{x}_{\alpha_i} - \alpha_i h\|}{\|h\|} \geqq \frac{\|\hat{x}_{\alpha_i} - \alpha_i h\|}{\|h\|}.$$

In view of assumption (4.4), MF is supported at $\hat{\alpha}h$ by a hyperplane H with normalized norm η satisfying $\eta' h < 0$. Hence

$$\|\hat{x}_{\alpha_i} - \alpha_i h\| = \text{mindist}\,[\alpha_i h, MF] \geqq \text{mindist}\,[\alpha_i h, H] = [\alpha_i - \hat{\alpha}]\eta' h,$$

so, from (C12),

$$\tau_{\alpha_i}(\hat{x}_{\alpha_i}) - \alpha_i \geqq [\alpha_i - \hat{\alpha}]\eta' h \|h\|^{-1}.$$

Consequently, from (C10),

$$\alpha_{i+1} - \alpha_i \geqq (1 - \varepsilon_2)(1 - \varepsilon_3)\eta' h \|h\|^{-1}[\alpha_i - \hat{\alpha}].$$

Subtracting each side from $\hat{\alpha}$ we obtain

$$(\hat{\alpha} - \alpha_{i+1}) \leqq [1 + (1 - \varepsilon_2)(1 - \varepsilon_3)\eta' h \|h\|^{-1}](\hat{\alpha} - \alpha_i).$$

Therefore

$$\text{(C13)} \qquad (\hat{\alpha} - \alpha_i) \leqq [1 + (1 - \varepsilon_2)(1 - \varepsilon_3)\eta' h \|h\|^{-1}]^i(\hat{\alpha} - \alpha_0),$$

which establishes (4.23).

Since $\eta' h < 0$ (from assumption (4.4)), (C13) shows that if stopping condition (4.21) in Algorithm 4.1 were omitted, so that the algorithm iterated indefinitely, then

$$\text{(C14)} \qquad \alpha_i \to \hat{\alpha}.$$

Taking account of (C7), $\alpha_i \uparrow \hat{\alpha}$ with $\alpha_i < \alpha_{i+1} < \hat{\alpha}$ for all $i \geqq 0$, which has proved the first part of (4.22).

To start proving that $\beta_i \downarrow \hat{\alpha}$, as required to complete the proof of (4.22), it will be shown that $\tilde{\beta}_{i+1}$ of step III is an upper-bound for $\hat{\alpha}$.

Assumption (4.6) states that $\Xi \subset MF$ where $\Xi = \zeta h + \{\theta \in \mathbb{R}^q\colon \|\theta\| \leqq k\}$. Let

$$K_i = \{x_i + a(\xi - x_i)\colon a \in [0, 1], \xi \in \Xi\}.$$

Then $K_i \subset MF$ since $x_i \in MF$ (from (4.13)), $\Xi \subset MF$ and MF is convex.

Consider

$$\text{(C15)} \qquad \tilde{\beta}_{i+1} = \min\{\alpha \in \mathbb{R}\colon \alpha h \in K_i\}.$$

Then, for $i \geqq 0$

$$\hat{\alpha} = \min\{\alpha \in \mathbb{R} : \alpha h \in MF\} \leqq \min\{\alpha \in R : \alpha h \in K_i\}$$
$$= \tilde{\beta}_{i+1} \leqq \min\{\alpha \in \mathbb{R} : \alpha h \in \Xi\} = \zeta - \frac{k}{\|h\|},$$

since $\Xi \subset K_i \subset MF$. Hence, from (4.12) and (C15),

$$\text{(C16)} \qquad \beta_0 = \zeta - \frac{k}{\|h\|} \geqq \tilde{\beta}_{i+1} \geqq \hat{\alpha} \quad \forall i \geqq 0.$$

The value of $\tilde{\beta}_{i+1}$ computed by step III of Algorithm 4.1 will next be shown to be that of (C15) and therefore, in view of (C16), to be an upper bound for $\hat{\alpha}$.

Consider first the case when $x_i \in \Xi$, i.e., when $\|\zeta h - x_i\| \leqq k$. Then, since Ξ is convex, $K_i = \Xi$ so $\tilde{\beta}_{i+1}$ of (C15) is given by

$$\tilde{\beta}_{i+1} = \min\{\alpha \in \mathbb{R} : \alpha h \in \Xi\} = \zeta - \frac{k}{\|h\|}.$$

Therefore the value of $\tilde{\beta}_{i+1}$ computed by step III (actually by (4.16)) of Algorithm 4.1 is that of (C15) when $\|\zeta h - x_i\| \leqq k$.

Now suppose $\|\zeta h - x_i\| > k$. Consider the situation represented in the linear subspace spanned by x_i and h, as shown in Fig. 4.1. It can easily be checked that the formulae of (4.17) specify θ_i and ψ_i. One case is that when the angle θ_i, between h and $-(x_i - \zeta h)$, and the angle ψ_i are related by $\cos(\theta_i) \leqq \cos(\psi_i)$. By considering Fig. 4.1 for that case it can be seen that then the value of $\tilde{\beta}_{i+1}$ obtained from (C15) is again $\tilde{\beta}_{i+1} = \zeta - k/\|h\|$ so the formula for $\tilde{\beta}_{i+1}$ given in (4.18) of Algorithm 4.1 is correct for the case when $\cos(\theta_i) \leqq \cos(\psi_i)$. The other case possible is when $\cos(\theta_i) > \cos(\psi_i)$. Then the situation is exactly that shown in Fig. 4.1. From Fig. 4.1, it is clear that (C15) gives $\tilde{\beta}_{i+1} = \zeta - k/(\cos(\psi_i - \theta_i)\|h\|)$, which verifies (4.19) for that case.

Hence, in every case, the value of $\tilde{\beta}_{i+1}$ computed by step III is that of (C15) and is therefore an upper bound for $\hat{\alpha}$.

Some more analysis is necessary before it can be shown that $\beta_i \downarrow \hat{\alpha}$.

Recall that $L = \{\alpha h : \alpha \in \mathbb{R}]$. Clearly,

$$\text{(C17)} \qquad \begin{aligned} 0 \leqq \operatorname{mindist}[x_i, L] &= \|x_i - \zeta h\|^2 - \frac{[(x_i - \zeta h)'h]^2}{\|h\|^2} \\ &\leqq \|x_i - \alpha_i h\|^2 \leqq (1+\varepsilon_2)^2 \|\hat{x}_{\alpha_i} - \alpha_i h\|^2 \leqq [(1+\varepsilon_2)(\hat{\alpha} - \alpha_i)\|h\|]^2 \end{aligned}$$

where the second inequality occurs because $\alpha_i h$ is in L but is not necessarily the nearest point to x_i in L, the third is a consequence of the fact that $x_i \in \varepsilon_2$-minpoint $[\alpha_i h, MF]$ and Definition 4.1, and where the fourth inequality arises because $\hat{\alpha} h$ belongs to MF and $\hat{x}_{\alpha_i}$ minimizes $\|x - \alpha_i h\|$ with respect to $x \in MF$.

Making use of (C17) in the second inequality below, we have

$$\text{(C18)} \qquad \begin{aligned} \|x_i - \zeta h\|^2 &= \|(x_i - \alpha_i h) - (\zeta h - \alpha_i h)\|^2 \\ &\geqq -2\|x_i - \alpha_i h\|(\zeta - \alpha_i)\|h\| + (\zeta - \alpha_i)^2\|h\|^2 \\ &\geqq -2(1+\varepsilon_2)(\hat{\alpha} - \alpha_i)(\zeta - \alpha_i)\|h\| + (\zeta - \alpha_i)^2\|h\|^2. \end{aligned}$$

Since (from (C14)) $\alpha_i \to \hat{\alpha}$, the last expression on the right of (C18) converges to $(\zeta - \hat{\alpha})^2\|h\|^2 > 0$ as $i \to \infty$. Hence there is an integer i_1 such that

$$\text{(C19)} \qquad \|x_i - \zeta h\|^2 \geqq \tfrac{1}{2}(\zeta - \hat{\alpha})^2\|h\|^2 > 0 \quad \forall i \geqq i_1$$

where the strict equality is from (C16). Consequently, by dividing (C17) through by $\|x_i - \zeta h\|^2$ and by making use of (C19) and (C14)

$$0 \leqq 1 - \frac{[(x_i - \zeta h)'h]^2}{\|x_i - \zeta h\|^2 \|h\|^2} \leqq \frac{[(1+\varepsilon_2)(\hat{\alpha} - \alpha_i)\|h\|]^2}{\|x_i - \zeta h\|^2} \to 0.$$

So

$$\frac{|(x_i - \zeta h)'h|}{\|x_i - \zeta h\| \, \|h\|} \to 1.$$

Hence, for θ_i of (4.17),

(C20) $$\theta_i \to \theta.$$

Now consider

(C21) $$\tilde{\beta}_{i+1} = \zeta - \frac{\|x_i - \zeta h\|}{\|h\|}.$$

Clearly,

$$\|x_i - \hat{\alpha} h\| \leqq \|x_i - \alpha_i h\| + \|\alpha_i h - \hat{\alpha} h\|$$

where, from (C14), $\alpha_i \to \hat{\alpha}$ and consequently, from (C17), $\|x_i - \alpha_i h\| \to 0$. Hence

(C22) $$\|x_i - \hat{\alpha} h\| \to 0.$$

Furthermore,

$$\|\hat{\alpha} h - \zeta h\| - \|x_i - \hat{\alpha} h\| \leqq \|x_i - \zeta h\| \leqq \|\hat{\alpha} h - \zeta h\| + \|x_i - \hat{\alpha} h\|.$$

So, from (C22) and since, from (C16), $\zeta - \hat{\alpha} > 0$,

(C23) $$\|x_i - \zeta h\| \to \|\hat{\alpha} h - \zeta h\| = (\zeta - \hat{\alpha})\|h\|.$$

Therefore, from (C21) and (C23)

(C24) $$\tilde{\beta}_i \to \hat{\alpha}.$$

In view of (4.17) and (C21), $\tilde{\beta}_{i+1} = \zeta - k/(\cos(\psi_i)\|h\|)$ whenever $\|x_i - \zeta h\| > k$. So, since (by (C20)), $\theta_i \to 0$,

(C25) if there exists an integer i_2 such that $\|x_i - \zeta h\| > k$ for all $i > i_2$, then $\zeta - k/(\cos(\psi_i - \theta_i)\|h\|) \to \hat{\alpha}$.

It will be shown next that $\beta_i \downarrow \hat{\alpha}$, by making use of (C25).

From (4.20) and (C16), $\beta_i = \min\{\beta_0, \tilde{\beta}_1, \cdots, \tilde{\beta}_i\} \geqq \hat{\alpha}$, for all $i \geqq 0$. Therefore

(C26) $\beta_i \downarrow \hat{\alpha}$ if $\tilde{\beta}_p = \hat{\alpha}$ for some p or if $\tilde{\beta}_i \to \hat{\alpha}$.

In view of (C16), there are two possible cases: (a) $\zeta - \hat{\alpha} = k/\|h\|$; (b) $\zeta - \hat{\alpha} > k/\|h\|$.

Consider case (a) first. Suppose there is a finite j so $\|x_j - \zeta h\| \leqq k$ or so both $\|x_j - \zeta h\| > k$ and $\cos(\theta_j) \leqq \cos(\psi_j)$. Then either (4.16) or (4.18) sets $\tilde{\beta}_{j+1} = \zeta - k/\|h\|$ so, since $\zeta - \hat{\alpha} = k/\|h\|$ in case (a), $\tilde{\beta}_{j+1} = \hat{\alpha}$. Therefore, from (C26), $\beta_i \downarrow \hat{\alpha}$. Suppose next that there is no finite j so $\|x_j - \zeta h\| \leqq k$ or so both $\|x_j - \zeta h\| > k$ and $\cos(\theta_j) \leqq \cos(\psi_j)$. Then, for all $i \geqq 0$, $\|x_i - \zeta h\| > k$ and (4.19) sets $\tilde{\beta}_{i+1} = \zeta - k/(\cos(\psi_i - \theta_i)\|h\|)$. Therefore, from (C25), $\tilde{\beta}_i \to \hat{\alpha}$ and, from (C26), $\beta_i \downarrow \hat{\alpha}$. Hence for case (a): $\beta_i \downarrow \hat{\alpha}$.

Now consider case (b). Then $\|\zeta h - \hat{\alpha} h\| = k + \delta$ for some $\delta > 0$ and

$$\|x_i - \zeta h\| \geqq \|\zeta h - \hat{\alpha} h\| - \|x_i - \hat{\alpha} h\| \geqq k + \delta - \|x_i - \hat{\alpha} h\|.$$

Hence, since (C22) reveals that $\|x_i - \hat{\alpha} h\| \to 0$, there is a finite integer i_3 such that

(C27) $$\|x_i - \zeta h\| > k + \tfrac{1}{2}\delta \quad \forall i > i_3.$$

Hence step III of Algorithm 4.1 executes (4.17) for all $i > i_3$ and, from (4.17), $\cos(\psi_i) < k/(k + \frac{1}{2}\delta)$, for all $i \geqq i_3$. Consequently, since (by (C20)), $\theta_i \to 0$, there is an integer $i_4 > i_3$ such that $\cos(\theta_i) > \cos(\psi_i)$ for all $i > i_4$. Therefore (4.19) sets $\tilde{\beta}_{i+1} = \zeta - k/(\cos(\psi_i - \theta_i)\|h\|)$, for all $i > i_4$. Subsequently, from (C25) and (C27), $\tilde{\beta}_i \to \hat{\alpha}$. Therefore, by (C26), $\beta_i \downarrow \hat{\alpha}$ in case (b).

This has shown that $\beta_i \geqq \hat{\alpha}$ and $\beta_i \downarrow \hat{\alpha}$ whether case (a) or case (b) occurs, which completes the proof of (4.22).

Since (4.22) reveals that $\alpha_i \uparrow \hat{\alpha}$ and $\beta_i \downarrow \hat{\alpha}$, it follows that $\hat{\alpha} \in (\alpha_{i+1}, \beta_{i+1})$ and $\beta_{i+1} - \alpha_{i+1} \to 0$. Hence stopping condition (4.21) will be satisfied for some finite i and, from (4.21), for that i, $|\hat{\alpha} - \alpha_{i+1}| \leqq \varepsilon_{11}|\hat{\alpha}| + \varepsilon_{12}$. Also, from (4.22), $\alpha_{i+1} < \hat{\alpha}$. So, for $\bar{\alpha}$ and $\bar{f}$ of (4.21), $\bar{\alpha} < \hat{\alpha}$, $|\hat{\alpha} - \bar{\alpha}| \leqq \varepsilon_{11}|\hat{\alpha}| + \varepsilon_{12}$ and $\bar{f} \in \varepsilon_2$-minpoints$_F$ $[\bar{\alpha} h, MF]$, which establishes post-conditions (4.24)–(4.25) and completes the proof of Theorem 4.1. □

Proof of Theorem 4.2. Consider Algorithm 4.2 with stopping condition (4.32) omitted. Some consequences of the convergence, assumed in the theorem, of x_i to $\hat{x}_\alpha$ = minpoint $[\alpha h, MF]$ will be established first.

Since f_0, $t_0 \in F$ and x_0, $y_0 \in MF$, and since F and MF are convex, it follows from (4.36) that $f_1 \in F$ and $x_1 \in MF$. In the same way it can be seen that

(C28) $f_i \in F$ and $x_i \in MF$, with $x_i = Mf_i$, $\forall i \geqq 0$.

Since x_i, $\hat{x}_\alpha \in MF$ and $\alpha h \notin MF$,

(C29) $$\|x_i - \alpha h\| \geqq \|\hat{x}_\alpha - \alpha h\| > 0 \quad \forall i \geqq 0.$$

Since it follows from (4.28) that $y_i \in \arg\min\{(x_i - \alpha h)'y \colon y \in MF\}$ and since $x_i \in MF$,

(C30) $$(x_i - \alpha h)'(y_i - x_i) \leqq 0 \quad \forall i \geqq 0.$$

It will be shown next that, since $x_i \to \hat{x}_\alpha$,

(C31) $$(x_i - \alpha h)'(y_i - x_i) \to 0.$$

For otherwise there is a finite integer i_1 and a $\delta \in \mathbb{R}_>$ such that

(C32) $$|(x_i - \alpha h)'(y_i - x_i)| > \delta \quad \forall i > i_1,$$

which yields a contradiction, as follows.

Suppose $\hat{w}$ minimizes $\|\{x_i + w[y_i - x_i]\} - \alpha h\|$ with respect to $w \in [0, 1]$ and $\tilde{w}$ minimizes it on $\mathbb{R}$. In view of (C30), $\tilde{w} \geqq 0$, so $\hat{w} = \min\{1, \tilde{w}\}$ and it is easy to check that

$$\|\{x_i + \hat{w}[y_i - x_i]\} - \alpha h\|^2 \leqq \|x_i - \alpha h\|^2 - \bar{\delta} \quad \forall i \geqq i_1,$$

where $\bar{\delta} = \min\{\delta, \delta^2/D^2\}$ and D is the diameter of MF (the point about D being that $\|y_i - x_i\| \leqq D$). Therefore, since $x_i + \hat{w}[y_i - x_i] \in MF$ (because MF is convex, $\hat{w} \in [0, 1]$ and x_i, $y_i \in MF$), $\|\hat{x}_\alpha - \alpha h\|^2 \leqq \|\{x_i + \hat{w}[y_i - x_i]\} - \alpha h\|^2 \leqq \|x_i - \alpha h\|^2 - \bar{\delta}$, for all $i > i_1$, which contradicts the fact that $x_i \to \hat{x}_\alpha$. Hence (C32) is false, which verifies (C31).

From (4.29)

$$x_i - z_i = -\frac{(y_i - x_i)'(x_i - \alpha h)}{\|x_i - \alpha h\|^2}(x_i - \alpha h),$$

so, in view of (C29), (C31) and the assumption that $x_i \rightarrow \hat{x}_\alpha$

(C33) $$z_i - x_i \rightarrow 0, \qquad z_i \rightarrow \hat{x}_\alpha.$$

Also, from (4.29) and (C30)-(C31) there is a finite integer i_2 such that

(C34) $$\|z_i - \alpha h\| = \frac{|(x_i - \alpha h + [y_i - x_i])'(x_i - \alpha h)|}{\|x_i - \alpha h\|} \leqq \|x_i - \alpha h\| \quad \forall i \geqq i_2$$

and

(C35) $$\frac{\|z_i - \alpha h\|}{\|x_i - \alpha h\|} \rightarrow 1.$$

Furthermore, there is a finite integer i_3 such that (from (C29) and (C34), (C35))

(C36) $$0 < \|z_i - \alpha h\| \leqq \|x_i - \alpha h\| \quad \forall i > i_3$$

(from (C31) and the fact that $(y_i - \alpha h)'(x_i - \alpha h) = (y_i - x_i)'(x_i - \alpha h) + \|x_i - \alpha h\|^2$)

(C37) $$(y_i - \alpha h)'(x_i - \alpha h) > 0 \quad \forall i > i_3$$

(since, from (4.8) of Lemma 4.1, $(\alpha h - \hat{x}_\alpha)'h < 0$ and since $x_i \rightarrow \hat{x}_\alpha$)

(C38) $$(\alpha h - x_i)'h < 0 \quad \forall i > i_3$$

and (from (C33) and (C38))

(C39) $$(\alpha h - z_i)'h < 0 \quad \forall i > i_3.$$

Assume now that stopping condition (4.32) has been satisfied, so that, from (4.30), $\|z_i - \alpha h\| \leqq \|x_i - \alpha h\|$ (since otherwise (4.32) could not have been reached) and conditions (4.33)-(4.35) have been satisfied. It will be shown that then $f_i \in \varepsilon_2$-minpoints$_F$ $[\alpha h, MF]$ and $x_i \in \varepsilon_2$-minpoint $[\alpha h, MF]$.

Clearly when conditions (4.33)-(4.35) are satisfied, it follows from (C29) and the fact that $\varepsilon_2 \in (0, 1)$ that the right-hand side of (4.35) is strictly positive. Therefore, in view of (4.34), $(y_i - \alpha h)'(x_i - \alpha h) > 0$, so, from (4.29) and the fact, from (C34), that $\|z_i - \alpha h\| \leqq \|x_i - \alpha h\|$,

(C40) $$z_i - \alpha h = \Psi_i(x_i - \alpha h) \quad \text{where } \Psi_i \in (0, 1].$$

It is easy to check that z_i belongs to the hyperplane H_i with normal $(\alpha h - x_i)$ that supports MF and has contact point y_i. Hence, by (C40), H_i also has normal $\alpha h - z_i$. Therefore

(C41) $$\|z_i - \alpha h\|^2 \leqq (z_i - \alpha h)'(x - \alpha h) \quad \forall x \in MF, \quad (z_i - \alpha h)'h > 0$$

where the second part follows from the first part of (4.34).

Since $\hat{x}_\alpha \in MF$, it follows from (C41) that $\|z_i - \alpha h\|^2 \leqq \|z_i - \alpha h\| \, \|\hat{x}_\alpha - \alpha h\|$, so

$$\|z_i - \alpha h\| \leqq \|\hat{x}_\alpha - \alpha h\| \leqq \|x_i - \alpha h\|.$$

Consequently, $\|x_i - \alpha h\| \leqq (1 + \varepsilon_2)$ mindist $[\alpha h, MF]$ if

$$\|x_i - \alpha h\| \leqq (1 + \varepsilon_2) \|z_i - \alpha h\|.$$

Hence, since (4.32) sets $\bar{f} = f_i \in F$ and $\bar{x} = x_i \in MF$ for the final i, $\bar{x}$ satisfies conditions (i) and (ii) of Definition 4.2 for a point in ε_2-minpoint $[\alpha h, MF]$ if stopping conditions (4.33)-(4.35) are satisfied. Next we will show that then $\bar{x}$ also satisfies condition (iii) of Definition 4.1.

In view of (C41) and the fact that $\|\hat{x}_\alpha - \alpha h\| \leqq \|x_i - \alpha h\|$,

$$\text{(C42)} \qquad \hat{x}_\alpha \in X_i = \{x \in \mathbb{R}^q \colon \|z_i - \alpha h\|^2 \leqq (z_i - \alpha h)'(x - \alpha h) \text{ and } \|x - \alpha h\|^2 \leqq \|x_i - \alpha h\|^2\}.$$

Since $\hat{x}_\alpha \in X_i$, it follows from (4.8) of Lemma 4.1 that

$$\text{(C43)} \qquad \tau_\alpha(\hat{x}_\alpha) \leqq \sup_{x \in X_i \cap F_i} \alpha + \frac{\|x - \alpha h\|^2}{h'(x - \alpha h)}$$

where

$$\text{(C44)} \qquad F_i = \{x \in \mathbb{R}^q \colon (\alpha h - x)'h < 0\}.$$

From (C29), $\|x_i - \alpha h\| \neq 0$. By (4.33), $\|x_i - \alpha h\| \leqq (1 + \varepsilon_2)\|z_i - \alpha h\|$ when stopping conditions (4.33)-(4.35) are satisfied. So, when those conditions are satisfied, $z_i - \alpha h \neq 0$ and $\mathbb{R}^q$ can be written as $\mathbb{R}^q = R[\langle\!\langle z_i - \alpha h \rangle\!\rangle] \oplus {}^{\perp}R[\langle\!\langle z_i - \alpha h \rangle\!\rangle]$. Consequently,

$$\text{(C45)} \qquad x - \alpha h = \bar{x}(x)\langle\!\langle z_i - \alpha h \rangle\!\rangle + \tilde{x}(x), \quad \forall x \in X_i,$$

$$\text{(C46)} \qquad h = \delta_i \langle\!\langle z_i - \alpha h \rangle\!\rangle + \tilde{h}_i$$

where $\bar{x}(x)$, $\delta_i \in \mathbb{R}$ and $\tilde{x}(x)$, $\tilde{h}_i$ are orthogonal to $z_i - \alpha h$.

Then X_i of (C42) can be written as

$$\text{(C47)} \qquad X_i = \{x \in \mathbb{R}^q \colon \bar{x}(x) \geqq \|z_i - \alpha h\| \text{ and } \bar{x}(x)^2 + \|\tilde{x}(x)\|^2 \leqq \|x_i - \alpha h\|^2\}.$$

In view of (C46) and the second part of (C41)

$$\text{(C48)} \qquad \delta_i = h'\langle\!\langle z_i - \alpha h \rangle\!\rangle > 0.$$

So, from (C47), $x \in X_i \Leftrightarrow (\bar{x}(x) \, \|\tilde{x}(x)\|)' \in W_i$ for the set W_i shown in Fig. 4.2.

From (C45)-(C46)

$$\text{(C49)} \qquad \begin{aligned} (x - \alpha h)'h &= \bar{x}(x)\delta_i + \tilde{x}(x)'\tilde{h}_i \geqq \bar{x}(x)\delta_i - \|\tilde{x}(x)\| \, \|\tilde{h}_i\| \\ &= (\bar{x}(x) \, \|\tilde{x}(x)\|)(\delta_i - \|\tilde{h}_i\|)' = \|x_i - \alpha h\| \, \|h\| \cos(\nu_i + \chi_i) \quad \forall x \in X_i \end{aligned}$$

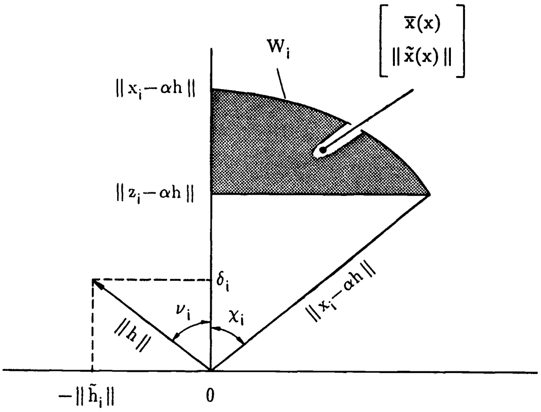


FIG. 4.2

for the angles ν_i and χ_i of Fig. 4.2. From (4.34), $\cos(\nu_i + \chi_i) > 0$ when stopping conditions (4.33)-(4.35) are satisfied. Hence, from (C49), then $(x - \alpha h)'h > 0$, for all $x \in X_i$, so (from (C44)) $X_i \cap F_i = X_i$. Then, from (C43),

$$\tau_\alpha(\hat{x}_\alpha) \leqq \sup_{x \in X_i} \alpha + \frac{\|x-\alpha h\|^2}{h'(x-\alpha h)} = \sup_{x \in X_i} \alpha + \frac{\|x-\alpha h\|}{\|h\|} \frac{\|x-\alpha h\| \; \|h\|}{h'(x-\alpha h)}$$

$$\leqq \alpha + \frac{\|x_i - \alpha h\|}{\|h\| \cos(\nu_i + \chi_i)} = \alpha + \frac{(\alpha h - x_i)'(y_i - \alpha h)}{(1-\varepsilon_2)(\alpha h - x_i)'h} = \alpha + \frac{[\tau_\alpha(x_i) - \alpha]}{(1-\varepsilon_2)}$$

where the second inequality is from (C42), (C49), the penultimate equality is from (4.35) and the final equality is from (4.7) and (4.34). Therefore

$$\tau_\alpha(\hat{x}_\alpha) - \tau_\alpha(x_i) \leqq \varepsilon_2[\tau_\alpha(\hat{x}_\alpha) - \alpha].$$

Furthermore, since, from (4.34), $(\alpha h - x_i)'h < 0$ it follows from Lemma 4.1 that $\tau_\alpha(x_i) \leqq \hat{\alpha}$. Consequently, x_i satisfies condition (iii) of Definition 4.1. Since it has been shown already that x_i satisfies conditions (i) and (ii) of that definition, it follows that $x_i \in \varepsilon_2$-minpoint $[\alpha h, MF]$ if stopping conditions (4.33)-(4.35) are satisfied. Consequently, if the algorithm terminates, $\bar{x} = x_i \in \varepsilon_2$-minpoint $[\alpha h, MF]$ and, from (C28), $\bar{f} = f_i \in \varepsilon_2$-minpoints$_F$ $[\alpha h, MF]$. Therefore the claim of Theorem 4.2 has been verified if termination occurs.

All that remains to be shown is that termination does actually occur.

Since $\alpha h \notin MF$, $\hat{x}_\alpha$ minimizes $\|x - \alpha h\|$ on MF and $\Xi \subset MF$,

$$\|\hat{x}_\alpha - \alpha h\| \leqq \min_{x \in \Xi} \|x - \alpha h\| = \left(1 - \frac{k}{\|\zeta h - \alpha h\|}\right) \|\zeta h - \alpha h\| < \|\zeta h - \alpha h\|.$$

Therefore, because it is being assumed that $\|x_i - \alpha h\| \downarrow \|\hat{x}_\alpha - \alpha h\|$, the condition $\|x_i - \alpha h\| \leqq \|\zeta h - \alpha h\|$ will be satisfied for all i sufficiently large.

Furthermore, (C33) reveals that $z_i - x_i \to 0$ so $\|x_i - \alpha h\| \leqq (1 + \varepsilon_2)\|z_i - \alpha h\|$ will be satisfied for some finite i. Therefore condition (4.33) will be satisfied for some finite i.

In view of (C38), $(\alpha h - x_i)'h < 0$ for all $i > i_3$. From (4.31) and (C35), $\chi_i \to 0$. Let $\cos(\nu) = h'(\hat{x}_\alpha - \alpha h)/(\|h\| \; \|\hat{x}_\alpha - \alpha h\|)$. From (4.8) and (C29), $\cos(\nu) > 0$. Then $h'(z_i - \alpha h)/(\|h\| \; \|z_i - \alpha h\|) \to \cos(\nu)$ since, from (C33), $z_i \to \hat{x}_\alpha$. Hence, from (4.31), $\cos(\nu_i) \to \cos(\nu)$. So, since $\chi_i \to 0$,

$$\text{(C50)} \qquad \cos(\nu_i + \chi_i) \to \cos(\nu) = \frac{h'(\hat{x}_\alpha - \alpha h)}{\|h\| \; \|\hat{x}_\alpha - \alpha h\|} > 0.$$

Therefore condition (4.34) will be satisfied for some finite i.

Finally, consider condition (4.35). By (C31), (C38) and the assumption that $x_i \to \hat{x}_\alpha$,

$$\frac{(\alpha h - x_i)'(y_i - \alpha h)}{(\alpha h - x_i)'h} = \frac{\|x_i - \alpha h\|^2 + (x_i - ah)'(y_i - x_i)}{(x_i - \alpha h)'h} \to \frac{\|\hat{x}_\alpha - \alpha h\|^2}{(\hat{x}_\alpha - \alpha h)'h},$$

so, in view of (C50),

$$\|h\| \left[\frac{(\alpha h - x_i)'(y_i - \alpha h)}{(\alpha h - x_i)'h}\right] \cos(\nu_i + \chi_i) \to \|\hat{x}_\alpha - \alpha h\|.$$

Thus the right-hand side of (4.35) converges to $\|\hat{x}_\alpha - \alpha h\|$. Clearly the left-hand side of (4.35) converges to $(1 - \varepsilon_2)\|\hat{x}_\alpha - \alpha h\|$. Since $(1 - \varepsilon_2) \in (0, 1)$, for some finite i the left-hand side is less than or equal to the right-hand side, so that condition (4.35) is satisfied.

Therefore stopping conditions (4.33)–(4.35) will all be satisfied after some finite number of iterations, which proves termination of the algorithm and completes the proof of Theorem 4.2. □

Appendix D. Proofs for § 5. A fact that is used often in the following is the result that

$$\text{(D1)} \qquad \overline{\text{vec}}\left[\sum_{i=1}^{l} \mu_i H_i - \lambda I_n\right] = H\mu - \lambda\iota,$$

which is a consequence of (5.1)–(5.2).

Proof of Theorem 5.1. Let Σ denote $\Sigma_{i=1}^{l}$ and $\overline{\Sigma}$ denote $\Sigma_{i=1}^{\bar{l}}$. Since the H_i are taken to be linearly dependent, a $\mu^{\#} \in \mathbb{B}^l$ can be determined such that $\Sigma\, (\mu^{\#})_i H_i = 0$. Consequently, $\lambda(\mu^{\#}) = 0$, so

$$\text{(D2)} \qquad \hat{\lambda} = \max\,\{\lambda(\mu) : \mu \in \mathbb{B}^l\} \geqq 0.$$

(i) In this case, $\hat{\bar{\lambda}} = \bar{\lambda}(\hat{\bar{\mu}}) > 0$. Since the span of the H_i is that of the $\bar{H}_i$, there is a $\overset{*}{\mu} \in R^l$ so that $\Sigma\, \overset{*}{\mu}_i H_i = \overline{\Sigma}\, \hat{\bar{\mu}}_i \bar{H}_i$. Therefore, $\lambda(\overset{*}{\mu}) = \lambda_{\min}\,(\Sigma\, \overset{*}{\mu}_i H_i) = \lambda_{\min}\,(\overline{\Sigma}\, \hat{\bar{\mu}}_i \bar{H}_i) = \bar{\lambda}(\hat{\bar{\mu}}) = \hat{\bar{\lambda}} > 0$. Hence there is a $\overset{*}{\mu} \in \mathbb{R}^l$ such that $\lambda(\overset{*}{\mu}) > 0$. Consequently, $\hat{\lambda} > 0$ when $\hat{\bar{\lambda}} > 0$.

(ii) Here $\hat{\bar{\lambda}} = 0$. Since $\hat{\bar{\mu}} \neq 0$ and the $\bar{H}_i$ are linearly independent, $\overline{\Sigma}\, \hat{\bar{\mu}}_i \bar{H}_i \neq 0$ and there is a nonzero μ, $\mathring{\mu}$ say, such that $\Sigma\, \mathring{\mu}_i H_i = \overline{\Sigma}\, \hat{\bar{\mu}}_i \bar{H}_i$. Then, much as in part (i), $\lambda(\mathring{\mu}) = \bar{\lambda}(\hat{\bar{\mu}}) = \hat{\bar{\lambda}} = 0$ so $\lambda(\langle\!\langle \mathring{\mu} \rangle\!\rangle) = 0$ and $\Sigma\, \langle\!\langle \mathring{\mu} \rangle\!\rangle_i H_i \neq 0$. Furthermore, since $\hat{\bar{\lambda}} = 0$ implies that $\lambda_{\min}\,(\overline{\Sigma}\, \bar{\mu}_i \bar{H}_i) \leqq 0$ for all $\bar{\mu} \in \mathbb{R}^l$ and since the spans of the H_i and $\bar{H}_i$ are the same, $\lambda_{\min}\,(\Sigma\, \mu_i H_i) \leqq 0$ for all $\mu \in \mathbb{R}^l$. Therefore $\hat{\lambda} \leqq 0$. Hence, in view of (D2), $\hat{\lambda} = 0$. Since $\lambda(\langle\!\langle \mathring{\mu} \rangle\!\rangle) = 0$, it follows that a $\hat{\mu}$ is $\langle\!\langle \mathring{\mu} \rangle\!\rangle$ and gives $\Sigma\, \hat{\mu}_i H_i \neq 0$.

(iii) Here $\hat{\bar{\lambda}} < 0$. Therefore $\bar{\lambda}(\bar{\mu}) < 0$ for all $\bar{\mu} \in \mathbb{B}^{\bar{l}}$. The $\bar{H}_i$ are linearly independent so $\overline{\Sigma}\, \bar{\mu}_i \bar{H}_i \neq 0$ for all $\bar{\mu} \in \mathbb{B}^{\bar{l}}$. Since the spans of the H_i and of the $\bar{H}_i$ are the same, this reveals that $\lambda(\mu) < 0$ for all $\mu \in \mathbb{B}^l$ such that $\Sigma\, \mu_i H_i \neq 0$. However, $\lambda(\mu^{\#}) = 0$ for the vector $\mu^{\#}$ defined initially, for which $\Sigma\, (\mu^{\#})_i H_i = 0$. Hence $\hat{\lambda} = 0$ and a $\hat{\mu}$ is any $\mu \in \mathbb{B}^l$ which gives $\Sigma\, \mu_i H_i = 0$. □

Proof of Theorem 5.2. Let $\overset{*}{\mu} = H^{\dagger}\iota$. Then

$$\lambda(\overset{*}{\mu}) - 1 = \lambda_{\min}\left(\sum_{i=1}^{l} \overset{*}{\mu}_i H_i - I_n\right) \geqq -\left\|\sum_{i=1}^{l} \overset{*}{\mu}_i H_i - I_n\right\| \geqq -\left\|\sum_{i=1}^{l} \overset{*}{\mu}_i H_i - I_n\right\|_{\mathcal{F}}$$

$$= -\left\|\overline{\text{vec}}\left(\sum_{i=1}^{l} \overset{*}{\mu}_i H_i - I_n\right)\right\| = -\|H\overset{*}{\mu} - \iota\| = -\|d\|$$

where the first equality is from (1.2), the first two inequalities and the second equality are from Lemma B1 of Appendix B, the third equality is from (D1) and the last equality is from (5.3)–(5.4). Therefore

$$\lambda(\overset{*}{\mu}) \geqq 1 - \|d\|.$$

Suppose $\|d\| < (\leqq)\ 1$. Since, from (5.3), $d = \iota - HH^{\dagger}\iota$, it follows that $\|d\| = \|\iota\|$ if $H^{\dagger}\iota = 0$. Now $\|\iota\| > 1$ (because it has been assumed in § 1 that $n \geqq 2$) and we have assumed here that $\|d\| \leqq 1$, so here $H^{\dagger}\iota \neq 0$. Thus, $\overset{*}{\mu} \neq 0$, because $\overset{*}{\mu} = H^{\dagger}\iota$, and, since $\bar{\mu}$ of Theorem 5.1 is given by $\bar{\mu} = \langle\!\langle \overset{*}{\mu} \rangle\!\rangle \in \mathbb{B}^l$, it follows that

$$\hat{\lambda} \geqq \lambda(\bar{\mu}) = \frac{\lambda(\overset{*}{\mu})}{\|\overset{*}{\mu}\|} \geqq \frac{1 - \|d\|}{\|\overset{*}{\mu}\|} > (\geqq) 0. \qquad \square$$

Proof of Theorem 5.3. (i) It is clear from Theorem 3.2 that $n^{-1}\iota \in \Gamma$. Hence, in view of (5.3),

$$n^{-1}d = D[n^{-1}\iota] \in D\Gamma.$$

Therefore the line $L = \{\alpha d\colon \alpha \in \mathbb{R}\}$ intersects the set $D\Gamma$. Since Γ is bounded [Theorem 3.1(ii)], its recession cone $0^{+}\Gamma = \{0\}$ [12, Thm. 8.4]. Therefore, since Γ is closed [Theorem 3.1(ii)], $D\Gamma$ is closed [12, Thm. 9.1] and is therefore compact. Consequently $L \cap D\Gamma$ is compact. Therefore, since $L \cap D\Gamma = \{\alpha d\colon \alpha d \in D\Gamma\}$, there exists a most negative value of α such that $\alpha d \in D\Gamma$, i.e., $\hat\alpha$ of (5.8) exists.

Since D of (5.4) is an orthogonal projector and $\|\gamma\| \leqq 1$ for all $\gamma \in \Gamma$ (by Theorem 3.2), $\|D\gamma\| \leqq 1$ for all $\gamma \in \Gamma$. Therefore $\|D\gamma + \alpha d\| > 0$ whenever $\|\alpha d\| > 1$, i.e., whenever $|\alpha| > \|d\|^{-1}$. So $\|D\gamma + \alpha d\| = 0$ implies that $|\alpha| \leqq \|d\|^{-1}$. From (5.8), $\hat\alpha$ is the most negative $\alpha \in \mathbb{R}$ such that $\|D\gamma + \alpha d\| = 0$ for some $\gamma \in \mathbb{R}$. Therefore $|\hat\alpha| \leqq \|d\|^{-1}$, as claimed.

From Theorem 3.2, $n^{-1}d + DM\Xi_2 \subset D\Gamma$, so $n^{-1}d \in D\Gamma$. Hence

$$\text{(D3)} \qquad \hat\alpha = \min\{\alpha\in\mathbb{R}\colon \alpha d \in D\Gamma\} \leqq n^{-1}.$$

Throughout the rest of the proof of part (i), suppose $H'\iota = 0$. Then $\iota \in {}^{\perp}R[H]$ so

$$\text{(D4)} \qquad R[H] \subset {}^{\perp}R[\iota] = R[M],$$

as a result of the definition of M in (3.13).

It will be shown next that $d \notin R[DM]$, by contradiction.

Now $(d \in R[DM]) \Rightarrow (D\iota \in R[DM]) \Rightarrow (D(\iota - M\theta) = 0$ for some $\theta) \Rightarrow (\iota - M\theta \in N[D] = R[H] \subset R[M]$ {by the projection property of D and by (D4)}$) \Rightarrow (\iota \in R[M])$. But $(\iota \in R[M])$ is false because, by (3.13), $0 \neq \iota \in {}^{\perp}R[M]$. Hence $d \notin R[DM]$, as claimed earlier.

Next $(\alpha d \in D\Gamma) \Rightarrow (\alpha d \in n^{-1}d + R[DM]$ {by (3.9), (3.11)}$) \Rightarrow (\{\alpha - n^{-1}\}d \in R[DM]) \Rightarrow (\alpha = n^{-1}$ {since $d \notin R[DM]$}$)$. Hence $\hat\alpha = n^{-1}$ when $H'\iota = 0$.

Consequently, if $H'\iota = 0$ then $\hat\alpha = n^{-1}$ else (by (D3)) $\hat\alpha \leqq n^{-1}$, which completes the proof of part (i).

(ii) Consider $x \in \mathbb{R}^{r}$. Decompose ι and x orthogonally as $\iota = \bar\iota + \tilde\iota$ and $x = \bar x + \tilde x$ where $\bar\iota, \bar x \in N[H']$ and $\tilde\iota, \tilde x \in R[H]$. Since it is assumed in the statement of Theorem 5.3 that $\iota \notin R[H]$, $\bar\iota \neq 0$. Furthermore, since the case with $\iota \notin N[H']$ is being considered, $\tilde\iota \neq 0$. Choose $\bar x = \bar\iota$ and $\tilde x = -[\|\bar\iota\|^2/\|\tilde\iota\|^2]\tilde\iota$. Then $\iota - x \in R[H]$ and $x \in {}^{\perp}R[\iota] = \{z \in \mathbb{R}^{r}\colon z'\iota = 0\}$. Consequently, from the definition of M in Theorem 3.2, $x = M\theta$ for some $\theta \in \mathbb{R}^{r-1}$. For D of (5.4), $N[D] = R[H]$. Therefore, since $\iota - x \in R[H]$, it follows that $D(\iota - x) = 0$. Hence $d = D\iota = Dx = DM\theta$, i.e., $d \in R[DM]$, as required.

(iii) This is an immediate consequence of parts (iv), (v), and (vi), which will be proved next.

(iv)–(vi) From Definition 5.1, $H\mu_{\tilde\alpha} + \tilde\alpha\iota \in \Gamma$. Now

$$H\mu_{\hat\alpha} + \hat\alpha\iota \in \Gamma \Rightarrow \sum_{i=1}^{l} (\mu_{\hat\alpha})_i H_i + \hat\alpha I_n \geqq 0 \Rightarrow \sum_{i=1}^{l} (\mu_{\hat\alpha})_i H_i \geqq -\hat\alpha I_n$$

$$\Rightarrow \lambda(\mu_{\hat\alpha}) \geqq -\hat\alpha,$$

where the first implication is from (D1) and the fact that, in view of Theorem 3.1(i), $x \in \Gamma$ implies that $\overline{\text{vec}}^{-1}(x) \in S^{n}_{\geqq}$. Hence

$$\text{(D5)} \qquad \lambda(\mu_{\hat\alpha}) \geqq -\hat\alpha.$$

Now, provided $\mu_{\hat\alpha} \neq 0$, $\langle\!\langle \mu_{\hat\alpha} \rangle\!\rangle \in \mathbb{B}^l$ and, from (D5),

$$\text{(D6)} \qquad \hat\lambda \geqq \lambda(\langle\!\langle \mu_{\hat\alpha} \rangle\!\rangle) = \lambda(\mu_{\hat\alpha}) \|\mu_{\hat\alpha}\|^{-1} \geqq -\hat\alpha \|\mu_{\hat\alpha}\|^{-1}.$$

Suppose $\hat\alpha < 0$. Then, by (D5), $\lambda(\mu_{\hat\alpha}) > 0$ so $\mu_{\hat\alpha} \neq 0$ and consequently (D6) implies most of part (iv), in particular that $\hat\lambda > 0$. The proof of the upper bound for $\hat\lambda$ of part (iv) follows.

Clearly, $\sum_{i=1}^l \hat\mu_i H_i - \hat\lambda I_n \geqq 0$, since $\hat\lambda = \lambda(\hat\mu)$, so $H\hat\mu - \hat\lambda\iota \in \operatorname{cone}(\Gamma)$ (by (D1) and Theorem 3.1(i)). Hence

$$\text{(D7)} \qquad \theta(H\hat\mu - \hat\lambda\iota) \in \Gamma \quad \text{for some } \theta \in \mathbb{R}_{\geqq}.$$

It follows from Theorem 3.2 that, for $\gamma \in \Gamma$, $n^{-1/2} \leqq \|\gamma\|$. Therefore, from (D7),

$$\text{(D8)} \qquad \theta \geqq \frac{n^{-1/2}}{\|H\hat\mu - \hat\lambda\|}.$$

Since it has been shown above that $\hat\lambda > 0$ in this case (for which $\hat\alpha < 0$), and since $\sum_{i=1}^l \hat\mu_i H_i - \hat\lambda I_n \geqq 0$,

$$\left(\sum_{i=1}^l \hat\mu_i H_i\right)_{jj} \geqq \left(\sum_{i=1}^l \hat\mu_i H_i - \hat\lambda I_n\right)_{jj} \geqq 0, \qquad j = 1, 2, \cdots, n.$$

This gives rise to the first inequality below, where the first equality is from (D1) and Lemma B1, and the second equality is from (D1) and Lemma B1,

$$\|H\hat\mu - \hat\lambda\iota\| = \left\| \sum_{i=1}^l \hat\mu_i H_i - \hat\lambda I_n \right\|_{\mathcal{I}} \leqq \left\| \sum_{i=1}^l \hat\mu_i H_i \right\|_{\mathcal{I}} = \|H\hat\mu\| \leqq \|H\| \, \|\hat\mu\| = \|H\|.$$

So, from (D8),

$$\theta \geqq (n^{1/2}\|H\|)^{-1}.$$

Since $\hat\alpha$ is the most negative number α such that $H\mu + \alpha\iota \in \Gamma$ for some $\mu \in \mathbb{R}^l$, and since, from (D7), $H(\theta\hat\mu) + (-\theta\hat\lambda)\iota \in \Gamma$, it follows that $\hat\alpha \leqq -\theta\hat\lambda$, i.e., that, since $\hat\alpha < 0$, $\hat\lambda \leqq -\hat\alpha/\theta \leqq -\hat\alpha n^{1/2}\|H\|$. This has completed the proof of part (iv) by establishing the upper bound on $\hat\lambda$ given there.

Suppose next that $\hat\alpha = 0$. Then

$$\text{(D9)} \qquad \sum_{i=1}^l \mu_i H_i \not> 0, \quad \forall \mu \in \mathbb{R}^l,$$

since otherwise

$\exists \mu \in \mathbb{R}^l$ such that $\sum_{i=1}^l \mu_i H_i > 0$
$\Rightarrow \exists(\mu, \alpha) \in \mathbb{R}^l \times \mathbb{R}_<$ such that $\sum_{i=1}^l \mu_i H_i + \alpha I_n \geqq 0$,
i.e., such that $H\mu + \alpha\iota \in \operatorname{cone}(\Gamma)$, with $H\mu + \alpha\iota \neq 0$ (since $\alpha \in \mathbb{R}_<$ and it is assumed that $\iota \notin R[H]$ in Theorem 5.3, which is being proved)
$\Rightarrow \exists(\theta, \mu, \alpha) \in \mathbb{R}_> \times \mathbb{R}^l \times \mathbb{R}_<$ such that $\theta(H\mu + \alpha\iota) \in \Gamma$ (since $0 \notin \Gamma$, by (3.10) of Theorem 3.2)
$\Rightarrow \exists(\mu, \alpha) \in \mathbb{R}^l \times \mathbb{R}_<$ such that $H\mu + \alpha\iota \in \Gamma$
$\Rightarrow \hat\alpha < 0$,

which contradicts the assumption that $\hat\alpha = 0$.

Now $\mu_{\hat\alpha} \neq 0$, because $0 \notin \Gamma$, $H\mu_{\hat\alpha} + \hat\alpha\iota \in \Gamma$ and because $\hat\alpha = 0$ here.

In view of (D9), $\lambda(\mu) \leqq 0$ for all $\mu \in \mathbb{R}^l$, so $\hat{\lambda} \leqq 0$. Since $\mu_{\hat{\alpha}} \neq 0$, (D6) applies and, because the case $\hat{\alpha} = 0$ is being considered here, reveals that $\hat{\lambda} \geqq \lambda(\langle\langle\mu_{\hat{\alpha}}\rangle\rangle) \geqq 0$, so $\hat{\lambda} = 0$ and $\langle\langle\mu_{\hat{\alpha}}\rangle\rangle$ maximizes λ on $\mathbb{B}^l$, which completes the proof of part (v).

Now suppose $\hat{\alpha} > 0$. Then

$$\text{(D10)} \qquad \sum_{i=1}^{l} \mu_i H_i \ngeqq 0 \quad \forall \mu \in \mathbb{R}^l - \{0\},$$

since otherwise

$\exists \mu \in \mathbb{R}^l - \{0\}$ such that $\sum_{i=1}^{l} \mu_i H_i \geqq 0$
$\Rightarrow \exists (\mu, \alpha) \in \mathbb{R}^l - \{0\} \times \mathbb{R}_{\leqq}$ such that $\sum_{i=1}^{l} \mu_i H_i + \alpha I_n \geqq 0$,
i.e., such that $H\mu + \alpha\iota \in$ cone $[\Gamma]$, with $H\mu + \alpha\iota \neq 0$ (because $H\mu \neq 0$ (since $\mu \neq 0$ and the columns of H are linearly independent because it has been assumed that the H_i are linearly independent) and since it has been assumed in Theorem 5.3 that $\iota \notin R[H]$)
$\Rightarrow \exists (\theta, \mu, \alpha) \in \mathbb{R}_> \times \mathbb{R}^l - \{0\} \times \mathbb{R}_{\leqq}$ such that $\theta(H\mu + \alpha\iota) \in \Gamma$, since $0 \notin \Gamma$
$\Rightarrow \exists (\mu, \alpha) \in \mathbb{R}^l - \{0\} \times \mathbb{R}_{\leqq}$ such that $H\mu + \alpha\iota \in \Gamma$
$\Rightarrow \hat{\alpha} \leqq 0$,

which contradicts the assumption that $\hat{\alpha} > 0$.

From (D10), $\lambda(\mu) < 0$ for all $\mu \in \mathbb{R}^l - \{0\}$. Hence $\hat{\lambda} < 0$, which proves part (vi) and completes the proof of Theorem 5.3. □

Proof of Theorem 5.4. Parts (i)–(iii) were established in the text preceding the theorem. The proof of part (iv) follows.

By Theorem 5.3(i), $D\Gamma$ is compact so $0^+ D\Gamma = \{0\}$ [12, Thm. 8.4]. Hence $STD\Gamma$ is also compact. From part (i), $\hat{\alpha}$ is the most negative number α such that $\alpha\tilde{d} \in STD\Gamma$. Hence $\hat{\alpha}\tilde{d} \in \partial STD\Gamma$. Furthermore, part (ii) reveals that $STD\Gamma$ has an interior. Therefore there exists at least one nonsingular (i.e., not containing all of $STD\Gamma$) hyperplane supporting $STD\Gamma$ and passing through $\hat{\alpha}\tilde{d}$ [14, Cor. 3.4.12]. Let η be its normal. Such a nonsingular supporting hyperplane does not contain any points in the interior of $STD\Gamma$ [14, Lemma 4.3.4]. Consequently, since $n^{-1}\tilde{d} \in \operatorname{int}(STD\Gamma)$ (from part (ii)),

$$\text{(D11)} \qquad (n^{-1}\tilde{d})'\eta < (\hat{\alpha}\tilde{d})'\eta.$$

From part (iii), $\hat{\alpha} < n^{-1}$. Therefore (D11) reveals that $\tilde{d}'\eta < 0$, which completes the proof of Theorem 5.4.

The proof of Theorem 5.5 will be facilitated by the following result.

LEMMA D1. *Suppose $d \neq 0$, $H'\iota \neq 0$, and $\alpha < \hat{\alpha}$. Let $\bar{\varepsilon} \in \mathbb{R}_>$ and let $\gamma \in \Gamma$ with $\|D\gamma - \alpha d\| \leqq (1 + \bar{\varepsilon})$ mindist $[\alpha h, D\Gamma]$ and $\|D\gamma - \alpha d\| \leqq \|n^{-1}d - \alpha d\|$. Furthermore, let $\mu = H^\dagger[\gamma - \alpha\iota]$. Then $\mu \neq 0$, $n^{-1} - \alpha > 0$ and*

$$\lambda(\langle\langle\mu\rangle\rangle) \geqq -\frac{|\alpha + (1 + \bar{\varepsilon})(\hat{\alpha} - \alpha)\|d\|\,|\,\|\iota\|\;\|H\|}{(n^{-1} - \alpha)\|HH^\dagger\iota\|^2}. \qquad □$$

Proof of Lemma D1. A lower bound on $\lambda(\mu)$ will be obtained first. In view of the definitions of μ as $H^\dagger[\gamma - \alpha\iota]$ in Lemma D1, of d in (5.3) and of D in (5.4),

$$\gamma - H\mu - \alpha\iota = D(\gamma - \alpha\iota) = D\gamma - \alpha d.$$

So, applying $\overline{\text{vec}}^{-1}$ throughout, remembering definition (5.1) of H,

$$\sum_{i=1}^{l} \mu_i H_i = X_\gamma - \alpha I_n - Z_i,$$

where $X_\gamma = \overline{\text{vec}}^{-1}(\gamma)$, $Z_i = \overline{\text{vec}}^{-1}(D\gamma - \alpha d)$. Here, since $\gamma \in \Gamma$, it follows from Theorem 3.1(i) that $X_\gamma \geqq 0$. In addition, from Lemma B1(i), $\|Z_i\| \leqq \|Z_i\|_{\mathcal{F}} = \|D\gamma - \alpha d\|$. Therefore

$$\sum_{i=1}^{l} \mu_i H_i \geqq -(\alpha + \|Z_i\|)I_n \geqq -(\alpha + \|D\gamma - \alpha d\|)I_n,$$

so a lower bound for $\lambda(\mu)$ is given by

$$\lambda(\mu) \geqq -(\alpha + \|D\gamma - \alpha d\|) \geqq -(\alpha + (1+\bar{\varepsilon})(\hat{\alpha} - \alpha)\|d\|). \tag{D12}$$

Here the last inequality occurs because it has been assumed in Lemma D1 that $\|D\gamma - \alpha d\| \leqq (1 + \bar{\varepsilon})$ mindist $[\alpha d, D\Gamma]$ and because

$$\text{mindist}[\alpha d, D\Gamma] \leqq (\hat{\alpha} - \alpha)\|d\|,$$

since $\hat{\alpha} d \in D\Gamma$ and $\alpha < \hat{\alpha}$. The above-mentioned lower bound on $\lambda(\mu)$ is given by the right-hand side of (D12).

Next a lower bound on $\|\mu\|$ will be established. In view of Theorem 3.2, and because $\gamma \in \Gamma$,

$$\gamma = n^{-1}\iota + M\theta \tag{D13}$$

for some $\theta \in \mathbb{R}^{r-1}$. Therefore, since it has been assumed in Lemma D1 that

$$\|D\gamma - \alpha d\| \leqq \|n^{-1}d - \alpha d\|,$$

and since $d = D\iota$,

$$\|D\{(n^{-1} - \alpha)\iota + M\theta\}\| \leqq \|D\{(n^{-1} - \alpha)\iota\}\|. \tag{D14}$$

Direct computation with (D14) {making use of the facts that $\iota' M = 0$ (from (3.13)), that $\alpha < \hat{\alpha}$ (assumed in Lemma D1), that $\hat{\alpha} < n^{-1}$ (from Theorem 5.4(iii), since it is assumed here that $H'\iota \neq 0$) and that $D = I_r - HH^\dagger$} reveals that

$$\iota' HH^\dagger M\theta \geqq 0. \tag{D15}$$

Furthermore, $\|HH^\dagger \iota\| > 0$ since $N[HH^\dagger] = {}^{\perp}R[H]$ and since $\iota \notin {}^{\perp}R[H]$ because it is assumed here that $H'\iota \neq 0$, i.e., that $\iota \notin N[H'] = {}^{\perp}R[H]$. Hence, because $HH^\dagger$ is a projector,

$$\iota' HH^\dagger \iota = \|HH^\dagger \iota\|^2 > 0. \tag{D16}$$

Now, from (D13) and the definition of μ as $H^\dagger[\gamma - \alpha\iota]$, $H\mu = HH^\dagger[\gamma - \alpha\iota] = HH^\dagger[(n^{-1} - \alpha)\iota + M\theta]$. So, in view of (D15)-(D16),

$$\iota' H\mu \geqq (n^{-1} - \alpha)\|HH^\dagger \iota\|^2 > 0, \tag{D17}$$

where $n^{-1} - \alpha > 0$ since the case $\alpha < \hat{\alpha}$ is being considered and since $\hat{\alpha} \leqq n^{-1}$ (by Theorem 5.4(iii)).

From (3.13), $\mathbb{R}^r = R[\iota] \oplus R[M]$. It follows that $H\mu = \xi \langle\!\langle \iota \rangle\!\rangle + M\theta$ for some $\xi \in \mathbb{R}$ and some $\theta \in \mathbb{R}^{r-1}$. From (D17), $\xi \geqq (n^{-1} - \alpha)\|HH^\dagger \iota\|^2 / \|\iota\|$. Furthermore, since $\iota' M = 0$ and since $\|\langle\!\langle \iota \rangle\!\rangle\| = 1$,

$$\|H\mu\| \geqq \|\xi \langle\!\langle \iota \rangle\!\rangle\| \geqq (n^{-1} - \alpha)\frac{\|HH^\dagger \iota\|^2}{\|\iota\|}.$$

Consequently the required lower bound for $\|\mu\|$ is provided by

$$\text{(D18)} \qquad \|\mu\| \geqq \frac{\|H\mu\|}{\|H\|} \geqq (n^{-1}-\alpha)\frac{\|HH^{\dagger}\iota\|^2}{\|\iota\|\ \|H\|} > 0.$$

Therefore, from (D12) and (D18)

$$\lambda(\langle\!\langle \mu \rangle\!\rangle) = \frac{\lambda(\mu)}{\|\mu\|} \geqq -\frac{|\alpha+(1+\bar{\varepsilon})(\hat{\alpha}-\alpha)\|d\|\,|\ \|\iota\|\ \|H\|}{(n^{-1}-\alpha)\|HH^{\dagger}\iota\|^2},$$

which completes the proof of Lemma D1. □

Proof of Theorem 5.5. Consider the application of Algorithm 4.1′ mentioned in Theorem 5.5. Theorem 4.1 reveals that

$$\text{(D19)} \qquad \alpha_i < \hat{\alpha} \leqq \beta_i \quad \forall i \geqq 0, \quad \alpha_i \uparrow \hat{\alpha}, \quad \beta_i \downarrow \hat{\alpha}.$$

Therefore, if $\hat{\alpha} > 0$ then $\alpha_{i+1} \geqq 0$ will occur after a finite number of iterations. If $\hat{\alpha} < 0$ then $\beta_{i+1} < 0$ will occur after a finite number of iterations. If $\hat{\alpha} = 0$ then, from (D19), $\alpha_{i+1} < 0$ and $\beta_{i+1} \geqq 0$ for all $i \geqq 0$ and, from (D19) and (4.21′) of Theorem 5.5, $\delta_{i+1} \rightarrow 0$. Hence the condition $\alpha_{i+1} < 0$, $\beta_{i+1} \geqq 0$, $\delta_{i+1} < \delta$ will be satisfied eventually. Thus Algorithm 4.1′, with stopping condition (4.21′) of Theorem 5.5 replacing stopping condition (4.21), will stop after a finite number of iterations whatever is the value of $\hat{\alpha}$.

If Algorithm 4.1′ stops owing to $\alpha_{i+1} \geqq 0$, then it follows from (D19) that $\hat{\alpha} > 0$ and consequently, from Theorem 5.3(iii), it will be known that $\hat{\lambda} < 0$. Similarly, if the algorithm stops owing to $\beta_{i+1} < 0$, then it will be known that $\hat{\lambda} > 0$, as claimed in Theorem 5.5.

Consider now the situation when Algorithm 4.1′ stops with the situation

$$\text{(D20)} \qquad \alpha_{i+1} < 0, \quad \beta_{i+1} \geqq 0, \quad \delta_{i+1} \leqq \delta.$$

Then it can be seen that

$$\text{(D21)} \quad -\frac{|\alpha_{i+1}+(1+\varepsilon_2)[\hat{\alpha}-\alpha_{i+1}]\|d\|\,|\ \|\iota\|\ \|H\|}{(n^{-1}-\alpha_{i+1})\|HH^{\dagger}\iota\|^2} \leqq \lambda(\langle\!\langle \bar{\mu} \rangle\!\rangle) \leqq \hat{\lambda} \leqq -\alpha_{i+1}n^{1/2}\|H\|,$$

as follows.

In view of (D19) and the situation being considered in (D20), Algorithm 4.1′ stops with $\alpha_{i+1} < \hat{\alpha}$ and $\bar{\gamma} \in \varepsilon_2\text{-minpoints}_{\Gamma}\,[\alpha_{i+1}, D\Gamma]$. Consequently, from Definition 4.1, it also ends with $\|D\bar{\gamma} - \alpha_{i+1}d\| \leqq (1+\varepsilon_2)\,\text{mindist}\,[\alpha_{i+1}d, D\Gamma]$ and $\|D\gamma - \alpha_{i+1}d\| \leqq \|n^{-1}d - \alpha_{i+1}d\|$. Therefore, from Lemma D1, the vector $H^{\dagger}[\bar{\gamma} - \alpha_{i+1}\iota]$ in (4.21′) of Theorem 5.5 is nonzero (so that $\bar{\mu}$ of Theorem 5.5 is defined) and the first inequality of (D21) is valid.

The second inequality occurs because $\langle\!\langle \bar{\mu} \rangle\!\rangle \in \mathbb{B}^l$.

From Theorem 5.3(iv)–(vi) and (D19)–(D20), if $\hat{\alpha} < 0$ then $\hat{\lambda} \leqq -\hat{\alpha}n^{1/2}\|H\| < -\alpha_{n+1}n^{1/2}\|H\|$ and if $\hat{\alpha} \geqq 0$ then $\hat{\lambda} \leqq 0$. Hence, either way, the final inequality of (D21) is satisfied.

From (D21), the definition of δ_{i+1} in Theorem 5.5 and from (D19), $\lambda(\langle\!\langle \bar{\mu} \rangle\!\rangle) \in [\hat{\lambda} - \delta_{i+1}, \hat{\lambda}]$ with $\delta_{i+1} \geqq 0$. Since the case when the algorithm stops with (D20) satisfied is being considered, $\lambda(\langle\!\langle \bar{\mu} \rangle\!\rangle) \in [\hat{\lambda} - \delta, \hat{\lambda}]$, which completes the proof of Theorem 5.5. □

Acknowledgments. The author thanks M. K. Mak for programming the algorithms and for producing the numerical results of § 6, and both M. K. Mak and I. M. Jaimoukha for reducing the number of mistakes in the paper.

REFERENCES

[1] J. C. ALLWRIGHT, LQP: *Dominant output feedbacks*, IEEE Trans. Automat. Control, 27 (1982), pp. 915–921.

[2] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

[3] R. FLETCHER, *Semi-definite matrix constraints in optimization*, SIAM J. Control Optim., 23 (1985), pp. 493–513.

[4] M. L. OVERTON, *On minimizing the maximum eigenvalue of a symmetric matrix*, in Linear Algebra in Signals, Systems and Control, B. N. Datta, ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1987.

[5] M. L. OVERTON AND R. S. WOMERSLEY, *On minimizing the spectral radius of a nonsymmetric matrix function—optimality conditions and duality theory*, SIAM J. Matrix Anal. Appl., to appear.

[6] O. TAUSSKY, *Positive-definite matrices*, in Inequalities, O. Shisha, ed., Academic Press, New York, 1967.

[7] L. BRICKMAN, *On the field of values of a matrix*, Proc. Amer. Math. Soc., 12 (1961), pp. 61–66.

[8] J. C. ALLWRIGHT, *Positive semi-definite matrices: Characterization via conical hulls and least-squares solution of a matrix equation*, SIAM J. Control Optim., 26 (1988), pp. 537–556; erratum and addendum (with K. G. Woodgate), submitted to SIAM.

[9] E. GILBERT, *An iterative method for computing the minimum of a quadratic form on a convex set*, SIAM J. Control, 4 (1966), pp. 61–80.

[10] J. E. HAUSER, *Proximity algorithms: Theory and implementation*, Memorandum UCB/ERL M86/53, Electronics Research Laboratory, University of California, Berkeley, CA, May 1986.

[11] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1968.

[12] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[13] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, North Oxford Academic Publishing, Oxford, England, 1983.

[14] J. STOER AND C. WITZGALL, *Convexity and Optimization in Finite Dimensions* I, Springer-Verlag, Berlin, New York, 1974.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 3, pp. 383-406, July 1989

PADÉ APPROXIMANTS TO MATRIX STIELTJES SERIES: CONVERGENCE AND RELATED PROPERTIES*

SANKAR BASU

Abstract. Following earlier work on Padé approximants to matrix Stieltjes series and their network theoretic relevance, it is shown that certain paradiagonal sequences of matrix Padé approximants to the series under consideration always converge. Interpretations of this result in terms of representation of impedances of RC distributed multiport networks are given. Matricial generalizations of the classical Hamburger and Stieltjes moment problems are discussed in this context. Matrix polynomials of the second kind orthogonal on the real line, which fall out as numerators of the matrix Padé approximants of certain orders, are singled out and their properties are studied.

Key words. Padé approximation, Stieltjes series, network theory

AMS(MOS) subject classification. 93

1. Introduction.

Consider a formal power series as in (1.1), where

$$T(s)=\sum_{k=0}^{\infty} T_k s^k; \tag{1.1}$$

each T_k is a real symmetric matrix of size $(p \times p)$. The rational matrix $Q_L(s)P_M^{-1}(s)$ (or $\tilde{P}_M^{-1}(s)\tilde{Q}_L(s)$), where $Q_L(s)$, $\tilde{Q}_L(s)$ and $P_M(s)$, $\tilde{P}_M(s)$ are $(p \times p)$ polynomial matrices of respective formal degrees L and M[1] is said to be a right matrix Padé approximant (or left matrix Padé approximant) to $T(s)$ if the first $(L + M + 1)$ terms of the Maclaurin's series expansion of $Q_L(s)P_M^{-1}(s)$ (or $\tilde{P}_M^{-1}(s)\tilde{Q}_L(s)$) matches with those of $T(s)$ in (1.1). In addition, the formal power series $T(s)$ in (1.1) is said to be a matrix Stieltjes series [1] if for each n the block Hankel matrices $H_n(T)$ and $H'_n(T)$ as given in (1.2) are positive definite and negative definite, respectively:

$$H_n(T)=\begin{bmatrix} T_0 & T_1\cdots T_n \\ T_1 & T_2 \\ \vdots & & \vdots \\ T_n & & \cdots T_{2n} \end{bmatrix}, \tag{1.2a}$$

$$H'_n(T)=\begin{bmatrix} T_1 & T_2\cdots T_n \\ T_2 & T_3 \\ \vdots & & \vdots \\ T_n & & \cdots T_{2n-1} \end{bmatrix}. \tag{1.2b}$$

Note that for matrix Stieltjes series the right and left matrix Padé approximants uniquely exist and are necessarily identical [1], [3]. Thus, the term matrix Padé approximant (MPA) of order $[L/M]$ will henceforth be used to denote $[L/M](s) = Q_L(s)P_M^{-1}(s) = \tilde{P}_M^{-1}(s)\tilde{Q}_L(s)$. The fact that the paradiagonal sequences of MPAs of order $[m - 1/m]$ and $[m - 1/m - 1]$ for $m = 1, 2, \cdots$ to a matrix Stieltjes series can be identified as the

* Received by the editors March 2, 1988; accepted for publication (in revised form) November 29, 1988. A revised version of this work was prepared while the author was visiting the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

† Department of Electrical Engineering and Computer Science, Stevens Institute of Technology, Hoboken, New Jersey 07030. Present address, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 (BASU@LIDS.MIT.edu).

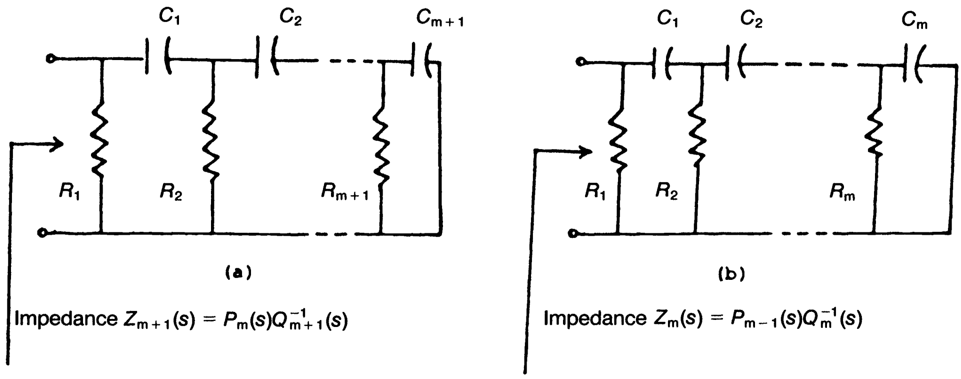
[1] The formal degree of a polynomial matrix is defined as the largest degree of its polynomial entries.

impedances of admittances of multiport electrical networks containing two types of elements (e.g., RC or RL) has been established in [1] via utilization of recently developed tools of matrix continued fraction expansion and the Cauchy index of a rational matrix. Also, the Padé approximation problem can be cast in terms of the partial realization problem as occurring in linear system theory [13], [19]. The positive definiteness of the block Hankel matrices $H_n(T)$ for all n then imply, in particular, that every point in the partial realization data is a "jump point" with jump size equal to 1 [13]. Thus, as has been shown via the tools of matrix continued fraction expansion [1], [15], as well as the Cauchy index of a rational matrix [1], the positive definiteness of $H_n(T)$ and $H'_n(T)$ can be viewed as the conditions which the partial realization data needs to satisfy so that the realized transfer function matrix is an impedance or admittance of an RC multiport (similar formulations for RL or LC impedances or admittances are also possible).

The question of convergence of the sequences $[m - 1/m]$ or $[m - 1/m - 1]$, $m = 1, 2, \cdots$ of MPAs, when $T(s)$ is a matrix Stieltjes series, however, has not been addressed in the literature. Although a discussion of this issue in the scalar (i.e., $p = 1$) case is available in [3], system theoretic interpretation of the results are not readily available. In the present paper it is shown by exploiting the network theoretic interpretations developed in [1] that the sequences of $[m - 1/m]$ and $[m - 1/m - 1]$, $m = 1, 2, \cdots$. MPAs to a matrix Stieltjes series do indeed converge uniformly in an open (bounded) region of the complex s-plane excluding the negative real axis.

Furthermore, since the sequences of $[m - 1/m]$ and $[m - 1/m - 1]$ MPAs can be viewed as the successive convergents of certain special types of matrix continued fractions [1], the convergence of the paradiagonal sequences of MPAs can also be interpreted as the convergence property of the related matrix continued fraction expansions. Since the continued fraction expansion just mentioned is, in fact, associated with a ladder realizable RC multiport (cf. Fig. 1 for $p = 1$), we essentially have the result that the sequence of RC multiport ladder impedances (or admittances) so derived from a matrix Stieltjes series is always convergent (even if the formal power series $T(s)$ in (1.1) is not). Thus, this latter result can be interpreted in terms of the important fact that the matrix Stieltjes series in (1.1), which may not necessarily converge (cf. [3] for examples in $p = 1$ case), can be meaningfully used to represent a nonrational impedance or admittance matrix associated with a multiport RC distributed transmission line [17].

Next it is shown that, as in the scalar case the positive definiteness of $H_n(T)$ in (1.2a) for all n guarantees the existence of a bounded nondecreasing real symmetric matrix valued measure $\sigma(x)$, $-\infty < x < \infty$, such that each of the T_k's in (1.2) can be



(a)

Impedance $Z_{m+1}(s) = P_m(s)Q_{m+1}^{-1}(s)$

(b)

Impedance $Z_m(s) = P_{m-1}(s)Q_m^{-1}(s)$

FIG. 1. $Z_{m+1}(s) > Z_m(s)$ *if* $\operatorname{Im} s = 0$, $\operatorname{Re} s > 0$.

viewed as the kth order moment associated with $\sigma(x)$. This result provides a direct matricial generalization of the classical Hamburger moment problem [6], [10]. If, in addition to (1.2a), the negative definiteness of $H'_n(T)$ in (1.2b) is also imposed, then the support of $\sigma(x)$ is shown to be restricted to the semi-infinite interval $0 < x < \infty$, thus providing a solution to the matrix version of classical Stieltjes moment problem [6], [10].

Furthermore, if the power series $T(s)$ in (1.1) is assumed to converge in a disc of radius R, then we show that the sequences of MPAs $[m-1/m](s)$ and $[m-1/m-1](s)$ indeed converge to $T(s)$. An integral representation of $T(s)$ in terms of the matrix-valued measure $\sigma(x)$, which coincides with the Cauer's representation [20] of RC (multiport) impedances, or in the scalar case with the closely related class of classical Stieltjes functions [3], is also derived in this context.

On the other hand, the "denominator" polynomial matrices associated with MPAs of order $[m-1/m]$ and $[m-1/m-1]$ have been shown in [1] to form sequences of polynomial matrices orthogonal on the real line. Various properties of these polynomials such as the three-term recurrence relation followed by them, properties of location their zeros and their relationship to the matricial Gauss quadrature formula were also derived in [1]. In this paper it is shown that while in [1] the orthogonality of the matrix polynomials was viewed in terms of certain vector space representations, by using the measure $\sigma(x)$ the orthogonality relationship can be seen more transparently in terms of an inner product in standard form.

More importantly, this approach also establishes the interesting fact that the sequences of "numerator" polynomial matrices of the $[m-1/m](s)$ and $[m-1/m-1](s)$ sequences of MPAs also satisfy certain orthogonality properties similar to those satisfied by the "denominators" of the paradiagonal sequences of MPAs in question. From a system theoretic standpoint this result is to be expected in view of the fact that the property of RC, RL, or LC impedance (or admittance) realizability remain invariant under the operation of inversion of the rational matrix concerned. The sequence of matrix orthogonal polynomials corresponding to the "numerator" sequences of MPAs are thus also found to provide matricial generalization of the orthogonal polynomials of the second kind discussed in the classical literature [6], [9].

In the rest of this section related previous research on specific aspects of the problem considered in the present paper will be briefly reviewed and comparisons to our approach to the problem will be made. The study of matrix Padé approximants, their relationships to continued fractions, various moment problems, and issues of convergence were initiated in [23], [24]. Both convergence of sequences of Padé type approximants to Stieltjes series [22] as well as the related moment problems [25] have been discussed in the mathematical literature in an (infinite-dimensional) operator theoretic setting by assuming that the T_k's in (1.1) are not just matrices but infinite-dimensional operators in Hilbert space. Convergence of Padé approximants to a formal power series of the matrix Stieltjes type has been previously considered in [21].

This paper deviates from those mentioned above in the following respects. First, our proofs are simpler, more elegant, and make use of elementary tools from linear algebra and complex function theory. This is so because we make full use of the finite-dimensional (i.e., matrix) nature of the problem considered. In fact, although the major results on the moment problems in [25] is known to be in error [26] in the infinite-dimensional case, a correct elementary discussion for the finite-dimensional problem is not known.

The convergence proof of [21] starts from a slightly different (albeit equivalent) definition of matrix Stieltjes series, where T_k's are assumed to be the moments associated with a nondecreasing symmetric matrix-valued measure at the very outset. This already amounts to assuming a solution to the corresponding moment problem referred to above, that is worked out in the present framework in § 3 of our paper. Furthermore, although the final results in [21] hold only for the finite-dimensional ease, their proofs hinge on powerful operator theoretic results (e.g., Naimakar's theorem linking method of moments for self-adjoint operators in Hilbert space). While our definition of matrix Stieltjes series is via the algebraic constraints imposed on the sign definiteness of $H_n(T)$ and $H'_n(T)$, our proof is more direct, elementary and does not make use of a solution to the moment problem at all.

Finally, the most contrasting aspect of the present contribution is that our discussions including the details of proofs are guided throughout by system theoretic intuition—an approach not adopted by earlier authors in the area.

2. Convergence proof of sequences of MPAs to matrix Stieltjes series. The major content of this section is the proof of the fact that the $[m-1/m](s)$ and $[m-1/m-1](s)$ sequences of MPAs converge uniformly to an analytic function in the domain $D(\Delta)$, where $D(\Delta)$ is any bounded domain of the complex plane at least at a distance Δ away from the negative real axis: $-D \leqq \operatorname{Re} s \leqq 0$, $\operatorname{Im} s = 0$. The strategy of our proof is to first show that the required convergence is attained for all fixed real positive values of s. This is achieved by establishing certain monotonicity and boundedness properties of the approximants that result as consequences of the RC realizability of $[m-1/m](s)$ and $[m-1/m-1](s)$, as shown in [1]. Uniform convergence in $D(\Delta)$ is then proved by essentially exploiting standard arguments on convergence continuation [5]. A mathematically equivalent procedure has been pursued in [3] for the scalar ($p = 1$) case without the use of network theoretic arguments.

The following notations will be used in the rest of the paper. If A is a real symmetric positive-definite matrix, then we will write $A > 0$. Also, the notations $A > B$ and $A \geqq B$ will be taken to mean that the real symmetric matrix $A - B$ is positive definite or nonnegative definite, respectively. Obvious variations of this notation with the symbols $>$ and $\geqq$ replaced by $<$ and $\leqq$ will also be used.

We first need the following theorem.

THEOREM 2.1. *The sequences of* $[m-1/m](s)$ *and* $[m-1/m-1](s)$ *approximants to a matrix Stieltjes series each, respectively, form an increasing and decreasing sequence of symmetric matrix fraction descriptions on the positive real axis, i.e., for all* $m = 1, 2, 3, \cdots$, *and for all s with* $\operatorname{Re} s > 0$ *and* $\operatorname{Im} s = 0$ *we have*

$$[m/m+1](s)-[m-1/m](s)>0, \tag{2.1}$$

$$[m/m](s)-[m-1/m-1](s)<0. \tag{2.1'}$$

Proof. The proof relies on the result [1, p. 211] that $[m-1/m]$ and $[m-1/m-1]$ approximants to the matrix Stieltjes series (1.1) can be obtained by truncating the matrix continued fraction expansion

$$T(s)=\left[B_1+\left[\frac{1}{s}B_2+\cdots\left[\frac{1}{s^{\lambda_k}}B_k+\frac{1}{s^{\lambda_{k+1}}}T_k(s)\right]^{-1}\right]^{-1}\right]^{-1} \tag{2.2}$$

where $\lambda_k = 0$ for k odd and $\lambda_k = 1$ for k even, the B_i, $i = 1, 2, \cdots$ are constant real symmetric positive definite matrices and $T_k(s)$ is a matrix Stieltjes series. In particular, it is shown in [1] that for $m = 1, 2, \cdots$ (2.3) and (2.3') hold true:

$$[m-1/m](s)=\left[B_1+\left[\frac{1}{s}B_2+\cdots+\left[\frac{1}{s}B_{2m}\right]^{-1}\right]^{-1}\right]^{-1}, \tag{2.3}$$

$$(2.3') \qquad [m-1/m-1](s)=\left[B_1+\left[\frac{1}{s}B_2+\cdots+[B_{2m-1}]^{-1}\right]^{-1}\right]^{-1}.$$

We shall prove (2.1) only; the proof for (2.1′) is analogous. Note first that due to (2.3), the approximant of order $[m/m+1]$ can be written as in (2.4):

$$(2.4) \qquad \begin{aligned}[m/m+1](s)=\Bigg[B_1+\Bigg[\frac{1}{s}B_2+\cdots\\ +\left[\frac{1}{s}B_{2m}+\left[B_{2m+1}+\left[\frac{1}{s}B_{2m+2}\right]^{-1}\right]^{-1}\right]^{-1}\Bigg]^{-1}\Bigg]^{-1}.\end{aligned}$$

Obviously, for Re $s>0$ and Im $s=0$ we have $B_i/s>0$ for all i. Since the sum as well as inverse of real symmetric positive-definite matrices is also real symmetric positive definite, we have $[B_{2m+1}+[B_{2m+2}/s]^{-1}]^{-1}>0$; consequently,

$$B_{2m}/s+[B_{2m+1}+[B_{2m+2}/s]^{-1}]^{-1}>B_{2m}/s$$

for Re $s>0$ and Im $s=0$. By using the result that if A and B are two real symmetric positive-definite matrices such that $A>B$, then $A^{-1}<B^{-1}$ [2, p. 86] it then follows that

$$(2.5) \qquad \left[\frac{1}{s}B_{2m}+\left[B_{2m+1}+\left[\frac{1}{s}B_{2m+2}\right]^{-1}\right]^{-1}\right]^{-1}<\left[\frac{1}{s}B_{2m}\right]^{-1}.$$

Repeating the process of adding the matrices B_i/s^{λ_i} and subsequently considering the inverses of the resulting matrices in the left- and right-hand sides of (2.5) for $i=2m-1, 2m-2, \cdots, 1$, where $\lambda_i=0$ when i is odd and $\lambda_i=1$ when i is even, it follows from (2.3) and (2.4) that $[m/m+1](s)>[m-1/m](s)$ for Re $s>0$, Im $s=0$. □

The physical implication of the above theorem is obviously clear in electrical network theoretic terms, when the $[m-1/m](s)$ and $[m-1/m-1](s)$ approximants to a Stieltjes series are interpreted as being the input impedance of RC ladder network, as depicted for the scalar case $p=1$, in Figs. 1(a) and 1(b), respectively. The monotonicity property of the sequences of approximants then trivially follow from the fact that for all real and positive values of s, the input impedances can be computed by replacing the capacitors by positive resistances.

The norm $\|x\|$ of a vector x will be defined as the well-known Euclidean norm, whereas the norm $\|\cdot\|$ of a matrix A will be defined as the spectral norm

$$\|A\|\triangleq \max\,[\,\|Ax\|;\ \|x\|=1].$$

We recall the following properties of the spectral norm $\|\cdot\|$ of a matrix A.

PROPERTY 2.1 [4]. *$\|A\|$ is equal to the largest singular value of A. In particular, if A is real then $\|A\|=\sqrt{(\lambda_m(A^tA))}$, where A^t is the transpose of the matrix A, and $\lambda_m(A^tA)$ denotes the largest eigenvalue of A^tA.*

PROPERTY 2.2. *For any real symmetric matrix A, $\|A\|=|\lambda_m(A)|$, where $\lambda_m(A)$ is the eigenvalue of A having largest absolute value. Thus, in particular, $\|\alpha A\|=|\alpha|\cdot\|A\|$, where α is any real number.*

Proof. The proof follows from the fact that $\lambda(A^tA)=\lambda(A^2)=\lambda^2(A)$, where $\lambda(A)$ is an eigenvalue (necessarily real) of A. □

PROPERTY 2.3. *If A, B, C are real symmetric positive- (nonnegative-) definite matrices such that $A=B+C$, then $\|A\|>\|B\|$ ($\|A\|\geqq\|B\|$).*

Proof. Since C is positive- (nonnegative-) definite from the Courant–Fisher min-max theorem (e.g., [2, p. 73]) it follows that $\lambda_m(A)=\lambda_m(B+C)>\lambda_m(B)$ (or

$\lambda_m(B + C) \geqq \lambda_m(B)$). The result then follows from Property 2.2 via the observation that eigenvalues of A and B are positive (nonnegative). □

PROPERTY 2.4. *If A and B are two real symmetric positive- (nonnegative-) definite matrices, and a and b are two real numbers such that $0 < a < 1$ and $0 < b < 1$, then $\|aA + bB\| < \|A + B\|$.*

Proof. Since $0 < (1 - a) < 1$ and $0 < (1 - b) < 1$, the matrices $(1 - a)A$ and $(1 - b)B$, and thus $\{(1 - a)A + (1 - b)B\}$, are real symmetric positive- (nonnegative-) definite. Thus, $\|A + B\| = \|(aA + bB) + \{(1 - a)A + (1 - b)B\}\| > \|aA + bB\|$ (or $\geqq \|aA + bB\|$, correspondingly). The last step follows by the use of Property 2.3 above. □

COROLLARY 2.1.1. *The* MPA*s to a matrix Stieltjes series satisfy $\|[m/m + 1](s)\| > \|[m - 1/m](s)\|$, for $m = 1, 2, \cdots$ and $\|[m/m](s)\| < \|[m - 1/m - 1](s)\|$ for $m = 1, 2, 3, \cdots$ for all real positive values of s.*

Proof. We first note that the MPAs to the matrix series of Stieltjes (1.1) are necessarily symmetric rational matrices. To substantiate this, note that if $Q_L(s)P_M^{-1}(s)$ is a right MPA of order $[L/M]$ to the series $T(s)$, then $P_M^{-t}(s)Q_L^t(s)$[2] is also a left MPA of order $[L/M]$ to the matrix series $T^t(s) = T(s)$. The last equality follows from the fact that in (1.1) $T_i^t = T_i$ for all $i = 0, 1, 2, \cdots$. However, this proves that both right and left approximants of order $[L/M]$ to the series $T(s)$ exist, and hence they must be equal [3], i.e., $Q_L(s)P_M^{-1}(s) = P_M^{-t}(s)Q_L^t(s)$. Thus the approximant of order $[L/M]$ is symmetric. (Alternatively, this result also follows from the representations (2.3) and (2.3′) of the sequences $[m - 1/m](s)$, $m = 1, 2, \cdots$ and $[m - 1/m - 1](s)$, $m = 1, 2, \cdots$ of approximants.) Also, it follows from Theorem 2.1 that if $[m/m + 1](s) - [m - 1/m](s) = P_{m1}(s)$ and $[m - 1/m - 1](s) - [m/m](s) = P_{m2}(s)$, then for all real positive values of s and for all $m = 1, 2, \cdots$, $P_{m1}(s)$ and $P_{m2}(s)$ are real symmetric positive-definite matrices. Since due to (2.3) and (2.4) the MPAs in the last two equalities are themselves real symmetric positive definite for all real and positive values of s, the required result follows from Property 2.3 of the spectral norm. □

Next, we consider the matrix continued fraction expansions of the MPAs $[m - 1/m](s)$ and $[m - 1/m - 1](s)$ to the matrix Stieltjes series $T(s)$. Since the MPAs just mentioned are known to be the impedance matrices of electrical networks consisting of positive resistors and capacitors only, the continued fraction expansions, due to results discussed in [1], can be expressed as the matrix partial fraction expansion:

$$[m-1/m](s) = \sum_{\nu=1}^{r} \frac{1}{1+\gamma_\nu s} A_\nu, \tag{2.6}$$

$$[m-1/m-1](s) = A'_0 + \sum_{\nu=1}^{r'} \frac{1}{1+\gamma'_\nu s} A'_\nu \tag{2.6$'$}$$

where in (2.6) and (2.6′), the constant matrices A_ν, A'_ν are all real symmetric nonnegative definite, and the constants γ_ν and γ'_ν are real and positive. Note that in the scalar case (i.e., if $p = 1$) (2.6) or (2.6′) can thus be interpreted as the input impedance of a circuit, as shown in Fig. 2, where $A_i = R_i$ and $\gamma_i = R_iC_i$. Similar interpretations are possible for $p > 1$. Expanding the right-hand side of (2.6) and (2.6′) in a power series around $s = 0$, and recalling the fact that the first $2m$ terms of the expansion for $[m - 1/m](s)$ and the

[2] Superscript t denotes the transpose of a real matrix.

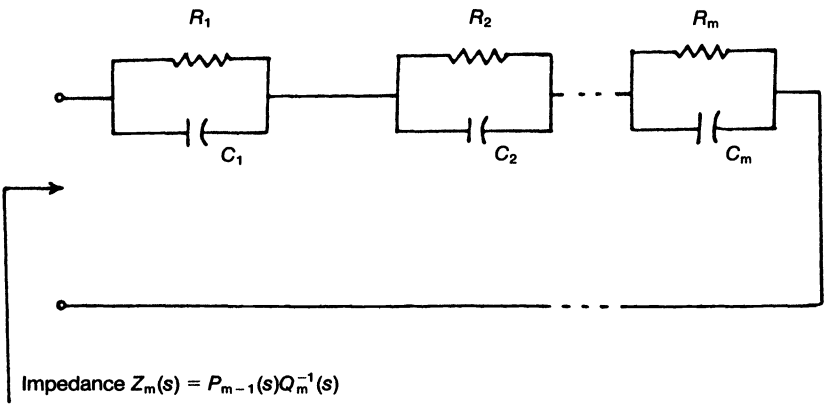


FIG. 2. $Z_m(s) < R_1 + R_2 + \cdots + R_m$ *if* $\operatorname{Im} s = 0, \operatorname{Re} s > 0$.

first $(2m - 1)$ terms in the expansion for $[m - 1/m - 1](s)$ must be identical with the given power series $T(s)$ in (1.1), it respectively follows that

$$T_k = \sum_{\nu=1}^{r} (-\gamma_\nu)^k A_\nu \quad \text{for } k = 0, 1, \cdots, 2m-1, \tag{2.7}$$

$$T_0 = \sum_{\nu=0}^{r'} A'_\nu, \tag{2.7'a}$$

$$T_k = \sum_{\nu=1}^{r'} (-\gamma_\nu)^k A'_\nu \quad \text{for } k = 1, 2, \cdots, 2(m-1). \tag{2.7'b}$$

We next state Theorem 2.2, which follows from the representations (2.6) and (2.6′) associated with the MPAs of respective orders $[m - 1/m](s)$ and $[m - 1/m - 1](s)$ to a matrix Stieltjes series.

If only $H_n(T)$ (but not $-H'_n(T)$) is positive definite for all n, then representations (2.6) and (2.6′), and thus (2.7) and (2.7′), still hold true. However, γ_ν's may then assume positive as well as negative real values.

THEOREM 2.2. *For all s with* $\operatorname{Re} s > 0$ *and* $\operatorname{Im} s = 0$, *the sequences of norms* $\|[m - 1/m](s)\|$ *and* $\|[m - 1/m - 1](s)\|$, $m = 1, 2, \cdots$, *of* MPA*s to a matrix Stieltjes series each possesses a uniform upper bound.*

Proof. Since $\gamma_\nu > 0$ and $\gamma'_\nu > 0$, we have $|(1 + \gamma_\nu s)^{-1}| < 1$ and $|(1 + \gamma'_\nu s)^{-1}| < 1$ for all real and positive values of s. Therefore, it follows from (2.6), (2.6′), Property 2.4 of spectral norm $\|\cdot\|$, and the triangle inequality for the spectral norm [4] that (2.8) and (2.8′), respectively, hold true for all real and positive values of s, and for each $m = 1, 2, \cdots$

$$\|[m-1/m](s)\| \leqq \left\| \sum_{\nu=1}^{r} A_\nu \right\|, \tag{2.8}$$

$$\|[m-1/m-1](s)\| \leqq \left\| \sum_{\nu=0}^{r'} A'_\nu \right\|. \tag{2.8'}$$

Furthermore, by considering (2.7) and (2.7′) with $k = 0$, it immediately follows from

(2.8) and (2.8′) that for Re $s > 0$, Im $s = 0$ and each $m = 1, 2, \cdots$, (2.9) and (2.9′) in the following hold:

$$\|[m-1/m](s)\| \leqq \|T_0\|, \tag{2.9}$$

$$\|[m-1/m-1](s)\| \leqq \|T_0\|. \tag{2.9′}$$

The fact that the sequences $\|[m-1/m](s)\|$ and $\|[m-1/m-1](s)\|$ for $m = 1, 2, \cdots$, each have uniform upper bounds for real positive s has, therefore, been established. □

In the scalar case, i.e., if $p = 1$ Theorem 2.2 admits an obvious physical interpretation when $[m-1/m](s)$ or $[m-1/m-1](s)$ is viewed as an impedance of the RC circuit, as in Fig. 2 (for $p = 1$), or equivalently, as in (2.6) or (2.6′). The uniform upper bound on the approximants is then provided by the sum of all resistors in the network. Similar interpretations are also possible when $p > 1$.

THEOREM 2.3. *For all real positive value of s, the sequences* $[m-1/m](s)$, $m = 1, 2, \cdots$, *as well as* $[m-1/m-1](s)$, $m = 1, 2, \cdots$, *of* MPA*s to a matrix Stieltjes series converge pointwise.*

Proof. Since a (strictly) monotone (increasing or decreasing), bounded sequence of real numbers necessarily converge, it follows from Corollary 2.1.1 and Theorem 2.2 that the sequences $\|[m-1/m](s)\|$, $m = 1, 2, \cdots$, and $\|[m-1/m-1](s)\|$, $m = 1, 2, \cdots$ are convergent for all real positive values of s. The required result then follows by noting [4] that convergence of the sequence of norms $\|[m-1/m](s)\|$, $m = 1, 2, \cdots$, is a sufficient condition for the matrix sequence $[m-1/m](s)$, $m = 1, 2, \cdots$, to converge. Similar arguments hold for $[m-1/m-1](s)$, $m = 1, 2, \cdots$. □

COROLLARY 2.3.1. *For any* i, j *with* $1 \leqq i, j \leqq p$ *and for any real positive value of s, the sequences* $[m-1/m]_{ij}(s)$ *and* $[m-1/m-1]_{ij}(s)$ $m = 1, 2, \cdots$ *of* ijth *entries of* MPA*s of respective orders* $[m/m-1]$ *and* $[m-1/m-1]$ *to a matrix Stieltjes series converge pointwise.*

Our next objective is to enlarge the domain of convergence of the sequence of approximants under consideration to a region $D(\Delta)$ larger than the positive real axis, where $D(\Delta)$ is any bounded region of the complex plane, which is at least at a distance Δ away from the negative real axis. The region $D(\Delta)$ is shown in Fig. 3.

We first need the following lemma.

LEMMA 2.4. *Assuming that* $T(s)$ *is a matrix Stieltjes series, if* $A_\nu^{(ij)}$, $A_\nu'^{(ij)}$ *are the respective* ijth *elements of the matrices* A_ν, A'_ν, *and* γ_ν, γ'_ν *are positive numbers as appearing in* (2.6) *and* (2.6′), *then the following inequalities hold true*:

$$\sum_{\nu=1}^{r} |A_\nu^{(ij)}| \leqq \sqrt{(T_0^{(ii)}T_0^{(jj)})}, \tag{2.10a}$$

$$\sum_{\nu=1}^{r} |A_\nu^{(ij)}|\gamma_\nu \leqq \sqrt{(T_1^{(ii)}T_1^{(jj)})}, \tag{2.10b}$$

$$\sum_{\nu=0}^{r'} |A_\nu'^{(ij)}| \leqq \sqrt{(T_0^{(ii)}T_0^{(jj)})}, \tag{2.10′a}$$

$$\sum_{\nu=0}^{r'} |A_\nu'^{(ij)}|\gamma_\nu \leqq \sqrt{(T_1^{(ii)}T_1^{(jj)})}. \tag{2.10′b}$$

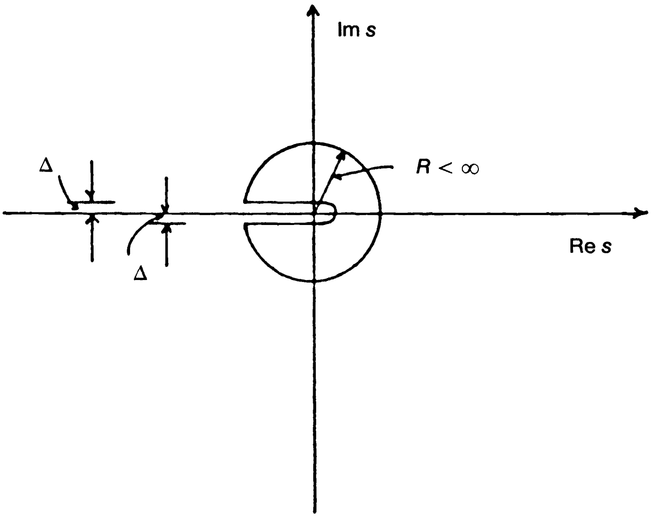


FIG. 3. *Region* $D(\Delta)$.

Proof. Only proofs for (2.10a) and (2.10b) will be given. Analogous proofs hold for (2.10′a) and (2.10′b). Consider the case $i = j$ first. Since the A_ν are real symmetric nonnegative-definite matrices, the diagonal elements $A_\nu^{(ii)}$ are necessarily nonnegative. Furthermore, considering the iith elements of the matrices in (2.7) with $k = 0$, and $k = 1$, we obtain, respectively, (2.11a) and (2.11b):

(2.11a) $$\sum_{\nu=1}^{r} |A_\nu^{(ii)}| = \sum_{\nu=1}^{r} A_\nu^{(ii)} = T_0^{(ii)},$$

(2.11b) $$\sum_{\nu=1}^{r} |A_\nu^{(ii)}|\gamma_\nu = \sum_{\nu=1}^{r} A_\nu^{(ii)}\gamma_\nu = |T_1^{(ii)}|.$$

Next, when $i \neq j$, the (2×2) principal minor of A_ν obtained by considering the ith row and jth column of A_ν are also nonnegative definite. Thus, it follows that $|A_\nu^{(ij)}| \leqq ((A_\nu^{(ii)}A_\nu^{(jj)}))^{1/2}$. The last inequality, along with an application of the well-known Cauchy–Schwartz inequality, yields (2.12a) and (2.12b):

(2.12a) $$\sum_{\nu=1}^{r} |A_\nu^{(ij)}| \leqq \sum_{\nu=1}^{r} \sqrt{(A_\nu^{(ii)}A_\nu^{(jj)})} \leqq \sqrt{\left[\left(\sum_{\nu=1}^{r} A_\nu^{(ii)}\right)\left(\sum_{\nu=1}^{r} A_\nu^{(jj)}\right)\right]},$$

(2.12b) $$\sum_{\nu=1}^{r} |A_\nu^{(ij)}|\gamma_\nu \leqq \sum_{\nu=1}^{r} \sqrt{(A_\nu^{(ii)}A_\nu^{(jj)}\gamma_\nu^2)} \leqq \sqrt{\left[\left(\sum_{\nu=1}^{r} A_\nu^{(ii)}\gamma_\nu\right)\left(\sum_{\nu=1}^{r} A_\nu^{(jj)}\gamma_\nu\right)\right]}.$$

Using (2.7) with $k = 0$ and $k = 1$, it easily follows that the right-hand sides of (2.12a) and (2.12b), respectively, are equal to $((T_0^{(ii)}T_0^{(jj)}))^{1/2}$ and $((T_1^{(ii)}T_1^{(jj)}))^{1/2}$. The inequalities (2.10a) and (2.10b) are thus established. □

THEOREM 2.5. *Each element of the sequences of* MPAs $[m - 1/m](s)$ *and* $[m - 1/m - 1](s)$ *for* $m = 1, 2, \cdots$ *to a matrix Stieltjes series is uniformly bounded in* $D(\Delta)$.

Proof. We first prove the result for the region $\{s;\ \operatorname{Re} s \geqq 0,\ s \in D(\Delta)\}$. Consider the ijth element $[m - 1/m]_{ij}(s)$ and $[m - 1/m - 1]_{ij}(s)$ of $[m - 1/m](s)$ and $[m - 1/m - 1](s)$, respectively. From (2.6) and (2.6′), respectively, it follows that

$$[m - 1/m]_{ij}(s) = \sum_{\nu=1}^{r} \frac{1}{1+\gamma_\nu s} A_\nu^{(ij)}, \tag{2.13}$$

$$[m - 1/m - 1]_{ij}(s) = A_0^{(ij)} + \sum_{\nu=1}^{r'} \frac{1}{1+\gamma'_\nu s} A_\nu'^{(ij)}. \tag{2.13′}$$

Clearly, since $\gamma_\nu, \gamma'_\nu > 0$, we have for $\operatorname{Re} s \geqq 0$ and arbitrary $\operatorname{Im} s$ that $|1 + \gamma_\nu s| \geqq 1$, $|1 + \gamma'_\nu s| \geqq 1$ for all ν. By making use of the last inequalities along with (2.13) and (2.13′), (2.14) and (2.14′), respectively, are obtained via the use of triangle inequality:

$$|[m - 1/m]_{ij}(s)| \leqq \sum_{\nu=1}^{r} \left|\frac{1}{1+\gamma_\nu s}\right| |A_\nu^{(ij)}| \leqq \sum_{\nu=1}^{r} |A_\nu^{(ij)}|, \tag{2.14}$$

$$|[m - 1/m - 1]_{ij}(s)| \leqq |A_0^{(ij)}| + \sum_{\nu=1}^{r'} \frac{1}{|1+\gamma'_\nu s|} |A_\nu^{(ij)}| = \sum_{\nu=0}^{r'} |A_\nu'^{(ij)}|. \tag{2.14′}$$

It then follows from (2.14) and (2.14′), via the use of (2.10a) and (2.10′a) in Lemma 2.4, that for $\operatorname{Re} s \geqq 0$ and for each $m = 1, 2, \cdots$ $|[m - 1/m]_{ij}(s)| \leqq ((T_0^{(ii)} T_0^{(jj)}))^{1/2}$ and $|[m - 1/m - 1]_{ij}(s)| \leqq ((T_0^{(ii)} T_0^{(jj)}))^{1/2}$. Thus the theorem has been proved, in particular, for all s in $\{s;\ \operatorname{Re} s \geqq 0,\ s \in D(\Delta)\}$.

In the following equations we consider values of s in the region $\{s;\ \operatorname{Re} s < 0,\ s \in D(\Delta)\}$. Note first that since the identity $|s|^2 |1 + \gamma s|^2 - |\operatorname{Im} s|^2 = (\gamma |s|^2 + \operatorname{Re} s)^2$ holds for any real γ, we have $|1 + \gamma_\nu s| \geqq |\operatorname{Im} s|/|s| < 1$ and $|1 + \gamma'_\nu s| \geqq |\operatorname{Im} s|/|s| < 1$ for each ν. Consequently, (2.15) and (2.15′) follow from (2.6) and (2.6′), respectively, for each $i, j = 1, 2, \cdots, p$:

$$|[m - 1/m]_{ij}(s)| \leqq \sum_{\nu=1}^{r} \frac{1}{|1+\gamma_\nu s|} |A^{(ij)}| \leqq \left(|s| \sum_{\nu=1}^{r} |A_\nu^{(ij)}|\right) \Big/ |\operatorname{Im} s|, \tag{2.15}$$

$$\begin{aligned} |[m - 1/m - 1]_{ij}(s)| &\leqq |A_0^{(ij)}| + \sum_{\nu=1}^{r'} \frac{1}{|1+\gamma'_\nu s|} |A_\nu'^{(ij)}| \\ &\leqq \left(|s| \sum_{\nu=0}^{r'} |A_\nu'^{(ij)}|\right) \Big/ |\operatorname{Im} s|. \end{aligned} \tag{2.15′}$$

Again invoking (2.10a) and (2.10′a) of Lemma 2.4 along with (2.15) and (2.15′), respectively, it follows that for all s with $\operatorname{Re} s < 0$ we have that

$$|[m - 1/m]_{ij}(s)| \leqq |s| \sqrt{(T_0^{(ii)} T_0^{(ij)})} / |\operatorname{Im} s|, \tag{2.16}$$

$$|[m - 1/m - 1]_{ij}(s)| \leqq |s| \sqrt{(T_0^{(ii)} T_0^{(ij)})} / |\operatorname{Im} s|. \tag{2.16′}$$

If $R_M < \infty$ is the radius of a circle, which completely encloses $D(\Delta)$ in the complex plane, then (2.16) and (2.16′) establish a uniform upper bound of $(R_M((T_0^{(ii)} T_0^{(ij)}))^{1/2}/\Delta)$ for the sequences $|[m - 1/m]_{ij}(s)|$ and $|[m - 1/m - 1]_{ij}(s)|$, $m = 1, 2, \cdots$ in $\{s;\ \operatorname{Re} s < 0,\ s \in D(\Delta)\}$. If $M = \max(R_M/\Delta, 1)$, then due to the result proved in the last paragraph, $M((T_0^{(ii)} T_0^{(ij)}))^{1/2}$ serves as a uniform upper bound on each of the sequences $|[m - 1/m]_{ij}(s)|$ and $|[m - 1/m - 1]_{ij}(s)|$, $m = 1, 2, \cdots$ in $D(\Delta)$. The theorem is thus proved. □

We are now in a position to prove the convergence of the sequences of rational matrices $[m-1/m](s)$ and $[m-1/m-1](s)$, $m=1, 2, \cdots$, by using standard techniques from complex function theory [5]. The result is summarized in the following theorem.

THEOREM 2.6. *The sequences of* MPAs *of order* $[m-1/m](s)$, $m=1, 2, \cdots$, *and* $[m-1/m-1](s)$, $m=1, 2, \cdots$, *to a matrix Stieltjes series converge uniformly in the region* $D(\Delta)$ *of the complex plane. Furthermore, the matrix-valued functions* $G(s)$ *and* $G'(s)$, *to which the two sequences, respectively, converge are both real symmetric* (*i.e.*, $\bar{G}(s)=G(\bar{s})$, $\bar{G}'(s)=G'(\bar{s})$)[3] *and analytic in* $D(\Delta)$.

Proof. The following discussion will be only in terms of the sequence

$$[m-1/m](s),$$

$m=1, 2, \cdots$. Analogous arguments hold for the sequence $[m-1/m-1](s)$, $m=1, 2, \cdots$. We shall establish the convergence of $[m-1/m](s)$, $m=1, 2, \cdots$, by showing that the sequence of ijth elements $[m-1/m]_{ij}(s)$, $m=1, 2, \cdots$, of $[m-1/m](s)$ converge.

Let $D(\Delta')$ be a region similar to $D(\Delta)$ but slightly larger and containing the closure of $D(\Delta)$. Then since $[m-1/m]_{ij}(s)$ is uniformly bounded in the closure of $D(\Delta')$, each subsequence of $[m-1/m]_{ij}(s)$ is normal in $D(\Delta')$, and thus contains another subsequence converging locally uniformly to some analytic function $D(\Delta')$. Since all these limit functions are the same on positive real axis, they are identical in $D(\Delta')$ due to analytic continuation. Hence, $[m-1/m]_{ij}(s)$ converges locally uniformly in $D(\Delta')$ to a limit function analytic on $D(\Delta')$, and in particular converges uniformly on $D(\Delta)$.

Finally, since $G(s)$ is holomorphic in $D(\Delta)$, which is symmetric with respect to the real axis, and the property of realness of $G(s)$ is inherited by the property of realness of $[m-1/m]_{ij}(s)$ for real values of s, it follows from the well-known Schwartz reflection principle that $G(s)$ is real symmetric, i.e., $\bar{G}(s)=G(\bar{s})$. □

Note that if $T(s)$ is a Hamburger series, i.e., if only $H_n(T)$ (but not $-H'_n(T)$) is positive definite for arbitrary n, then property of uniform boundedness, as proved in Theorems 2.5 and 2.6, still holds true when $D(\Delta)$ is replaced by the bounded, disconnected, two-component domain

$$D_I(\Delta)=\{s;\ |\mathrm{Im}\ s|>\Delta,\ |s|<R<\infty\}.$$

Consequently, the first paragraph in the proof of Theorem 2.6 applies, and we may assert that there exists a subsequence of the sequence of MPAs that converge uniformly everywhere in $D_I(\Delta)$ to a real symmetric function analytic in $D_I(\Delta)$. However, since in this case γ_ν's are not necessarily positive, Theorems 2.2 and 2.3 do not apply and consequently, the pointwise convergence of the sequence of MPAs for real positive values of s cannot be established.

3. Matricial Hamburger and Stieltjes moment problem and related results.

In this section we undertake the solution of the matricial version of the classical Hamburger or Stieltjes moment problem. More specifically, the following result, stated in Theorem 3.1, will be proved. An integral representation of the functions $G(s)$ and $G'(s)$ of Theorem 2.6, when the Stieltjes series (1.1) has a nonzero radius of convergence is also derived in this connection.

[3] The bar "‾" denotes complex conjugate.

We first need the following definition.

DEFINITION. A real symmetric matrix-valued function $\sigma(x)$ of a real variable x will be said to be *nondecreasing* (increasing) if the matrix $\sigma(x_1) - \sigma(x_2)$ is nonnegative (positive) definite, whenever $x_1 > x_2$.

THEOREM 3.1. (a) *If the block Hankel matrices $H_n(T)$ in* (1.2a) *are positive definite for all nonnegative integer values of n, then there exists a nondecreasing matrix measure $\sigma(x)$ such that the matricial Stieltjes integral representation* (3.1) *for T_k holds true*:

$$(3.1) \qquad (-1)^k T_k = \int_{-\infty}^{\infty} x^k \, d\sigma(x), \qquad k=0,1,2,\cdots.$$

(b) *Furthermore, if in addition to the conditions stated in part* (a) *the block Hankel matrices $H_n'(T)$ in* (1.2b) *are negative definite for all nonnegative integer values of n, i.e., if $T(s)$ as in* (1.1) *is a matrix Stieltjes series then the lower limit of the integral in* (3.1) *can be replaced by zero.*

Note that the Riemann–Stieltjes integral over a matrix measure, as in (3.1), was first introduced and their properties studied by Wiener and Masani in [7] in the context of multivariate stochastic process.

Before embarking on a proof of Theorem 3.1, the matricial version of Gauss quadrature formula proved in [1, Thm. 3.3] will be recalled in a notation compatible with the present discussion.

THEOREM 3.2 [1]. *If $H_n(T)$ is positive definite for all nonnegative integer values of n, then for any fixed integer $m > 0$, there exist real symmetric nonnegative-definite $(p \times p)$ matrices A_ν and real numbers γ_ν, each depending on m*[4], *such that*

$$(3.2) \qquad (-1)^k T_k = \sum_{\nu=1}^{r} A_\nu \gamma_\nu^k \quad \textit{for } k=0,1,\cdots(2m-1)$$

where $r = mp$. Furthermore, if $H_n'(T)$ is negative definite for all n, then the γ_ν's are necessarily positive.

Note that when both $H_n(T)$ and $(-H_n'(T))$ are positive definite, (3.2) follows from (2.6) and then by observing that the coefficient of s^k in the power series expansion of $[m - 1/m](s)$ around $s = 0$ is T_k for $k = 0, 1, \cdots, (2m - 1)$, thus establishing the matricial Gauss quadrature formula (3.2) via electrical network theoretic arguments (in fact, (3.2) is identical to (2.7)). However, when only $H_n(T)$ but not $(-H_n'(T))$ is positive definite, the network interpretations of $[m - 1/m](s)$ in (2.6) cannot be given and a detailed proof of (3.2) as worked out in [1] is called for.

DEFINITION [7]. A matrix-valued function $\sigma(x)$ will be said to be of *bounded variation* in $[a, b]$ if $\sum_{\nu=1}^{k} \|\sigma(x_\nu) - \sigma(x_{\nu-1})\|$ is bounded for any partition $a = x_0 < x_1 \cdots < x_k = b$ of the interval $[a, b]$.

LEMMA 3.3. *If $\sigma(x)$ is a nondecreasing real symmetric matrix-valued function such that $M \geqq \sigma(x) \geqq 0$ for all $x \in [a, b]$, then each element of $\sigma(x)$ is bounded for all $x \in [a, b]$. Furthermore, $\sigma(x)$ as well as each of its entries are of bounded variation in $[a, b]$.*

Proof. Let $\sigma^{(ij)}(x)$ denote the ijth entry of the matrix $\sigma(x)$. Since $M \geqq \sigma(x) \geqq 0$, it follows from Property 2.3 of the spectral norm that $\|\sigma(x)\| \leqq \|M\|$ for all $x \in [a, b]$. If e_j denotes the jth column of the $(p \times p)$ identity matrix, then

$$\left(\sum_{i=1}^{p} |\sigma^{(ij)}(x)|^2 \right)^{1/2} = \|\sigma(x)e_j\| \leqq \|\sigma(x)\| \leqq \|M\|.$$

[4] To avoid clutter in notation this dependence is not reflected explicitly in (3.2).

Consequently, $|\sigma^{(ij)}(x)| < \|M\|$ for all $i = 1, 2, \cdots, p$. Since j is chosen arbitrarily, each element of the matrix $\sigma(x)$ is bounded by $\|M\|$. This result along with the nondecreasing character of $\sigma(x)$ implies that [7, Lemma 4.2(b)] the functions $\sigma^{(ij)}(x)$, $i \neq j$ are each functions of bounded variation. However, for all j, $\sigma^{(jj)}(x)$ is nondecreasing since $\sigma(x)$ is so, and furthermore $0 \leqq \sigma^{(jj)}(x) \leqq \|M\|$ (the first inequality follows from nonnegativeness of $\sigma(x)$). Thus, $\sigma^{(jj)}(x)$ is also of bounded variation in $[a, b]$. Consequently, each entry of $\sigma(x)$ is of bounded variation in $[a, b]$, which is a necessary and sufficient condition for $\sigma(x)$ to be of bounded variation in $[a, b]$ (cf. [7, Lemma 4.2(a)]). □

Remark. We note that if $\sigma(x)$ is a real symmetric matrix-valued function of bounded variation in $[a, b]$, then due to Lemma 4.2(a) of [7] $\sigma^{(ij)}(x)$ is of bounded variation for all i, j. Consequently, if $f(x)$ is any continuous function in $[a, b]$ then $\int_a^b f(x)\, d\sigma^{(ij)}(x)$ exists [9] and, consequently, due to Lemma 4.8 of [7] the matricial Riemann–Stieltjes integral $\int_a^b f(x)\, d\sigma(x)$ also exists.

LEMMA 3.4. *If $\sigma(x)$ is any real symmetric nondecreasing matrix-valued function of bounded variation in $[a, b]$, then*

$$\left\| \int_a^b f(x)\, d\sigma(x) \right\| \geqq \left\| \int_a^b g(x)\, d\sigma(x) \right\| \tag{3.3}$$

where $f(x)$ and $g(x)$ are continuous scalar functions such that $f(x) \geqq g(x) \geqq 0$ for all x in the interval of integration.

Proof. Consider the function $h(x) = f(x) - g(x)$ defined in $[a, b]$. The existence of the integrals in (3.3) and of $\int_a^b h(x)\, d\sigma(x)$ then immediately follow from the remark preceding the present lemma. Furthermore, we also have

$$\int_a^b h(x)\, d\sigma(x) = \int_a^b f(x)\, d\sigma(x) - \int_a^b g(x)\, d\sigma(x). \tag{3.4}$$

Since $f(x) \geqq g(x) \geqq 0$, the functions $f(x)$, $g(x)$, and $h(x)$ are all nonnegative in $[a, b]$. Consequently, due to the nondecreasing character of the matrix-valued measure $\sigma(x)$, it trivially follows from the definition of the matricial Riemann–Stieltjes integrals that each of the integrals in (3.4) is real symmetric nonnegative definite. The present lemma then follows from Property 2.3 of spectral norm. □

Our strategy for proof of Theorem 3.1 is, in fact, a matricial generalization of a technique elaborated in [6] in the scalar context.

Proof of Theorem 3.1. Let $\gamma_1 \leqq \gamma_2 \leqq \cdots \leqq \gamma_r$ be an ordering of the γ_ν's in (3.2) and consider the real symmetric nonnegative definite matrix-valued function $\sigma_m(x)$ defined over $-\infty < x < +\infty$ as in (3.5):

$$\begin{aligned} \sigma_m(x) &= 0 \quad \text{for } x < \gamma_1 \\ &= \sum_{\nu=1}^{\mu} A_\nu \quad \text{for } \gamma_\mu \leqq x < \gamma_{\mu+1} \\ &= \sum_{\nu=1}^{r} A_\nu \quad \text{for } x \geqq \gamma_r. \end{aligned} \tag{3.5}$$

Then the following properties of $\sigma_m(x)$ are clear.

(P1) $\sigma_m(x)$ is real symmetric nonnegative definite for all real x. Furthermore, if $x_1 > x_2$ then $\sigma_m(x_1) \geqq \sigma_m(x_2)$, i.e., $\sigma_m(x)$ is a nondecreasing matrix-valued function of x.

(P2) From (3.2) with $k = 0$ and (3.5) it follows that $(T_0 - \sigma_m(x))$ is real symmetric nonnegative definite for all real x and for all $m > 0$.

(P3) Since due to (P1), (P2), and Lemma 3.3, $\sigma(x)$ is of bounded variation, the matricial Riemann-Stieltjes integral $\int_{-\infty}^{\infty} x_k \, d\sigma_m(x)$ exists elementwise [9] and is equal to $\sum_{\nu=1}^{r} \gamma_\nu^k A_\nu$, which, due to (3.2), is equal to $(-1)^k T_k$ for $k = 0, 1, \cdots (2m - 1)$.

It thus follows from (P1) and (P2) above and from Theorem A1 in the Appendix that a subsequence $\sigma_{m_i}(x)$; $i = 1, 2, \cdots$ of the sequence $\sigma_m(x)$, $m = 1, 2, \cdots$ converges to a real symmetric nonnegative-definite matrix-valued function $\sigma(x)$, which is also nondecreasing. Therefore, it follows from (P3) above that

$$(-1)^k T_k = \int_{-\infty}^{\infty} x^k \, d\sigma_{m_i}(x) \quad \text{for all } m_i \geqq \frac{1}{2}(k+1). \tag{3.6}$$

The following considerations then hold for any choice of finite real numbers a, b with $a < -1$, $1 < b$ and $m_i \geqq k + 1$.

Thus, from (3.6) and the triangle inequality for spectral norm [4], we have

$$\begin{aligned} \left\| (-1)^k T_k - \int_a^b x^k \, d\sigma(x) \right\| &= \left\| \int_{-\infty}^{+\infty} x^k \, d\sigma_{m_i}(x) - \int_a^b x^k \, d\sigma(x) \right\| \\ &\leqq \left\| \int_{-\infty}^{a} x^k \, d\sigma_{m_i}(x) \right\| + \left\| \int_a^b x^k \, d\sigma_{m_i}(x) - \int_a^b x^k \, d\sigma(x) \right\| \\ &\quad + \left\| \int_b^{\infty} x^k \, d\sigma_{m_i}(x) \right\| . \end{aligned} \tag{3.7}$$

However,

$$\begin{aligned} \left\| \int_{-\infty}^{a} x^k \, d\sigma_{m_i}(x) \right\| &= \left\| (-1)^k \int_{-\infty}^{a} |x^k| \, d\sigma_{m_i}(x) \right\| \\ &= \left\| \int_{-\infty}^{a} |x|^k \, d\sigma_{m_i}(x) \right\| \\ &\leqq \left\| \frac{1}{|a|^{k+2}} \int_{-\infty}^{a} x^{2k+2} \, d\sigma_{m_i}(x) \right\| \\ &\leqq \frac{1}{|a|^{k+2}} \left\| \int_{-\infty}^{+\infty} x^{2k+2} \, d\sigma_{m_i}(x) \right\| \\ &= \frac{1}{|a|^{k+2}} \|T_{2k+2}\| \end{aligned} \tag{3.8}$$

where in (3.8) the first equality follows from the fact that for $x < 0$, $x^k = (-1)^k |x|^k$; the second equality from Property 2.2 of $\|\cdot\|$ with $\alpha = -1$; whereas the first inequality follows from the fact that if $x \leqq a < -1$ then $x^{2k+2}/|a|^{k+2} \geqq |x|^k > 0$ in conjunction with Lemma 3.4; the second inequality from Properties 2.2 and 2.3 of spectral norm $\|\cdot\|$; and the last equality from (3.6) above as a consequence of the choice $m_i \geqq k + 1$. It can also be shown in an analogous fashion that

$$\left\| \int_b^{\infty} x^k \, d\sigma_{m_i}(x) \right\| \leqq \frac{1}{|b|^{k+2}} \|T_{2k+2}\| . \tag{3.9}$$

Thus, from (3.7) it is possible to assert that (3.10) in the following holds for $a < -1$, $1 < b$, and $m_i \geqq (k+1)$:

$$\begin{aligned} &\left\| (-1)^k T_k - \int_a^b x^k \, d\sigma(x) \right\| \\ &\quad \leqq \left\| \int_a^b x^k \, d\sigma_{m_i}(x) - \int_a^b x^k \, d\sigma(x) \right\| + \|T_{2k+2}\| \left(|a|^{-(k+2)} + |b|^{-(k+2)} \right). \end{aligned} \tag{3.10}$$

From Theorem A2 in the Appendix it follows that the first terms in the right-hand side of (3.10) goes to zero as $m_i \to \infty$. Thus, for all $k = 0, 1, 2, \cdots$,

$$\left\| (-1)^k T_k - \int_a^b x^k \, d\sigma(x) \right\| \leqq \|T_{2k+2}\| \left(|a|^{-(k+2)} + |b|^{-(k+2)} \right). \tag{3.11}$$

Furthermore, as $a \to -\infty$ and $b \to \infty$, (3.11) yields $\|(-1)^k T_k - \int_{-\infty}^{\infty} x^k \, d\sigma(x)\| = 0$, thus [4] proving that (3.1) holds for all $k = 0, 1, 2, \cdots$.

Part (b) of the theorem follows by observing that in Theorem 3.2 if $H_n'(T)$ is negative definite for all n, then γ_ν's are necessarily positive, which in turn implies that $\sigma_m(x)$, as defined in (3.5), and thus $\sigma(x)$, is zero for all negative x. □

Note that if $U_k = (-1)^k T_k$ and $H_n(U)$ and $H_n'(U)$ are the Hankel matrices obtained by replacing the T_k's in (1.2a), (1.2b) by the corresponding U_k's, then it follows via straightforward algebraic manipulation that $H_n(T) > 0$ if and only if $H_n(U) > 0$, whereas $H_n'(T) < 0$ if and only if $H_n'(U) > 0$. The solutions to matricial versions of Hamburger and Stieltjes moment problems then follow in a more conventional form as stated in [10] in the scalar case from this observation.

Note that the following result can be viewed as a matricial generalization of the well-known scalar result [10] that the nondecreasing Stieltjes measure $\sigma(x)$ must, in fact, have infinitely many points of increase.

PROPERTY 3.5. *For any nonzero* $(1 \times p)$ *real constant vector* v, *the function* $v^t\sigma(x)v$ *of* x *must have infinite number of points of increase.*

Proof. Assume that the result is false, i.e., there exists some v such that $v^t\sigma(x)v$ can be viewed as a linear combination of a finite number N of step functions, occurring at, say, $\alpha_1, \alpha_2, \cdots, \alpha_N$. Consider the polynomial $p(x)$, as in (3.12a). Then (3.12b) follows from (3.1) and (1.2a):

$$p(x) = \prod_{i=1}^{N} (x - \alpha_i) = \sum_{k=0}^{N} a_k x^k, \tag{3.12a}$$

$$\int_{-\infty}^{\infty} p^2(x) \, d\sigma(x) = A_N^t H_{N+1}(T) A_N \tag{3.12b}$$

where A_N^t is the block row matrix $(a_0 I, a_1 I, \cdots, a_N I)$ and I is the $(p \times p)$ identity matrix. Since not all a_i's are zero, A_N is of rank p, thus implying, in view of positive definiteness of $H_{N+1}(T)$, that

$$v^t \left(\int_{-\infty}^{\infty} p^2(x) \, d\sigma(x) \right) v > 0. \tag{3.13}$$

However, by recalling the definition of Riemann–Stieltjes integrals over a matrix measure, it follows from the fact that $v^t\sigma(x)v$ is a linear combination of step functions that the left-hand side of (3.13) is exactly equal to zero, which is a contradiction. □

We next assume that the matrix Stieltjes series $T(s)$ in (1.1), which so far has been considered only as a formal power series, to have a radius of convergence R. Then by using the representation (3.1) we can further prove the following theorem.

THEOREM 3.6. *If $T(s)$ in* (1.1) *has a nonzero radius of convergence R and the associated Hankel matrix $H_n(T)$ satisfies $H_n(T) > 0$ for all n, then*

(i) *$\sigma(x) =$ constant for $|x| > R^{-1}$;*

(ii) *for all s in $|s| < R$ we may write*

$$T(s)=\int_{-\infty}^{\infty}\frac{1}{1+sx}\,d\sigma(x)=\int_{-R^{-1}}^{R^{-1}}\frac{1}{1+sx}\,d\sigma(x); \tag{3.14}$$

(iii) *The sequences of* MPA*s* $[m-1/m](s)$ *and* $[m-1/m-1](s)$, $m=1, 2, \cdots$, *converge uniformly to the expression* (3.14). *In particular, if $T(s)$ is a matrix Stieltjes series, then the limit functions $G(s)$ and $G'(s)$ of Theorem* 2.6 *are both given by* (3.14);

(iv) *If, in addition, $H'_n(T) < 0$ for all n, i.e., $T(s)$ is a matrix Stieltjes series, then the lower limits in the integrals in* (3.11) *can be replaced by zero.*

Note that in the last case (3.14) coincides with the integral representation of RC impedances known as Cauer's representation in classical network theory [20].

Proof of Theorem 3.6 uses Proposition 4.1, which, however, has been included in § 4 for an improved categorization of results of similar nature.

Proof. (i) Let $P_m(s)$ be the denominator polynomial matrix associated with the right MPA of order $[m-1/m]$ to $T(s)$, and $\hat{P}_m(s)$ be the corresponding "inverse" polynomial matrix as defined in (4.1). Define r_m via $r_m^{-1} = \max(|\hat{\alpha}_m|, |\hat{\beta}_m|)$, where $\hat{\alpha}_m$ and $\hat{\beta}_m$ are as described in Proposition 4.1, from which it also follows that $r_{m+1} \leqq r_m$ for all $m = 1, 2, \cdots$. Then $[m-1/m](s)$ is analytic in $|s| < r_m$ and thus, its power series expansion around $s=0$ converges in $|s| < r_m$. Furthermore, as $m \to \infty$ this latter expansion coincides with $T(s)$ in (1.1), which is assumed to have a radius of convergence R. Consequently, $R < r_m$ and thus, $|\hat{\alpha}_m| < R^{-1}$, $|\hat{\beta}_m| < R^{-1}$ for $m = 1, 2 \cdots$. Next, since for any fixed m, the γ_ν's in (2.6) are (a subset of) the zeros of $\det \hat{P}_m(s)$, the latter conclusion yields that $|\gamma_\nu| < R^{-1}$ for all ν and m. Thus, it follows from (3.5) that for all m, $\sigma_m(x) =$ constant if $|x| > R^{-1}$, which in turn imply that $\sigma(x) =$ constant if $|x| > R^{-1}$.

(ii) The following considerations hold for real-valued s with $|s| < R$. Define $\phi_n(x) = \sum_{k=0}^{n}(-sx)^k$ and $\psi(x) = (1-|sx|)^{-1}$. Clearly, then for all x in $-R^{-1} \leqq x \leqq R^{-1}$ we have $|\phi_n(x)| < \psi(x)$ and $\phi_n(x) \to (1+sx)^{-1}$ as $n \to \infty$. Furthermore, since $\psi(x)$ is continuous and $\sigma(x)$ is of bounded variation (cf. proof of Theorem 3.1) in $-R^{-1} \leqq x \leqq R^{-1}$, it follows from [7] that the matricial Riemann-Stieltjes integral of $\psi(x)$ with respect to $d\sigma(x)$ over the interval $-R^{-1} \leqq x \leqq R^{-1}$ exists. By applying the dominated convergence theorem of the theory of functions of a real variable to the sequences formed from the respective entries of matrices, it then follows that

$$\int_{-R^{-1}}^{R^{-1}}\phi_n(x)\,d\sigma(x)\to\int_{-R^{-1}}^{R^{-1}}(1+sx)^{-1}\,d\sigma(x)\quad \text{as } n\to\infty.$$

Thus, the proof of (3.14) for real values of s follows from (1.1) and (3.15), in which use of (3.1) along with the fact that $\sigma(x) =$ constant for $|x| > R^{-1}$ have been made:

$$\sum_{k=0}^{n} T_k s^k=\sum_{k=0}^{n}\int_{-R^{-1}}^{R^{-1}}(-sx)^k\,d\sigma(x)=\int_{-R^{-1}}^{R^{-1}}\phi_n(x)\,d\sigma(x). \tag{3.15}$$

The validity of (3.14) for complex values of s then follows from the principle of analytic continuation by noting that both $T(s)$ in (1.1) and the extreme right-hand side of (3.14) are analytic in $|s| < R$.

Part (iii) follows from the fact that the limit functions to which the sequences $[m-1/m](s)$ and $[m-1/m-1](s)$ of MPAs converge and $T(s)$ are each holomorphic in $|s| < R$, in which they have identical power series expansion, namely, (1.1).

(iv) Finally, if $T(s)$ is a matrix Stieltjes series, then all γ_ν's are positive; thus due to (3.5), $\sigma(x) =$ constant for $x < 0$. Consequently, the lower limit of the integrals in (3.14) can be replaced by zero. □

The following comment is in order with respect to item (iii) of the above theorem. In the scalar case it has been shown that even if $R = 0$, the integral representation (3.14) for the limit functions remains valid if the coefficients of the power series further satisfy the so-called "Carleman criterion" [3]. An extension of this result in the matrix case is not pursued here (see, e.g., [21] and references therein).

4. Matrix orthogonal polynomials of the second kind. The fact that the sequence of inverse polynomial matrices, constituting the "denominators" of MPAs to a matrix Stieltjes series form a sequence of matrix orthogonal polynomials has already been pointed out in [1]. However, in [1] the orthogonality relation was viewed as an algebraic relation, i.e., in terms of orthogonality of vector spaces. Presently, it will be shown that this relation can be interpreted as an orthogonality relation with respect to the matrix-valued measure $\sigma(x)$ developed in the previous section. Certain other results as natural generalizations of the scalar theory such as the orthogonal polynomials of the second kind and their properties follow as consequences of this discussion.

Consider the set of "inverse" polynomial matrix $\hat{P}_m(s)$, as in (4.1), where $P_m(s)$ is the "denominator" polynomial matrix associated with the $[m-1/m](s)$ MPAs for $T(s)$, and $H_n(T)$ in (1.2a) for each n is positive definite (for the purpose of the present section, no restriction is imposed on $H'_n(T)$):

$$\hat{P}_m(s) = s^m P_m(s^{-1}) \quad \text{for all } m. \tag{4.1}$$

Then the following results hold true.

PROPOSITION 4.1. *If $\hat{\alpha}_m$ and $\hat{\beta}_m$ are, respectively, the largest and smallest zeros of* $\det \hat{P}_m(s)$, *then $\hat{\alpha}_{m+1} \geqq \hat{\alpha}_m$, $\hat{\beta}_{m+1} \leqq \hat{\beta}_m$ for all $m = 1, 2, \cdots$.*

Proof. As shown in [1], the zeros of $\det \hat{P}_m(s)$ are the eigenvalues of the block tridiagonal matrix in (4.2a), where $C_k = D_k^{-1}K_kD_k$, $C_0 = D_0^{-1}T_1$, $\lambda_k = D_{k-1}^{-1}D_k$, and D_k's are real symmetric positive definite, whereas the K_kD_k's are real symmetric matrices:

$$\begin{bmatrix} C_0 & \lambda_1 & & \\ I & C_0 & \ddots & \\ & \ddots & \ddots & \lambda_{m-1} \\ & & I & C_{m-1} \end{bmatrix}, \tag{4.2a}$$

$$Z_{m-1} = \begin{bmatrix} T_1 & D_1^{1/2} & & \\ D_1^{1/2} & K_1D_1 & \ddots & \\ & \ddots & \ddots & D_{m-1}^{1/2} \\ & & D_{m-1}^{1/2} & K_{m-1}D_{m-1} \end{bmatrix}. \tag{4.2b}$$

Thus, the zeros of det $\hat{P}_m(s)$ are also eigenvalues of the real symmetric block tridiagonal matrix Z_{m-1} in (4.2b), where $D_k^{1/2}$ stands for the Hermitian square root of D_k. It then follows from the Courant–Fisher theorem [2] that

$$\hat{\alpha}_m = \max\{x^t Z_{m-1} x; \|x\| = 1\}, \tag{4.3a}$$

$$\hat{\alpha}_{m+1} = \max\{y^t Z_m y; \|y\| = 1\} \tag{4.3b}$$

where x and y are column vectors of size mp and $(m+1)p$, respectively. Since from (4.2b) we have that

$$Z_m = \left[\begin{array}{c|c} Z_{m-1} & D_m^{1/2} \\ \hline D_m^{1/2} & K_m D_m \end{array}\right] \tag{4.4}$$

it follows from (4.3a) that $\hat{\alpha}_m$ can also be considered as the maximum value of $y^t Z_m y$ subject to the restriction that $\|y\| = 1$ and that the last p elements of y are zero. Thus, $\hat{\alpha}_{m+1} \geqq \hat{\alpha}_m$. The result $\hat{\beta}_{m+1} \leqq \hat{\beta}_m$ also follows from similar arguments if $\hat{\beta}_m$ and $\hat{\beta}_{m+1}$ are expressed as the minimum values of the quadratic forms in (4.3). □

Note that in the scalar case, i.e., if $p = 1$ the above argument also leads to the interlacing property of zeros of $\hat{P}_m(s) = \det \hat{P}_m(s)$ and $\hat{P}_{m+1}(s) = \det \hat{P}_{m+1}(s)$, whereas in the matrix case interlacing properties of this type are not known to hold.

PROPOSITION 4.2. *If $H_n(T)$, as given in* (1.2a), *is positive definite for all n, then the matrical Stieltjes integral*

$$\int_{-\infty}^{\infty} \hat{P}_\mu^t(x)\, d\sigma(x) \hat{P}_\nu(x) \tag{4.5}$$

is positive definite when $\mu = \nu$ and is a zero matrix when $\mu \neq \nu$, where $\sigma(x)$ is the real symmetric nondecreasing matrix-valued function of the real variable x, as appearing in Theorem 3.1.

Proof. Let $P_m(s) = \sum_{k=0}^m p_k^{(m)} s^k$ and consequently, $\hat{P}_m(s) = \sum_{k=0}^m p_{m-k}^{(m)} s^k$, where $p_k^{(m)}$'s are real $(p \times p)$ matrices. Also, note that since $\sigma(x)$ is real symmetric and $P_\mu(x)$ as well as $P_\nu(x)$ are real-valued matrices for real x, it is enough to prove the result for $\nu \geqq \mu$. The case of $\nu < \mu$ then follows by considering the transpose of (4.5). It follows via the use of (3.1) in a straightforward manner that

$$\int_{-\infty}^{\infty} \hat{P}_\mu^t(x)\, d\sigma(x) \hat{P}_\nu(x) = [M_\mu^t \overbrace{|0| \cdots |0|}^{\nu-\mu}] H_\nu(T) M_\nu \tag{4.6}$$

where $H_\nu(T)$ is as defined in (1.2a) and M_ν is defined as the $p \times (\nu+1)p$ matrix $M_\nu = [p_\nu^{(\nu)t} | p_{\nu-1}^{(\nu)t} | \cdots | p_1^{(\nu)t} | I]^t$. However, it also follows from the normal equations (equation (3.1) in [1]) defining the right MPAs that $H_\nu(T) M_\nu = [0|0| \cdots |0| D_\nu^t]^t$, where D_ν is a real symmetric positive-definite matrix of size $(p \times p)$. Therefore, due to (4.6) we have that $\int_{-\infty}^{\infty} \hat{P}_\mu^t(x)\, d\sigma(x) \hat{P}_\nu(x)$ is equal to D_ν when $\nu = \mu$ and is equal to zero when $\nu \neq \mu$. The proposition is thus proved. □

PROPOSITION 4.3. *If $P(s)$ is any $(p \times p)$ polynomial matrix such that each of its elements are of degree strictly less than m, then*

$$\int_{-\infty}^{\infty} P(x)\, d\sigma(x) \hat{P}_m(x) = 0, \tag{4.7a}$$

$$\int_{-\infty}^{\infty} \hat{P}_m^t(x)\, d\sigma(x) P(x) = 0. \tag{4.7b}$$

Proof. Since implicit in the definition of right MPA [1] is the fact that $p_0^{(m)} = P_m(0) = I$, i.e., $\hat{P}_m(x)$ is monic for all m, it follows that $P'(s)$ can be written as

$$P'(s) = a_{m-1}\hat{P}_{m-1}(s) + a_{m-2}\hat{P}_{m-2}(s) + \cdots + a_0\hat{P}_0(s)$$

where a_i's are constant $(p \times p)$ matrices. Then (4.7a) follows from Proposition 4.2. Analogous arguments hold for (4.7b). □

Next, for all $m = 0, 1, \cdots$ define the matrix polynomial $\hat{Q}_{m-1}(s)$ of degree $(m-1)$ (where $\hat{P}(s)$ is the inverse polynomial matrix corresponding to $P_m(s)$, as given in (4.1)) via the relation

$$\hat{Q}_{m-1}(s) = \int_{-\infty}^{\infty} d\sigma(x)[(\hat{P}_m(s) - \hat{P}_m(x))/(s-x)]. \tag{4.8}$$

The following properties of $\hat{Q}_{m-1}(s)$ are then imminent.

PROPOSITION 4.4. *For any $m = 1, 2, \cdots$ if $P_m(s)$ is the denominator polynomial matrix associated with the right* MPA *of order $[m - 1/m]$ to the series $T(s)$, which satisfies the condition $H_n(T) > 0$ for all n, then $Q_{m-1}(s)$ defined via $Q_{m-1}(s) = s^{m-1}\hat{Q}_{m-1}(s^{-1})$ is, in fact, the numerator polynomial matrix of the right* MPA *of order $[m - 1/m]$ to $T(s)$. Furthermore, the identity* (4.9) *holds true for all $m = 1, 2, \cdots$*

$$\int_{-\infty}^{\infty} d\sigma(x)[(s\hat{P}_m(s) - x\hat{P}_m(x))/(s-x)] = s\hat{Q}_{m-1}(s). \tag{4.9}$$

Proof. It follows from $\hat{P}_m(s) = \sum_{k=0}^{m} p_{m-k}^{(m)} s^k$ and (4.8) that

$$\begin{aligned} \hat{Q}_{m-1}(s) &= \int_{-\infty}^{\infty} d\sigma(x) \sum_{k=0}^{m} [(s^k - x^k)/(s-k)] p_{m-k}^{(m)} \\ &= \int_{-\infty}^{\infty} d\sigma(x) \sum_{k=1}^{m} \left(\sum_{i=0}^{k-1} x^i s^{k-1-i} \right) p_{m-k}^{(m)} = \sum_{k=1}^{m} \sum_{i=0}^{k-1} T_i p_{m-k}^{(m)} s^{k-1-i} \end{aligned} \tag{4.10}$$

where the last equality follows via the use of (3.1). Furthermore, we then also have

$$Q_{m-1}(s) = s^{m-1}\hat{Q}_{m-1}(s^{-1}) = \sum_{k=1}^{m} \sum_{i=0}^{k-1} T_i p_{m-k}^{(m)} s^{m-k+i} = \sum_{j=0}^{m-1} \left(\sum_{h=0}^{j} T_h p_{j-h}^{(m)} \right) s^j \tag{4.11}$$

where the last equality follows by a straightforward rearrangement of the indices of the double sum. Since from (4.11) it follows that $Q_{m-1}(s) - T(s)P_m(s) = o(s^{2m})$, i.e., the coefficients of s^k for $k = 0, 1, \cdots, (2m-1)$ are all zero, the polynomial matrix $Q_{m-1}(s)$ is indeed the "numerator" associated with the right MPA of order $[m - 1/m]$ corresponding to the formal power series $T(s)$.

By following a sequence of steps analogous to that used in the derivation of (4.11) above, it can also be shown that

$$\begin{aligned} \int_{-\infty}^{\infty} d\sigma(x)[(s\hat{P}_m(s) - x\hat{P}_m(x))/(s-x)] &= \sum_{k=0}^{m} \sum_{i=0}^{k} T_i s^{k-i} p_{m-k}^{(m)} \\ &= \sum_{j=0}^{m} \left(\sum_{h=0}^{j} T_h p_{j-h}^{(m)} \right) s^{m-j} \end{aligned} \tag{4.12}$$

where the first equality follows from straightforward algebraic manipulation and a use of (3.1), whereas the second equality involves a rearrangement of indices of the double sum. The result in (4.9) then follows by noting that due to the normal equations [1] defining the right MPAs the term in (4.12) with $j = m$ is zero, i.e., $\sum_{h=0}^{m} T_h p_{m-h}^{(m)} = 0$ and from (4.11) $s\hat{Q}_{m-1}(s) = \sum_{j=0}^{m=1} \sum_{h=0}^{j} T_h p_{j-h}^{(m)} s^{m-j}$. □

Note that in view of the properties elaborated on in the following equation, the sequence of matrix polynomials $\hat{Q}_{m-1}(s)$, $m = 1, 2, \cdots$, can be regarded as the natural generalization of sequence of scalar polynomials of the second kind treated in the classical literature [10].

The fact that the sequence of matrix polynomials $\hat{P}_m(s)$, $m = 0, 1, 2 \cdots$, satisfies the recurrence relation (4.13) has been shown in [1], i.e., (4.13) holds for $m = 1, 2, \cdots$

$$\hat{P}_{m+1}(s) = \hat{P}_m(s)(sI - C_m) - \hat{P}_{m-1}(s)\lambda_m \tag{4.13}$$

where C_m and λ_m are real $(p \times p)$ matrices such that $C_m = D_m^{-1} K_m D_m$ and $\lambda_m = D_{m-1}^{-1} D_m$ with D_m for all m are real symmetric positive definite, and $K_m D_m$ for all m are real symmetric matrices.

PROPOSITION 4.5. *The sequence of polynomials* $\hat{Q}_m(s)$, $m = 0, 1, 2, \cdots$ *satisfies the same recurrence relations as* $\hat{P}_m(s)$. *More specifically, following three term recurrence relation holds true*:

$$\hat{Q}_{m+1}(s) = \hat{Q}_m(s)(sI - C_m) - \hat{Q}_{m-1}(s)\lambda_m, \qquad m = 0, 1, \cdots, \tag{4.14}$$

with $\hat{Q}_{-1}(s) = 0$, $\hat{Q}_0(s) = T_0$, *and* C_m *and* λ_m, *as in the context of* (4.13).

Proof. We substract (4.13) with $s = s$ from (4.13) with $s = x$. By considering the (left) Stieltjes integral of the resulting equation with respect to the matrix measure $d\sigma(x)$, the recurrence relation (4.14) follows by observing equations (4.8) and (4.9). Finally, the facts that $\hat{Q}_{-1}(s) = 0$ and $\hat{Q}_0(s) = T_0$ follow obviously from (4.8) and that $\hat{P}_0(s)$ is monic. □

The following result shows that the zeros of $\hat{Q}_{m-1}(s)$, enjoy properties similar to those of the zeros of $\hat{P}_m(s)$, as discussed in [1].

PROPOSITION 4.6. *If* $H_n(T) > 0$ *for all* n, *then*

(i) *All zeros of* det $\hat{Q}_{m-1}(s)$ *are real*;

(ii) *If* β_j *is a zero of* det $\hat{Q}_m(s)$ *of multiplicity* n, *there exists a set of exactly* n *linearly independent sets of* $(1 \times p)$*vectors* $\{v_j^1, v_j^2, \cdots, v_j^n\}$ *such that* $v_j^i \hat{Q}_m(s) = 0$, $i = 1, 2, \cdots, n$;

(iii) *Any zero of* det $\hat{Q}_{m-1}(s)$ *cannot be of multiplicity larger than* p;

(iv) *Invariant factors in the Smith canonical form for* $\hat{Q}_{m-1}(s)$ *cannot have zeros of multiple order*.

Since the proof of the above proposition is essentially a consequence of the recurrence relation (4.13) and follows in exactly the same way as that of the corresponding properties of the sequence of matrix polynomials $\hat{P}_m(s)$, as elaborated on in Theorem 3.1 and Corollaries 3.1 and 3.2 of [1], it will be omitted for the sake of brevity.

The three-term recurrence relation (4.14) connecting successive members of the "denominator" sequence of matrix polynomials, when coupled with the corresponding recurrence relation for the "numerator" sequence (4.13) discussed in [1], provides a fast recursive algorithm for computing the paradiagonal sequence of MPAs to a matrix Stieltjes series. We note that similar recursion for the problem of computing matrix Padé approximants in general has been discussed in [14]. If, in addition to $H_n(T) > 0$, we also have $H_n'(T) < 0$ for all n, i.e., $T(s)$ is a matrix Stieltjes series, then it follows from the impedance or admittance interpretation of $[m - 1/m](s)$ that zeros of det $\hat{Q}_{m-1}(s)$ are also negative.

5. Conclusion. The present work can be viewed as a continuation of [1]. While algebraic properties of the sequences of matrix Padé approximants of certain orders to a matrix Stieltjes series were investigated in [1], the present work is concerned with the relevant analytic and convergence properties of paradiagonal sequences of MPAs. Although our exposition has been in terms of the sequences $[m-1/m]$ and $[m/m]$ of MPAs, in general it is possible to derive analogous results for any paradiagonal sequence $[m+j/m]$, $j \geqq -1$. However, the network theoretic interpretations of the results are then lost.

By using network theoretic interpretations of Padé approximants to a matrix Stieltjes series of certain orders, it has been shown that the sequences of these MPAs always converge uniformly in an open bounded region of the complex plane excluding the negative real axis. Thus, a formal matrix Stieltjes series can be used to meaningfully represent a class of RC-distributed multiports in terms of an equivalent circuit. This result, which to the best of our knowledge has not appeared anywhere, is indeed interesting in view of the fact that the criteria for realizability of nonrational positive functions in terms of interconnections of (infinite number of) conventional lumped elements is not known [17].

Solutions to the matricial versions of classical Hamburger and Stieltjes moments problems are obtained, and as a consequence of this discussion an integral representation for the RC-distributed multiport impedance, which in fact is closely related to the Cauer's representation for RC-impedances, is obtained when the associated Stieltjes series is assumed to be convergent in a disc of finite radius. This representation is also found to be a direct matricial generalization of the well-known Stieltjes function in classical scalar literature [10].

The sequence of "numerator" polynomial matrices of MPAs of certain orders to a matrix Stieltjes series are shown to be a natural generalization of scalar orthogonal polynomials of second kind, and their properties studied by making reference to the corresponding results for "denominator" sequences, i.e., the matrix polynomials of the first kind elaborated on in [1]. Thus, the present discussions along with those in [1] is believed to provide a more complete theory of orthogonal polynomial matrices on the real line, analogous to the theory of orthogonal polynomial matrices on the unit circle discussed in [11], [12]. Finally, the relevance of orthogonal polynomials of the former kind in the context of scattering theory is also noted in [18].

It must be noted that under the present framework all results of §§ 2 and 3 (except Property 3.5), including their proofs, remain valid if $H_n(T)$ and $-H'_n(T)$ in (1.2) are assumed nonnegative definite. This is so primarily due to the fact that the MPAs of order $[m-1/m]$ and $[m/m]$ can still be interpreted as impedance, or admittance, matrices of RC networks, even under this broader assumption [15] (the "McMillan degree" which is the number of capacitors in a minimal realization in such a case can be less than mp, while under the restricted assumption adopted throughout this paper it is exactly mp; but this is of no consequence to our presentation). However, the orthogonality properties of $\hat{P}_m(s)$ and $\hat{Q}_m(s)$ discussed in § 4, and Proposition 4.2 in particular, are affected if the strict positive definiteness of $H_n(T)$ and $-H'_n(T)$ are relaxed.

From the standpoint of applications, it may be mentioned that although the present work primarily deals with connections of Padé approximations to matrix Stieltjes series and their interpretations in terms of distributed RC-multiport networks, in view of their relationship with problems such as inverse scattering [18], AR modeling of stationary stochastic processes [16], etc., the potential for utilizing the results developed here in other areas of signal and system theory cannot be ruled out.

Appendix. In this Appendix we prove the matricial version of two classical scalar theorems known as Helly's theorems [9]. We note that similar results have been derived via alternate techniques in [8] in a different context.

THEOREM A1. *Let $\sigma_m(x)$, $m = 1, 2, \cdots$, be a sequence of nondecreasing real symmetric nonnegative definite matrix-valued functions defined for real values of x. If there exists a constant real symmetric nonnegative-definite matrix M_0 such that $M_0 - \sigma_m(x)$ is nonnegative definite for real x and for all $m = 0, 1, \cdots$, then there is a subsequence of the sequence $\sigma_m(x)$, $m = 1, 2, \cdots$, which converges to a real symmetric nonnegative-definite matrix-valued function $\sigma(x)$, which is nondecreasing.*

Proof. Let $f_m(x) = \|\sigma_m(x)\|$, where $\|\cdot\|$ denotes the spectral norm of a matrix. Since $\sigma_m(x)$ is nondecreasing, if $x_1 > x_2$ then $\sigma_m(x_1) - \sigma_m(x_2)$ is nonnegative definite. Thus, due to nonnegative definiteness of $\sigma_m(x)$ and Property 2.3, $\|\sigma_m(x_1)\| \geqq \|\sigma_m(x_2)\|$, i.e., $f_m(x_1) \geqq f_m(x_2)$. Consequently, $f_m(x)$ is a nondecreasing scalar function of x.

Furthermore, since $M_0 - \sigma_m(x)$ is nonnegative definite, it follows from Property 2.3 that $f_m(x) = \|\sigma_m(x)\| \leqq \|M_0\|$, for real x and all m. Thus, the scalar sequence $f_m(x)$, $m = 1, 2, \cdots$, is uniformly bounded. Therefore, by invoking a weak version of (scalar) Helley's theorem (see, e.g., [6]), it follows that a subsequence of the sequence $f_m(x) = \|\sigma_m(x)\|$, $m = 1, 2, \cdots$, converges to a bounded nondecreasing function $f(x)$. However, since the convergence of the sequence of norms $\|\cdot\|$ of a matrix sequence implies the convergence of the matrix sequence itself [4], it follows that the corresponding subsequence of the sequence $\sigma_m(x)$, $m = 1, 2, \cdots$, converges to $\sigma(x)$ with $\|\sigma(x)\| = f(x)$. The rest of the desired properties of $\sigma(x)$ follow from the corresponding properties of $\sigma_m(x)$.

THEOREM A2. *Let $\sigma_m(x)$, $m = 1, 2, \cdots$, be a sequence of nondecreasing real symmetric nonnegative-definite matrix-valued functions defined for all x in the compact interval $[a, b]$ of the real axis such that $M_0 \geqq \sigma_m(x)$ for all m, where M_0 is a constant real symmetric nonnegative-define matrix. Let $\sigma(x)$ be the limit function to which the above sequence converges for all x in $[a, b]$. Then for a continuous scalar-valued function $g(x)$ defined over $[a, b]$, (A1) holds true:*

$$\text{(A1)} \qquad \lim_{m \to \infty} \int_a^b g(x)\, d\sigma_m(x) = \int_a^b g(x)\, d\sigma(x).$$

Furthermore, an extension of the result holds when $a \to -\infty$ and $b \to \infty$ as in the scalar case [9].

Proof. First, since $0 \leqq \sigma_m(x) \leqq M_0$ for all $x \in [a, b]$ and for all $m = 1, 2, \cdots$ and $\sigma_m(x)$ is nondecreasing, the matrix-valued functions $\sigma_m(x)$ as well as the scalar functions $\sigma_m^{(ij)}(x)$, where $\sigma_m^{(ij)}(x)$ is the ijth element of $\sigma_m(x)$, due to Lemma 3.3, are of bounded variation in $[a, b]$. Consequently, $\sigma(x)$ is also of bounded variation in $[a, b]$. Since $g(x)$ is continuous in $[a, b]$, it is uniformly continuous in $[a, b]$. Therefore, for any $\varepsilon > 0$ it is possible to consider a partition $\{x_0, x_1, \cdots, x_k\}$ of $[a, b]$ such that

$$\text{(A2)} \qquad |g(x') - g(x'')| < \varepsilon \quad \text{for all } x', x'' \in [x_{\nu-1}, x_\nu], \quad 1 \leqq \nu \leqq k.$$

If $\zeta_\nu \in [x_{\nu-1}, x_\nu]$, then by using mean value theorem of scalar Stieltjes integrals [9]

$$\text{(A3)} \qquad \int_{x_{\nu-1}}^{x_\nu} g(x)\, d\sigma^{(ij)}(x) - g(\zeta_\nu)\Delta\sigma^{(ij)}(x_\nu) = [g(\zeta'_\nu) - g(\zeta_\nu)]\Delta\sigma^{(ij)}(x_\nu)$$

for some $\zeta'_\nu \in [x_{\nu-1}, x_\nu]$, where $\Delta\sigma^{(ij)}(x_\nu) = \sigma^{(ij)}(x_\nu) - \sigma^{(ij)}(x_{\nu-1})$, $\sigma^{(ij)}(x)$ being the ijth element of the matrix $\sigma(x)$. Note that the existence of the integrals in the left-hand side is guaranteed since $\sigma(x)$ is a function of bounded variation (cf. remark preceding Lemma 3.4).

Summing (A3) over ν we obtain via the use of triangle inequality

$$\text{(A4)}\qquad \begin{aligned}\left|\int_a^b g(x)\, d\sigma^{(ij)}(x)-\sum_{\nu=1}^{k} g(\zeta_\nu)\,\Delta\sigma^{(ij)}(x_\nu)\right| &\leqq \sum_{\nu=1}^{k} |g(\zeta'_\nu)-g(\zeta_\nu)|\,|\Delta\sigma^{(ij)}(x_\nu)|\\ &<\varepsilon \sum_{\nu=1}^{k} |\Delta\sigma^{(ij)}(x_\nu)| \leqq \varepsilon V\end{aligned}$$

where $V<\infty$ is the total variation of the function $\sigma^{(ij)}(x)$ over the interval $[a, b]$. Proceeding similarly as above with $\int_a^b g(x)\, d\sigma_m^{(ij)}(x)$ instead of $\int_a^b g(x)\, d\sigma^{(ij)}(x)$, it follows that

$$\text{(A5)}\qquad \left|\int_a^b g(x)\, d\sigma_m^{(ij)}(x)-\sum_{\nu=1}^{k} g(\zeta_\nu)\,\Delta\sigma_m^{(ij)}(x_\nu)\right|<\varepsilon V$$

where $\Delta\sigma_m^{(ij)}(x_\nu)=\sigma_m^{(ij)}(x_\nu)-\sigma_m^{(ij)}(x_{\nu-1})$. Thus, from (A4), (A5), and triangle inequality, it follows that

$$\text{(A6)}\qquad \begin{aligned}&\left|\int_a^b g(x)\, d\sigma^{(ij)}(x)-\int_a^b g(x)\, d\sigma_m^{(ij)}(x)\right|\\ &\qquad\leqq 2\varepsilon V+\sum_{\nu=1}^{k} |g(\zeta_\nu)|\,|\Delta\sigma^{(ij)}(x_\nu)-\Delta\sigma_m^{(ij)}(x_\nu)|.\end{aligned}$$

Since the second term in the right-hand side of (A6) goes to zero as $m\to\infty$, we have essentially proved that

$$\int_a^b g(x)\, d\sigma^{(ij)}(x)=\lim_{m\to\infty}\int_a^b g(x)\, d\sigma_m^{(ij)}(x).$$

Extension of the proof when $a\to-\infty$ and $b\to\infty$ is identical to the scalar case [9] and is not repeated for brevity.

Acknowledgment. The assistance of Professor C. Suffel, Department of Mathematics, Stevens Institute of Technology, Hoboken, New Jersey, with the proof of Theorem 3.6 is gratefully acknowledged.

REFERENCES

[1] S. Basu and N. K. Bose, *Matrix Stieltjes series and network models*, SIAM J. Math. Anal., 14 (1982), pp. 209-222.

[2] E. F. Beckenbach and R. Bellman, *Inequalities*, Springer-Verlag, Berlin, New York, 1961.

[3] G. A. Baker, *Essentials of Padé Approximants*, Academic Press, New York, 1975.

[4] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Academic Press, New York, 1985.

[5] E. Hille, *Analytic Function Theory*, Vol. II, Ginn, Boston, MA, 1962.

[6] T. S. Chihara, *An Introduction to Orthogonal Polynomials*, Gordon and Breach, New York, 1978.

[7] N. Wiener and P. Masani, *The prediction theory of multivariate stochastic processes, Part* I, Acta. Math., 98 (1957), pp. 111-150.

[8] V. P. Potopov, *The multiplicative structure of J-contractive matrix functions*, Trans. Amer. Math. Soc., Ser. 2, 15 (1960), pp. 131-243.

[9] L. F. Natanson, *Theory of Functions of a Real Variable*, Vol. II, English Translation, Frederick Ungar, New York, 1960.

[10] N. I. Akhiezer, *The Classical Moment Problem*, Oliver and Boyd, London, 1965.

[11] D. Delsarte, Y. V. Genin, and Y. V. Kamp, *Orthogonal polynomial matrices on the unit circle*, IEEE Trans. Circuits Systems, 25 (1978), pp. 149–160.

[12] D. C. Youla and N. Kazanjian, *Bauer type factorization of positive matrices and theory of matrix orthogonal polynomials on unit circle*, IEEE Trans. Circuits Systems, 25 (1978), pp. 57–69.

[13] R. E. Kalman, *On partial realization, transfer functions and canonical forms*, ACTA Polytech. Scand. Math. Comput. Sci. Ser., 29 (1979), pp. 9–32.

[14] A. Bultheel, *Recursive algorithms for matrix Padé problems*, Math. Comp., 35 (1980), pp. 875–892.

[15] K. Morimoto, N. Matsumoto, and S.-I. Takahashi, *Matrix Padé approximants and multi-port networks*, Electron. Commun. Japan, 61A (1978), pp. 28–36.

[16] G. Cybenko, *Restrictions of normal operators Padé approximation and autoregressive time series*, SIAM J. Appl. Math., 15 (1984), pp. 753–767.

[17] B. Belevitch, *On the realizability of non-rational positive real functions*, Circuit Theory Appl., 1 (1973), pp. 17–30.

[18] J. S. Geronimo, *Scattering theory and matrix orthogonal polynomials on the real line*, Circuits Systems Signal Process., 1 (1984), pp. 471–494.

[19] W. B. Gragg and A. Lindquist, *On the partial realization problem*, Linear Algebra Appl., 50 (1983), pp. 277–319.

[20] W. Cauer, *The Poisson integral for functions with positive real parts*, Bull. Amer. Math. Soc., 38 (1932), pp. 713–717.

[21] Björn von Sydow, *Matrix-Valued Padé Approximation and Gaussian Quadrature*, in Det Kongelige Norse Videnskabers Selskab, Skrifter no. 1-1983, Padé Approximants and Continued Fractions, Universitetsforlaget Trondheim, 1983, pp. 117–127.

[22] F. J. Narcowich, *R-Operators II. On the approximation of certain operator-valued functions and the Hermitian moment problem*, Indiana Univ. Math. J., 26 (1977), pp. 483–513.

[23] J. Zinn-Justin, *Strong interaction dynamics with Padé approximants*, Phys. Lett., C1 (1971), pp. 55–102.

[24] J. L. Basdevant, D. Bessis, and J. Zinn-Justin, *Padé approximants in strong interactions: two-body Pion and Kaon systems*, Cimento A (11), 60 (1969), pp. 185–237.

[25] J. S. MacNerney, *Hermitian moment sequences*, Trans. Amer. Math. Soc., 103 (1962), pp. 45–81.

[26] ———, *Hermitian moment sequences—correction*, Notices Amer. Math. Soc., 25 (1978), pp. A-427.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 3, pp. 407–423, July 1989

008

THE RANKS OF EXTREMAL POSITIVE SEMIDEFINITE MATRICES WITH GIVEN SPARSITY PATTERN*

J. W. HELTON†, S. PIERCE‡, AND L. RODMAN§

Abstract. Let P be a symmetric set of ordered pairs of integers from 1 to n, and define $M^+(P)$ to be the closed cone of all positive semidefinite Hermitian matrices whose (i, j) entry is zero whenever $i \neq j$ and (i, j) is not in P. The extreme points of $M^+(P)$ are considered. In some special cases, the maximum rank that such an extreme point can have is calculated.

Key words. extremal matrix, order of a graph

AMS(MOS) subject classifications. primary 05C50; secondary 05B20, 15A57

1. Introduction.

Let P be a symmetric set of ordered pairs of integers from 1 to n and define $M^+(P)$ to be the closed cone of all positive semidefinite Hermitian matrices whose (i, j) entry is zero whenever $i \neq j$ and (i, j) is not in P. Two cases will be considered: (1) the matrices are over the field $\mathbf{C}$ of complex numbers; (2) the matrices are over the real numbers $\mathbf{R}$. Later, P will be naturally interpreted as an undirected graph with vertices $1, \cdots, n$.

We say a matrix A in $M^+(P)$ is *extremal* if $A = B + C$ for $B, C \in M^+(P)$ implies that B and C are scalar multiples of A. Let $J^+(P) \subset M^+(P)$ be the set of matrices that are extremal points of $M^+(P)$. We say that P has *order* k if k is the maximum of ranks of matrices in the set $J^+(P)$. The problem of characterizing the orders of graphs P is important in several areas.

For one thing, it is related to the "positive completion problem" [DG], [GJSW] for matrices in the sense that we might think of the positive completion problem as a strictly easier one. Thus progress on the order problem is probably essential to making progress on the positive completion problem.

In this paper we put forward some general remarks and ideas that allow us to establish orders of new classes of graphs.

This paper is built on [AHMR] and heavily uses (in Part I) some ideas from the earlier paper [M].

The paper has three parts (in addition to the introduction and preliminaries sections). The first two address the theme of the relationship of the order problem to techniques of Gaussian elimination for sparse matrices that are traditional in numerical analysis. This subject is devoted to doing the Cholesky decomposition L^TDL of a sparse matrix with the smallest number of algebraic operations. The fundamental problem is that for a given sparse matrix M when we perform a Gaussian elimination step to make a particular entry zero, usually several entries in M that are zero are made nonzero. This phenomenon

* Received by the editors May 19, 1988; accepted for publication December 21, 1988.

† Department of Mathematics, University of California at San Diego, La Jolla, California 92093. The work of this author was partially supported by grants from the National Science Foundation, the Office of Naval Research, and the Air Force Office of Scientific Research.

‡ Department of Mathematical Sciences, San Diego State University, San Diego, California 92182 (Q300085@CALSTATE.BITNET). The work of this author was partially supported by a National Science Foundation grant.

§ Department of Mathematics, Arizona State University, Tempe, Arizona 85287, and School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv, Israel. The work of this author was partially supported by a National Science Foundation grant. The work was done while this author, with support from the Office of Naval Research, visited the University of California at San Diego, La Jolla, California.

is known as *fill in*, generally, and the amount of it is generally very dependent on the order in which one performs Gaussian elimination on the matrix. A major branch of sparse matrix analysis is devoted to how we perform Gaussian elimination on matrices of a particular sparsity pattern so as to minimize fill-in (see [P], [R], [RT], [GL]).

In Part I we analyze a "divide-and-conquer" technique for the order problem and positive completion problems. We show that if these problems can be solved for certain submatrices of a given matrix, then they can be solved for the full matrix (provided that the submatrices interact in a certain way). This is very similar to classical use of divide-and-conquer methods in the Cholesky decomposition for the case where there is no Gaussian elimination fill in. While our results are simple, the matrix theoretic content as opposed to the graph theoretic content of all clear results on the order problem [PPS], [M] are consequences of a few simple principles.

Use of graph theory techniques developed (by Rose) for the Cholesky decomposition were introduced to the order problem by Grone et al. [GJSW]. Next Paulsen, Power, and Smith [PPS] used Rose's methods more directly to give a very elegant proof of the [GJSW] theorem (that is close to the one we have here). They also found intriguing connections with the completely positive maps that occur in operator theory.

While Part I analyzes the behavior of order in situations that correspond to Gaussian elimination having no fill-in, Part II begins to treat cases with fill-in. There is a natural measure $\alpha(G)$ of the minimum amount of fill-in producible by Gaussian elimination on a sparsity pattern G (see § 5). We conjecture (for real matrices) that

$$\text{order}\,(G) \leqq \alpha(G) + 1.$$

In other words, that fill-in puts an upper bound on order. Indeed we suspect (on the basis of numerous examples) that there is some relation between fill-in and order that we have not yet uncovered. Section 5 gives conjectures and examples.

Part III goes in a different direction and merely computes the orders of several special classes of graphs.

All graphs G in this paper are finite, undirected, simple (i.e., without multiple edges), and without edges of the form (v, v), for a vertex v of G. The set of vertices of a graph G is denoted $V(G)$, and the set of edges is denoted $E(G)$.

For a given nonempty set $S \subset V(G)$, denote by $G(S)$ the graph obtained from G by deleting all the vertices *not* in S together with all adjoining edges. So, $V(G(S)) = S$ and $(i, j) \in E(G(S))$ if and only if $i \neq j$, $(i, j) \in E(S)$ and $(i, j) \in E(G)$.

As a graph is obviously preserved under natural graph isomorphisms, the statements and proofs will be given modulo graph isomorphisms.

The k-dimensional vector spaces $\mathbf{R}^k$ and $\mathbf{C}^k$ over the field of reals and the field of complexes, respectively, will be represented as spaces of column vectors with k components, with the standard inner products in $\mathbf{R}^k$ and $\mathbf{C}^k$.

2. Preliminaries. For the reader's convenience we state here some results that will be used frequently. All these results were obtained in [AHMR], and Theorem 2.1 in [PPS] as well.

THEOREM 2.1. *A graph G has order* 1 *if and only if G is a chordal, or triangulated graph, i.e., for any cycle* $(v_1, v_2), (v_2, v_3), \cdots, (v_{p-1}, v_p), (v_p, v_1) \in E(G), p \geqq 4$ *there is a chord* $(v_i, v_j) \in E(G)$, *where* $1 \leqq i < j \leqq p$ *and* $2 \leqq j - i \leqq p - 2$.

The class of chordal graphs is an important class that appears in diverse problems (see, e.g., [G] for more information about chordal graphs and the problems in which they play a central role).

THEOREM 2.2. *Let G be a loop with $n \geqq 3$ vertices, i.e.,*

$$E(G)=\langle(v_1,v_2),(v_2,v_3),\cdots,(v_{n-1},v_n),(v_n,v_1)\rangle$$

for some ordering $v_1,\cdots,v_n$ of the vertices of G. Then the order of G (over $\mathbf{R}$) is $n-2$.

Introduce the partial order $\leqq_v$ on the set of all (finite, undirected, simple, without edges of the form (v, v)) graphs $G_1 \leqq_v G_2$ if (and only if) G_1 is isomorphic to $G_2(S)$ for some set $S \subset V(G_2)$.

THEOREM 2.3. *If $G_1 \leqq_v G_2$, then* order $(G_1) \leqq$ order (G_2).

We remark that another natural partial order $\leqq_e$ on the set of graphs $G_1 \leqq_e G_2$ if and only if $V(G_1) = V(G_2)$ and $E(G_1) \subset E(G_2)$ generally does not imply any regularity between the orders. Indeed, let the graphs G_1, G_2, G_3, be defined by

$$V(G_i)=\{1,2,3,4\}, \qquad i=1,2,3,$$

$$E(G_1)=\{(1,2),(2,3),(3,4)\}, \qquad E(G_2)=E(G_1)\cup(1,4),$$

$$E(G_3)=E(G_2)\cup(1,3).$$

Then $G_1 \leqq_e G_2 \leqq_e G_3$ but

$$\text{order}\,(G_1)=\text{order}\,(G_3)=1, \qquad \text{order}\,(G_2)=2.$$

We shall often implicitly use the rather obvious fact that if $G_1,\cdots,G_p$ are the connected components of G, then order (G) = max order (G_i).

Part 1. Divide and Conquer Using Cut Sets and Cliques

3. Cut sets and cliques. A *clique* of a graph G is a subset $S \subset V(G)$ such that every pair (i, j) with $i, j \in S$, $i \neq j$ belongs to $E(G)$. A *cut set* of G is a subset $S \subset V(G)$ with the property that the graph $G(V(G)\backslash S)$ is not connected.

In this section we study the orders of graphs in terms of cut sets and cliques.

THEOREM 3.1. *Let $S \subset V(G)$ be a cut set, and assume that S is a clique. Then*

$$\text{order}\, G=\max\,(\text{order}\, G(S_1\cup S),\cdots,\text{order}\, G(S_r\cup S)), \tag{3.1}$$

where $G(S_1),\cdots,G(S_r)$ are all the (nonempty) connected components of $G(V(G)\backslash S)$.

This theorem appeared (at least implicitly) in [M]; we shall provide an independent proof.

Proof. As the inequality $\geqq$ in (3.1) follows from Theorem 2.3, we have only to prove $\leqq$.

Without loss of generality we may assume $r = 2$ (otherwise, use an easy induction on r). Reorder the vertices in G so that $S_1 = \{1,\cdots,p\}$; $S_2 = \{p+1,\cdots,q\}$; $S = \{q+1,\cdots,n\}$. Take $M \in M^+(G)\backslash\{0\}$. Then M has the form

$$M=\begin{bmatrix} A_1 & 0 & Q_1 \\ 0 & A_2 & Q_2 \\ Q_1^* & Q_2^* & Q \end{bmatrix}.$$

As M is positive definite,

$$\text{range}\, Q_1 \subset \text{range}\, A_1.$$

So for some matrix W we have $Q_1 = A_1 W$. Form the matrix

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

Then

$$M=E^*\begin{bmatrix} A_1 & 0 & 0 \\ 0 & A_2 & Q_2 \\ 0 & Q_2^* & \tilde{Q} \end{bmatrix}E,$$

where $\tilde{Q} = Q - W^*A_1W$. Since S is a clique, the matrix

$$\tilde{M}=\begin{bmatrix} A_1 & 0 & 0 \\ 0 & A_2 & Q_2 \\ 0 & Q_2^* & \tilde{Q} \end{bmatrix}$$

belongs to $M^+(G)$. Write

$$\begin{bmatrix} A_2 & Q_2 \\ Q_2^* & \tilde{Q} \end{bmatrix}=\sum_j \tilde{T}_j,$$

where $\tilde{T}_j \in M^+(G(S_2 \cup S))$, and rank $\tilde{T}_j \leqq$ order $(G(S_2 \cup S))$. Now

$$\begin{aligned} M&=E^*\begin{bmatrix} A_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}E+\sum_j E^*\begin{bmatrix} 0 & 0 \\ 0 & \tilde{T}_j \end{bmatrix}E \\ &=\begin{bmatrix} A_1 & 0 & Q_1 \\ 0 & 0 & 0 \\ Q_1^* & 0 & Q_1^*A_1Q_1 \end{bmatrix}+\sum_j\begin{bmatrix} 0 & 0 \\ 0 & \tilde{T}_j \end{bmatrix}. \end{aligned} \tag{3.2}$$

The matrix

$$N=\begin{bmatrix} A_1 & Q_1 \\ Q_1^* & Q_1^*A_1Q_1 \end{bmatrix}$$

belongs to $M^+(G(S_1 \cup S))$ and hence is the sum of matrices from $M^+(G(S_1 \cup S))$ of ranks $\leqq$ order $G(S_1 \cup S)$. Substituting this sum into (3.2), we represent M as a sum of matrices from $M^+(G)$, the ranks of which do not exceed

$$\max(\text{order } G(S_1 \cup S), \text{order } G(S_2 \cup S)). \qquad \square$$

Some remarks concerning Theorem 3.1 are in order.

Remark 3.2. If we relax the assumptions and requirements that S becomes a clique after addition of just one edge, then the disparity between the right- and left-hand sides of (3.1) can be arbitrarily large, as the following example shows. Let G be the loop with n vertices $\{1, \cdots, n\}$ (so $E(G) = \{(1,2),(2,3), \cdots, (n,1)\}$). Assuming, for instance, that n is even, let $S = \{1, n/2 + 1\}$. Obviously, S will become a clique after addition of one edge. However, by Theorem 2.2 we have (in the real case)

$$\text{order } G = n-2, \qquad \text{order } G(S_1 \cup S) = \text{order } G(S_2 \cup S) = n/2 - \tfrac{1}{2},$$

where $G(S_1)$ and $G(S_2)$ are the connected components of $G(V(G)\backslash S)$.

Remark 3.3. The following example shows that the right-hand side in (3.1) cannot be replaced by

$$\max(\text{order } G(S_1), \cdots, \text{order } G(S_r)). \tag{3.3}$$

Let $V(G) = \{1,2,3,4,5\}$; $E(G) = \{(1,2),(2,3),(3,4),(4,5),(5,2)\}$; $S = \{2\}$. In this example the left-hand side of (3.1) is at least 2 (use Theorem 2.1 to verify this), while (3.3) is equal to 1 by the same Theorem 2.1.

The following particular case of Theorem 3.1 deserves special attention.

COROLLARY 3.4. *Let $v \in V(G)$ be such that the adjacent set*

$$\operatorname{adj}(v) = \{u \in V(G) \mid (u,v) \in E(G)\}$$

is a clique. Then

$$\operatorname{order} G = \operatorname{order} G(V(G) \setminus \{v\}).$$

For the proof just observe that adj (v) is a cut set, and apply Theorem 3.1.

The condition of an adjacent set being a clique appears naturally in the analysis of Gaussian elimination of sparse matrices (see [R], [G]).

As applications of Theorem 3.1 we can compute orders of many classes of graphs. Consider one such class as an example.

COROLLARY 3.5. *Let G be a graph, and suppose $V(G) = S_1 \cup \cdots S_p$, where the nonempty sets S_j have the following properties*:

(i) *For every $i \neq j$ either $S_i \cap S_j = \emptyset$ or $S_i \cap S_j$ consists of two vertices u and v with $(u, v) \in E(G)$*;

(ii) *Every edge in G belongs to some $G(S_i)$*;

(iii) *The induced graph $\hat{G}$ defined by $V(\hat{G}) = \{1, \cdots, p\}$, $(i,j) \in E(\hat{G}) \Leftrightarrow S_i \cap S_j \neq \emptyset$ is a forest, i.e., $\hat{G}$ has no loops.*

Then

$$\operatorname{order}(G) = \max_{1 \leqq i \leqq p} \operatorname{order}(G(S_i)).$$

Proof. Induction on the number p. We can assume that $\hat{G}$ has at least one edge (otherwise, everything is trivial). Then $\hat{G}$ must have a pendant, i.e., a vertex with precisely one adjacent edge. Say 1 is a pendant, and $(1, 2) \in E(\hat{G})$. The hypotheses of the corollary easily imply that the intersection $S_1 \cap S_2$ is a cut set (consisting of two vertices), which is a clique. We are now done in view of Theorem 3.1 and the induction hypothesis. □

The following result gives an upper bound for orders of graphs containing cliques.

THEOREM 3.6. *Let G be a graph with n vertices and let $S \subset V(G)$, $S \neq V(G)$ be such that $G(S)$ is a clique. Then*

$$\operatorname{order}(G) \leqq n - |S|. \tag{3.4}$$

Here $|S|$ is the cardinality of the finite set S.

Theorem 3.6 is contained in [M]. We shall include a proof anyway.

Proof. Let $X \in M^+(G)$ be of rank greater than $n - |S|$. Enumerate all the vertices so that $S = \{1, \cdots, k\}$. Then

$$\operatorname{range} X \cap \operatorname{span}\{e_1, \cdots, e_k\} \neq \{0\},$$

where e_j is the jth unit coordinate vector (with 1 in the jth place and zeros elsewhere). Pick $x \in \operatorname{range} X \cap \operatorname{span}\{e_1, \cdots, e_k\}$, $x \neq \{0\}$, and put $R = xx^*$. Then $R \in M^+(G)$ and range $R \subset$ range X. It is easy to see (for example, by writing the linear transformations X and R in an orthonormal basis consisting of eigenvectors of X) that $X - \varepsilon R$ is positive semidefinite for $\varepsilon > 0$ sufficiently small. Now

$$X = \tfrac{1}{2}(X - \varepsilon R) + \tfrac{1}{2}(X + \varepsilon R),$$

which shows that X is not extremal in $M^+(G)$ unless rank $X = 1$. However, the possibility that rank $X = 1$ is excluded because $n > |S|$. □

It is of interest to characterize those graphs for which equality holds in (3.4). Some information along these lines will be given later (Theorem 7.1).

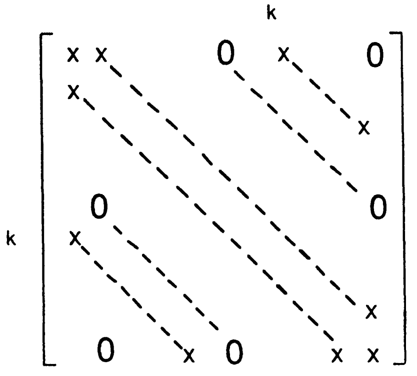


FIG. 1.

We now indicate another sparsity pattern (see Fig. 1) whose graph contains a "ladder," as in Fig. 2, and a large loop. We are able, however, to use the divide-and-conquer method to split the computation of order into two more canonical looking graphs. Thus the problem is worth considerable study, and even some compromises are justified.

For large n, consider the tridiagonal pattern whose graph is the line from 1 to n. Fix a k, preferably close to n. Let G be the graph that includes the line from 1 to n and in addition whose edges include $(1, k), (2, k+1), \cdots, (n-k+1, n)$. Thus we obtain a pattern as indicated in Fig. 1. For n and k large, the matrix is still rather sparse and in a sense is not "too far" from the pattern that yields the n-loop. A careful drawing of the graph yields the shape in Fig. 2. (Note that we must assume $k > n - k + 1$.)

We now approach the problem with an obvious compromise. We lose only one zero in the matrix in Fig. 1 if we join vertex k and vertex $n - k + 1$ with an edge. Call the graph G modified by this addition $\tilde{G}$, and note that these two vertices form a cut set that is a clique for $\tilde{G}$. This cut set splits $\tilde{G}$ into two graphs:

(1) A loop with $2k - n$ vertices;

(2) The ladder in Fig. 2 with k and $n - k + 1$ joined.

The loop is standard and has order $2k - n - 2$. The graph in (2) is something like a loop of double thickness and analyzing it suggests an interesting line of questions. Perhaps known methods for studying loops will generalize.

Finally, in the special case $k = n - 1$, we observe that G contains an $(n-1)$-loop and hence in the real case has order at least $n - 3$. Since G itself is not a loop, the order of G is exactly $n - 3$.

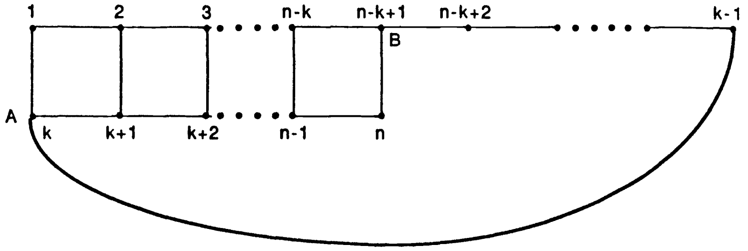


FIG. 2.

4. Completion problems. We study here implications of the "divide-and-conquer" technique for various completion problems.

Let G be a graph with n vertices, and call its vertices $\{1, 2, \cdots, n\}$. A *partial Hermitian matrix* subordinate to G is, by definition, an $n \times n$ array of complex numbers and question marks such that the (i, j) entry is ? if and only if $i \neq j$ and (i, j) is not in $E(G)$, and that is Hermitian symmetric in the usual sense: if $(i, j) \in E(G)$ (so that the (i, j) entry is a complex number a_{ij}), then the (j, i) entry is $\overline{a_{ij}}$. We say that an $n \times n$ Hermitian matrix B is a *completion* of an $n \times n$ partial Hermitian matrix A subordinate to G if the (i, j) entry of B coincides with the (i, j) entry of A whenever the latter is *not* a question mark. In informal terms, B is obtained from A by replacing all ?'s by some complex numbers in the Hermitian way. Various completion problems have been studied in [DG], [GJSW], [JR], [EGL], [PPS]; see also [H, Chap. 8].

We consider three classes of completion problems:

(1) For a given integer k, $0 \leqq k \leqq n$, and a given partial Hermitian matrix A subordinate to G, determine if A admits a completion with precisely k nonpositive eigenvalues (counted with multiplicities).

(2) Same as (1) with replacement of "nonpositive" by "negative."

(3) For a given partial Hermitian matrix A subordinate to G determine the completions B of A for which $\lambda_{\min}(B)$ is maximal among all completions of A. Here $\lambda_{\min}(B)$ is the minimal eigenvalue of B.

Given a partial Hermitian matrix A subordinate to G, and a subgraph F (so $F \leqq_v G$), the naturally defined *restriction* $A \mid F$ is a partial Hermitian matrix subordinate to F.

In the rest of this section G is assumed to be a graph with a cut set S that is a clique. The connected components of $G(V(G) \backslash S)$ will be denoted $G_1, \cdots, G_p$.

THEOREM 4.1. *Let A be a partial Hermitian matrix subordinate to G. If each restriction $A \mid V(G_j) \cup S$ admits a completion B_j, whose size is equal to the cardinality of $V(G_j) \cup S$, with precisely k_j nonpositive eigenvalues ($j = 1, \cdots, p$), then A admits a completion with precisely* $\max_{1 \leqq j \leqq p} k_j$ *nonpositive eigenvalues.*

Proof. Order the n vertices of G so that

$$S = \{1, \cdots, n_1\}, \quad G_1 = \{n_1 + 1, \cdots, n_2\} \cdots, \quad G_p = \{n_p + 1, \cdots, n\}.$$

Write

$$B_j = \begin{bmatrix} B_{j1} & B_{j2} \\ B_{j2}^* & B_{j3} \end{bmatrix},$$

where B_{j1} is $n_1 \times n_1$. Since S is a clique, B_{j1} is independent of j. Consider the partial Hermitian matrix

$$\hat{B} = \begin{bmatrix} B_{11} & B_{12} & B_{22} \cdots B_{p2} \\ B_{12}^* & B_{13} & ? \cdots ? \\ B_{22}^* & ? & B_{23} \cdots ? \\ \vdots & \vdots & \vdots \\ B_{p2}^* & ? & ? \cdots B_{p3} \end{bmatrix}.$$

The matrix $\hat{B}$ is subordinate to the graph $\hat{G}$ obtained from G by adding all edges of the type (q_1, q_2), where (q_1, q_2) is not in $E(G)$ and

$$q_1, q_2 \in \{1, \cdots, n_1, n_j + 1, \cdots, n_{j+1}\}$$

for some $1 \leqq j \leqq p$ (by definition, $n_{p+1} = n$). We check easily that $\hat{G}$ is chordal. Now application of Theorem 1 in [JR] finishes the proof. (For the reader's convenience, we quote this theorem. Given a partial Hermitian matrix A subordinate to a chordal graph G, there is a completion B of A such that the number of nonpositive eigenvalues of B coincides with the maximum number of nonpositive eigenvalues of any restriction $A|_V$, where V is a clique of G.) □

Theorem 4.1 provides the best result in the following sense. If A admits a completion with precisely k_0 nonpositive eigenvalues, and k_0 is maximal among all completions of A, then no completion of $A | V(G_j) \cup S$ can have more than k_0 nonpositive eigenvalues. This follows from the interlacing inequalities between eigenvalues of a Hermitian matrix and eigenvalues of its principal submatrices.

The case when all k_j are zero is of special interest.

COROLLARY 4.2. *Let A be as in Theorem 4.1. If each restriction $A | V(G_j) \cup S$ admits a positive definite completion, then A itself admits a positive definite completion.*

For the second class of completion problems we have the following result.

THEOREM 4.3. *Let A be a partial Hermitian matrix subordinate to G. If each restriction $A | V(G_j) \cup S$ admits a nonsingular completion with precisely k negative eigenvalues, then A admits a nonsingular completion with precisely $\max_{1 \leqq j \leqq p} k_j$ negative eigenvalues.*

Observe that nonsingularity is required in Theorem 4.3 (in contrast to Theorem 4.1).

The proof is the same as that of Theorem 4.1.

Note that, under the hypotheses of Theorem 4.3, the existence of a (not necessarily nonsingular) completion of A with precisely $\max_{1 < j < p} k_j$ negative eigenvalues follows from Theorem 4.1 (applied to $A + \varepsilon I$ for a small positive ε).

We now pass to the third class of completion problems.

THEOREM 4.4. *Let A be as in Theorem 4.1. Let λ_j be the maximum of $\lambda_{\min}(B_j)$ taken over the set of all completions B_j of $A | V(G_j) \cup S$. Then there is a completion B of A for which*

$$\lambda_{\min}(B) = \min\{\lambda_1, \cdots, \lambda_p\}.$$

Theorem 4.4 follows immediately from Corollary 4.2 by subtracting a suitable multiple of I from A.

Finally, we remark that all results of this section are true in the real case as well (i.e., A is assumed to be real, and only real completions are allowed).

Part II. Gaussian Elimination Fill-In and the Order Problem

5. Gaussian elimination for sparse positive semidefinite matrices. Let G be an (undirected) connected graph. We say that $\hat{G}$ is obtained by *one-step elimination* from G if for some vertex v in G the graph $\hat{G}$ is obtained by removing v and all its adjacent edges and by adding edges (x, y) for all pairs of vertices x, y different from v in G such that $(x, v), (y, v)$ are edges in G but (x, y) is not. Thus, $\hat{G}$ has one vertex less than G.

The one-step elimination procedure is the basic step in symmetric Gauss elimination and has been studied extensively from the graph-theoretic point of view (see [R], [RT], [G], [GL]).

For a given graph G, let $\alpha(G) = \min$ {total number of edges added to $E(G)$ in consecutive one-step eliminations starting with G and ending in a one-vertex graph}, the minimum being taken over all orderings of the vertices.

It follows from Theorem 2.1, combined with another characterization of chordal graphs (see, e.g., [G]) that order $(G) = 1$ if and only if $\alpha(G) = 0$. Thus, it is of interest to find the relations between order (G) and $\alpha(G)$. We have the following conjecture.

CONJECTURE 5.1. *There is a universal constant C such that*

$$\text{order}\,(G) \leqq \alpha(G) + C \tag{5.1}$$

in the real case, and

$$\text{order}\,(G) \leqq 2\alpha(G) + C \tag{5.2}$$

in the complex case, for all graphs G.

The stronger conjecture is the following.

CONJECTURE 5.2. *The inequalities* (5.1) *and* (5.2) *are valid with* $C = 1$.

We could not prove either conjecture. In this part we verify Conjecture 5.2 for some graphs and prove that in some sense Conjecture 5.2 cannot be improved (Corollary 6.2).

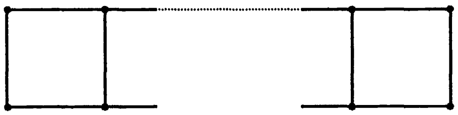
Two simple remarks are in order.

Remark 5.3. Let G be a disjoint union of k copies of the graph G_0. Then order G = order G_0. However, $\alpha(G) = k\alpha(G_0)$. This shows that there are no constants C_1, C_2 with $C_1 > 0$ such that

$$C_1\alpha(G) + C_2 \leqq \text{order}\, G$$

for all graphs G.

Remark 5.4. Let G be the ladder graph with $2m$ vertices:



Then $\alpha(G) = 2m - 1$ (see [R]). On the other hand, a repeated application of Theorem 3.1 (with the cut sets consisting of two vertices) shows that the order of G coincides with the order of the four-vertex loop, which is two (over **R** as well as over **C**; see Theorem 7.1 in [AHMR]). This shows that the difference $\alpha(G)$ − order (G) can be arbitrarily large, even for connected graphs (the graph in Remark 5.1 was disconnected).

For a given graph G define $\beta(G)$ to be the minimal number of edges necessary to add to G in order to obtain a chordal graph. It is not difficult to see that $\alpha(G) = \beta(G)$. Indeed, the inequality $\beta(G) \leqq \alpha(G)$ is obvious. To prove the opposite, let $\hat{G}$ be the chordal graph obtained from G by adding $\beta(G)$ new edges.

Let $N : V(\hat{G}) \to \{1, \cdots, n\}$ be the enumeration of vertices given by a perfect elimination scheme for $\hat{G}$ (see [R], [G], [GL] for a definition and properties of this notion). Use the ordering N to do consecutive one-step eliminations. Then it is necessary to put $\beta(G)$ new edges in G in this procedure so $\alpha(G) \leqq \beta(G)$.

PROPOSITION 5.5. *If* $G_1 \leqq_v G_2$, *then* $\alpha(G_1) \leqq \alpha(G_2)$.

Proof. Clearly, $\beta(G_1) \leqq \beta(G_2)$. Now use the fact that $\alpha(G_j) = \beta(G_j)$ for $j = 1, 2$. □

We can now prove the following.

THEOREM 5.6. *If* order $(G) \leqq 3$, *then*

$$\text{order}\,(G) \leqq \alpha(G) + 1 \tag{5.3}$$

in the real case.

Proof. If order $(G) = 1$, then, as we have observed already, $\alpha(G) = 0$, and (5.3) holds.

Assume order $(G) = 2$. Then G is not chordal (Theorem 2.1) and hence $\alpha(G) \geqq 1$; so (5.3) holds again. Assume now order $(G) = 3$.

Then G contains (in the sense of $\leqq_v$-partial order) a minimal graph G_0 of order three. The list of all possible graphs G_0 (there are 16 of them) given by Theorem 8.2 in [AHMR] shows that $\alpha(G_0) \geqq 2$. By Proposition 5.3, $\alpha(G) \geqq 2$. □

The argument used in the proof of Theorem 5.4 also gives the following statement. As defined in [AHMR], a graph G is called a *k-block* if order $(G) = k$ and any graph strictly $\leqq_v$-contained in G has order less than k. This notion depends on the choice of the field (**R** or **C**).

THEOREM 5.7. *Fix positive integers k and C. Then in the real case the inequality*

$$\text{order}\,(G) \leqq \alpha(G) + C$$

holds for all graphs G of order k if and only if it holds for all k-blocks.

An analogous statement is valid in the complex case concerning the inequality (5.2).

In connection with Theorem 5.5, observe that for every fixed k the number of k-blocks is finite [AHMR, Cor. 4.4]). Thus, in principle we could decide if (5.1) or (5.2) holds for all k-blocks using a finite procedure.

We conclude this section with a simple example.

Example 5.1. Let G be a loop with n vertices. Then it is easy to see that $\alpha(G) = n - 3$. Combining with Theorem 2.2, we see that (in the real case)

$$\text{order}\,(G) = \alpha(G) + 1.$$

6. Fully bipartite graphs. In this section we compute orders of a large class of fully bipartite graphs, thereby giving another illustration of Conjecture 5.2 in the real, as well as in the complex case.

The fully bipartite graph $G(n, m)$ on $n + m$ vertices is defined as follows (here m, n are positive integers):

$$V(G) = \{1, \cdots, m+n\};$$

$(i, j) \in E(G(m, n))$ if and only if $i \neq j$ and precisely one of the indices i and j is in the set $\{1, \cdots, n\}$ (so that the other index is in the set $\{n + 1, \cdots, m + n\}$). We shall assume $n \leqq m$, and (to avoid known cases) $n \geqq 2$. An easy inspection (see [R]) shows that

$$\alpha(G) = \frac{n(n-1)}{2}. \tag{6.1}$$

THEOREM 6.1. *For special n and m as indicated, we have*

$$\text{order}\, G(n,m) = \begin{cases} m & \text{if } \dfrac{n^2-n}{4} < m \leqq \dfrac{n^2-n+2}{2}, \\ \dfrac{n^2-n+2}{2} & \text{if } m \geqq \dfrac{n^2-n+2}{2}, \end{cases}$$

in the real case, and

$$\text{order}\, G(n,m) = \begin{cases} m & \text{if } (n^2-n)/3 < m \leqq n^2-n+1, \\ n^2-n+1 & \text{if } m \geqq n^2-n+1. \end{cases}$$

Comparing with (6.1) we immediately get Corollary 6.2.

COROLLARY 6.2. *In the real case for $m > (n^2 - n)/4$ we have*

$$\text{order}\, G(n,m) \leqq \alpha(G(n,m)) + 1,$$

and equality holds for $m = (n^2 - n + 2)/2$. *In the complex case for* $m > (n^2 - n)/3$ *we have*

$$\text{order } G(n,m) \leqq 2\alpha(G(n,m)) + 1,$$

and equality holds for $m = n^2 - n + 1$.

Thus, Corollary 6.2 confirms Conjecture 5.2 and shows that in a certain sense this conjecture is best possible.

The rest of this section will be devoted to the proof of Theorem 6.1.

First, we need a simple lemma. For a real $p \times p$ matrix $A = (a_{ij})$, let

$$\operatorname{diag} A = (a_{11}, a_{22}, \cdots, a_{pp})^T \in \mathbf{R}^p$$

be the diagonal part of A.

LEMMA 6.3. *Let* $n \geqq 2$ *be an integer, and let* $n \leqq m \leqq (n^2 - n + 2)/2$. *Then there exists a linearly independent orthogonal set* $Y(1), \cdots, Y(n)$ *in* $\mathbf{R}^m$ *such that the vectors* (*the superscript "T" denotes transposition*)

$$\operatorname{diag}(Y(i)Y(j)^T + Y(j)Y(i)^T) \in \mathbf{R}^m \qquad (i \neq j)$$

span the linear space

$$F = \{(x_1, \cdots, x_m)^T \in \mathbf{R}^m \mid x_1 + \cdots + x_m = 0\}.$$

Proof. The proof is by induction on n. The case $n = 2$ is trivial. Pick m' such that

$$n-1 \leqq m' \leqq ((n-1)^2 - (n-1) + 2)/2 \text{ and}$$

$$m' \geqq m - (n-1), \qquad m' \leqq m-1.$$

Suppose vectors $\hat{Y}(1), \cdots, \hat{Y}(n-1)$ in $\mathbf{R}^{m'}$ with the desired properties are constructed already. Then put

$$Y(1) = \begin{bmatrix} \hat{Y}(1) \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad Y(2) = \begin{bmatrix} \hat{Y}(2) \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \cdots, Y(m-m') = \begin{bmatrix} \hat{Y}(m-m') \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix},$$

$$Y(i) = \begin{bmatrix} \hat{Y}(i) \\ 0 \end{bmatrix} \quad \text{for } m - m' < i < n-1; \qquad Y(n) = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_{m'} \\ 1 \\ \vdots \\ 1 \end{bmatrix},$$

where the numbers $\alpha_1, \cdots, \alpha_{m'}$ satisfy the equation

$$\begin{pmatrix} \hat{Y}(1)^T \\ \hat{Y}(2)^T \\ \vdots \\ \hat{Y}(n-1)^T \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \\ \vdots \\ -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \longrightarrow m - m'.$$

The vectors $Y(1), \cdots, Y(n-1)$ are obviously linearly independent; $Y(n)$ cannot be in the linear span of $Y(1), \cdots, Y(n-1)$ because $Y(n) \neq 0$ and

$$Y(n) \perp \operatorname{span}\{Y(1), \cdots, Y(n-1)\}. \qquad \square$$

Next, recall the notion of a representation of the graph $G(n, m)$ introduced (for any undirected graph) and studied in [AHMR]. A function $Y\colon \{1, \cdots, n+m\} \to \mathbf{R}^k$ is called a *k-dimensional representation* of $G(n, m)$ if the following properties hold:

(i) The set $Y(1), \cdots, Y(n)$ is orthogonal, and the set $Y(n+1), \cdots, Y(n+m)$ is orthogonal (note that the vectors $Y(j)$ need not be nonzero);

(ii) The vectors $Y(1), \cdots, Y(n+m)$ span $\mathbf{R}^k$.

If Y is a k-dimensional representation of $G(n,m)$, then the $n \times n$ matrix $A_Y = [Y(1) \cdots Y(n)]^T[Y(1) \cdots Y(n)]$ has rank k and belongs to $M^+(G(n, m))$. The following fact has been established in Corollary 3.2 of [AHMR].

PROPOSITION 6.4. *The order of $G(n, m)$ in the real case coincides with the maximal k, for which there is a k-dimensional representation Y of $G(n, m)$ with the additional property that*

$$\dim \operatorname*{span}_{\substack{1 \leqq i \neq j \leqq n \\ or \\ n+1 \leqq i \neq j \leqq n+m}} \{Y(i)Y(j)^T + Y(j)Y(i)^T\} = \frac{k^2+k-2}{2} \tag{6.2}$$

A representation Y that satisfies (6.2) will be called *extremal*.

We now prove Theorem 6.1 in the real case. First, by putting

$$Y(i) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} i\text{th place}, \quad i = 1, \cdots, n,$$

$$Y(n+1) = \begin{bmatrix} 1 \\ -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad Y(n+2) = \begin{bmatrix} 1 \\ 1 \\ -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$Y(n+3) = \begin{bmatrix} 1 \\ 1 \\ 2 \\ -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, Y(n+4) = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 4 \\ -1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \cdots, Y(2n-1) = \begin{bmatrix} 1 \\ 1 \\ 2 \\ \vdots \\ 2n-3 \\ -1 \end{bmatrix},$$

$$Y(j) = 0 \quad \text{for } 2n \leqq j \leqq n+m$$

we obtain an extremal n-dimensional representation of $G(n, m)$. Thus, the order k of $G(n, m)$ is at least n. On the other hand, let Y be a k-dimensional extremal representation of $G(n, m)$.

The number of nonzero vectors in the set $Y(1), \cdots, Y(n)$ is at most $\min(k, n) = n$; the number of nonzero vectors in the set $Y(n+1), \cdots, Y(n+m)$ is at most $\min(k, m)$. Comparing with (6.2), we obtain

$$\frac{k^2+k-2}{2} \leqq \frac{n(n-1)}{2} + \frac{\min(k,m)(\min(k,m)-1)}{2}. \tag{6.3}$$

This inequality implies easily that

$$k \leqq (n^2-n-2)/2. \tag{6.4}$$

On the other hand, if $m > (n^2 - n)/4$, then $k \leqq m$. Indeed, assuming by contradiction that $k \geqq m + 1$, we have with the help of (6.3):

$$(m+1)^2+(m+1)-2 \leqq k^2+k-2 \leqq n^2-n+m^2-m,$$

which contradicts $m > (n^2 - n)/4$.

Now consider the case when $m = (n^2 - n + 2)/2$. The function

$$Y: \{1, \cdots, n+m\} \to R^m,$$

where $Y(1), \cdots, Y(n)$ is taken from Lemma 6.3 with $m = (n^2 - n + 2)/2$ and

$$Y(n+1) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad Y(n+2) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \cdots, Y(n+m) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{6.5}$$

is an irreducible representation of $G(n, (n^2 - n + 2)/2, (n^2 - n + 2)/2)$. Thus, $k \geqq m$. Together with (6.4) this shows

$$\text{order } G\left(\frac{n^2-n+2}{2}, \frac{n^2-n+2}{2}\right) = \frac{n^2-n+2}{2}.$$

As by Theorem 2.3

$$\text{order } G(n,m) \geqq \text{order } G\left(n, \frac{n^2-n+2}{2}\right)$$

for all $m > (n^2 - n + 2)/2$, the inequality (6.4) shows also that

$$\text{order } G(n,m) = \frac{n^2-n+2}{2}$$

for all $m \geqq ((n^2 - n + 2)/2)$.

Now consider the case when $m < ((n^2 - n + 2)/2)$.

To prove the theorem in the real case it remains for us to exhibit an irreducible m-dimensional representation of $G(n, m)$. To this end we define $Y(n + i)$ as in (6.5) for $i = 1, \cdots, m$, and $Y(j)$ as in Lemma 6.3 for $j = 1, \cdots, n$. This completes proof of Theorem 6.1 in the real case.

The proof of Theorem 6.1 in the complex case proceeds along similar lines. First, we prove Lemma 6.5.

LEMMA 6.5. *Let $n \geqq 2$ be an integer, and let $n \leqq m \leqq n^2 - n + 1$. Then there exists a linearly independent orthogonal set $Y(1), \cdots, Y(n)$ in $\mathbf{C}^m$ such that the vectors*

$$\operatorname{diag}(Y(i)Y(j)^*) \in \mathbf{C}^m \qquad (i \neq j)$$

span (over $\mathbf{C}$*) the linear space*

$$F = \{(z_1, \cdots, z_m)^T \in \mathbf{C}^m \mid z_1 + \cdots + z_m = 0\}.$$

Proof. The proof proceeds by induction on n. Suppose first that $n = 2$; so $2 \leqq m \leqq 3$. For $m = 2$, put

$$Y(1) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad Y(2) = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

For $m = 3$, put

$$Y(1) = \begin{bmatrix} 1 \\ 1 \\ i \end{bmatrix}, \qquad Y(2) = \begin{bmatrix} 1 \\ i \\ 1-i \end{bmatrix}.$$

Suppose the lemma is already proved with n replaced by $n - 1$. Let m' be such that

$$n - 1 \leqq m' \leqq (n-1)^2 - (n-1) + 1$$

and

$$1 \leqq m - m' \leqq 2n - 2.$$

By the induction hypothesis, there exist $\hat{Y}(1), \cdots, \hat{Y}(n-1) \in \mathbf{C}^{m'}$ with properties as in Lemma 6.5. Put

$$Y(1) = \begin{bmatrix} \hat{Y}(1) \\ 1 \\ i \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad Y(2) = \begin{bmatrix} \hat{Y}(2) \\ 0 \\ 0 \\ 1 \\ i \\ 0 \\ \vdots \\ 0 \end{bmatrix}; \cdots; Y(p) = \begin{bmatrix} \hat{Y}(p) \\ 0 \\ \vdots \\ 1 \\ i \end{bmatrix}, \quad \text{or } Y(p) = \begin{bmatrix} \hat{Y}(p) \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

depending on if $m - m'$ is even or odd; here $p = ((m - m' + 1)/2)$. Put

$$Y(j) = \begin{bmatrix} \hat{Y}(j) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{for } p < j \leqq n-1,$$

$$Y(n) = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_{m'} \\ 1 \\ \vdots \\ 1 \end{bmatrix}.$$

(The vectors $Y(1), \cdots, Y(n)$ are m-dimensional.) The numbers $\alpha_1, \cdots, \alpha_m$ are determined to satisfy the equations

$$\begin{pmatrix} \hat{Y}(1)^* \\ \hat{Y}(2)^* \\ \vdots \\ \hat{Y}(p-1)^* \\ \hat{Y}(p)^* \\ \hat{Y}(p+1)^* \\ \vdots \\ \hat{Y}(n-1)^* \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{m'} \end{pmatrix} = \begin{bmatrix} -1+i \\ -1+i \\ -1+i \\ x \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

where $x = -1 + i$ if $m - m'$ is even and $x = -1$ if $m - m'$ is odd. The vectors $Y(1), \cdots, Y(n)$ form a linearly independent and orthogonal set in $\mathbf{C}^m$. Furthermore,

$$\operatorname{diag}(Y(n)Y(j)^*) = (*, \cdots, *, 1, -i, 0, \cdots, 0) \tag{6.6}$$

for $j = 1, \cdots, p-1$, where 1 and $-i$ appear in the positions $m' + 2j - 1$ and $m' + 2j$, respectively (the stars in the right-hand side of (6.6) denote entries of no immediate interest to us). Also,

$$\operatorname{diag}(Y(n)Y(p)^*) = (*, \cdots, *, 1, -i)$$

or $(*, \cdots, *, 1)$, depending on if $m - m'$ is even or odd. These equalities together with the properties of $\hat{Y}(1), \cdots, \hat{Y}(n-1)$ (assumed by induction) show that

$$\operatorname{diag}(Y(k)Y(j)^*), \qquad k \neq j, \quad 1 \leqq k, j \leqq n$$

span the subspace F.

The proof of Theorem 6.1 is again based on the notion of a complex representation of $G(n, m)$ (this notion was introduced and studied in [AHMR]). The definition of a complex representation Y is the same as in the real case with the only modification that $Y: \{1, \cdots, n+m\} \rightarrow \mathbf{C}^k$. The analogue of Proposition 6.4 runs as follows (see [AHMR, Cor. 3.2]).

PROPOSITION 6.6. *The order of $G(n, m)$ in the complex case coincides with the maximal k, for which there is a complex k-dimensional representational Y of $G(n, m)$ with*

$$\dim \operatorname*{span}_{\substack{1 \leqq i \neq j \leqq n \\ or \\ n+1 \leqq i \neq j \leqq n+m}} \{Y(i)Y(j)^*\} = k^2 - 1. \tag{6.7}$$

(*The dimension here is the dimension of a vector space over C.*)

A complex representation Y with the property (6.7) will be called extremal. Return to the proof of Theorem 6.1 in the complex case. Let k be the order of $G(n, m)$ over $\mathbf{C}$. As in the real case, there is an extremal n-dimensional complex representation of $G(n, m)$, so $k \geqq n$. On the other hand, let Y be a k-dimensional (complex) extremal representation of $G(n, m)$. Arguing as in the proof of the real case and using Corollary 3.2 of [AHMR] again, we obtain

$$k^2 - 1 \leqq n(n-1) + \min(k, m)(\min(k, m) - 1). \tag{6.8}$$

This inequality easily implies

$$k \leqq n^2 - n + 1.$$

Also, if $m > (n^2 - n)/3$, then $k \leqq m$ (we prove this by contradiction assuming $k \geqq m + 1$ and using (6.8)). Now we finish the proof as in the real case using Lemma 6.5. □

Part III. Graphs of Large Order

7. Graphs with largest order. Theorem 3.6 will be used to provide the following description of graphs with the largest order (relative to the number of vertices).

THEOREM 7.1. *The order of a graph G with exactly n vertices (where $n \geqq 3$) is $\leqq n - 2$. Moreover, in the real case when $n \geqq 4$*, order $G = n - 2$ *if and only if G is a loop (see Theorem* 2.2 *for the definition of a loop).*

Proof. The first statement is just a particular case of Theorem 3.6 (because we can assume that G has at least one edge; otherwise, everything is trivial). Assuming the matrices are over R, the "if" part of the second statement is just Theorem 2.2.

Now let A be an extremal element in $M^+(G)$, and let rank $A = n - 2$ (we continue to consider the real case). By Theorem 3.6, G has no 3-cliques, i.e., triangles. On the other hand, Corollary 3.2 in [AHMR] shows that the number of edges in G is at most n.

Next, we show that G does not have pendants, i.e., vertices with degree 1. Indeed, assume that the first vertex has degree 1. Let A be an extremal matrix in $M^+(G)$ of rank $n - 2$. We can suppose that the first column of A is nonzero (otherwise, the first row and first column of A are zeros, and by deleting the first row and column we reduce the problem to the case of $(n - 1) \times (n - 1)$ matrices). Let a_1 be the first column of A. Put $R = a_1 a_1^*$. Because the first vertex has degree 1, we have $R \in M^+(G)$. Then for small $\varepsilon > 0$ we have $A - \varepsilon R \in M^+(G)$ (cf. the proof of Theorem 3.6), and hence the equality

$$A = \tfrac{1}{2}(A - \varepsilon R) + \tfrac{1}{2}(A + \varepsilon R)$$

contradicts the extremality of A (unless A is a scalar multiple of R; however, this case is excluded because rank $A = n - 2$ and $n \geqq 4$).

Furthermore, it is easy to see that G must be connected (because the order of a disconnected graph is the maximum of the orders of its connected components). Using the fact that the sum of the degrees of the vertices is twice the number of edges, and the absence of pendants, we note that the degree of each vertex in G is precisely two, so G must be a loop. □

8. Graphs with six vertices. As an application of Theorem 7.1 we can describe the orders of all graphs with at most six vertices.

Only the real case will be considered in this section.

We shall exclude from consideration chordal graphs (as their order is 1), disconnected graphs (as their order is the maximum among the orders of their connected components) and loops (as their order is given by Theorem 2.2).

THEOREM 8.1. *Let G be a connected nonchordal graph with at most six vertices and that is not a loop. Then the order of G is* 2 *except when G either $\leqq_v$-contains a loop with five vertices or G is one of the following six graphs:*

(1) *The fully bipartite graph $G(3, 3)$ (see* § 6).

(2) *The fully bipartite graph $G(3, 3)$ with precisely one edge added (the place of the added edge is immaterial because of graph isomorphism).*

(3) $G(3, 3)$ *with precisely one edge removed.*

(4) $G(3, 3)$ *with precisely one edge added and one edge removed, where the added and the removed edges are adjacent to the same vertex;*

(5) $V(G) = \{1, 2, 3, 4, 5, 6\}$; $E(G) = \{(i, j) \mid 1 \leqq i \neq j \leqq 6;\ 1 < |j - i| < 5\}$.

(6) *The graph as in* (5) *with the edge* (1, 6) *added.*

In all the exceptional cases the order of G is 3.

Proof. By Theorem 7.1 the order of G is either 2 or 3. If $G \leqq_v$- contains a 5-loop, then order $(G) = 3$ by Theorems 2.3 and 2.2. Graphs (1)–(6) have order 3 by Theorem 8.2 of [AHMR] (all of them are 3-blocks). Conversely, assume that G has order 3. Then G must $\leqq_v$-contain a 3-block. By Theorem 7.1 the only 3-block with five or fewer vertices is the 5-loop, while Theorem 8.2 of [AHMR] provides a list of all 3-blocks with six vertices, which are precisely the six graphs listed in the theorem. □

Acknowledgment. We thank P. DeWilde for pointing out that the sparsity pattern in Fig. 1 and similar double-banded patterns occur frequently in VLSI circuit analysis.

REFERENCES

[AHMR] J. Agler, J. W. Helton, S. McCullough, and L. Rodman, *Positive definite matrices with a given sparsity pattern*, Linear Algebra Appl., 107 (1988), pp. 101–149.

[DG] H. Dym and I. Gohberg, *Extensions of band matrices with band inverses*, Linear Algebra Appl., 36 (1981), pp. 1–24.

[EGL] R. Ellis, I. Gohberg, and D. C. Lay, *Invertible selfadjoint extensions of band matrices and their entropy*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 483–500.

[G] M. C. Golumbic, *Algorithmic Graph Theory and Perfect Graphs*, Academic Press, New York, 1980.

[GJSW] R. Grone, C. R. Johnson, E. M. de Sa, and H. Wolkowitz, *Positive definite completions of partial Hermitian matrices*, Linear Algebra Appl., 58 (1984), pp. 102–124.

[GL] A. George and J. W. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice Hall, Englewood Cliffs, NJ, 1981.

[H] J. W. Helton (with J. A. Ball, C. R. Johnson, and J. N. Palmer), *Operator theory, analytic functions, matrices and electrical engineering*, CBMS Vol. 68, American Mathematical Society, Providence, RI, 1987.

[JR] C. R. Johnson and L. Rodman, *Inertia possibilities for completions of partial Hermitian matrices*, Linear and Multilinear Algebra, 16 (1984), pp. 179–195.

[M] S. McCullough, *2-chordal graphs*, Operator Theory: Adv. Appl., 35 (1988), pp. 133–192.

[P] S. Parter, *The use of linear graphs in Gauss elimination*, SIAM Rev., 3 (1961), pp. 119–130.

[PPS] V. I. Paulsen, S. C. Power, and R. R. Smith, *Schur products and matrix completions*, preprint.

[R] D. Rose, *A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations*, in Graph Theory and Computing, R. Read, ed., Academic Press, New York, 1973, pp. 183–217.

[RT] D. J. Rose and R. E. Tarjan, *Algorithmic aspects of vertex elimination*, in Proc. 7th Annual ACM Symposium on the Theory of Computing, Association for Computing Machinery, New York, 1975, pp. 245–254.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 3, pp. 424–428, July 1989

009

UPDATING THE TRIANGULAR FACTORIZATION OF A MATRIX*

J. L. NAZARETH†

Abstract. When one column of a square nonsingular matrix, say B^0, is replaced by another, the Bartels–Golub update of the LU factors of B^0 forms a product form representation that is suitable for the efficient solution of associated systems of linear equations. However, it does not update the factors so as to obtain (even implicitly) the LU factors of the new matrix. In this note an exceedingly simple modification of the Bartels–Golub technique that yields the true LU factors is described, and it is shown to be mathematically equivalent to the update (specifically the partial pivoting version) recently proposed by Fletcher and Matthews [*Math. Programming*, 30 (1984), pp. 267–284]. The other commonly used update, Forrest–Tomlin, is quite evidently a mathematical variant of the Bartels–Golub technique, so this result enables a common ground for the three major updating techniques of linear programming to be established.

Key words. LU factors, updating triangular factorization, Bartels–Golub update, Fletcher–Matthews update

AMS(MOS) subject classifications. 65, 90, 49

1. Introduction. Consider an $n \times n$ nonsingular matrix B^0 and its triangular factorization in the form

$$P^0B^0 = L^0U^0, \tag{1.1}$$

where P^0 is a permutation matrix, L^0 is a unit lower triangular matrix, and U^0 is an upper triangular matrix. When a single column of B^0 is replaced, yielding a new nonsingular matrix, say B, the updating technique of Bartels and Golub [1] may be used to revise the factorization. It develops a *product form representation* that is suitable for the efficient solution of associated systems of linear equations. However, it does not update the factors in (1.1) so as to obtain (even implicitly) a new triangular factorization of the form

$$PB = LU, \tag{1.2}$$

where again P is a permutation matrix, L is a unit lower triangular matrix, and U is an upper triangular matrix.

In this note we show that an exceedingly simple modification of the Bartels–Golub update does indeed yield the true LU factors. We use the well-known uniqueness property of LU factorization to show *mathematical equivalence* to the recent updating technique (specifically its partial pivoting version) proposed by Fletcher and Matthews [2]. The other main update of linear programming is that of Forrest and Tomlin [3], which is quite evidently a mathematical variant of Bartels–Golub updating. By establishing here the connection between the Bartels–Golub and Fletcher–Matthews updates, we are therefore able to provide a common ground for the three main updating techniques of linear programming.

2. Derivation of the update.

2.1. Background on the Bartels–Golub update. It is convenient to carry out the development in terms of a 4×4 matrix. The reader will then have no difficulty whatsoever in extending the results to the more general case of an $n \times n$ matrix. Without loss of generality, we assume that the new matrix B is obtained by removing the *first* column

* Received by the editors May 25, 1987; accepted for publication (in revised form) December 21, 1988.

† Department of Pure and Applied Mathematics, Washington State University, Pullman, Washington 99164 (nazareth@wsumath.bitnet).

of B^0, advancing subsequent columns by one position, and inserting the new replacing column in the last position. (For descriptive purposes, we shall employ explicit inverses. An actual computation should be organized differently, using backsubstitution.)

With the above assumptions, we see that (1.1) gives

$$(L^0)^{-1}(P^0B)=H^{(0)}, \tag{2.1}$$

where $H^{(0)}$ is *an upper Hessenberg* matrix, as depicted in Fig. 2.1. (For intermediate matrices, we shall use parentheses around the superscript.) Since B^0 is assumed to be nonsingular, we have $h_{k+1,k}^{(0)} \neq 0$, $1 \leqq k \leqq n-1$. In Bartels–Golub updating with partial pivoting, elementary matrices are used to eliminate successive subdiagonal elements, thereby yielding an upper triangular matrix. To establish our notation, let us consider the first step that eliminates the element in position (2, 1) of $H^{(0)}$, namely,

$$\hat{H}^{(1)}=P_{1,1'}H^{(0)}, \tag{2.2a}$$

$$\bar{H}^{(1)}=\Gamma_1\hat{H}^{(1)}, \tag{2.2b}$$

where $P_{1,1'}$ is an elementary permutation matrix, either the identity matrix (when $|h_{11}^{(0)}| \geqq |h_{21}^{(0)}|$ so that no interchange of rows is performed), or $P_{1,2}$, the elementary permutation matrix that interchanges the first and second rows of $H^{(0)}$. Γ_1 denotes an elementary lower triangular matrix (also depicted in Fig. 2.1), with $\gamma = -\hat{h}_{21}^{(1)}/\hat{h}_{11}^{(1)}$. The process can be continued in the obvious way to eliminate subsequent subdiagonal elements, and details may be found in Bartels and Golub [1].

Suppose that interchanges were permitted throughout the procedure but *none were actually needed* to maintain stability (or, in the more general case, to reduce fill-in), i.e., $P_{k,k'} = I$, $k = 1, 2, 3$. The Bartels–Golub update would then take the form

$$\Gamma_3\Gamma_2\Gamma_1(L^0)^{-1}(P^0B)=U, \tag{2.2c}$$

where U is upper triangular. This may be expressed as $P^0B = (L^0\Gamma_1^{-1}\Gamma_2^{-1}\Gamma_3^{-1})U$. Noting that Γ_k^{-1}, $k=1, 2, 3$ is also lower triangular and using the definitions $L \equiv L^0\Gamma_1^{-1}\Gamma_2^{-1}\Gamma_3^{-1}$ and $P \equiv P^0$, we have the update of the (true) triangular factorization, namely, (1.2). Clearly, the novelty of techniques that seek the true triangular factors must arise when interchanges occur, i.e., when $P_{k,k'} \neq I$ for some k.

Let us therefore return to (2.2a)–(2.2b) and now assume that an interchange was performed, so $P_{1,1'} = P_{1,2}$. The elimination step (2.2b) completes the first iteration of Bartels–Golub updating, which proceeds to the processing of the second column of $\bar{H}^{(1)}$, in order to eliminate the element in position (3, 2), and so on. In the modification now to be described, (2.2b) still represents only an intermediate stage of this first iteration.

$$L^0=\begin{bmatrix}1&0&0&0\\x&1&0&0\\x&x&1&0\\x&x&x&1\end{bmatrix},\quad H^{(0)}=\begin{bmatrix}x&x&x&x\\x&x&x&x\\0&x&x&x\\0&0&x&x\end{bmatrix},\quad \Gamma_1=\begin{bmatrix}1&0&0&0\\\gamma&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix},$$

$$M^{(1)}=\begin{bmatrix}x&x&0&0\\x&x&0&0\\x&x&1&0\\x&x&x&1\end{bmatrix},\quad U^{(1)}=\begin{bmatrix}x&x&0&0\\0&x&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix},\quad H^{(1)}=\begin{bmatrix}x&x&x&x\\0&x&x&x\\0&x&x&x\\0&0&x&x\end{bmatrix}.$$

FIG. 2.1

2.2. A simple modification to obtain the true LU factors. So far there has been no departure whatsoever from Bartels–Golub updating. Using (2.1), (2.2a)–(2.2b), and the assumption that an interchange was performed, we have

$$P^0B=(L^0P_{1,2}\Gamma_1^{-1})\bar{H}^{(1)}. \tag{2.3}$$

Now consider the matrix

$$M^{(1)}\equiv L^0P_{1,2}\Gamma_1^{-1}.$$

This is obviously lower triangular in all but the first two columns, as depicted in Fig. 2.1. The elements $m_{11}^{(1)}$ and $m_{12}^{(1)}$ of the matrix $M^{(1)}$ are given by

$$m_{11}^{(1)}=l_{12}^0+l_{11}^0(\hat{h}_{21}^{(1)}/\hat{h}_{11}^{(1)})=(\hat{h}_{21}^{(1)}/\hat{h}_{11}^{(1)})=(h_{11}^{(0)}/h_{21}^{(0)}), \tag{2.4a}$$

$$m_{21}^{(1)}=l_{22}^0+l_{21}^0(\hat{h}_{21}^{(1)}/\hat{h}_{11}^{(1)})=1+l_{21}^0(\hat{h}_{21}^{(1)}/\hat{h}_{11}^{(1)})=1+l_{21}^0(h_{11}^{(0)}/h_{21}^{(0)}). \tag{2.4b}$$

The final expressions in (2.4a)–(2.4b) use the fact that L^0 is *unit* lower triangular and $P_{1,1'}=P_{1,2}$ in (2.2a). Observe that $m_{11}^{(1)}=0$ implies that $m_{21}^{(1)}\neq 0$, so clearly both elements cannot be simultaneously zero. This observation, along with nonsingularity of the matrix $M^{(1)}$, implies that we can perform the triangular factorization with partial row pivoting restricted to the first *two* rows and ensure that the process does *not* break down in *exact* arithmetic because of a *zero* pivot. This factorization of $M^{(1)}$ is given by

$$P_{1,\bar{1}}M^{(1)}=L^{(1)}U^{(1)}, \tag{2.5}$$

where $P_{1,\bar{1}}$ is either the identity matrix (when $|m_{11}^{(1)}|\geqq|m_{21}^{(1)}|$ so no row interchange is performed) or the elementary permutation matrix $P_{1,2}$ (when the first two rows are interchanged). $L^{(1)}$ is unit lower triangular, $U^{(1)}$ is upper triangular of the form depicted in Fig. 2.1, and obviously both matrices are nonsingular.

Therefore, from (2.3) and (2.5),

$$P^0B=M^{(1)}\bar{H}^{(1)}=P_{1,\bar{1}}L^{(1)}(U^{(1)}\bar{H}^{(1)}).$$

Define $H^{(1)}\equiv U^{(1)}\bar{H}^{(1)}$, and observe that $H^{(1)}$ is also upper Hessenberg with $h_{21}^{(1)}=0$. This too is depicted in Fig. 2.1. Therefore

$$P_{1,\bar{1}}P^0B=L^{(1)}H^{(1)}. \tag{2.6}$$

This completes the first cycle of the iterative procedure summarized by the example of Fig. 2.1. The objective has been to obtain a factorization (2.6) that is of the required form (1.2) insofar as the first column is concerned.

2.3. Completing the procedure. We can now repeat the above procedure for the second column of $H^{(1)}$ in a completely analogous manner, which we give for completeness:

$$(L^{(1)})^{-1}P_{1,\bar{1}}P^0B=H^{(1)},\quad \hat{H}^{(2)}=P_{2,2'}H^{(1)},\quad \bar{H}^{(2)}=\Gamma_2\hat{H}^{(2)},$$

where $P_{2,2'}$ is either the identity matrix (in which case we may proceed to the next iteration) or the elementary permutation matrix $P_{2,3}$ that interchanges the second and third rows of $H^{(1)}$, and Γ_2 is defined analogously to Γ_1:

$$P_{1,\bar{1}}P^0B=(L^{(1)}P_{2,2'}\Gamma_2^{-1})\bar{H}^{(2)}=M^{(2)}\bar{H}^{(2)},$$

with the definition $M^{(2)}\equiv L^{(1)}P_{2,2'}\Gamma_2^{-1}$:

$$P_{1,\bar{1}}P^0B=P_{2,\bar{2}}L^{(2)}(U^{(2)}\bar{H}^{(2)}),$$

where $P_{2,\bar{2}}M^{(2)}=L^{(2)}U^{(2)}$, with $L^{(2)}$ unit lower triangular and $U^{(2)}$ upper triangular.

Therefore

$$P_{2,\bar{2}}P_{1,\bar{1}}P^{0}B = L^{(2)}H^{(2)},$$

where $H^{(2)} \equiv U^{(2)}\bar{H}^{(2)}$ and $h_{32}^{(2)} = 0$.

Finally, after one more such iteration,

$$P_{3,\bar{3}}P_{2,\bar{2}}P_{1,\bar{1}}P^{0}B = L^{(3)}H^{(3)}, \tag{2.7}$$

where $L^{(3)}$ is unit lower triangular, and $H^{(3)}$ is now upper triangular. With the definitions $P \equiv P_{3,\bar{3}}P_{2,\bar{2}}P_{1,\bar{1}}P^{0}$, $L \equiv L^{(3)}$, and $U \equiv H^{(3)}$, we have the required factorization (1.2).

The extension to the update of an $n \times n$ matrix is straightforward.

3. Mathematical equivalence to the Fletcher–Matthews update with partial pivoting. When B is nonsingular and P is prescribed, it is well known and easy to establish that the LU factorization (1.2) is unique. Equivalence of the updating technique of § 2 and the partial pivoting version of the Fletcher–Matthews update is a direct consequence. All that must be established is that the permutation matrix P is the same in both cases. Let us again consider the first iteration, with the argument at subsequent iterations being completely analogous. Obviously the two updates do not differ when no interchange is performed in (2.2a), so let us consider the second possibility discussed in § 2.2. In this case it follows from (2.4a)–(2.4b) that the comparison of the two elements that determines the choice for $P_{1,\bar{1}}$ can be stated as

$$|h_{11}^{(0)}| \geqq |h_{21}^{(0)} + l_{21}^{0}h_{11}^{(0)}|. \tag{3.1}$$

This is precisely expression (2.19) in Fletcher and Matthews [2] with the appropriate transposition of notation. Analogous arguments show that the tests determining $P_{2,\bar{2}}$ and $P_{3,\bar{3}}$ are identical, and hence the matrix P is the same in the two developments (see also expression (2.16) in [2]).

We again emphasize that we have been concerned in this short note with establishing a *mathematical* relationship between the Bartels–Golub and Fletcher–Matthews updates. Viewed from this perspective, the derivation and procedure given by Fletcher and Matthews [2] can be seen to be a particular *reformulation* of a simpler mathematical algorithm that underlies it. The reformulation is designed to improve numerical characteristics, in particular:

(1) To enhance *efficiency* by observing, at the first iteration, that (2.5) can be reformulated as $P_{1,\bar{1}}L^{0}P_{1,2}\Gamma_{\bar{1}}^{-1}(U^{(1)})^{-1} = L^{(1)}$, where $(U^{(1)})^{-1}$ is an elementary matrix of the same form as $U^{(1)}$. The first two columns of $L^{(1)}$ are therefore linear combinations of the first two columns of L^{0}. Once $P_{1,1'}$ and $P_{1,\bar{1}}$ are chosen (for example, as discussed in §§ 2.1 and 2.2), then explicit expressions for these linear combinations can be easily written down. Analogous statements hold at subsequent iterations.

(2) To enhance *numerical stability* through the use of other tactics (backed by error analysis) to determine $P_{1,1'}$ and $P_{1,\bar{1}}$ (at the first iteration) again based on the elements $h_{11}^{(0)}$, $h_{21}^{(0)}$, and l_{21}^{0}. Analogous statements can be made at subsequent iterations.

These two reformulations represent important contributions of Fletcher and Matthews [2]. In particular, a judicious choice of pivot strategy in the factorization (2.5), and more generally in

$$P_{k,\bar{k}}M^{(k)} = L^{(k)}U^{(k)}, \tag{3.2}$$

enables Fletcher and Matthews [2] to carry out the update so that it is generally well behaved. Note, however, when partial pivoting is restricted at each step to rows k and

$k + 1$, we *cannot* place an a priori bound on the size of elements of $L^{(k)}$, in contrast to Bartels–Golub updating. To see this, let us return to the first iteration. The pivot choice is restricted to $m_{11}^{(1)}$ and $m_{21}^{(1)}$ in (2.5). Thus the elements of $L^{(1)}$ in positions (3, 1) and (4, 1) are of the form $m_{i1}^{(1)}/m_{11}^{(1)}$, $i = 3, 4$ or $m_{i1}^{(1)}/m_{21}^{(1)}$, $i = 3, 4$, and they could be large. We can, however, bound elements of $L^{(k)}$ by permitting partial pivoting in other rows when factorizing $M^{(k)}$ in (3.2). The price we pay is that $U^{(k)}$ is no longer necessarily a simple matrix with just one off-diagonal element, and there is an associated increase in the cost of factorizing $M^{(k)}$ and forming $H^{(k)} = U^{(k)}\bar{H}^{(k)}$. For example, at the first iteration, suppose $P_{1,\bar{1}}$ were taken to be $P_{1,4}$. Then $U^{(1)}$ would be a full upper triangular matrix. However, it will only be necessary to invoke this more general pivoting strategy under extreme circumstances, and it is clear that many improved strategies are possible. These permit some limited growth in elements, as used in various implementations of Bartels–Golub updating (see, in particular, Reid [4]).

We may also note that our development falls within a more conventional framework and permits more direct use of standard error analysis techniques following Wilkinson [5]. Although it is only speculation at this point, it may also be helpful when devising extensions to preserve sparsity, in the hope of making Fletcher–Matthews updating more competitive for large-scale applications. This is currently an open question.

Acknowledgments. This paper grew out of a discussion between the author and Dr. Michael Saunders, who should share credit for the main idea.

REFERENCES

[1] R. H. Bartels and G. H. Golub, *The simplex method of linear programming using the* LU *decomposition*, Comm. ACM, 12 (1969), pp. 266–268.

[2] R. Fletcher and S. P. J. Matthews, *Stable modification of explicit* LU *factors for simplex updates*, Math. Programming, 30 (1984), pp. 267–284.

[3] J. J. H. Forrest and J. A. Tomlin, *Updating triangular factors of the basis to maintain sparsity in the product form simplex method*, Math. Programming, 2 (1972), pp. 263–278.

[4] J. K. Reid, *A sparsity-exploiting version of the Bartels–Golub decomposition for linear programming bases*, AERE Harwell Report CSS 20, Harwell, U.K., 1975.

[5] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 4, pp. 429–434, October 1989

MAJORIZATION AND SINGULAR VALUES II*

R. B. BAPAT†

Abstract. Let $B, D_j, E_j, j = 1, 2, \cdots, k$, be $n \times n$ complex matrices. It is shown that

$$\sigma\left(\sum_{j=1}^{k} D_j B E_j^*\right) \prec_w \sigma(B) \circ \delta$$

where δ is any vector with components $\delta_1 \geqq \cdots \geqq \delta_n$ that weakly majorizes both the following vectors:

$$\sigma(\sum D_j D_j^*)^{1/2} \circ \sigma(\sum E_j E_j^*)^{1/2} \quad \text{and} \quad \sigma(\sum D_j^* D_j)^{1/2} \circ \sigma(\sum E_j^* E_j)^{1/2}.$$

Here $\sigma(\cdot)$ denotes the vector of singular values arranged in nonincreasing order, $\prec_w$ denotes weak majorization, and $\circ$ indicates Schur (entrywise) multiplication. The result unifies several known results concerning majorization statements for singular values.

Key words. majorization, singular values, Schur product

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

1. Introduction.

Let M_n denote the space of $n \times n$ complex matrices.

For any $A \in M_n$ we denote by

$$\sigma_1(A) \geqq \cdots \geqq \sigma_n(A) \geqq 0$$

the singular values of A that by definition are the nonnegative square roots of the eigenvalues of AA^*. Also, we set

$$\sigma(A) = (\sigma_1(A), \cdots, \sigma_n(A))^t$$

where t denotes transpose.

If $A \in M_n$ is a Hermitian matrix, then

$$\lambda_1(A) \geqq \cdots \geqq \lambda_n(A)$$

denotes the eigenvalues of A and again we define

$$\lambda(A) = (\lambda_1(A), \cdots, \lambda_n(A))^t.$$

If $A = ((a_{ij}))$ and $B = ((b_{ij}))$ are $m \times n$ matrices, then their Schur product (also known as the Hadamard product) is defined as $A \circ B = ((a_{ij} b_{ij}))$.

If $x \in R^n$, then $x_{[1]} \geqq \cdots \geqq x_{[n]}$ denotes the components of x arranged in nonincreasing order and we define

$$x_\downarrow = (x_{[1]}, \cdots, x_{[n]})^t.$$

If $x, y \in R^n$, then x is said to be majorized by y, $x \prec y$, if

$$\sum_{i=1}^{m} x_{[i]} \leqq \sum_{i=1}^{m} y_{[i]}, \qquad m = 1, 2, \cdots, n-1 \tag{1}$$

* Received by the editors June 22, 1988; accepted for publication (in revised form) December 15, 1988. This paper was presented at the Mathematical Sciences Lecture Series on Matrix Spectral Inequalities at Johns Hopkins University, Baltimore, Maryland, June 20–24, 1988. The principal lecturer for the 1988 Series was Professor Robert C. Thompson.

† Indian Statistical Institute, New Delhi, 110016, India.

and

$$\sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} y_{[i]}.$$

If $x, y \in R^n$, then x is said to be weakly majorized by y, $x \prec_w y$, if the inequalities (1) hold for $m = 1, 2, \cdots, n$.

If $\alpha \in R^n$, then $\Delta(\alpha)$ denotes the diagonal matrix

$$\Delta(\alpha) = \begin{bmatrix} \alpha_1 & & \\ & \ddots & \\ & & \alpha_n \end{bmatrix}.$$

We can now state the main result of this paper.

THEOREM 1. *Let $B, D_j, E_j \in M_n$, $j = 1, 2, \cdots, k$. Then*

$$\sigma(\sum D_j B E_j^*) \prec_w \sigma(B) \circ \delta_\downarrow \tag{2}$$

where δ is any vector which weakly majorizes both $\lambda(\sum D_j D_j^*)^{1/2} \circ \lambda(\sum E_j E_j^*)^{1/2}$ and $\lambda(\sum D_j^* D_j)^{1/2} \circ \lambda(\sum E_j^* E_j)^{1/2}$.

In Theorem 1 all the summations extend over $j = 1, 2, \cdots, k$ and this convention will be followed in the rest of the paper unless specified otherwise.

Before proving Theorem 1 we note two of its main consequences.

We first show that the main result of [2] follows easily from Theorem 1.

THEOREM 2 [2, Thm. 2]. *Let $B, D_j, E_j \in M_n$, $j = 1, 2, \cdots, k$. Then*

$$\sigma(\sum D_j B E_j^*) \prec_w \sigma(B) \circ \delta_\downarrow$$

where δ is a vector that weakly majorizes all the vectors: $\lambda(\sum D_j D_j^*)$, $\lambda(\sum D_j^* D_j)$, $\lambda(\sum E_j E_j^*)$, and $\lambda(\sum E_j^* E_j)$.

Proof. If δ satisfies the given hypothesis, then it weakly majorizes the vectors

$$\frac{1}{2}\{\lambda(\sum D_j D_j^*) + \lambda(\sum E_j E_j^*)\} \quad \text{and} \quad \frac{1}{2}\{\lambda(\sum D_j^* D_j) + \lambda(\sum E_j^* E_j)\}.$$

By the arithmetic mean-geometric mean inequality it follows that δ weakly majorizes both

$$\lambda(\sum D_j D_j^*)^{1/2} \circ \lambda(\sum E_j E_j^*)^{1/2}, \quad \text{and} \quad \lambda(\sum D_j^* D_j)^{1/2} \circ \lambda(\sum E_j^* E_j)^{1/2}.$$

Now the result follows from Theorem 1. □

In a recent paper, Ando, Horn, and Johnson [1] have given a basic majorization inequality for singular values of a Schur product. The inequality unifies a number of previously known results concerning eigenvalues and singular values of Schur products. It turns out that the basic inequality of [1] is a special case of Theorem 1 obtained by restricting the matrices D_j, E_j to be diagonal. This is shown next.

We will need the following notation. If Z is an $m \times n$ matrix, then $c_1(Z) \geqq c_2(Z) \geqq \cdots \geqq c_n(Z) \geqq 0$ will denote the Euclidean lengths of the columns of Z.

THEOREM 3 [1, Thm. 1]. *Let m, n be positive integers, let $q = \min\{m, n\}$, and let A, B be $m \times n$ matrices. Then*

$$\sum_{i=1}^{k} \sigma_i(A \circ B) \leqq \sum_{i=1}^{k} c_i(X) c_i(Y) \sigma_i(B), \qquad k = 1, 2, \cdots, q$$

for any $r \times m$ matrix X and any $r \times n$ matrix Y such that $A = X^* Y$.

Proof. As in [1] we first note that the general (nonsquare) case of the theorem follows from the case $m = n$ by augmenting nonsquare matrices to square ones with rectangular blocks of zeros. Thus, without loss of generality, we assume that $m = n$.

Define diagonal matrices

$$D_j = \begin{bmatrix} d_{j1} & & \\ & \ddots & \\ & & d_{jn} \end{bmatrix}, \quad E_j = \begin{bmatrix} e_{j1} & & \\ & \ddots & \\ & & e_{jn} \end{bmatrix}, \quad j = 1, 2, \cdots, r$$

by setting

$$d_{ji} = \overline{x_{ji}}, \quad e_{ji} = \overline{y_{ji}}, \quad i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, r.$$

Then it can be verified that

$$A \circ B = \sum_{j=1}^{r} D_j B E_j^*.$$

Also,

$$\sum D_j D_j^* = \sum D_j^* D_j = \begin{bmatrix} \sum_j |x_{j1}|^2 & & \\ & \ddots & \\ & & \sum_j |x_{jn}|^2 \end{bmatrix}$$

and

$$\sum E_j E_j^* = \sum E_j^* E_j = \begin{bmatrix} \sum_i |y_{i1}|^2 & & \\ & \ddots & \\ & & \sum_i |y_{in}|^2 \end{bmatrix}.$$

Thus

$$\begin{aligned} \lambda(\sum D_j D_j^*)^{1/2} \circ \lambda(\sum E_j E_j^*)^{1/2} &= \lambda(\sum D_j^* D_j)^{1/2} \circ \lambda(\sum E_j^* E_j)^{1/2} \\ &= (c_1(X)c_1(Y), \cdots, c_n(X)c_n(Y))^t \end{aligned}$$

and the result follows from Theorem 1. □

For several further consequences of Theorems 2 and 3 we refer to [1] and [2].

2. Proof of the main result. We first state several assertions that we need. Most of these are either well known or are easy to prove. A reference is given in each case for convenience.

The next result gives the familiar singular value decomposition of a matrix.

LEMMA 4 [3, p. 498]. *Let $A \in M_n$. Then there exist $n \times n$ unitary matrices U, V such that*

$$A = U \Delta(\sigma(A)) V^*.$$

LEMMA 5 [3, p. 249, 243]. *Let X, $Y \in M_n$. Then*

(a) $\sigma(XY) \prec_w \sigma(X) \circ \sigma(Y)$;

(b) $\sigma(X + Y) \prec_w \sigma(X) + \sigma(Y)$.

LEMMA 6 [1, p. 352]. *Let $X \in M_n$. Then the matrix*

$$\begin{bmatrix} \sigma_1(X) I_n & X \\ X^* & \sigma_1(X) I_n \end{bmatrix}$$

is positive semidefinite.

LEMMA 7 [1, p. 351]. *Suppose the matrix*

$$\begin{bmatrix} X & Y \\ Y^* & Z \end{bmatrix}$$

is positive semidefinite. Then there exists a matrix W with $\sigma_i(W) \leqq 1$ for all i, such that $Y = X^{1/2}WZ^{1/2}$.

LEMMA 8 [3, p. 228]. *Let $A \in M_n$. Then*

$$(|a_{11}|, \cdots, |a_{nn}|)' \prec_w \sigma(A).$$

In particular, $|\operatorname{tr} A| \leqq \sum_{i=1}^{n} \sigma_i(A)$.

We now proceed to give the proof of the main result. The proof broadly follows the same steps as those used by Ando, Horn, and Johnson [1] to prove their main result. However, there are several differences in the details of the proofs.

We first establish a result that is much weaker than (2).

LEMMA 9. *Let B, D_j, $E_j \in M_n$, $j = 1, 2, \cdots, k$ and let δ be a vector that weakly majorizes $\lambda(\sum D_jD_j^*)^{1/2} \circ \lambda(\sum E_jE_j^*)^{1/2}$. Then*

$$\sigma(\sum D_jBE_j^*) \prec_w \sigma_1(B)\delta_{\downarrow}.$$

Proof. By Lemma 6 the matrix

$$C = \begin{bmatrix} \sigma_1(B)I_n & B \\ B^* & \sigma_1(B)I_n \end{bmatrix}$$

is positive semidefinite, and hence so is the matrix

$$\sum \begin{bmatrix} D_j & 0 \\ 0 & E_j \end{bmatrix} C \begin{bmatrix} D_j^* & 0 \\ 0 & E_j^* \end{bmatrix} = \begin{bmatrix} \sigma_1(B)\sum D_jD_j^* & \sum D_jBE_j^* \\ \sum E_jB^*D_j^* & \sigma_1(B)\sum E_jE_j^* \end{bmatrix}.$$

By Lemma 7 there exists $W \in M_n$ with $\sigma_i(W) \leqq 1$, $i = 1, 2, \cdots, n$ such that

$$\sum D_jBE_j^* = \sigma_1(B)(\sum D_jD_j^*)^{1/2}W(\sum E_jE_j^*)^{1/2}.$$

Now by Lemma 5(a),

$$\begin{aligned} \sigma(\sum D_jBE_j^*) &\prec_w \sigma_1(B)\lambda(\sum D_jD_j^*)^{1/2} \circ \sigma(W) \circ \lambda(\sum E_jE_j^*)^{1/2} \\ &\prec_w \sigma_1(B)\lambda(\sum D_jD_j^*)^{1/2} \circ \lambda(\sum E_jE_j^*)^{1/2}, \end{aligned}$$

since $\sigma_i(W) \leqq 1$, $i = 1, 2, \cdots, n$.

Hence

$$\sigma(\sum D_jBE_j^*) \prec_w \sigma_1(B)\delta_{\downarrow}. \qquad \square$$

We now introduce notation. For any integer r, $1 \leqq r \leqq n$, K_r will denote the $n \times n$ matrix

$$\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}.$$

The next step in the proof is to establish the result when $B = K_r$ for some r.

LEMMA 10. *Let D_j, $E_j \in M_n$, $j = 1, 2, \cdots, k$; let δ be a vector satisfying the hypothesis given in Theorem 1; and let r be an integer, $1 \leqq r \leqq n$. Then*

$$\sigma(\sum D_jK_rE_j^*) \prec_w \sigma(K_r) \circ \delta_{\downarrow} = (\delta_{[1]}, \cdots, \delta_{[r]}, 0, \cdots, 0)'. \tag{3}$$

Proof. We first observe that if D_j, E_j are replaced by UD_j, VE_j, respectively, for each j, where U, V are unitary matrices, then the eigenvalues of

$$\sum D_jD_j^*, \quad \sum D_j^*D_j, \quad \sum E_jE_j^*, \quad \sum E_j^*E_j$$

remain unchanged. Thus, using the singular value decomposition (Lemma 4), we assume, without loss of generality, that

$$\sum D_jK_rE_j^* = \Delta(\sigma(\sum D_jK_rE_j^*)). \tag{4}$$

Let s be fixed, $1 \leqq s \leqq n$. By Lemma 9, we have

$$\begin{aligned} \sum_{i=1}^{s} \sigma_i(\sum D_jK_rE_j^*) &\leqq \sum_{i=1}^{s} \sigma_1(K_r)\delta_{[i]} \\ &= \sum_{i=1}^{s} \delta_{[i]}. \end{aligned} \tag{5}$$

Also, in view of (4), we have

$$\begin{aligned} \sum_{i=1}^{s} \sigma_i(\sum D_jK_rE_j^*) &= \sum_{i=1}^{s} \sigma_i(\sum D_jK_rE_j^*K_s) \\ &= \operatorname{tr} \sum D_jK_rE_j^*K_s \\ &= \sum \operatorname{tr} D_jK_rE_j^*K_s \\ &= \sum \operatorname{tr} K_rE_j^*K_sD_j \\ &= \operatorname{tr} K_r \sum E_j^*K_sD_j \\ &\leqq \sum_{i=1}^{n} \sigma_i(K_r \sum E_j^*K_sD_j) \quad \text{(by Lemma 8)} \\ &\leqq \sum_{i=1}^{n} \sigma_i(K_r)\sigma_i(\sum E_j^*K_sD_j) \quad \text{(by Lemma 5(a))} \\ &= \sum_{i=1}^{r} \sigma_i(\sum E_j^*K_sD_j) \\ &\leqq \sum_{i=1}^{r} \sigma_1(K_s)\delta_{[i]} \quad \text{(by Lemma 9)} \\ &= \sum_{i=1}^{r} \delta_{[i]}. \end{aligned} \tag{6}$$

Combining (5) and (6), we have

$$\sum_{i=1}^{s} \sigma_i(\sum D_jK_rE_j^*) \leqq \sum_{i=1}^{\min(r,s)} \delta_{[i]}$$

and this establishes (3). □

We can now complete the proof of the main result.

Proof of Theorem 1. Just as we remarked at the beginning of the proof of Lemma 10, the eigenvalues of $\sum D_jD_j^*$, etc. also remain unaffected if each D_j, E_j is replaced by

D_jU, E_jV for some unitary matrices U, V. Therefore using the singular value decomposition of B if necessary, we assume, without loss of generality, that

$$B=\Delta(\sigma(B)).$$

It is easily verified that

$$B=\sum_{r=1}^{n}(\sigma_r(B)-\sigma_{r+1}(B))K_r$$

where we set $\sigma_{n+1}(B)=0$. Thus

$$\sum D_jBE_j^*=\sum_{r=1}^{n}(\sigma_r(B)-\sigma_{r+1}(B))\sum D_jK_rE_j^*.$$

By Lemma 5(b),

$$\begin{aligned}\sigma(\sum D_jBE_j^*)\prec_w&\sum_{r=1}^{n}\sigma\{(\sigma_r(B)-\sigma_{r+1}(B))\sum D_jK_rE_j^*\}\\&=\sum_{r=1}^{n}(\sigma_r(B)-\sigma_{r+1}(B))\sigma(\sum D_jK_rE_j^*),\end{aligned}$$

since $\sigma_r(B)-\sigma_{r+1}(B)\geqq 0$, $r=1,2,\cdots,n$.

Now by Lemma 10

$$\begin{aligned}\sigma(\sum D_jBE_j^*)\prec_w&\sum_{r=1}^{n}(\sigma_r(B)-\sigma_{r+1}(B))\sigma(K_r)\circ\delta_\downarrow\\&=\sum_{r=1}^{n}(\sigma_r(B)-\sigma_{r+1}(B))(\delta_{[1]},\cdots,\delta_{[r]},0,\cdots,0)^t\\&=\sigma(B)\circ\delta_\downarrow.\end{aligned}$$

This completes the proof of the theorem. □

REFERENCES

[1] 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.

[2] R. B. Bapat, *Majorization and singular values*, Linear and Multilinear Algebra, 21 (1987), pp. 211–214.

[3] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 4, pp. 435–445, October 1989

G-INVARIANT HERMITIAN FORMS AND G-INVARIANT ELLIPTICAL NORMS*

CHI-KWONG LI† AND NAM-KIU TSING‡

Abstract. Let $\mathscr{V}$ be a finite-dimensional inner-product space over $\mathbb{C}$ or $\mathbb{R}$, and let G be a subgroup of the group of unitary operators on $\mathscr{V}$. The Hermitian forms and elliptical norms on $\mathscr{V}$ that are invariant under the operators in G are studied. The results are then applied to matrix spaces to obtain characterizations of Hermitian forms or elliptical norms that are invariant under unitary similarities, unitary equivalences, unitary congruences, or unitary row equivalences.

Key words. G-invariant, Hermitian form, elliptical norm, unitary similarity, congruence

AMS(MOS) subject classifications. 15A60, 15A63, 20H20

1. Introduction. Let $\mathbb{F}$ be the complex field $\mathbb{C}$ or the real field $\mathbb{R}$, and let $\mathbb{F}_{m\times n}$ be the linear space of all $m \times n$ matrices over $\mathbb{F}$. If $A \in \mathbb{F}_{m\times n}$, we use A^* to denote the conjugate transpose of A (or simply the transpose A^t of A if $\mathbb{F} = \mathbb{R}$). By a *norm* on $\mathbb{F}_{m\times n}$, we mean a function $\|\cdot\| : \mathbb{F}_{m\times n} \rightarrow \mathbb{R}$ that satisfies the following:

(a) $\|A\| > 0$ if $A \neq 0$,
(b) $\|cA\| = |c| \cdot \|A\|$ if $c \in \mathbb{F}$, and
(c) $\|A + B\| \leqq \|A\| + \|B\|$.

Denote by $U_n(\mathbb{F})$ the group of all $n \times n$ unitary or orthogonal matrices according to $\mathbb{F} = \mathbb{C}$ or $\mathbb{F} = \mathbb{R}$. A norm $\|\cdot\|$ on $\mathbb{F}_{n\times n}$ is said to be *unitary similarity invariant* (u.s.i.) if for any $A \in \mathbb{F}_{n\times n}$,

$$\|UAU^*\| = \|A\| \quad \text{for all } U \in U_n(\mathbb{F}). \tag{1.1}$$

A norm $\|\cdot\|$ on $\mathbb{F}_{m\times n}$ is said to be *unitarily invariant* (u.i.) if for any $A \in \mathbb{F}_{m\times n}$,

$$\|UAV\| = \|A\| \quad \text{for all } U \in U_m(\mathbb{F}) \text{ and } V \in U_n(\mathbb{F}). \tag{1.2}$$

A norm on a linear space $\mathscr{V}$ is *elliptical* if it is induced by an inner product. Several authors [1], [5], [6] have used special methods to study structures of elliptical norms that satisfy (1.1) or (1.2). In this note, we use group representation theory to develop general techniques that solve a more general class of problems. To describe the theory, let $\mathscr{V}$ be a linear space over $\mathbb{F}$ equipped with an inner product $\langle \cdot, \cdot \rangle$. (If $\mathscr{V}$ is $\mathbb{F}^n$ or $\mathbb{F}_{m\times n}$, we let $\langle \cdot, \cdot \rangle$ be the standard inner product on $\mathscr{V}$.) Suppose G is a subgroup of the group of unitary operators on $\mathscr{V}$. We say that a norm $\|\cdot\|$ on $\mathscr{V}$ is *G-invariant* if for any $v \in \mathscr{V}$,

$$\|g(v)\| = \|v\| \quad \text{for all } g \in G.$$

Evidently, if we let $\mathscr{V} = \mathbb{F}_{n\times n}$ and let G be the collection of all the unitary operators $T : \mathbb{F}_{n\times n} \rightarrow \mathbb{F}_{n\times n}$ defined by

$$T(A) = UAU^* \quad \text{for all } A \in \mathbb{F}_{n\times n}$$

* Received by the editors July 5, 1988; accepted for publication (in revised form) February 14, 1989.

† Department of Mathematics, The College of William and Mary, Williamsburg, Virginia 23185 (cklixx@wmmvs.bitnet).

‡ Department of Algebra, Combinatorics and Analysis, Auburn University, Auburn, Alabama 36849. Present address, Systems Research Center and Electrical Engineering Department, University of Maryland, College Park, Maryland 20742 (tsing@cacse.src.umd.edu).

where $U \in U_n(\mathbb{F})$, then G-invariant means unitary similarity invariant. Similarly, if we let $\mathscr{V} = \mathbb{F}_{m \times n}$ and let G be the collection of all the unitary operators $T: \mathbb{F}_{m\times n} \to \mathbb{F}_{m \times n}$ defined by

$$T(A) = UAV \quad \text{for all } A \in \mathbb{F}_{m\times n}$$

where $U \in U_m(\mathbb{F})$ and $V \in U_n(\mathbb{F})$, then G-invariant means unitarily invariant.

Recall that a *Hermitian* (or *symmetric* if $\mathbb{F} = \mathbb{R}$) *form* on $\mathscr{V}$ is a function $H: \mathscr{V} \times \mathscr{V} \to \mathbb{F}$ that satisfies the following:

$$\text{(a)} \qquad H(x,y) = \begin{cases} \overline{H(y,x)} & \text{if } \mathbb{F} = \mathbb{C}, \\ H(y,x) & \text{if } \mathbb{F} = \mathbb{R}, \end{cases}$$

$$\text{(b)} \qquad H(x+y,z) = H(x,z) + H(y,z), \quad \text{and}$$

$$\text{(c)} \qquad H(cx,y) = c \cdot H(x,y)$$

for all x, y, $z \in \mathscr{V}$ and $c \in \mathbb{F}$. If H further satisfies

$$\text{(d)} \qquad H(x,x) \geqq 0 \quad \text{with equality if and only if } x = 0,$$

then H is an inner product on $\mathscr{V}$. The purpose of this note is to characterize G-invariant elliptical norms and *G-invariant Hermitian forms* on $\mathscr{V}$, i.e., Hermitian forms H on $\mathscr{V}$ that for any x, $y \in \mathscr{V}$ satisfy

$$\text{(1.3)} \qquad H(x,y) = H(g(x), g(y)) \quad \text{for all } g \in G.$$

Using group representation theory, we prove our basic theorems in § 3. The results are then specialized to different matrix spaces and different subgroups of unitary operators in § 4. Using our technique, we not only give alternative proofs of old results, such as characterizing u.s.i. elliptical norms or u.i. elliptical norms, but we also determine the structures of the elliptical norms that are

(i) *Unitary congruence invariant* (u.c.i.), i.e., norms $\|\cdot\|$ that for any $A \in \mathbb{C}_{n\times n}$ satisfy

$$\text{(1.4)} \qquad \|UAU^t\| = \|A\| \quad \text{for all } U \in U_n(\mathbb{C});$$

(ii) *Unitary row equivalence invariant* (u.r.e.i.), i.e., norms $\|\cdot\|$ that for any $A \in \mathbb{F}_{m\times n}$ satisfy

$$\text{(1.5)} \qquad \|UA\| = \|A\| \quad \text{for all } U \in U_m(\mathbb{F}).$$

2. Preliminaries.

Let $\mathscr{V}$ and G be a vector space and a group, respectively, as described in § 1. For any subset S of $\mathscr{V}$ we use sp S to denote the linear span of S. Define

$$G(S) = \{g(s): g \in G, s \in S\}.$$

A subspace $\mathscr{W}$ of $\mathscr{V}$ is called *G-invariant* if $G(\mathscr{W}) \subset \mathscr{W}$. If, in addition, $\mathscr{W}$ is nonzero and does not contain any proper nonzero G-invariant subspaces, then $\mathscr{W}$ is called *G-irreducible*.

Regarding G as the image of a unitary representation of certain group, we have the following well-known results in group representation theory (e.g., see [4, pp. 112–129]).

THEOREM 2.1. *The vector space $\mathscr{V}$ can be decomposed into a direct sum*

$$\mathscr{V} = \mathscr{W}_1 \oplus \cdots \oplus \mathscr{W}_k$$

where $\mathscr{W}_1, \cdots, \mathscr{W}_k$ are mutually orthogonal G-irreducible subspaces.

THEOREM 2.2 (Schur's lemma). *Let $\mathcal{W}_1$ and $\mathcal{W}_2$ be G-irreducible subspaces of V. Suppose $L : \mathcal{W}_1 \to \mathcal{W}_2$ is linear and satisfies*

$$gL = Lg \quad \text{for all } g \in G.$$

Then L is either zero or is invertible.

We need the following characterization of *G*-irreducible subspaces in our discussion.

THEOREM 2.3. *Let $\mathcal{W}$ be a nonzero subspace of $\mathcal{V}$. Then $\mathcal{W}$ is G-irreducible if and only if*

$$\operatorname{sp} G(x) = \mathcal{W} \quad \text{for all nonzero } x \in \mathcal{W}.$$

3. *G*-invariant Hermitian forms and *G*-invariant elliptical norms. It is well known that every Hermitian form H on $\mathcal{V}$ corresponds to a unique Hermitian operator h on $\mathcal{V}$ such that

$$H(x,y) = \langle h(x), y\rangle \quad \text{for all } x, y \in \mathcal{V}.$$

Writing $\mathcal{V}$ as in Theorem 2.1, we have the following characterization of *G*-invariant Hermitian forms.

THEOREM 3.1. *Let $\mathcal{V} = \mathcal{W}_1 \oplus \cdots \oplus \mathcal{W}_k$, where $\mathcal{W}_1, \cdots, \mathcal{W}_k$ are mutually orthogonal G-irreducible subspaces, and let* $H : \mathcal{V} \times \mathcal{V} \to \mathbb{F}$ be a function. *Then H is a G-invariant Hermitian form if and only if there exist $a_1, \cdots, a_k \in \mathbb{R}$ and linear maps $h_{ij} : \mathcal{W}_i \to \mathcal{W}_j$ for all $1 \leqq i < j \leqq k$, such that*

$$gh_{ij} = h_{ij}g \quad \text{for all } g \in G, \tag{3.1}$$

and

$$H(x,y) = \sum_{i=1}^{k} a_i\langle x_i, y_i\rangle + \sum_{1 \leqq i<j \leqq k} (\langle h_{ij}(x_i), y_j\rangle + \langle x_j, h_{ij}(y_i)\rangle) \tag{3.2}$$

for all $x = x_1 + \cdots + x_k$, $y = y_1 + \cdots + y_k \in \mathcal{V}$ where $x_i, y_i \in \mathcal{W}_i$ for $i = 1, \cdots, k$.

Proof. Suppose H satisfies (3.2), where all h_{ij} ($1 \leqq i < j \leqq k$) satisfy (3.1). Then by direct checking, we see that H is a *G*-invariant Hermitian form.

Conversely, suppose H is a *G*-invariant Hermitian form on $\mathcal{V}$. From the decomposition of $\mathcal{V} = \mathcal{W}_1 \oplus \cdots \oplus \mathcal{W}_k$, where $\mathcal{W}_1, \cdots, \mathcal{W}_k$ are mutually orthogonal, H can be decomposed into Hermitian forms H_i on $\mathcal{W}_i$ ($i = 1, \cdots, k$) and linear maps $h_{ij} : \mathcal{W}_i \to \mathcal{W}_j$ ($1 \leqq i < j \leqq k$) such that

$$H(x,y) = \sum_{i=1}^{k} H_i\langle x_i, y_i\rangle + \sum_{1 \leqq i<j \leqq k} (\langle h_{ij}(x_i), y_j\rangle + \langle x_j, h_{ij}(y_i)\rangle) \tag{3.3}$$

for all $x = x_1 + \cdots + x_k$, $y = y_1 + \cdots + y_k \in \mathcal{V}$ where $x_i, y_i \in \mathcal{W}_i$ for $i = 1, \cdots, k$. Let $h_i : \mathcal{W}_i \to \mathcal{W}_i$ be the Hermitian operator associated with H_i, let a_i be the largest eigenvalue of h_i, and let $x_i \in \mathcal{W}_i$ be a corresponding unit eigenvector. Then for all $g \in G$, since g is unitary, $g(x_i)$ is also a unit vector. As $\mathcal{W}_i$ and H are *G*-invariant, by (3.3),

$$\begin{aligned}\langle h_i(g(x_i)), g(x_i)\rangle &= H_i(g(x_i), g(x_i)) = H(g(x_i), g(x_i)) \\ &= H(x_i, x_i) = \langle h_i(x_i), x_i\rangle = a_i.\end{aligned}$$

Thus $g(x_i)$ is also a unit eigenvector of h_i corresponding to the eigenvalue a_i. As a result, we have $G(x_i) \subset E(a_i)$, where $E(a_i)$ denotes the eigenspace of h_i in $\mathcal{W}_i$ corresponding to the eigenvalue a_i. Since $\mathcal{W}_i$ is G-irreducible, we have

$$\mathcal{W}_i = \operatorname{sp} G(x_i) \subset E(a_i) \subset \mathcal{W}_i.$$

This implies $E(a_i) = \mathcal{W}_i$; i.e., $h_i(x) = a_i x$ for all $x \in \mathcal{W}_i$. Hence

$$H_i(x_i, y_i) = a_i \langle x_i, y_i \rangle \quad \text{for all } x_i, y_i \in \mathcal{W}_i.$$

Now let $x_i \in \mathcal{W}_i$ and $y_j \in \mathcal{W}_j$, where $1 \leqq i < j \leqq k$. Since $\mathcal{W}_i$, $\mathcal{W}_j$, and H are G-invariant, by (3.3), we have

$$\langle g^* h_{ij} g(x_i), y_j \rangle = \langle h_{ij} g(x_i), g(y_j) \rangle = H(g(x_i), g(y_j))$$
$$= H(x_i, y_j) = \langle h_{ij}(x_i), y_j \rangle$$

for all $g \in G$. Since $x_i \in \mathcal{W}_i$ and $y_j \in \mathcal{W}_j$ are arbitrary, we have

$$g^* h_{ij} g = h_{ij} \quad \text{for all } g \in G.$$

Hence

$$h_{ij} g = g h_{ij} \quad \text{for all } g \in G. \qquad \square$$

Using techniques similar to those employed in the proof of Theorem 3.1, we may characterize the linear maps $T : \mathcal{V} \to \mathcal{V}$ that satisfy

$$gT = Tg \quad \text{for all } g \in G.$$

In particular, if $\mathbb{F} = \mathbb{C}$ and $\mathcal{V}$ is G-irreducible, then T is a scalar map by Theorem 2.2. Applying this result, we can readily get the conclusions about the maps h_i in Theorem 3.1 when $\mathbb{F} = \mathbb{C}$. Our proof is valid, however, for $\mathbb{F} = \mathbb{R}$ as well. Moreover, we may use Theorem 2.2 to get more information on the linear maps h_{ij} described in the statement of Theorem 3.1.

Since a norm on $\mathcal{V}$ is G-invariant and elliptical if and only if it is induced by a positive-definite G-invariant Hermitian form, we may use Theorem 3.1 to get a characterization of G-invariant elliptical norms on $\mathcal{V}$.

4. Applications to matrix spaces. In the following examples of matrix spaces, we apply the results in the preceding sections to characterize the Hermitian forms and elliptical norms that are G-invariant for various subgroups G of unitary operators.

4.1. Unitary similarity invariant. Let $\mathcal{V}$ be $\mathbb{F}_{n \times n}$ and let G be the collection of all the unitary operators $T : \mathbb{F}_{n \times n} \to \mathbb{F}_{n \times n}$ defined by

$$T(A) = UAU^* \quad \text{for all } A \in \mathbb{F}_{n \times n}$$

where $U \in U_n(\mathbb{F})$. As mentioned in § 1, the G-invariant concept reduces to the unitary similarity invariant (u.s.i.) concept. We consider two cases.

Case 1. $\mathbb{F} = \mathbb{C}$. Let

$$\mathcal{W}_1 = \{cI_n : c \in \mathbb{C}\}$$

and

$$\mathcal{W}_2 = \{A \in \mathbb{C}_{n \times n} : \operatorname{tr} A = 0\}.$$

Then $\mathbb{C}_{n \times n} = \mathcal{W}_1 \oplus \mathcal{W}_2$ and $A = A_1 + A_2$, where

$$A_1 = (\operatorname{tr} A) I/n \in \mathcal{W}_1, \qquad A_2 = A - (\operatorname{tr} A) I/n \in \mathcal{W}_2$$

for all $A \in \mathbb{C}_{n\times n}$. Note that $\mathscr{W}_1$ and $\mathscr{W}_2$ are mutually orthogonal, $\mathscr{W}_1$ is G-irreducible, and $\mathscr{W}_2$ is G-invariant. By a result of Tam [10],

$$\mathrm{sp}\,\{UAU^*: U\in U_n(\mathbb{C})\} = \mathscr{W}_2$$

for any nonzero $A \in \mathscr{W}_2$. Hence $\mathscr{W}_2$ is also G-irreducible by Theorem 2.3. Since dim $\mathscr{W}_1$ is always less than dim $\mathscr{W}_2$, any linear map from $\mathscr{W}_1$ to $\mathscr{W}_2$ cannot be invertible. As a result, we have the following theorem.

THEOREM 4.1.1. (a) *Let $H: \mathbb{C}_{n\times n}\times \mathbb{C}_{n\times n} \to \mathbb{C}$ be a function. Then H is a Hermitian form satisfying*

$$(4.1) \qquad H(UAU^*, UBU^*) = H(A,B)$$

for all $U \in U_n(\mathbb{C})$ and all A, $B \in \mathbb{C}_{n\times n}$ if and only if there exist α, $\beta \in \mathbb{R}$ such that

$$H(A,B) = \alpha(\mathrm{tr}\, A)(\mathrm{tr}\, B^*) + \beta\, \mathrm{tr}\,(AB^*)$$

for all A, $B \in \mathbb{C}_{n\times n}$.

(b) *Let $\|\cdot\|: \mathbb{C}_{n\times n} \to \mathbb{R}$ be a function. Then $\|\cdot\|$ is an elliptical norm on $\mathbb{C}_{n\times n}$ satisfying*

$$\|UAU^*\| = \|A\|$$

for all $U \in U_n(\mathbb{C})$ and $A \in \mathbb{C}_{n\times n}$ if and only if there exist α, $\beta \in \mathbb{R}$ such that $n\alpha + \beta > 0$, $\beta > 0$, and

$$\|A\|^2 = \alpha|\mathrm{tr}\, A|^2 + \beta\, \mathrm{tr}\,(AA^*) \quad \textit{for all } A\in\mathbb{C}_{n\times n}.$$

Proof. (a) From the discussion preceding the statement of the theorem, and by Theorems 3.1 and 2.2, we see that H is a Hermitian form satisfying (4.1) for all $U \in U_n(\mathbb{C})$ and A, $B \in \mathbb{C}_{n\times n}$ (i.e., is G-invariant) if and only if there exist a, $b \in \mathbb{R}$ such that for all A, $B \in \mathbb{C}_{n\times n}$,

$$\begin{aligned} H(A,B) &= a\langle(\mathrm{tr}\, A)I/n, (\mathrm{tr}\, B)I/n\rangle + b\langle A-(\mathrm{tr}\, A)I/n, B-(\mathrm{tr}\, B)I/n\rangle \\ &= (a-b)(\mathrm{tr}\, A)(\mathrm{tr}\, B^*)/n + b(\mathrm{tr}\,(AB^*)). \end{aligned}$$

Taking $\alpha = (a-b)/n$ and $\beta = b$, we get the result.

(b) As mentioned in § 3, $\|\cdot\|$ is a G-invariant elliptical norm if and only if it is induced by some positive-definite G-invariant Hermitian form H. In the proof of (a), H is positive definite if and only if a and b are positive. This is equivalent to $n\alpha + \beta > 0$ and $\beta > 0$, because $\alpha = (a-b)/n$ and $\beta = b$. □

We remark that Theorem 4.1.1 has also been obtained by Bhatia and Holbrook [1, Cor. 2.3].

Case 2. $\mathbb{F} = \mathbb{R}$. In this case, we have

$$\mathbb{R}_{n\times n} = \mathscr{W}_1 \oplus \mathscr{W}_2 \oplus \mathscr{W}_3$$

where

$$\mathscr{W}_1 = \{cI_n : c\in\mathbb{R}\}, \qquad \mathscr{W}_2 = \{A\in\mathbb{R}_{n\times n}: A^t = -A\},$$

and

$$\mathscr{W}_3 = \{A\in\mathbb{R}_{n\times n}: A^t = A, \mathrm{tr}\, A = 0\}.$$

Every $A \in \mathbb{R}_{n\times n}$ can be written as $A = A_1 + A_2 + A_3$, where

$$A_1 = (\mathrm{tr}\, A)I/n \in \mathscr{W}_1, \qquad A_2 = (A - A^t)/2 \in \mathscr{W}_2,$$

and

$$A_3 = (A + A^t)/2 - (\operatorname{tr} A)I/n \in \mathscr{W}_3.$$

Clearly, $\mathscr{W}_1$, $\mathscr{W}_2$, and $\mathscr{W}_3$ are mutually orthogonal and G-invariant. That $\mathscr{W}_1$ is G-irreducible is obvious. For any nonzero A, $B \in \mathscr{W}_2$, we can find U, $V \in U_n(\mathbb{R})$ such that

$$UAU^t = Y_1 \oplus \cdots \oplus Y_m \oplus X, \qquad VBV^t = Z_1 \oplus \cdots \oplus Z_m \oplus X$$

where m is the largest integer less than $(n+1)/2$,

$$Y_i = \begin{bmatrix} 0 & a_i \\ -a_i & 0 \end{bmatrix} \quad \text{and} \quad Z_i = \begin{bmatrix} 0 & b_i \\ -b_i & 0 \end{bmatrix}$$

are such that A and B have singular values $a_1, a_1, a_2, a_2, \cdots$, and $b_1, b_1, b_2, b_2, \cdots$, respectively, and X is void if n is even and is (0) if n is odd (see [3, § 4.4]). Therefore if $T : \mathbb{R}_{n \times n} \to \mathbb{R}_{n \times n}$ is defined by

$$T(C) = V^t UCU^t V \quad \text{for all } C \in \mathbb{R}_{n \times n},$$

then

$$\begin{aligned} \langle T(A), B \rangle &= \langle UAU^t, VBV^t \rangle \\ &= 2(a_1 b_1 + a_2 b_2 + \cdots + a_m b_m) > 0. \end{aligned}$$

This implies that the orthogonal complement of $G(A)$ in $\mathscr{W}_2$ is the zero space. Hence sp $G(A) = \mathscr{W}_2$. Since A $(\neq 0)$ in $\mathscr{W}_2$ is arbitrary, by Theorem 2.3, $\mathscr{W}_2$ is G-irreducible.

Now consider any nonzero A, $B \in \mathscr{W}_3$. Let $a_1, \cdots, a_n$ and $b_1, \cdots, b_n$ be the eigenvalues of A and B, respectively. As $(a_1, \cdots, a_n)$, $(b_1, \cdots, b_n) \neq (0, \cdots, 0)$, and $\sum_{i=1}^n a_i = \sum_{i=1}^n b_i = 0$, for a suitable permutation σ of the set $\{1, \cdots, n\}$, we have $\sum_{i=1}^n a_i b_{\sigma(i)} \neq 0$. Let U, $V \in U_n(\mathbb{R})$ be such that $UAU^t = \operatorname{diag}(a_1, \cdots, a_n)$ and $VBV^t = \operatorname{diag}(b_{\sigma(1)}, \cdots, b_{\sigma(n)})$. Then

$$\langle T(A), B \rangle = \langle UAU^t, VBV^t \rangle = \sum_{i=1}^n a_i b_{\sigma(i)} \neq 0$$

where $T : \mathbb{R}_{n \times n} \to \mathbb{R}_{n \times n}$ is defined in the same way as above. Using an argument similar to the one used in the case $\mathscr{W}_2$, we conclude that $\mathscr{W}_3$ is G-irreducible.

Since dim $\mathscr{W}_1$ and dim $\mathscr{W}_2$ are less than dim $\mathscr{W}_3$, no linear map from $\mathscr{W}_1$ to $\mathscr{W}_3$ or from $\mathscr{W}_2$ to $\mathscr{W}_3$ can be invertible. Also, dim $\mathscr{W}_1$ = dim $\mathscr{W}_2$ only when $n = 2$. For this particular case of $n = 2$, we may take

$$U = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \in U_2(\mathbb{R})$$

and consider $g_U \in G$ defined by

$$g_U(A) = UAU^t \quad \text{for all } A \in \mathbb{R}_{2 \times 2}.$$

If $h : \mathscr{W}_1 \to \mathscr{W}_2$ is a linear map that satisfies $g_U h = h g_U$, then

$$g_U h(I) = h g_U(I) = h(I).$$

Since $g_U(A) \neq A$ for all nonzero $A \in \mathscr{W}_2$, we then have $h(I) = 0$. Hence, h is not invertible. Using Theorem 2.2, we can conclude that, if $1 \leqq i < j \leqq 3$ and $h_{ij} : \mathscr{W}_i \to \mathscr{W}_j$ is a linear map that satisfies $h_{ij} g = g h_{ij}$ for all $g \in G$, then $h_{ij} = 0$.

Combining the preceding observations, we have the following result. As the proof is similar to that of Theorem 4.1.1, we omit the details.

THEOREM 4.1.2. (a) *Let $H:\mathbb{R}_{n\times n}\times\mathbb{R}_{n\times n}\to\mathbb{R}$ be a function. Then H is a symmetric form satisfying*

$$H(UAU^t,UBU^t)=H(A,B)$$

for all $U\in U_n(\mathbb{R})$ and A, $B\in\mathbb{R}_{n\times n}$ if and only if there exist α, β, $\gamma\in\mathbb{R}$ such that

$$H(A,B)=\alpha(\operatorname{tr}A)(\operatorname{tr}B)+\beta\operatorname{tr}(AB^t)+\gamma\operatorname{tr}(AB)$$

for all A, $B\in\mathbb{R}_{n\times n}$.

(b) *Let $\|\cdot\|:\mathbb{R}_{n\times n}\to\mathbb{R}$ be a function. Then $\|\cdot\|$ is an elliptical norm on $\mathbb{R}_{n\times n}$ satisfying*

$$\|UAU^t\|=\|A\|$$

for all $U\in U_n(\mathbb{R})$ and all $A\in\mathbb{R}_{n\times n}$ if and only if there exist α, β, $\gamma\in\mathbb{R}$ such that $\beta>|\gamma|$, $n\alpha+\beta+\gamma>0$, and

$$\|A\|^2=\alpha(\operatorname{tr}A)^2+\beta\operatorname{tr}(AA^t)+\gamma\operatorname{tr}(A^2)$$

for all $A\in\mathbb{R}_{n\times n}$.

We remark that we may consider $\mathscr{V}$ to be the real linear space H_n of all $n\times n$ Hermitian matrices and obtain characterizations of u.s.i. symmetric forms and u.s.i. elliptical norms on H_n that are similar to those in Theorem 4.1.1 (see [6, Thm. 4.2]). Also, we may consider $\mathscr{V}$ to be the real linear space $S_n(\mathbb{R})$ of all $n\times n$ real symmetric matrices or the space $K_n(\mathbb{R})$ of all real skew-symmetric matrices. By arguments similar to those in the proof of Theorem 4.1.2, we can deduce that

(a) The u.s.i. symmetric forms H on $S_n(\mathbb{R})$ are exactly those of the form

$$H(A,B)=\alpha(\operatorname{tr}A)(\operatorname{tr}B)+\beta\operatorname{tr}(AB)\quad\text{for all } A,B\in\mathbb{R}_{n\times n},$$

where α and β are real;

(b) The u.s.i. elliptical norms $\|\cdot\|$ on $S_n(\mathbb{R})$ are exactly those of the form

$$\|A\|^2=\alpha(\operatorname{tr}A)^2+\beta\operatorname{tr}(A^2)\quad\text{for all } A\in\mathbb{R}_{n\times n}$$

where $n\alpha+\beta>0$ and $\beta>0$.

Similarly, we have

(a) The u.s.i. symmetric forms H on $K_n(\mathbb{R})$ are exactly those of the form

$$H(A,B)=\alpha\operatorname{tr}(AB)\quad\text{for all } A,B\in\mathbb{R}_{n\times n}$$

where α is real;

(b) The u.s.i. elliptical norms $\|\cdot\|$ on $K_n(\mathbb{R})$ are exactly those of the form

$$\|A\|^2=\alpha\operatorname{tr}(AA^t)\quad\text{for all } A\in\mathbb{R}_{n\times n}$$

where $\alpha>0$.

4.2. Unitarily invariant. Let $\mathscr{V}$ be $\mathbb{F}_{m\times n}$ and let G be the collection of all the unitary operators $T:\mathbb{F}_{m\times n}\to\mathbb{F}_{m\times n}$ defined by

$$T(A)=UAV\quad\text{for all } A\in\mathbb{F}_{m\times n},$$

where $U\in U_m(\mathbb{F})$ and $V\in U_n(\mathbb{F})$. Then G-invariant norms are unitarily invariant (u.i.) norms. This class of norms has been studied by many authors, and many interesting results have been obtained (e.g., see [2, Chap. 3], [5], [7, Chap. 10], [8], [9] and the references therein).

Now suppose $A, B \in \mathbb{F}_{m \times n}$ have singular values $a_1 \geqq \cdots \geqq a_k$ ($\geqq 0$) and $b_1 \geqq \cdots \geqq b_k$ ($\geqq 0$), respectively (here $k = \min\{m, n\}$). Then (see [3, Ex. 7.4.13])

$$\begin{aligned}\max\{|\langle UAV, B\rangle| : U \in U_m(\mathbb{F}), V \in U_n(\mathbb{F})\} \\ = \max\{|\operatorname{tr}(UAVB^*)| : U \in U_m(\mathbb{F}), V \in U_n(\mathbb{F})\} \\ = \sum_{i=1}^{k} a_i b_i,\end{aligned}$$

which is greater than zero if both A, B are nonzero. Using arguments similar to that in § 4.1, we see that $\mathbb{F}_{m \times n}$ itself is G-irreducible. By Theorem 3.1, we have (cf. [5, Thm. 2.2]) the following result.

THEOREM 4.2. (a) *Let $H : \mathbb{F}_{m \times n} \times \mathbb{F}_{m \times n} \to \mathbb{F}$ be a function. Then H is a Hermitian form satisfying*

$$H(UAV, UBV) = H(A, B)$$

for all $A, B \in \mathbb{F}_{m \times n}$ and $U \in U_m(\mathbb{F})$, $V \in U_n(\mathbb{F})$ if and only if there exist $\alpha \in \mathbb{R}$ such that

$$H(A, B) = \alpha \operatorname{tr}(AB^*) \quad \textit{for all } A, B \in \mathbb{F}_{m \times n}.$$

(b) *Let $\|\cdot\| : \mathbb{F}_{m \times n} \to \mathbb{R}$ be a function. Then $\|\cdot\|$ is an elliptical norm on $\mathbb{F}_{m \times n}$ satisfying*

$$\|UAV\| = \|A\|$$

for all $A \in \mathbb{C}_{n \times n}$ and $U \in U_m(\mathbb{F})$, $V \in U_n(\mathbb{F})$ if and only if there exists some $\alpha > 0$ such that

$$\|A\|^2 = \alpha \operatorname{tr}(AA^*) \quad \textit{for all } A \in \mathbb{F}_{m \times n}.$$

4.3. Unitary congruence invariant. Let $\mathscr{V}$ be $\mathbb{C}_{n \times n}$ and let G be the collection of all the unitary operators $T : \mathbb{C}_{n \times n} \to \mathbb{C}_{n \times n}$ defined by

$$T(A) = UAU^t \quad \text{for all } A \in \mathbb{C}_{n \times n}.$$

Then G-invariant means unitary congruence invariant (u.c.i.). In fact, if we regard $A \in \mathbb{C}_{n \times n}$ as a bilinear form $A(\cdot, \cdot)$ on $\mathbb{C}^n$, such that

$$A(x, y) = xAy^t \quad \text{for all } x, y \in \mathbb{C}^n,$$

then UAU^t represents the same bilinear form $A(\cdot, \cdot)$ with respect to a new orthonormal basis.

Now $\mathbb{C}_{n \times n} = \mathscr{W}_1 \oplus \mathscr{W}_2$, where

$$\mathscr{W}_1 = \{A \in \mathbb{C}_{n \times n} : A^t = A\}, \qquad \mathscr{W}_2 = \{A \in \mathbb{C}_{n \times n} : A^t = -A\},$$

and for any $A \in \mathbb{C}_{n \times n}$, $A = A_1 + A_2$ where

$$A_1 = \frac{A + A^t}{2} \in \mathscr{W}_1, \qquad A_2 = \frac{A - A^t}{2} \in \mathscr{W}_2.$$

Clearly, $\mathscr{W}_1$ and $\mathscr{W}_2$ are mutually orthogonal and are G-invariant. If $A, B \in \mathscr{W}_1$, we can find $U, V \in U_n(\mathbb{C})$ such that

$$UAU^t = \operatorname{diag}(a_1, \cdots, a_n), \qquad VBV^t = \operatorname{diag}(b_1, \cdots, b_n)$$

where $a_1 \geqq \cdots \geqq a_n$ $(\geqq 0)$ and $b_1 \geqq \cdots \geqq b_n$ $(\geqq 0)$ are the singular values of A and B, respectively (e.g., see [3, § 4.4]). Suppose $T: \mathbb{C}_{n\times n} \to \mathbb{C}_{n\times n}$ is defined by

$$T(X) = V^tUXU^tV \quad \text{for all } X \in \mathbb{C}_{n\times n}.$$

Then

$$\langle T(A), B\rangle = \sum_{i=1}^{n} a_i b_i,$$

which is greater than zero if both A, B are nonzero. Therefore, $\mathscr{W}_1$ is G-irreducible.

If $A \in \mathscr{W}_2$, then there is $U \in U_n(\mathbb{C})$ such that

$$UAU^t = Y_1 \oplus \cdots \oplus Y_m \oplus X$$

where m is the largest integer less than $(n+1)/2$,

$$Y_i = \begin{bmatrix} 0 & a_i \\ -a_i & 0 \end{bmatrix}$$

are such that A has singular values $a_1, a_1, a_2, a_2, \cdots$, and X is void if n is even and is (0) if n is odd (see [3, § 4.4]). This situation is the same as that of $\mathscr{W}_2$ in Case 2 of § 4.1. Therefore $\mathscr{W}_2$ in the present example is G-irreducible.

Since dim $\mathscr{W}_1 >$ dim $\mathscr{W}_2$, no linear map from $\mathscr{W}_1$ to $\mathscr{W}_2$ can be invertible. Using Theorems 3.1 and 2.2, we have Theorem 4.3.

THEOREM 4.3. (a) *Let $H: \mathbb{C}_{n\times n} \times \mathbb{C}_{n\times n} \to \mathbb{C}$ be a function. Then H is a Hermitian form satisfying*

$$H(UAU^t, UBU^t) = H(A,B)$$

for all $U \in U_n(\mathbb{C})$ and A, $B \in \mathbb{C}_{n\times n}$ if and only if there exist α, $\beta \in \mathbb{R}$ such that

$$H(A,B) = \alpha \operatorname{tr}(AB^*) + \beta \operatorname{tr}(A\bar{B})$$

for all A, $B \in \mathbb{C}_{n\times n}$.

(b) *Let $\|\cdot\|: \mathbb{C}_{n\times n} \to \mathbb{R}$ be a function. Then $\|\cdot\|$ is an elliptical norm on $\mathbb{C}_{n\times n}$ satisfying*

$$\|UAU^t\| = \|A\|$$

for all $U \in U_n(\mathbb{C})$ and $A \in \mathbb{C}_{n\times n}$ if and only if there exist α, $\beta \in \mathbb{R}$ such that $\alpha > |\beta|$ and

$$\|A\|^2 = \alpha \operatorname{tr}(AA^*) + \beta(\operatorname{tr}(A\bar{A}))$$

for all $A \in \mathbb{C}_{n\times n}$.

Because the proof of the above is similar to that of Theorem 4.1.1, we omit the details.

We do not consider the u.c.i. concept on $\mathbb{R}_{n\times n}$ because it reduces to the u.s.i. concept, which has been studied in Case 2 of § 4.1. Nevertheless, we may consider the u.c.i. concept on the complex linear space $S_n(\mathbb{C})$ of all $n \times n$ complex symmetric matrices or the space $K_n(\mathbb{C})$ of all $n \times n$ complex skew-symmetric matrices. In both cases, we can show without difficulty that a u.c.i. Hermitian form must be a real multiple of the usual inner product, and a u.c.i. elliptical norm must be a positive multiple of the Frobenius norm.

4.4. Unitary row equivalence invariant. In our last example, let $\mathscr{V}$ be $\mathbb{F}_{m\times n}$ and let G be the collection of all the unitary operators $T:\mathbb{F}_{m\times n}\to\mathbb{F}_{m\times n}$ defined by

$$T(A)=UA \quad \text{for all } A\in\mathbb{F}_{m\times n}$$

where $U\in U_m(\mathbb{F})$. Then G-invariant means unitary row equivalence invariant (u.r.e.i.).

Now for each $i=1,\cdots,n$, let

$$\mathscr{W}_i=\{A\in\mathbb{F}_{m\times n}\text{: the } j\text{th column of } A \text{ is zero if } j\neq i\}.$$

It is routine to check that $\mathbb{F}_{m\times n}=\mathscr{W}_1\oplus\cdots\oplus\mathscr{W}_n$ and $\mathscr{W}_1,\cdots,\mathscr{W}_n$ are mutually orthogonal, G-irreducible subspaces. For any $A\in\mathbb{F}_{m\times n}$, if $i=1,\cdots,n$, let $x_i\in\mathbb{C}^m$ be the ith column of A and let $A_i\in\mathbb{F}_{m\times n}$ be such that the ith column of A_i is x_i and all the other columns are zero. Then

$$A=A_1+\cdots+A_n$$

and $A_i\in\mathscr{W}_i$ for all i. For $i=1,\cdots,n$, let $p_i:\mathscr{W}_i\to\mathbb{C}^n$ be the projection of the ith column of A_i onto $\mathbb{C}^m$. We need the following lemma to prove Theorem 4.4.

LEMMA. *Let $\mathscr{W}_1,\cdots,\mathscr{W}_n$ be defined as above. Suppose $1\leqq i<j\leqq n$ and suppose $h_{ij}:\mathscr{W}_i\to\mathscr{W}_j$ is a linear map. Then h_{ij} satisfies*

$$gh_{ij}=h_{ij}g \quad \textit{for all } g\in G \tag{4.2}$$

if and only if there exist scalars $k_{ij}\in\mathbb{F}$ such that

$$h_{ij}(A_i)=k_{ij}(p_j^{-1}p_i)(A_i) \quad \textit{for all } A_i\in\mathscr{W}_i. \tag{4.3}$$

Proof. Note that h_{ij} satisfies (4.2) if and only if the linear map $(p_jh_{ij}p_i^{-1}):\mathbb{C}^m\to\mathbb{C}^m$ satisfies

$$U(p_jh_{ij}p_i^{-1})(x)=(p_jh_{ij}p_i^{-1})U(x) \tag{4.4}$$

for all $U\in U_m(\mathbb{F})$ and $x\in\mathbb{C}^m$. Since condition (4.4) is equivalent to

$$(p_jh_{ij}p_i^{-1})=ke$$

where $k\in\mathbb{F}$ and e is the identity map on $\mathbb{C}^m$, the result follows. □

THEOREM 4.4. (a) *Let $H:\mathbb{F}_{m\times n}\times\mathbb{F}_{m\times n}\to\mathbb{F}$ be a function. Then H is a Hermitian form satisfying*

$$H(UA,UB)=H(A,B) \tag{4.5}$$

for all $U\in U_m(\mathbb{F})$ and A, $B\in\mathbb{F}_{m\times n}$ if and only if there exists a Hermitian matrix $K\in\mathbb{F}_{n\times n}$ such that

$$H(A,B)=\operatorname{tr}(AKB^*) \quad \textit{for all } A,B\in\mathbb{F}_{m\times n}.$$

(b) *Let $\|\cdot\|:\mathbb{F}_{m\times n}\to\mathbb{R}$ be a function. Then $\|\cdot\|$ is an elliptical norm on $\mathbb{F}_{m\times n}$ satisfying*

$$\|UA\|=\|A\| \tag{4.6}$$

for all $U\in U_m(\mathbb{F})$ and $A\in\mathbb{F}_{m\times n}$ if and only if there exists a positive definite Hermitian matrix $K\in\mathbb{F}_{n\times n}$ such that

$$\|A\|^2=\operatorname{tr}(AKA^*) \quad \textit{for all } A\in\mathbb{F}_{m\times n}.$$

Proof. (a) From the discussion preceding the lemma, and by Theorem 3.1 and the lemma, we see that H is a Hermitian form satisfying (4.5) if and only if there exist $k_{ii} \in \mathbb{R}$ (for $i = 1, \cdots, n$) and $k_{ij} \in \mathbb{F}$ (for $1 \leqq i < j \leqq n$) such that

$$(4.7) \quad H(A,B) = \sum_{i=1}^{n} k_{ii}\langle A_i, B_i\rangle + \sum_{1 \leqq i < j \leqq n} (k_{ij}\langle p_j^{-1}p_i(A_i), B_j\rangle + \overline{k_{ij}}\langle A_j, p_j^{-1}p_i(B_i)\rangle)$$

for all $A = A_1 + \cdots + A_n$, $B = B_1 + \cdots + B_n$, where $A_i, B_i \in \mathscr{W}_i$ for all i. Let $K = (k_{ij}) \in \mathbb{F}_{n \times n}$, where $k_{ji} = \overline{k_{ij}}$ for all $i < j$. Since $\langle A_i, B_j\rangle$ equals the (j, i) entry of the matrix B^*A, the expression (4.7) equals

$$\operatorname{tr}(KB^*A) = \operatorname{tr}(AKB^*).$$

(b) Note that $\|\cdot\|$ is a norm that satisfies (4.6) if and only if it is induced by a positive-definite Hermitian form H satisfying (4.5). Because the Hermitian form H in (a) is positive definite if and only if the matrix K is positive definite, the result follows. □

Using the same method, similar results on unitary column equivalence can be obtained.

Acknowledgments. The paper was finished when the authors were visiting the Johns Hopkins University during June and July, 1988. The authors wish to thank Professor Roger Horn for his generous help during the visit. Thanks are also due to the staff of the Johns Hopkins University for their warm hospitality.

REFERENCES

[1] R. Bhatia and J. A. R. Holbrook, *Unitary invariance and spectral variation*, Linear Algebra Appl., 95 (1987), pp. 43–68.

[2] I. C. Gohberg and M. G. Krein, *Introduction to the Theory of Linear Non-Self-Adjoint Operators*, AMS Transl. Math. Monographs 18, Providence, RI, 1969.

[3] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[4] A. A. Kirillov, *Elements of the Theory of Representations*, Springer-Verlag, Berlin, New York, 1976.

[5] C. K. Li and N. K. Tsing, *On the unitarily invariant norms and some related results*, Linear Multilinear Algebra, 20 (1987), pp. 107–119.

[6] ———, *Norms that are invariant under unitary similarities and the C-numerical radii*, Linear and Multilinear Algebra, 24 (1989), pp. 209–222.

[7] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.

[8] L. Mirsky, *Symmetric gauge functions and unitarily invariant norms*, Quart. J. Math., 11 (1960), pp. 50–59.

[9] J. von Neumann, *Some matrix-inequalities and metrization of matrix-space*, Tomsk Univ. Rev., 1 (1937), pp. 286–300.

[10] 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. 10, No. 4, pp. 446–456, October 1989

BACKWARD ERROR ANALYSIS FOR A POLE ASSIGNMENT ALGORITHM*

CHRISTOPHER L. COX† AND WILLIAM F. MOSS†

Abstract. Of the six or so pole assignment algorithms currently available, several have been claimed to be numerically stable, but no proofs have been published to date. It is shown, by performing a backward error analysis, that one of these algorithms, due to Petkov, Christov, and Konstantinov [*IEEE Trans. Automat. Control*, AC-29 (1984), pp. 1045–1048] is numerically stable.

Key words. backward error analysis, pole assignment, numerical stability

AMS(MOS) subject classifications. 65G05, 93B55, 93D15

1. Introduction. A single-input time-invariant linear control system has the form

$$\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{b}u \tag{1.1}$$

with $\mathbf{x}, \mathbf{b} \in R^n$, $A \in R^{n\times n}$, and $u \in R^1$. The function $\mathbf{x}(t)$ is known as the system state and u is called a control function, to be chosen to control the evolution of $\mathbf{x}(t)$. One way of choosing the function u is through the linear feedback relation

$$u = -\mathbf{k}^T\mathbf{x}$$

where $\mathbf{k} \in R^n$ is known as the gain vector. Equation (1.1) becomes

$$\dot{\mathbf{x}} = (A - \mathbf{b}\mathbf{k}^T)\mathbf{x}$$

which is known as the closed-loop system. Stabilization can be achieved by proper specification of the *poles* $\lambda_1, \cdots, \lambda_n$, which are eigenvalues of $A - \mathbf{b}\mathbf{k}^T$. It is well known [K] that if the linear system (1.1) is completely controllable, that is if $[\mathbf{b}, A\mathbf{b}, \cdots, A^{n-1}\mathbf{b}] \in R^{n\times n}$ has rank n, then the eigenvalues can be chosen or *assigned* at will and a unique $\mathbf{k}$ exists so that $A - \mathbf{b}\mathbf{k}^T$ has these eigenvalues. More background information on this problem can be found in Russell [R].

At this point the control theory problem enters the realm of numerical linear algebra. Given $A \in R^{n\times n}$ and $\mathbf{b} \in R^n$, we seek $\mathbf{k} \in R^n$ so that $A - \mathbf{b}\mathbf{k}^T$ has the eigenvalues $\lambda_1, \cdots, \lambda_n$. A number of pole assignment algorithms have appeared in the literature. At least three methods in print are based on orthogonal plane rotations to reduce the system matrix to triangular form, including Varga [V]; Miminis and Paige [MP]; and Petkov, Christov, and Konstantinov [PCK]. Proofs of numerical stability have **not** been published for any of these algorithms; however, the authors have a verbal communication from one of the authors of [MP] that their algorithm has been shown to be numerically stable. Two other papers that discuss numerical stability of algorithms for time-invariant linear systems are [P] and [D]. A broad overview of the matrix theory for linear control systems, supplemented by an extensive bibliography, is presented in [B].

In this paper we present a detailed backward error analysis of the pole assignment algorithm introduced by Petkov, Christov, and Konstantinov; we will refer to this as the PCK algorithm. The algorithm can assign eigenvalues that are distinct or repeated, and can handle complex conjugate pairs of eigenvalues using real arithmetic. Herein we will restrict our analysis to the real case.

* Received by the editors April 27, 1987; accepted for publication (in revised form) March 2, 1989.

† Department of Mathematical Sciences, Clemson University, Clemson, South Carolina 29634-1907 (CLCOX@CLEMSON and CLEMFIR@CLEMSON, respectively).

Before stating our main theorem, a brief outline of the PCK algorithm is in order; the details can be found in § 2. Throughout this paper we will assume that the pair $(A, \mathbf{b})$ is completely controllable. The first step is to reduce the problem to a canonical form. An orthogonal similarity transformation is used to reduce $A - \mathbf{b}\mathbf{k}^T$ to the form $A^{(0)} - \mathbf{b}^{(0)}\mathbf{k}^{(0)T}$, where $A^{(0)}$ is unreduced, upper Hessenberg, $\mathbf{b}^{(0)} = b_1^{(0)}\mathbf{e}_1$, $b_1^{(0)} \neq 0$, and $\mathbf{e}_i$ denotes the ith standard basis vector.

The second step applies a procedure that we call PCK deflation n times. PCK deflation is closely related to deflation for Hessenberg matrices (see [W]); that is, given an eigenvalue of an unreduced, upper Hessenberg matrix of order n, an unreduced, upper Hessenberg matrix of order $n - 1$ can be produced possessing the remaining $n - 1$ eigenvalues. Each application of PCK deflation assigns one eigenvalue and finds one component of an orthogonally transformed gain vector. The first application of PCK deflation begins with the pair $(A^{(0)}, \mathbf{b}^{(0)})$ and finds an orthogonal $Q^{(0)}$ and a unique scalar α_1 so that $Q^{(0)T}\mathbf{b}^{(0)} = [\beta_1, b_1^{(1)}, 0, \cdots, 0]^T$, $b_1^{(0)} \neq 0$, and for any $n-1$ vector $\mathbf{k}^{(1)}$, $(Q^{(0)T}A^{(0)}Q^{(0)} - Q^{(0)T}\mathbf{b}^{(0)}[\alpha_1, \mathbf{k}^{(1)T}])\mathbf{e}_1 = \lambda_1\mathbf{e}_1$. The $n - 1$ vector $\mathbf{k}^{(1)}$ is determined by assigning the remaining $n - 1$ eigenvalues to $A^{(1)} - \mathbf{b}^{(1)}\mathbf{k}^{(1)T}$, where $A^{(1)}$ is the lower right $n - 1 \times n - 1$ block of $Q^{(0)T}A^{(0)}Q^{(0)}$ and $\mathbf{b}^{(1)} = b_1^{(1)}\mathbf{e}_1$. Since $Q^{(0)}$ can be constructed so that $A^{(1)}$ is unreduced, upper Hessenberg, PCK deflation can be applied to the pair $(A^{(1)}, \mathbf{b}^{(1)})$. This process continues until a 1×1 system matrix is reached.

The third step transforms the result back to the original coordinate system.

Throughout this paper $\|\cdot\| = \|\cdot\|_2$, $\|\cdot\|_F$ denotes the Frobenius norm, and u denotes the unit rounding error.

Our numerical stability result is given in the following theorem.

THEOREM 1.1. *Let* $A \in R^{n\times n}$, $\mathbf{b} \in R^n$, $\lambda_1, \cdots, \lambda_n \in R$, *and let the pair* $(A, \mathbf{b})$ *be completely controllable. Let* $\mathbf{k}$ *denote the gain vector computed by the* PCK *algorithm. Then there exist* $\Delta A \in R^{n\times n}$ *and* $\Delta\mathbf{b} \in R^n$ *with* $\|\Delta A\|/\|A\| = \mathbf{O}(n^3u)$, $\|\Delta\mathbf{b}\|/\|\mathbf{b}\| = \mathbf{O}(n^2u)$, *so that* $A + \Delta A - (\mathbf{b} + \Delta\mathbf{b})\mathbf{k}^T$ *has eigenvalues* $\lambda_1, \cdots, \lambda_n$.

In § 2 we explain the PCK algorithm in detail and in § 3 we prove Theorem 1.1 by performing a backward error analysis in three parts. In Proposition 3.1 we show that the computed result from PCK deflation is the exact result for a matrix whose relative difference from the original matrix is on the order of the machine unit. In Proposition 3.2 we show that the PCK algorithm is numerically stable if we start with the above-mentioned canonical form. Finally, Theorem 1.1 is proved using Proposition 3.2 and Wilkinson's results on the error analysis of Householder transformations.

2. The PCK algorithm.

In this section we describe the three steps in the PCK algorithm.

Step One. The PCK algorithm first transforms the pair $(A, \mathbf{b})$ into a canonical form via an orthogonal similarity transformation. Let $P_0 \in R^{n\times n}$ be a Householder transformation constructed so that $P_0\mathbf{b} = -\text{sign}\,(b_1)\|\mathbf{b}\|\mathbf{e}_1$ (see [GV, pp. 38–43]). Next, construct a product of $n - 2$ Householder transformations $W := P_1P_2\cdots P_{n-2}$ so that $A^{(0)} := W^T(P_0AP_0)W$ is upper Hessenberg (see [GV, pp. 38–43]). Set $P := P_0W$. Then $A^{(0)} = P^TAP$ and the form of W implies that $\mathbf{b}^{(0)} := P^T\mathbf{b} = P_0\mathbf{b} = b_1^{(0)}\mathbf{e}_1$ with $b_1^{(0)} = -\text{sign}\,(b_1)\|\mathbf{b}\|\mathbf{e}_1$. Now $P^T[\mathbf{b}, A\mathbf{b}, \cdots, A^{n-1}\mathbf{b}] = [\mathbf{b}^{(0)}, A^{(0)}\mathbf{b}^{(0)}, \cdots, (A^{(0)})^{n-1}\mathbf{b}^{(0)}]$ has rank n so that the pair $(A^{(0)}, \mathbf{b}^{(0)})$ is completely controllable. Let H be upper Hessenberg, then the pair $(H, \beta\mathbf{e}_1)$ is completely controllable if and only if H is unreduced and $\beta \neq 0$ (see [MP]). Consequently, $A^{(0)}$ is unreduced. It is sufficient now to find the unique gain $\mathbf{k}^{(0)}$ so that $A^{(0)} - \mathbf{b}^{(0)}\mathbf{k}^{(0)T}$ has eigenvalues $\lambda_1, \cdots, \lambda_n$, since then, with $\mathbf{k} = P\mathbf{k}^{(0)}$, $P^T[A^{(0)} - \mathbf{b}^{(0)}\mathbf{k}^{(0)T}]P = A - \mathbf{b}\mathbf{k}^T$ has the same eigenvalues.

PCK Deflation. In Step Two of the PCK algorithm, PCK deflation is applied $n-1$ times. Before outlining this step, we describe a simplified version of PCK deflation in detail. At the end of this section we show how this simplified version was modified in [PCK] to reduce the operation count. As mentioned in § 1, PCK deflation is related to deflation for unreduced, upper Hessenberg matrices. In general, PCK deflation requires as input an eigenvalue to be assigned and a pair $(H, \beta \mathbf{e}_1)$, where H is unreduced, upper Hessenberg and $\beta \neq 0$. In Step Two, PCK deflation is applied first to the eigenvalue λ_1 and the pair $(A^{(0)}, \mathbf{b}^{(0)})$ and this is where we begin.

If $\mathbf{v}^{(0)} \neq \mathbf{0}$ is an eigenvector of $A^{(0)} - \mathbf{b}^{(0)}\mathbf{k}^{(0)T}$ corresponding to λ_1, then $\mathbf{v}^{(0)}$ must satisfy

$$(2.1) \qquad [\mathbf{e}_2, \cdots, \mathbf{e}_n]^T(A^{(0)} - \lambda_1 I)\mathbf{v}^{(0)} = \mathbf{0}$$

because $A^{(0)}$ and $A^{(0)} - \mathbf{b}^{(0)}\mathbf{k}^{(0)T}$ are identical except for row one. Since $A^{(0)}$ is unreduced, upper Hessenberg, $v_n^{(0)}$ cannot be zero, so that once $v_n^{(0)} \neq 0$ is chosen, (2.1) can be solved for $\mathbf{v}^{(0)}$ by backward substitution. Next, for $i = 1, \cdots, n-1$ use Givens rotations $J_i(n-i, n-i+1)$ in the $(n-i, n-i+1)$-plane to transform $\mathbf{v}^{(0)}$ so that

$$\mathbf{v}^{(i)} := J_i(n-i, n-i+1) \cdots J_1(n-1, n)\mathbf{v}^{(0)} = [v_1^{(i)}, \cdots, v_{n-i}^{(i)}, 0, \cdots, 0]^T$$

where $v_{n-i}^{(i)} \geqq |v_n^{(0)}| > 0$. For $i = 1, \cdots, n-1$, set $D^{(i)} := J_i D^{(i-1)} J_i^T$ with $D^{(0)} := A^{(0)}$ and define $Q^{(0)T} := J_{n-1} \cdots J_1$. Then $Q^{(0)T}\mathbf{v}^{(0)} = v_1^{(n-1)}\mathbf{e}_1$ and

$$Q^{(0)T}A^{(0)}Q^{(0)} = D^{(n-1)}.$$

Now according to (2.1) there is a unique scalar α_1 so that $(A^{(0)} - \lambda_1 I)\mathbf{v}^{(0)} = v_1^{(n-1)}\alpha_1\mathbf{b}^{(0)}$. Transforming, we have $(D^{(n-1)} - \lambda_1 I)\mathbf{e}_1 = \alpha_1 Q^{(0)T}\mathbf{b}^{(0)}$, where $Q^{(0)T}\mathbf{b}^{(0)} = J_{n-1}\mathbf{b}^{(0)} = [\beta_1, \mathbf{b}^{(1)T}]^T$, with $\mathbf{b}^{(1)} := b_1^{(1)}\mathbf{e}_1$, an $n-1$ vector, and $b_1^{(1)} = -\mathbf{b}_1^{(0)}v_2^{(n-2)}/\{[\mathbf{v}_1^{(n-2)}]^2 + [\mathbf{v}_2^{(n-2)}]^2\}^{1/2} \neq 0$. If $|\beta_1| \geqq |b_1^{(1)}|$, we find α_1 from $d_{11}^{(n-1)} - \lambda_1 = \alpha_1\beta_1$, otherwise from $d_{12}^{(n-1)} = \alpha_1 b_1^{(1)}$. Solving for α_1 in this fashion is crucial to our proof of Theorem 1.1. For any $n-1$ vector $\mathbf{k}^{(1)}$, $(D^{(n-1)} - Q^{(0)T}\mathbf{b}^{(0)}[\alpha_1, k^{(1)T}])\mathbf{e}_1 = \lambda_1\mathbf{e}_1$ and hence $D^{(n-1)} - Q^{(0)T}\mathbf{b}^{(0)}[\alpha_1, \mathbf{k}^{(1)T}]$ is block 2×2 upper triangular. The $n-1$ vector $\mathbf{k}^{(1)}$ is determined by assigning the remaining $n-1$ eigenvalues to $A^{(1)} - \mathbf{b}^{(1)}\mathbf{k}^{(1)T}$, where $A^{(1)}$ denotes the lower right $n-1 \times n-1$ block of $D^{(n-1)}$.

If $n > 2$ $D^{(1)}$ has fill-in in the $(n, n-2)$ entry due to the form of $A^{(0)}$ and for $i = 2, \cdots, n-2$, $D^{(i)}$ should have fill-in in the $(n-i+2, n-i)$ and $(n-i+1, n-i-1)$ entries due to the form of $D^{(i-1)}$; however, $(A^{(0)} - \lambda_1 I)\mathbf{v}^{(0)} = v_1^{(n-1)}\alpha_1\mathbf{b}^{(0)}$ implies that $(D^{(i)} - \lambda_1 I)\mathbf{v}^{(i)} = v_1^{(n-1)}\alpha_1\mathbf{b}^{(0)}$. It follows that $d_{n-i+2,n-i}^{(i)}v_{n-i}^{(i)} = 0$ so that $d_{n-i+2,n-i}^{(i)} = 0$, since $v_{n-i}^{(i)} \geqq |v_n^{(0)}| > 0$. Thus $D^{(i)}$ has fill-in only in the $(n-i+1, n-i-1)$ entry for $i = 2, \cdots, n-2$. In particular, $\mathbf{D}^{(n-2)}$ has fill-in only in the $(3, 1)$ entry as does $D^{(n-1)}$ due to the form of J_{n-1}. Since

$$(D^{(n-1)} - \lambda_1 I)\mathbf{v}^{(n-1)} = v_1^{(n-1)}\alpha_1 Q^{(0)T}\mathbf{b}^{(0)}, \qquad d_{31}^{(n-1)}v_1^{(n-1)} = 0,$$

which implies that $d_{31}^{(n-1)} = 0$, since $v_1^{(n-1)} \geqq |v_n^{(0)}| > 0$. Consequently, $D^{(n-1)}$ is upper Hessenberg.

To see that $A^{(1)}$ is unreduced, define $\mathbf{k}^{(0)T} := [\alpha_1, \mathbf{k}^{(1)T}]Q^{(0)T}$, and $C^{(0)} := A^{(0)} - \mathbf{b}^{(0)}\mathbf{k}^{(0)T}$. Now $C^{(0)}$ is unreduced, upper Hessenberg so that the pairs $(C^{(0)}, \mathbf{b}^{(0)})$ and $(Q^{(0)T}C^{(0)}Q^{(0)}, Q^{(0)T}\mathbf{b}^{(0)})$ are completely controllable. But $Q^{(0)T}C^{(0)}Q^{(0)}$ is 2×2 block upper triangular with a 1×1 upper left block equal to λ_1. Let $C^{(1)}$ denote the lower right $n-1 \times n-1$ block of $Q^{(0)T}C^{(0)}Q^{(0)}$. It follows that the pair $(C^{(1)}, \mathbf{b}^{(1)})$ is completely controllable and so $C^{(1)}$ must be unreduced, but then so must $A^{(1)} = C^{(1)} + \mathbf{b}^{(1)}\mathbf{k}^{(1)T}$, since $A^{(1)}$ and $C^{(1)}$ differ only in their first rows.

Step Two continues with the application of PCK deflation to the pair $(A^{(1)}, \mathbf{b}^{(1)})$. We summarize Step Two as follows.

Step Two. For $i = 1, \cdots, n-2$, apply PCK deflation to the pair $(A^{(i-1)}, b^{(i-1)})$ and the eigenvalue λ_i to find the scalar α_i and an $n - i + 1 \times n - i + 1$ orthogonal matrix $Q^{(i-1)}$ so that for any $n - i$ vector $\mathbf{k}^{(i)}$

$$Q^{(i-1)T}A^{(i-1)}Q^{(i-1)} - [Q^{(i-1)T}\mathbf{b}^{(i-1)}][\alpha_i, \mathbf{k}^{(i)T}]$$

has the form

$$\left[\begin{array}{c|c} s_i & \\ \hline 0 & \\ \hline \mathbf{0} & \end{array}\right] = \left[\begin{array}{c|c} \gamma_{11}^{(i)} & \\ \hline \gamma_{21}^{(i)} & \\ \hline \mathbf{0} & A^{(i)} \end{array}\right] - \left[\begin{array}{c} \beta_i \\ \hline \\ \mathbf{b}^{(i)} \end{array}\right] \left[\begin{array}{c|c} \alpha_i & \mathbf{k}^{(i)T} \end{array}\right]$$

where $A^{(i)}$ is an $n - i \times n - i$ unreduced, upper Hessenberg matrix, $\mathbf{b}^{(i)} = b_1^{(i)}\mathbf{e}_1$ is an $n - i$ vector with $b_1^{(i)} \neq 0$, and β_i, $\gamma_{11}^{(i)}$, and $\gamma_{21}^{(i)}$ are scalars.

Apply PCK deflation to the pair $(A^{(n-2)}, b^{(n-2)})$ and the eigenvalue λ_{n-1} to find the scalar α_{n-1} and a 2×2 orthogonal matrix Q_{n-2} so that for any scalar α_n

$$Q^{(n-2)T}A^{(n-2)}Q^{(n-2)} - [Q^{(n-2)T}\mathbf{b}^{(n-2)}][\alpha_{n-1}, \alpha_n]$$

has the form

$$\left[\begin{array}{c|c} s_{n-1} & \\ \hline 0 & \end{array}\right] = \left[\begin{array}{c|c} \gamma_{11}^{(n-1)} & \\ \hline \gamma_{21}^{(n-1)} & \gamma_{22}^{(n-1)} \end{array}\right] - \left[\begin{array}{c} \beta_{n-1} \\ \beta_n \end{array}\right] \left[\begin{array}{cc} \alpha_{n-1} & \alpha_n \end{array}\right].$$

Step Three. Set $\mathbf{k}^{(n-1)} = \alpha_n := (\gamma_{22}^{(n-1)} - s_n)/\beta_n$. Transform back to obtain the gain vector $\mathbf{k}^{(0)}$. For $i = n-2, \cdots, 0$, set

$$\mathbf{k}^{(i)} = Q_i \begin{bmatrix} \alpha_{i+1} \\ \mathbf{k}^{(i+1)} \end{bmatrix} \in R^{n-i}.$$

Finally, set $\mathbf{k} = P\mathbf{k}^{(0)} \in R^n$.

Operation Count. Step One requires about $5n^3/3$ flops using the Moler concept of flops (see [GV]), while Step Three requires about $2n^2$ flops. One application of PCK deflation to a matrix of order m requires about $11m^3/2$ flops. This takes into account $m^2/2$ flops for the backward substitution to find the eigenvector $\mathbf{v}^{(0)}$, $4(m-1)$ flops for the Givens rotations, and $5m^2$ flops for $m-1$ updates of the form JHJ^T, where J is a Givens rotation and H is an $m \times m$ upper Hessenberg matrix except for one fill-in element. As was pointed out in [PCK], the cost of the computation of the transformed eigenvectors $\mathbf{v}^{(i)}$ can be reduced. First, compute $v_{m-1}^{(0)}$ and then for $i = 1, \cdots, m-2$, compute $v_{m-i-1}^{(i-1)}$ and $J_i(m-i, m-i+1)$ and update $v_{m-i}^{(i-1)}$ and $v_{m-i+1}^{(i)}$. In this way the $m^2/2$ flops required to find $\mathbf{v}^{(0)}$ can be replaced by $3(m-1)$ flops, a 9 percent deduction in the work. Using this version of PCK deflation, Step Two requires about $5m^3/3$ flops. Henceforth, the term PCK deflation will refer to this $5m^2$ flop version which we now state in detail.

PCK Deflation. Given $D^{(0)} \in R^{m \times m}$ unreduced, upper Hessenberg, $\mathbf{b}^{(0)} = b_1^{(0)}\mathbf{e}_1 \in R^m$ with $b_1^{(0)} \in 0$, and a real eigenvalue λ, compute scalars $\alpha_1, \beta_1, b_1^{(1)}$, and an orthogonal matrix $Q \in R^{m \times m}$.

```
        Choose $v_m^{(0)} \neq 0$.
(2.2)   $v_{m-1}^{(0)} := (\lambda - d_{mm}^{(0)}) v_m^{(0)} / d_{m,m-1}^{(0)}$
        For $i = 1, \cdots, m-2$      (do not execute if $m = 2$)
(2.3)       $v_{m-i-1}^{(i-1)} :=$
            $[(\lambda - d_{m-i,m-i}^{(i-1)}) v_{m-i}^{(i-1)} - d_{m-i,m-i+1}^{(i-1)} v_{m-i+1}^{(i-1)}] / d_{m-i,m-i-1}^{(i-1)}$
            $v_{m-i-1}^{(i)} = v_{m-i-1}^{(i-1)}$
            Construct a Givens rotation $J_i = J_i(m-i, m-i+1)$ in the $(m-i,$
            $m-i+1)$
            coordinate plane so that
                $\mathbf{e}_{m-i}^T J_i \mathbf{v}^{(i-1)} =: v_{m-i}^{(i)} > 0$
                $\mathbf{e}_{m-i+1}^T J_i \mathbf{v}^{(i-1)} =: v_{m-i+1}^{(i)} = 0$
            For $k = m-i+2, \cdots, m$      (do not execute if $m > m-i+2$)
                $v_k^{(i)} := 0$
            repeat
            $D^{(i)} := J_i D^{(i-1)} J_i^T$
        repeat
```

Comment: all components of $v^{(m-2)}$ have been computed.

```
        Construct $J_{m-1} = J_{m-1}(1, 2)$ so that $\mathbf{v}^{(m-1)} := J_{m-1}\mathbf{v}^{(m-2)}$ satisfies
            $v_1^{(m-1)} > 0$ and $v_2^{(m-1)} = 0$
        $D^{(m-1)} := J_{m-1} D^{(m-2)} J_{m-1}^T$
        $Q^{(0)T} := J_{m-1} \cdots J_1$
        $\beta_1 := \mathbf{e}_1^T J_{m-1} \mathbf{b}^{(0)}$
        $b_1^{(1)} := \mathbf{e}_2^T J_{m-1} \mathbf{b}^{(0)}$
        If $(|\beta_1| \geqq |b_1^{(1)}|)$ then
            $\alpha_1 := (d_{11}^{(m-1)} - \lambda)/\beta_1$
        else
            $\alpha_1 := d_{21}^{(m-1)} / b_1^{(1)}$
        end if.
```

3. Backward error analysis.

In this section we will use the notation of Wilkinson and his error analysis of Givens rotations (see [W, pp. 131–141]). Following Wilkinson, we simplify bounds of the form $(1-u)^r \leqq 1 + \varepsilon \leqq (1+u)^r$, where u denotes the unit rounding error, by assuming that $ru < 0.1$; it then follows that $|\varepsilon| < r(1.06)u$ (see also [DB, p. 52]). PCK deflation takes as data an $m \times m$ matrix $D^{(0)}$, an m vector $\mathbf{b}^{(0)}$, a real eigenvalue λ_1, and a scalar normalization $v_m^{(0)}$. As output, it provides an orthogonal transformation $Q^{(0)}$; the updates $Q^{(0)T}D^{(0)}Q^{(0)}$ and $Q^{(0)T}\mathbf{b}^{(0)}$; and the first component, α_1, of the transformed gain vector. Proposition 3.1 is a perturbation result for PCK deflation.

PROPOSITION 3.1. *For a positive integer* $m \geqq 2$ *consider* PCK *deflation applied to the floating data* $D^{(0)}$, $\mathbf{b}^{(0)}$, λ, *and* $v_m^{(0)}$, *where* $D^{(0)} \in R^{m \times m}$ *is unreduced, upper Hessenberg, and* $\mathbf{b}^{(0)} = b_1^{(0)}\mathbf{e}_1$ *with* $b_1^{(0)} \neq 0$. *Let* $\hat{J}_1, \cdots, \hat{J}_{m-1} \in R^{m \times m}$ *denote the computed, approximate Givens rotations and let* $D^{(k)} := fl(\hat{J}_k D^{(k-1)} \hat{J}_k^T)$, $k = 1, \cdots, m-1$. *Assume that* $12(m-1)u < 0.1$ *and set* $d := 1.06$.

Then there exists an unreduced, upper Hessenberg perturbation $\tilde{D}^{(0)}$, *with* $\|\tilde{D}^{(0)} - D^{(0)}\|_F \leqq (5m-4)du\|D^{(0)}\|_F$, *and a perturbation* $\tilde{v}_m^{(0)}$, *with* $|\tilde{v}_m^{(0)} - v_m^{(0)}| \leqq 2(m-1)du|v_m^{(0)}|$, *so that* PCK *deflation applied in exact arithmetic to* $\tilde{D}^{(0)}$, $\mathbf{b}^{(0)}$, λ_1, *and* $\tilde{\mathfrak{v}}_m^{(0)}$ *yields Givens rotations* $J_1, \cdots, J_{m-1} \in R^{m\times m}$, *with* $\| J_k - \hat{J}_k \| \leqq 3du$, *and updates* $\tilde{D}^{(k)} = J_k \tilde{D}^{(k-1)} J_k^T$, *with* $\|\tilde{D}^{(k)} - D^{(k)}\|_F \leqq \{5m - 4 + 12k\} du \| D^{(0)}\|_F$ *for* $k = 1, \cdots, m-1$. If $Q^{(0)T} := J_{m-1} \cdots J_1$ and $\hat{Q}^{(0)T}$ *denotes the computed accumulation of the transformations* $\hat{J}_1, \cdots, \hat{J}_{m-1}$, *then* $\|Q^{(0)} - \hat{Q}^{(0)}\|_F \leqq 6\sqrt{m}(m-1)\,du$.

Proof. Let $v_j^{(i)}$ denote the computed components of the transformed eigenvectors in PCK deflation. Using the Wilkinson model of floating point arithmetic (see [W, pp. 112–116]), we have that

$$v_{m-1}^{(0)} = (\lambda - d_{mm}^{(0)}) v_m^{(0)} / [d_{m,m-1}^{(0)}(1+\delta_{-1})] \tag{3.1a}$$

with $|\delta_{-1}| \leqq 3du$ and for $k = 0, \cdots, m-3$

$$\begin{aligned} v_{m-k-2}^{(k+1)} &= v_{m-k-2}^{(k)} \\ &= [(\lambda - d_{m-k-1,m-k-1}^{(k)}) v_{m-k-1}^{(k)} - d_{m-k-1,m-k}^{(k)} v_{m-k}^{(k)}(1+\gamma_k)]/ \\ &\qquad [d_{m-k-1,m-k-2}^{(k)}(1+\delta_k)] \end{aligned} \tag{3.1b}$$

where $|\delta_k| \leqq 4du$ and $|\gamma_k| \leqq 3du$. Now for $i = 1, \cdots, m-1$, define the $m \times m$ Givens rotation $J_i = J_i(m-i, m-i+1)$ by $c_i = v_{m-i}^{(i-1)}/\mathrm{SQR}(i)$ and $s_i = v_{m-i+1}^{(i-1)}/\mathrm{SQR}(i)$, where $\mathrm{SQR}(i) := \{[v_{m-i}^{(i-1)}]^2 + [v_{m-i+1}^{(i-1)}]^2\}^{1/2}$ so that

$$\begin{bmatrix} c_i & s_i \\ -s_i & c_i \end{bmatrix} \begin{bmatrix} v_{m-i}^{(i-1)} \\ v_{m-i+1}^{(i-1)} \end{bmatrix} = \begin{bmatrix} \mathrm{SQR}(i) \\ 0 \end{bmatrix}$$

and let $\hat{J}_i$ denote the corresponding computed, approximate Givens rotation with computed cosine and sine $\hat{c}_i$ and $\hat{s}_i$. Then $|\hat{c}_i - c_i| \leqq 3du|c_i|$, $|\hat{s}_i - s_i| \leqq 3du|s_i|$, $\|J_i - \hat{J}_i\| \leqq 3du$, and $v_{m-i}^{(i)} = \mathrm{SQR}(i)(1+\varepsilon_i)$ with $|\varepsilon_i| \leqq 2du$. In addition, if $Q^{(0)T} := J_{m-1} \cdots J_1$ and $\hat{Q}^{(0)T}$ denotes the accumulation of the corresponding computed transformations, then $\|Q^{(0)} - \hat{Q}^{(0)}\|_F \leqq 6\sqrt{m}(m-1)du$ (see [W]). Next, for $i = 0, \cdots, m-1$, define vectors $\tilde{\mathbf{v}}^{(i)}$ by

$$\begin{aligned} \tilde{\mathbf{v}}^{(m-1)} &= \mathbf{v}^{(m-1)} = [v_1^{(m-1)}, 0, \cdots, 0]^T, \\ \tilde{\mathbf{v}}^{(m-2)} &= [v_1^{(m-2)}(1+\varepsilon_{m-1}), v_2^{(m-2)}(1+\varepsilon_{m-1}), 0, \cdots, 0]^T, \\ \tilde{\mathbf{v}}^{(m-3)} &= [v_1^{(m-2)}(1+\varepsilon_{m-1}), v_2^{(m-3)}(1+\varepsilon_{m-1})(1+\varepsilon_{m-2}), \\ &\qquad v_3^{(m-3)}(1+\varepsilon_{m-1})(1+\varepsilon_{m-2}), 0, \cdots, 0]^T, \\ \tilde{\mathbf{v}}^{(0)} &= \begin{bmatrix} v_1^{(m-2)}(1+\varepsilon_{m-1}) \\ v_2^{(m-3)}(1+\varepsilon_{m-1})(1+\varepsilon_{m-2}) \\ v_3^{(m-4)}(1+\varepsilon_{m-1})(1+\varepsilon_{m-2})(1+\varepsilon_{m-3}) \\ \cdots \\ v_{m-2}^{(1)}(1+\varepsilon_{m-1})\cdots(1+\varepsilon_2) \\ v_{m-1}^{(0)}(1+\varepsilon_{m-1})\cdots(1+\varepsilon_1) \\ v_m^{(0)}(1+\varepsilon_{m-1})\cdots(1+\varepsilon_1) \end{bmatrix}. \end{aligned} \tag{3.2}$$

It follows that $\tilde{\mathbf{v}}^{(m-1)} = \mathbf{v}^{(m-1)} = J_{m-1}\tilde{\mathbf{v}}^{(m-2)} = J_{m-1}J_{m-2}\tilde{\mathbf{v}}^{(m-3)} = \cdots = J_{m-1}J_{m-2}\cdots J_1\tilde{\mathbf{v}}^{(0)}$, and that for $k = 0, \cdots, m-3$

$$\begin{aligned} \tilde{v}_{m-k}^{(k)} &= v_{m-k}^{(k)}(1+\varepsilon_{m-1}) \cdots (1+\varepsilon_{k+1}) \\ \tilde{v}_{m-k-1}^{(k)} &= v_{m-k-1}^{(k)}(1+\varepsilon_{m-1}) \cdots (1+\varepsilon_{k+1}) \\ \tilde{v}_{m-k-2}^{(k+1)} &= \tilde{v}_{m-k-2}^{(k)} = v_{m-k-2}^{(k+1)}(1+\varepsilon_{m-1}) \cdots (1+\varepsilon_{k+2}). \end{aligned} \tag{3.3}$$

The first equation in (3.3) with $k = 0$ implies that $|\tilde{v}_m^{(0)} - v_m^{(0)}| \leqq 2(m-1)du|v_m^{(0)}|$.

PCK deflation has the property that for $k = 1, \cdots, m-3$

$$(3.4) \qquad d^{(k)}_{m-k-1,m-k-1} = d^{(0)}_{m-k-1,m-k-1} \quad \text{and} \quad d^{(k)}_{m-k-1,m-k-2} = d^{(0)}_{m-k-1,m-k-2}.$$

Multiplying both sides of (3.1a) by $(1+\varepsilon_{m-1}) \cdots (1+\varepsilon_1)$ and using (3.3) with $k = 0$, we find

$$(3.5a) \qquad \tilde{v}^{(0)}_{m-1} = (\lambda - d^{(0)}_{mm})\tilde{v}^{(0)}_m / [d^{(0)}_{m,m-1}(1+\delta_{-1})].$$

Similarly, for $k = 0, \cdots, m-3$, multiplying both sides of (3.1b) by $(1+\varepsilon_{m-1}) \cdots (1+\varepsilon_{k+2})$ and using (3.4) and (3.3), we have

$$\begin{aligned} \tilde{v}^{(k+1)}_{m-k-2} &= \tilde{v}^{(k)}_{m-k-2} \\ (3.5\text{b}) \qquad &= [(\lambda - d^{(0)}_{m-k-1,m-k-1})\tilde{v}^{(k)}_{m-k-1} - d^{(k)}_{m-k-1,m-k}\tilde{v}^{(k)}_{m-k}(1+\gamma_k)] / \\ &\qquad [d^{(0)}_{m-k-1,m-k-2}(1+\beta_k)] \end{aligned}$$

where $1+\beta_k = (1+\delta_k)(1+\varepsilon_{k+1})$ and hence $|\beta_k| \leqq 6du$. Now define $\tilde{d}^{(0)}_{m,m-1} := d^{(0)}_{m,m-1}(1+\delta_{-1})$ and for $i = 1, \cdots, m$, define $\tilde{d}^{(0)}_{1i} := d^{(0)}_{1i}$ and $\tilde{d}^{(0)}_{ii} := d^{(0)}_{ii}$. Then (3.5a) becomes

$$\tilde{v}^{(0)}_{m-1} = (\lambda - \tilde{d}^{(0)}_{mm})\tilde{v}^{(0)}_m / \tilde{d}^{(0)}_{m,m-1}.$$

Next, define $\tilde{d}^{(0)}_{m-1,m-2} := d^{(0)}_{m-1,m-2}(1+\beta_0)$ and $\tilde{d}^{(0)}_{m-1,m} := d^{(0)}_{m-1,m}(1+\gamma_0)$. Then from (3.5b) with $k = 0$ we obtain

$$\tilde{v}^{(1)}_{m-2} = \tilde{v}^{(0)}_{m-2} = [(\lambda - \tilde{d}^{(0)}_{m-1,m-1})\tilde{v}^{(0)}_{m-1} - \tilde{d}^{(0)}_{m-1,m}\tilde{v}^{(0)}_m] / \tilde{d}^{(0)}_{m-1,m-2}.$$

To continue this process we have

$$\begin{aligned} d^{(1)}_{m-2,m-1} &= fl(d^{(0)}_{m-2,m-1}\hat{c}_1 + d^{(0)}_{m-2,m}\hat{s}_1) \\ &= d^{(0)}_{m-2,m-1}(1+\zeta_{m-2,m-1})c_1 + d^{(0)}_{m-2,m}(1+\sigma_{m-2,m})s_1, \end{aligned}$$

where $|\zeta_{m-2,m-1}|$, $|\sigma_{m-2,m-1}| \leqq 5du$. Define $\tilde{d}^{(0)}_{m-2,m-3} := d^{(0)}_{m-2,m-3}(1+\beta_1)$, $\tilde{d}^{(1)}_{m-2,m-1} := d^{(1)}_{m-2,m-1}(1+\gamma_1)$, $\tilde{d}^{(0)}_{m-2,m-1} := d^{(0)}_{m-2,m-1}(1+\zeta_{m-2,m-1})(1+\gamma_1)$, and $\tilde{d}^{(0)}_{m-2,m} := d^{(0)}_{m-2,m}(1+\sigma_{m-2,m})(1+\gamma_1)$. Then from (3.5b) with $k = 1$ we obtain

$$\tilde{v}^{(2)}_{m-3} = \tilde{v}^{(1)}_{m-3} = [(\lambda - \tilde{d}^{(0)}_{m-2,m-2})\tilde{v}^{(1)}_{m-2} - \tilde{d}^{(1)}_{m-2,m-1}\tilde{v}^{(1)}_{m-1}] / \tilde{d}^{(0)}_{m-2,m-3}.$$

If $m \geqq 5$, then for each k, $k = 2, \cdots, m-3$, we can start from

$$\begin{aligned} &d^{(k)}_{m-k-1,m-k} \\ &\quad = fl(d^{(k-1)}_{m-k-1,m-k}\hat{c}_k + d^{(k-1)}_{m-k-1,m+1}\hat{s}_k) \\ &\quad = d^{(0)}_{m-k-1,m-k}(1+\zeta_{m-k-1,m-k})c_k + d^{(k-1)}_{m-k-1,m-k+1}(1+\sigma_{m-k-1,m-k+1})s_k, \end{aligned}$$

where $|\zeta_{m-k-1,m-k}|$, $|\sigma_{m-k-1,m-k+1}| \leqq 5du$, and define $\tilde{d}^{(0)}_{m-k-1,m-k-2}$ and $\tilde{d}^{(0)}_{m-k-1,m-k+j}$ for $j = 0, \cdots, k$, so that

$$|\tilde{d}^{(0)}_{m-k-1,m-k-2} - d^{(0)}_{m-k-1,m-k-2}| \leqq (5m-12)du|d^{(0)}_{m-k-1,m-k-2}|,$$

$|\tilde{d}^{(0)}_{m-k-1,m-k+j} - d^{(0)}_{m-k-1,m-k+j}| \leqq (5m-12)du|d^{(0)}_{m-k-1,m-k+j}|$. Substituting in (3.5b), we find

$$\begin{aligned} \tilde{v}^{(k+1)}_{m-k-2} &= \tilde{v}^{(k)}_{m-k-2} \\ &= [(\lambda - \tilde{d}^{(0)}_{m-k-1,m-k-1})\tilde{v}^{(k)}_{m-k-1} - \tilde{d}^{(k)}_{m-k-1,m-k}\tilde{v}^{(k)}_{m-k}] / \tilde{d}^{(0)}_{m-k-1,m-k-2}. \end{aligned}$$

This process defines an $m \times m$ unreduced, upper Hessenberg $\tilde{D}^{(0)}$ satisfying the entrywise bound $|\tilde{D}^{(0)} - D^{(0)}| \leqq (5m-4)du|D^{(0)}|$ (which holds for $m \geqq 2$).

It follows that starting with $\tilde{D}^{(0)}$ and $\tilde{v}_m^{(0)}$, PCK deflation in exact arithmetic produces the m vectors $\tilde{\mathbf{v}}^{(0)}, \cdots, \tilde{\mathbf{v}}^{(m-2)}, \tilde{\mathbf{v}}^{(m-1)} = \mathbf{v}^{(m-1)}$, the $m \times m$ Givens rotations $J_1, \cdots, J_{m-1}$, and the updates $\tilde{D}^{(k)} = J_k \tilde{D}^{(k-1)} J_k^T$. The estimate $\|\tilde{D}^{(0)} - D^{(0)}\|_F \leqq (5m-4)du\|D^{(0)}\|_F$, together with Wilkinson's bound on Givens updates yields $\|\tilde{D}^{(k)} - D^{(k)}\|_F \leqq (5m - 4 + 12k)du\|D^{(0)}\|_F$ for $k = 1, \cdots, m-1$. By examining [W, pp. 131–141], it can be seen that in this proof bounds of the form $(1-u)^r \leqq 1 + \varepsilon \leqq (1+u)^r$ arise with $r \leqq 12(m-1)$ so that d can be taken to be 1.06, since we have assumed that $12(m-1)u < 0.1$. □

The next step in the proof of Theorem 1.1 is to prove that this theorem holds when the problem is in the canonical form mentioned in § 1.

PROPOSITION 3.2. *Let $A^{(0)} \in R^{n \times n}$ be unreduced, upper Hessenberg and let $\mathbf{b}^{(0)} = b_1^{(0)}\mathbf{e}_1$, $b_1^{(0)} \neq 0$. Let $\mathbf{k}^{(0)}$ be the gain vector computed by the* PCK *algorithm applied to the pair $(A^{(0)}, \mathbf{b}^{(0)})$. Then there exists Z, ΔZ, $\Delta A^{(0)} \in R^{n \times n}$ and $\Delta \mathbf{b}^{(0)} \in R^n$, with Z orthogonal and $\|\Delta Z\|_F \leqq 4n^2u + O(u^2)$, $\|\Delta A^{(0)}\|_F / \|A^{(0)}\|_F \leqq 16n^3u + O(u^2)$, $\|\Delta \mathbf{b}^{(0)}\| / \|\mathbf{b}^{(0)}\| \leqq 5n^2u + O(u^2)$, so that*

$$A^{(0)} + \Delta A^{(0)} - (\mathbf{b}^{(0)} + \Delta \mathbf{b}^{(0)})\mathbf{k}^{(0)T} = (Z + \Delta Z)\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}(Z + \Delta Z)^{-1} \tag{3.6}$$

where $\lambda_1, \cdots, \lambda_n$ are the eigenvalues to be assigned.

Proof by induction on n. We begin by using Proposition 3.1 with $m = n$, $\lambda = \lambda_1$, $D^{(0)} = A^{(0)}$, and the machine unit u sufficiently small. Let $Q^{(0)T} = J_{n-1} \cdots J_1$, $[\beta_1, \mathbf{b}^{(1)T}]^T := fl(\hat{J}_{n-1}\mathbf{b}^{(0)})$ and $[\tilde{\beta}_1, \tilde{\mathbf{b}}^{(1)T}]^T := Q^{(0)T}\mathbf{b}^{(0)} = J_{n-1}\mathbf{b}^{(0)}$. In the proof of Proposition 3.1, it is shown that $\beta_1 = \tilde{\beta}_1(1 + \varepsilon_1)$, and $\mathbf{b}_1^{(1)} = \tilde{\mathbf{b}}_1^{(1)}(1 + \varepsilon_2)$ with $|\varepsilon_1|$, $|\varepsilon_2| \leqq 4du$.

Now set $D^{(n-1)} = R + E$, where R is the upper Hessenberg part of $D^{(n-1)}$ and E is zero except for fill-in due to roundoff on the second subdiagonal of $D^{(n-1)}$. Since $\tilde{D}^{(n-1)}$ is upper Hessenberg, $\|E\|_F \leqq \|\tilde{D}^{(n-1)} - D^{(n-1)}\|_F \leqq (17n - 16)du\|A^{(0)}\|_F$. We will find scalars α_1 and η with $|\eta| \leqq 2du$, and a matrix F that is zero except in the $(1, 1)$ and $(2, 1)$ entries so that $\|F\|_F \leqq (34n - 22)du\|A^{(0)}\|_F$ and for any $n-1$ vector $\mathbf{k}^{(1)}$,

$$R + F - \begin{bmatrix} \beta_1(1+\eta) \\ \mathbf{b}^{(1)} \end{bmatrix}[\alpha_1, \mathbf{k}^{(1)T}] = \begin{bmatrix} \lambda_1 & \\ \mathbf{0} & R^{(1)} - \mathbf{b}^{(1)}\mathbf{k}^{(1)T} \end{bmatrix}. \tag{3.7}$$

Here, $R^{(1)}$ denotes the lower right $n-1$ by $n-1$ block of R.

In case $|b_1^{(1)}/\beta_1| \leqq 1$, PCK deflation computes α_1 in (3.7) by

$$\alpha_1 = fl[(d_{11}^{(n-1)} - \lambda_1)/\beta_1],$$

hence

$$\alpha_1 = ((d^{(n-1)})_{11} - \lambda_1)/[\beta_1(1+\eta)] \tag{3.8}$$

for some η with $|\eta| \leqq 2du$. Set $f_{11} := 0$. Then equality holds for the $(2, 1)$ entry in (3.7). We now show that $|f_{21}| \leqq (34n - 22)du\|A^{(0)}\|_F$. PCK deflation applied in exact arithmetic to the pair $(\tilde{D}^{(0)}, \mathbf{b}^{(0)})$ and the eigenvalue λ_1 yields a unique scalar $\tilde{\alpha}_1$ so that $(\tilde{D}^{(n-1)} - \lambda_1 I)\mathbf{e}_1 = \tilde{\alpha}_1[\tilde{\beta}_1, \tilde{\mathbf{b}}^{(1)T}]^T$, where $\tilde{\mathbf{b}}^{(1)} = \tilde{b}_1^{(1)}\mathbf{e}_1$ with $\tilde{b}_1^{(1)} \neq 0$. Consequently, $\tilde{d}_{11}^{(n-1)} - \lambda_1 = \tilde{\alpha}_1\tilde{\beta}_1$ and $\tilde{d}_{21}^{(n-1)} = \tilde{\alpha}_1\tilde{b}_1^{(1)}$ from which we obtain the identity

$$\lambda_1 = \tilde{d}_{11}^{(n-1)} - \tilde{\beta}_1\tilde{d}_{21}^{(n-1)}/\tilde{b}_1^{(1)}. \tag{3.9}$$

Using (3.8) and (3.9), we find that

$$f_{21}=b_1^{(1)}[d_{11}^{(n-1)}-\tilde{d}_{11}^{(n-1)}]/\beta_1(1+\eta)+b_1^{(1)}\tilde{\beta}_1\tilde{d}_{21}^{(n-1)}/[\tilde{b}_1^{(1)}\beta_1(1+\eta)]-d_{21}^{(n-1)}.$$

Since $|d_{ij}^{(n-1)}-\tilde{d}_{ij}^{(n-1)}|\leqq\|\tilde{D}^{(n-1)}-D^{(n-1)}\|_F$, it follows from Proposition 3.1 that $|f_{21}|\leqq(34n-22)du\|A^{(0)}\|_F$. The case $|\beta_1/b_1^{(1)}|<1$ can be treated similarly, except that now $\alpha_1=fl(d_{21}^{(n-1)}/b_1^{(1)})$. We note that our proof fails if α_1 is not computed in this way. The PCK algorithm proceeds by assigning the remaining eigenvalues $\lambda_2,\cdots,\lambda_n$ to the pair $(R^{(1)},\mathbf{b}^{(1)})$. Denote the resulting gain vector by $\mathbf{k}^{(1)}$. By the induction hypothesis we have $n-1\times n-1$ matrices $Z^{(1)}$, $\Delta Z^{(1)}$, $\Delta R^{(1)}$, and an $n-1$ vector $\Delta\mathbf{b}^{(1)}$ with $Z^{(1)}$ orthogonal and $\|\Delta Z^{(1)}\|_F\leqq 4(n-1)^2u+O(u^2)$, $\|\Delta R^{(1)}\|_F/\|R^{(1)}\|_F\leqq 16(n-1)^3u+O(u^2)$, $\|\Delta\mathbf{b}^{(1)}\|/\|\mathbf{b}^{(1)}\|\leqq 5(n-1)^2u+O(u^2)$, and

(3.10)

$$R^{(1)}+\Delta R^{(1)}-(\mathbf{b}^{(1)}+\Delta\mathbf{b}^{(1)})\mathbf{k}^{(1)T}=(Z^{(1)}+\Delta Z^{(1)})\begin{bmatrix}\lambda_2 & \\ O & \lambda_n\end{bmatrix}(Z^{(1)}+\Delta Z^{(1)})^{-1}.$$

Now set $W:=\operatorname{diag}(1,Z^{(1)})$, $\Delta W:=\operatorname{diag}(0,\Delta Z^{(1)})$, $H:=\operatorname{diag}(0,\Delta R^{(1)})$, and

$$G=\begin{bmatrix}0 & \mathbf{0}\\ \alpha_1\Delta\mathbf{b}^{(1)} & \mathbf{0}\end{bmatrix}.$$

From (3.7) and (3.10) we have

$$\begin{aligned}R+F+H+G-\begin{bmatrix}\beta_1(1+\eta)\\ \mathbf{b}^{(1)}+\Delta\mathbf{b}^{(1)}\end{bmatrix}[\alpha_1,\mathbf{k}^{(1)^T}]\\ =\tilde{D}^{(n-1)}+L-[Q^{(0)T}\mathbf{b}^{(0)}+\mathbf{f}][\alpha_1,\mathbf{k}^{(1)^T}]\\ =(W+\Delta W)\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ \mathbf{O} & & \lambda_n\end{bmatrix}(W+\Delta W)^{-1}\end{aligned}\tag{3.11}$$

where $L=D^{(n-1)}-\tilde{D}^{(n-1)}-E+F+H+G$ and

$$\mathbf{f}=\begin{bmatrix}\beta_1-\tilde{\beta}_1+\beta_1\eta\\ \mathbf{b}^{(1)}-\tilde{\mathbf{b}}^{(1)}+\Delta\mathbf{b}^{(1)}\end{bmatrix}.$$

Next, we obtain estimates for $\|L\|_F$ and $\|\mathbf{f}\|$. First,

$$\begin{aligned}\|G\|_F&=\|\alpha_1\Delta\mathbf{b}^{(1)}\|=|\alpha_1|[5(n-1)^2u+O(u^2)]\|\mathbf{b}^{(1)}\|\\ &=|\alpha_1b_1^{(1)}|[5(n-1)^2u+O(u^2)]\leqq\|\tilde{D}^{(n-1)}\|_F[5(n-1)^2u+O(u^2)]\\ &=\|A^{(0)}\|_F[5(n-1)^2u+O(u^2)].\end{aligned}$$

Here we have used (3.7). Also, $\|H\|_F=\|\Delta R^{(1)}\|_F\leqq[16(n-1)^3u+O(u^2)]\|R^{(1)}\|_F\leqq[16(n-1)^3u+O(u^2)]\|R\|_F\leqq[16(n-1)^3u+O(u^2)]\|A^{(0)}\|_F$. It follows that $\|L\|_F\leqq[\{68n-54+16(n-1)^3u+5(n-1)^2\}du+O(u^2)]\|A^{(0)}\|_F$. Finally, $\|\mathbf{f}\|\leqq\{4\sqrt{2}du+[5(n-1)^2u+O(u^2)]\}\|\mathbf{b}^{(0)}\|$.

The computed gain vector is found from

$$\mathbf{k}^{(0)}=fl(\hat{J}_1\cdots fl(\hat{J}_{n-2}fl(\hat{J}_{n-1}[\alpha_1,\hat{\mathbf{k}}^{(1)T}]^T))\cdots).$$

Now given $\mathbf{x} \in R^n$, there exists $\Delta J_i \in R^{n\times n}$ so that $\|\Delta J_i\| \leqq 5\sqrt{2}du$ and $fl(\hat{J}_i\mathbf{x}) = (J_i + \Delta J_i)\mathbf{x}$. It follows that there exists $\Delta Q^{(0)} \in R^{n\times n}$ so that

$$\|\Delta Q^{(0)}\|_F \leqq (n-1)5\sqrt{2}du + O(u^2)$$

and $\mathbf{k}^{(0)} = (Q^{(0)} + \Delta Q^{(0)})[\alpha_1, \hat{\mathbf{k}}^{(1)T}]^T$. Setting $(Q^{(0)} + \Delta Q^{(0)})^{-T} = Q^{(0)} + \Delta U$, it follows that

$$\|\Delta U\|_F \leqq \|\Delta Q^{(0)}\|_F/(1 - \|\Delta Q^{(0)}\|_F) \leqq (n-1)5\sqrt{2}du + O(u^2).$$

Multiplying (3.11) on the right by $(Q^{(0)} + \Delta Q^{(0)})^T$ and on the left by $Q^{(0)} + \Delta U$ and collecting terms we have (3.6) with $\Delta A^{(0)} = Q^{(0)}\tilde{D}^{(n-1)}\Delta Q^{(0)T} + Q^{(0)}LQ^{(0)T} + Q^{(0)}L\Delta Q^{(0)T} + \Delta U\tilde{D}^{(n-1)}Q^{(0)T} + \Delta U\tilde{D}^{(n-1)}\Delta Q^{(0)T} + \Delta ULQ^{(0)T} + \Delta UL\Delta Q^{(0)T}$, $\Delta\mathbf{b}^{(0)} = Q^{(0)}\mathbf{f} + \Delta UQ^{(0)T}\mathbf{b}^{(0)} + \Delta U\mathbf{f}$, $Z := Q^{(0)}W$, and $\Delta Z = \Delta UW + Q\Delta W + \Delta U\Delta W$. Finally, for $n \geqq 2$, $\|\Delta A^{(0)}\|_F/\|A^{(0)}\|_F \leqq 2(n-1)5\sqrt{2}du + (68n-54)du + 16(n-1)^3u + 5(n-1)^2u + O(u^2) \leqq 16n^3u + O(u^2)$. $\|\Delta\mathbf{b}^{(0)}\|/\|\mathbf{b}^{(0)}\| \leqq 4\sqrt{2}du + 5(n-1)^2u + (n-1)5\sqrt{2}du \leqq 5n^2$; and $\|\Delta' Z\|_F \leqq (n-1)5\sqrt{2}du + 4(n-1)^2u + O(u^2) \leqq 4n^2u + O(u^2)$. □

We now have the tools to prove Theorem 1.1.

Proof of Theorem 1.1. Let $A \in R^{n\times n}$ and $\mathbf{b} \in R^n$ and let the pair $(A, \mathbf{b})$ be completely controllable. We use the notation of the first step of the PCK algorithm in § 2. Our statements about the error analysis for Householder transformations are based on [W, pp. 152–162] (see also [GV, p. 41]). Let $\hat{P}_0$ denote the computed Householder transformation and define $\mathbf{b}^{(0)} := fl(\hat{P}_0\mathbf{b})$ and $B := fl(\hat{P}_0A\hat{P}_0)$. Then $\mathbf{b}^{(0)} = P_0(\mathbf{b} + \mathbf{e})$ and $B = P_0(A + E)P_0$, where $\|\mathbf{e}\| \leqq n^2du\|\mathbf{b}\|$ and $\|E\| \leqq n^2du\|A\|$. Let $W = P_1\cdots P_{n-2}$ be as in § 2 and let $A^{(0)}$ denote the computed upper Hessenberg matrix. Then $A^{(0)} = W^T(B + F)W$, where $\|F\| \leqq \sqrt{2}n^2du\|B\|$. Set $P := P_0W$. Then $A^{(0)} = P^T(A + G)P$, where $\|G\| \leqq 3n^2du\|A\|$ and $\|A^{(0)}\| = \|A\|(1 + \mathbf{O}(u))$. Since P^TAP is unreduced, upper Hessenberg so is $A^{(0)}$ for u sufficiently small. We also have $\mathbf{b}^{(0)} = W^TP_0(\mathbf{b} + \mathbf{e}) = P^T(\mathbf{b} + \mathbf{e})$. Following Wilkinson's analysis, we can derive the following estimates. If $\hat{P}_k$ is the computed Householder transformation, then given $\mathbf{y} \in R^n$ there exists $\Delta P_k \in R^{n\times n}$ with $\|\Delta P_k\| \leqq [15 + 2(n - k + 1)]du$ so that $fl(\hat{P}_k\mathbf{y}) = (P_k + \Delta P_k)\mathbf{y}$. Due to the form of the Householder transformations $P_0, \cdots, P_{n-2}$, it follows that there exists $\Delta P \in R^{n\times n}$ with $\|\Delta P\| \leqq (n^2 + 18n - 19)du$ so that $\mathbf{k} := fl(\hat{P}\mathbf{k}^{(0)}) = (P + \Delta P)\mathbf{k}^{(0)}$.

We now apply Proposition 3.2 to the pair $(A^{(0)}, \mathbf{b}^{(0)})$. Substitute $A^{(0)} = P^T(A + G)P$, $\mathbf{b}^{(0)} = P^T(\mathbf{b} + \mathbf{e})$, and $\mathbf{k}^{(0)} = (P + \Delta P)^{-1}\mathbf{k}$ into (3.6) and define $T := PZ$ and ΔT by $T + \Delta T := (P + \Delta P)^{-T}(Z + \Delta Z)$. We have

$$A + \Delta A - (\mathbf{b} + \Delta\mathbf{b})\mathbf{k}^T = (T + \Delta T)\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}(T + \Delta T)^{-1} \tag{3.12}$$

and the theorem follows. Also, T is orthogonal and $\|\Delta T\| = \mathbf{O}(n^2u)$, $\|\Delta A\|/\|A\| = \mathbf{O}(n^3u)$, $\|\Delta\mathbf{b}\|/\|\mathbf{b}\| = \mathbf{O}(n^2u)$. Here we have used $\|A^{(0)}\|/\|A\| = 1 + \mathbf{O}(u)$. Also $(P + \Delta P)^{-T}P^T = I + \Delta V$, where $\Delta V = -P(\Delta P)^T[I + P(\Delta P)^T]^{-1}$ and $P(P + \Delta P)^T = I + \Delta W$ with $\|\Delta V\|$, $\|\Delta W\| \leqq (n^2 + 18n - 19)du$. □

REFERENCES

[B] S. Barnett, *Matrices, polynomials, and linear time-invariant systems*, IEEE Trans. Automat. Control, 18 (1973), pp. 1–10.

[DB] G. Dahlquist and A. Björck, *Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[D] P. M. Van Dooren, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 111–129.

[GV] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[K] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[MP] G. S. Miminis and C. C. Paige, *An algorithm for pole assignment of time invariant linear systems*, Internat. J. Control, 35 (1982), pp. 341–354.

[P] C. C. Paige, *Properties of numerical algorithms related to computing controllability*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 130–138.

[PCK] P. H. R. Petkov, N. D. Christov, and M. M. Konstantinov, *A computational algorithm for pole assignment of linear single-input systems*, IEEE Trans. Automat. Control, AC–29 (1984), pp. 1045–1048.

[R] D. L. Russell, *Mathematics of Finite-Dimensional Control Systems*, Marcel Dekker, New York 1979.

[V] A. Varga, *A Schur method for pole assignment*, IEEE Trans. Automat. Control, AC–26 (1981), pp. 517–519.

[W] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 4, pp. 457–480, October 1989

004

THE COMPUTATION OF GENERALIZED CROSS-VALIDATION FUNCTIONS THROUGH HOUSEHOLDER TRIDIAGONALIZATION WITH APPLICATIONS TO THE FITTING OF INTERACTION SPLINE MODELS*

CHONG GU†, DOUGLAS M. BATES†, ZEHUA CHEN†, AND GRACE WAHBA†

Abstract. An efficient algorithm for computing the GCV (generalized cross-validation) function for the general cross-validated regularization/smoothing problem is provided. This algorithm is based on the Householder tridiagonalization, similar to Elden's [*BIT*, 24 (1984), pp. 467–472] bidiagonalization and is appropriate for problems where no natural structure is available, and the regularization/smoothing problem is solved (exactly) in a reproducing kernel Hilbert space. It is particularly appropriate for certain multivariate smoothing problems with irregularly spaced data, and certain remote sensing problems, such as those that occur in meteorology, where the sensors are arranged irregularly.

The algorithm is applied to the fitting of interaction spline models with irregularly spaced data and two smoothing parameters, and favorable timing results are presented. The algorithm may be extended to the computation of certain GML (generalized maximum likelihood) functions. Application of the GML algorithm to a problem in numerical weather forecasting, and to a broad class of hypothesis testing problems, is noted.

Key words. computation of GCV functions, interaction splines, Householder tridiagonalization, distributed truncation

AMS(MOS) subject classifications. 41A13, 41A63, 41A65, 60G60, 62J07, 65F20, 65U05

1. Introduction. Generalized cross validation is generally recognized to be an effective method of automatically choosing smoothing parameters in various regularization problems. Applications have been found in remote sensing problems, ridge regression, univariate and multivariate smoothing spline regression, partial spline models, penalized GLIM models, penalized likelihood estimation, penalized log-density and log-hazard estimation, etc. (see for example, [16], [11], [12], [23], [25], [41], [34], [27], [26], [28]). Some of the theoretical properties of this method are well known (see [32], [30], [29], [21], [9], [24]).

The general regularization problem we consider can be written as follows. The data y_i are modeled by

$$y_i = L_i f + \varepsilon_i, \qquad i = 1, \cdots, n$$

where $f \in H$, some Hilbert space with the property that L_i's are bounded linear functionals, and ε_i's are uncorrelated zero mean random errors with common variance. The method of regularization seeks f_λ in H to minimize

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - L_i f)^2 + \lambda \| P_1 f \|_H^2 \tag{1.1}$$

where P_1 is the orthogonal projection in H onto a subspace H_1 of codimension M, with $M \ll n$. Smoothing problems are included if the L_i's are point evaluations, and remote sensing problems can be modeled by choosing L_i's as integrals. For example, if $S = [0, 1]$ and $H = W_2^m[0, 1]$, we may take $\|P_1 f\|^2 = \int_0^1 (f^{(m)}(s))^2 \, ds$.

* Received by the editors May 19, 1988; accepted for publication (in revised form) January 5, 1989. This research was partially supported by the Air Force Office of Scientific Research under grant AFOSR-87-0171, the National Science Foundation under grant ATM-8410373, the National Aeronautics and Space Administration under contract NAG5-316, and the Wisconsin Alumni Research Foundation.

† Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin 53706 (wahba@stat.wisc.edu.).

The generalized cross validation (GCV) estimate $\hat{\lambda}$ for λ is the minimizer of $V(\lambda)$ given by

$$V(\lambda)=\frac{(1/n)\|(\mathbf{I}-\mathbf{A}(\lambda))\mathbf{y}\|^2}{[(1/n)\operatorname{tr}(\mathbf{I}-\mathbf{A}(\lambda))]^2} \tag{1.2}$$

where $\mathbf{A}(\lambda)$ is the $n \times n$ matrix satisfying

$$\begin{pmatrix} L_1 f_\lambda \\ \vdots \\ L_n f_\lambda \end{pmatrix}=\mathbf{A}(\lambda)\mathbf{y}.$$

The problem we are concerned with is efficiently computing the minimizer of $V(\lambda)$ for large n.

In many examples (see, e.g., [35], [25], [26]), it is convenient to approximate the minimizer f_λ of (1.1) by a linear combination of some basis functions $\{B_l\}_{l=1}^p$, and to minimize (1.1) in the span of $\{B_l\}_{l=1}^p$. Generally, to avoid losing information at this stage, p should be chosen to be fairly large. Letting

$$f_\lambda=\sum_{l=1}^{p} c_l B_l,$$

the problem becomes the following. Find $\mathbf{c}=(c_1, \cdots, c_p)^T$ to minimize

$$\frac{1}{n}\|\mathbf{y}-\tilde{\mathbf{X}}\mathbf{c}\|^2+\lambda\mathbf{c}^T\mathbf{J}\mathbf{c} \tag{1.3}$$

where $\tilde{X}_{il}=(L_i B_l)$ and $J_{ij}=\langle P_1 B_i, P_1 B_j\rangle_H$, where $\langle \cdot, \cdot\rangle_H$ is the inner product in H. If $H=W_2^m$, then B-splines are a popular choice for the B_l's. Let $\mathbf{c}_\lambda$ be the minimizer of (1.3), and $\mathbf{L}$ be the Cholesky factor of $\mathbf{J}$ (so $\mathbf{L}^T\mathbf{L}=\mathbf{J}$) and $\mathbf{u}_\lambda=\mathbf{L}\mathbf{c}_\lambda$ while $\mathbf{X}=\tilde{\mathbf{X}}\mathbf{L}^{-1}$; then (1.3) becomes: minimize

$$\frac{1}{n}\|\mathbf{y}-\mathbf{X}\mathbf{u}\|^2+\lambda\mathbf{u}^T\mathbf{u},$$

which gives

$$\mathbf{u}_\lambda=(\mathbf{X}^T\mathbf{X}+n\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y},$$

and

$$\mathbf{A}(\lambda)=\mathbf{X}(\mathbf{X}^T\mathbf{X}+n\lambda\mathbf{I})^{-1}\mathbf{X}^T=\mathbf{I}-n\lambda(\mathbf{X}\mathbf{X}^T+n\lambda\mathbf{I})^{-1}. \tag{1.4}$$

We see that this reduces to the standard ridge regression formulation (see [16]).

In the ridge regression case, when $\mathbf{A}(\lambda)$ is of the form of (1.4), there are at least four different strategies that we know of for minimizing $V(\lambda)$, which we will discuss in § 2. In this paper, we will be mainly concerned with the efficient minimization of $V(\lambda)$ when the variational problem (1.1) is solved directly in H rather than in span $\{B_l\}$. The minimizer of (1.1) is known to be in a particular $(n+M)$-dimensional subspace of H, and the form of $\mathbf{A}(\lambda)$ will be slightly different. In some applications (we will give a concrete example below), it is more appropriate to solve directly in H. Frequently, these applications involve H as a space of functions of several variables, where the data functionals exhibit irregular patterns, and good approximating subspaces of reasonable dimension much less than $n+M$ are not readily apparent. The form of $\mathbf{I}-\mathbf{A}(\lambda)$ in this case is known (see [20], [33], [36]). To establish notation, we briefly sketch the derivation of $\mathbf{I}-\mathbf{A}(\lambda)$ in this case.

From [20] (see also [33]), the exact minimizer of (1.1) may be written

$$f_\lambda = \sum_{i=1}^{n} c_i \xi_i + \sum_{\nu=1}^{M} d_\nu \phi_\nu \tag{1.5}$$

where $\xi_i = P_1 \eta_i$, η_i being the representer of L_i in H, $L_i f = \langle \eta_i, f \rangle_H$; the $\{\phi_\nu\}_{\nu=1}^{M}$ span the null space of $\|P_1 f\|_H^2$ and $\mathbf{c} = (c_1, \cdots, c_n)^T$ and $\mathbf{d} = (d_1, \cdots, d_M)^T$ satisfy

$$(\tilde{\Sigma} + n\lambda \mathbf{I})\mathbf{c} + \mathbf{S}\mathbf{d} = \mathbf{y}, \qquad \mathbf{S}^T\mathbf{c} = \mathbf{0} \tag{1.6}$$

where

$$\tilde{\Sigma}_{ij} = (\langle \xi_i, \xi_j \rangle) \tag{1.7}$$

and

$$\mathbf{S}_{n \times M} = (L_i \phi_\nu). \tag{1.8}$$

We remark that if H_1 has a reproducing kernel $Q(s,t), s, t \in S$, then $\xi_i(s) = L_{i(t)}Q(s, t)$, where $L_{i(t)}$ means L_i applied to what follows as a function of t, and $\langle \xi_i, \xi_j \rangle = L_{i(s)}L_{j(t)}Q(s, t)$.

Letting the QR decomposition of S be

$$\mathbf{S} = (\mathbf{F}_1 \mathbf{F}_2)\begin{pmatrix} \mathbf{R} \\ \mathbf{O} \end{pmatrix},$$

a series of standard calculations (see, e.g., [36]), we obtain

$$\mathbf{I} - \mathbf{A}(\lambda) = n\lambda \mathbf{F}_2(\mathbf{F}_2^T \tilde{\Sigma} \mathbf{F}_2 + n\lambda \mathbf{I})^{-1} \mathbf{F}_2^T$$

and if we let $\Sigma = \mathbf{F}_2^T \tilde{\Sigma} \mathbf{F}_2$ and $\mathbf{z} = \mathbf{F}_2^T \mathbf{y}$, we have

$$\mathbf{V}(\lambda) = \frac{(1/n)\mathbf{z}^T(\Sigma + n\lambda \mathbf{I})^{-2}\mathbf{z}}{[(1/n)\operatorname{tr}(\Sigma + n\lambda \mathbf{I})^{-1}]^2}. \tag{1.9}$$

The present work centers on strategies for minimizing $V(\lambda)$ of the form (1.9), while existing results in the ridge regression case are based on a form similar to (1.9), with Σ replaced by $\mathbf{X}^T\mathbf{X}$.

Some strategies for the GCV computation are readily applicable to the minimization of the generalized maximum likelihood (GML) function

$$M(\lambda) = \frac{(1/n)\mathbf{z}^T(\Sigma + n\lambda \mathbf{I})^{-1}\mathbf{z}}{\det(\Sigma + n\lambda \mathbf{I})^{-1/(n-M)}}. \tag{1.10}$$

The GML estimates are useful in a certain general class of hypothesis testing problems ([10], [38]), and when the Bayesian model behind the estimate f_λ is true. GML is not recommended for the general regularization problems since it is not robust to deviations from the exact Bayes model (see [36]).

In § 2, we will review some of the computational strategies in the ridge regression case. In § 3 we will present an algorithm based on the Householder tridiagonalization of the matrix Σ for the efficient minimization of $V(\lambda)$ and $M(\lambda)$ of (1.9) and (1.10) when Σ is not sparse. A distributed truncation strategy is proposed in § 4 that may speed up the tridiagonalization process when Σ is rank-deficient. In § 5, we will illustrate our algorithm using an interaction spline model (see [2], [3], [37], [8]) as an example. In this example more than one smoothing parameter appears, making efficient computation even more crucial. At the end of the paper, we offer some remarks on other applications and further studies in § 6.

2. A brief review of existing results for ridge regression. We briefly describe some of the existing algorithms for minimizing

$$V(\lambda)=\frac{(1/n)\|(\mathbf{I}-\mathbf{A}(\lambda))\mathbf{y}\|^2}{[(1/n)\operatorname{tr}(\mathbf{I}-\mathbf{A}(\lambda))]^2}, \tag{2.1}$$

in the case

$$\mathbf{A}(\lambda)=\mathbf{X}(\mathbf{X}^T\mathbf{X}+n\lambda\mathbf{I})^{-1}\mathbf{X}^T \tag{2.2}$$

where

$$\operatorname{tr}\mathbf{A}(\lambda)=p-n\lambda\operatorname{tr}(\mathbf{X}^T\mathbf{X}+n\lambda\mathbf{I})^{-1},$$

and **X** is the so-called design matrix of size $n\times p$.

To our knowledge, basically four different strategies have been proposed. We list them as follows.

- *Singular value decomposition approach* (SVD). *Do the singular value decomposition of* $\mathbf{X}=\mathbf{UDV}$, *where* **U**, **V** *are orthogonal and* **D** *is diagonal. Then* $V(\lambda)$ *can be represented as a rational function of* λ. *Do a grid search to find the minimizing* λ. This approach was originally proposed-by Golub, Heath, and Wahba [16], and has been refined by Bates and Wahba [5], and implemented by Bates et al. [6] in GCVPACK.

- *Cholesky decomposition approach* (CD). *For each trial value of* λ, *do a Cholesky decomposition of* $n\lambda\mathbf{I}+\mathbf{X}^T\mathbf{X}=\mathbf{C}^T\mathbf{C}$. *Compute the numerator of the* GCV *function from the Cholesky factor* **C**, *and compute the denominator from its inverse* $\mathbf{C}^{-1}$. *Do a grid search or a golden-section search to find the minimizing* λ. A very efficient algorithm has been developed by Hutchinson and deHoog [19] following this approach, making use of the special band structure in **X** that is available in some applications like one-dimensional smoothing spline regression.

- *Bidiagonalization approach* (BD). *Do the Householder bidiagonalization in the first part of the singular value decomposition algorithm, but do not iterate to the final diagonal form. Evaluate the* GCV *function for each trial value of* λ *from the bidiagonal matrix* $\mathbf{B}=\mathbf{U}^T\mathbf{X}\mathbf{V}^T$, *where* **U** *and* **V** *are orthogonal.* This approach was proposed by Elden in [14].

- *Monte-Carlo approximation.* In some very large sparse linear systems like those that appear in image processing and tomography problems, the solutions are usually obtained through iterative methods, hence the trace term in the denominator of the GCV function is very difficult to evaluate. In an unique work by Girard [15], however, a nice simple Monte-Carlo approximation to the trace term is provided. In this method, *for each trial value of* λ, *the regularized least square system is solved by some iterative methods and automatically we will get the numerator of the* GCV *function. Then passing a pure noise vector as data through the same iteration steps will result in a Monte-Carlo approximation to the denominator of the* GCV *function. The approximate* GCV *values are then compared over trial values to find the minimizing* λ.

2.1. Comparison of the basic strategies. Now we briefly discuss the pros and cons of the existing algorithms. Our main concern is the *speed* of the algorithms. To be specific in later discussions, we call the Cholesky decomposition approach described above the *direct Cholesky decomposition* approach (DCD henceforth). It is worth keeping in mind that in most applications of the generalized cross validation technique, *p is often of the same order as n.*

The SVD approach explores the singular value structure that makes possible the explicit expression of $V(\lambda)$ as a rational function of λ. The explicit expression is useful for theoretical analysis, and the singular values are useful in model diagnostics. For the sole purpose of computing the GCV estimates of λ, however, the explicit diagonal structure is *not* necessary. At the same time, the singular values are very expensive, as observed by many numerical experimenters (see, e.g., [6]).

One alternative to the SVD approach is to evaluate GCV function at each trial value of λ directly from (2.1) and (2.2). The standard technique is through the Cholesky decomposition of $n\lambda\mathbf{I}+\mathbf{X}^T\mathbf{X}=\mathbf{C}^T\mathbf{C}$ to compute the numerator of (2.1), and to compute the denominator from $\mathbf{C}^{-1}$, which requires about the same amount of calculation as the Cholesky decomposition. When the design matrix $\mathbf{X}$ is *banded*, each evaluation can be done in $O(p)$ flops, thus this approach is very efficient (see, e.g., [19]). See also [28] for its use in nonlinear settings. When $\mathbf{X}$ is a general matrix, though, each evaluation will take $(1/3)p^3$ flops (for both $\mathbf{C}$ and $\mathbf{C}^{-1}$), so the grid search of DCD in general cases is very expensive.

In understanding the SVD and DCD approaches, we see two extreme trends in their strategies. SVD performs an expensive one-time structural exploration on the design matrix $\mathbf{X}$ before doing the grid search on λ; the cost for the grid search that follows is then negligible. DCD directly does the grid search based on the raw design matrix; its performance is then determined largely by the property of the raw design matrix. Elden's BD approach [14] provides a nice middle strategy for the general dense matrix $\mathbf{X}$. Instead of exploring the unnecessary expensive singular value structure, BD stops at the less expensive bidiagonal form. It is obvious that the bidiagonal structure provides the band structure that makes the evaluation of GCV function very efficient through Cholesky decomposition. (However, Elden's original treatment of $\mathbf{B}$ is not through CD.) BD is thought to be superior to SVD.

It must be emphasized that the above comparison is solely based on the leading terms of the flop counts. For small and moderately sized problems, the other terms can be as big as the leading term, and thus affect the relative performance of the algorithms.

The above three approaches are all based on suitable matrix decompositions. However, in some very large sparse ill-posed linear systems, the regularized solutions to the systems are usually obtained through iterative methods such as conjugate gradient methods or successive overrelaxation (SOR) methods, without explicit decomposition of the corresponding matrices, hence the trace of the influence matrix $\mathbf{A}(\lambda)$ cannot be obtained through the above methods or any variants. Girard's ([15]) Monte-Carlo approximation to the trace term is the only method yet known to apply the GCV technique of choosing smoothing parameters to such systems. These systems are common in image processing problems and computerized tomography (CT) problems. On the other hand, in problems where the direct matrix decomposition approach is appropriate, the Monte-Carlo approximation is unnecessary.

3. Proposed algorithm. To evaluate (1.5), where Σ, or equivalently $\mathbf{X}^T\mathbf{X}$, is directly available, we propose the following algorithm for computing generalized cross validation and maximum likelihood estimates of λ.

ALGORITHM 1.

Step a. Tridiagonalize Σ by

$$\mathbf{U}^T\Sigma\mathbf{U}=\mathbf{T}$$

where $\mathbf{U}$ is orthogonal and $\mathbf{T}$ is tridiagonal.

Step b. Form $\mathbf{x}=\mathbf{U}^T\mathbf{z}$. The GCV and likelihood functions become

$$V(\lambda)=\frac{(1/n)\mathbf{x}^T(n\lambda\mathbf{I}+\mathbf{T})^{-2}\mathbf{x}}{[(1/n)\operatorname{tr}(n\lambda\mathbf{I}+\mathbf{T})^{-1}]^2} \tag{3.1}$$

and

$$M(\lambda)=\frac{(1/n)\mathbf{x}^T(n\lambda\mathbf{I}+\mathbf{T})^{-1}\mathbf{x}}{\det(n\lambda\mathbf{I}+\mathbf{T})^{-1/(n-M)}}. \tag{3.2}$$

Step c. For each trial value of λ, do a Cholesky decomposition of $(n\lambda\mathbf{I}+\mathbf{T})$. Compute the GCV and likelihood functions from the Cholesky factors. Do a grid search on λ to find the minimum.

Step a can be done by successively applying the Householder transformation, taking about $(2/3)n^3$ flops. $\mathbf{U}$ can be stored in factored form in the strict lower triangle of Σ (see [17, pp. 276–277]). A strategy for speeding up this step by appropriate truncation will be presented in § 4.

Step b can be done with n^2 flops, using the LINPACK routine *dqrsl* in an appropriate way, from the factored $\mathbf{U}$. See [13, Chap. 9].

For step c, we can find routines for the Cholesky decomposition of banded positive definite matrix and related routines from LINPACK [13, Chap. 4]. The only tricky part in evaluating (3.1) and (3.2) is to compute the denominator of (3.2), for which we will use the formula presented by Elden [14] as illustrated later in this section. The overall operations needed for the evaluations are of order $O(n)$. In most cases, the likelihood function and the GCV function are unimodal as λ varies, so the golden-section search might be used for a grid search on λ. Generally speaking, 20–30 evaluations will suffice in most applications.

3.1. Compute the denominator of the GCV function. Given the Cholesky decomposition $(n\lambda\mathbf{I}+\mathbf{T})=\mathbf{C}^T\mathbf{C}$, where

$$\mathbf{C}=\begin{bmatrix} a_1 & b_1 & & & \\ & a_2 & b_2 & & \\ & & \ddots & \ddots & \\ & & & a_{n-M-1} & b_{n-M-1} \\ & & & & a_{n-M} \end{bmatrix}$$

is upper bidiagonal. The calculation of the denominator of the GCV function is derived by Elden in [14], as illustrated below.

We need to calculate $\operatorname{tr}(\mathbf{C}^{-1}\mathbf{C}^{-T})$. Denote the ith row of $\mathbf{C}^{-1}$ by $\mathbf{c}_i^T$. We have $\operatorname{tr}(\mathbf{C}^{-1}\mathbf{C}^{-T})=\sum\|\mathbf{c}_i\|^2$. From

$$\mathbf{C}^{-T}\mathbf{C}^{-T}=(\mathbf{c}_1,\mathbf{c}_2,\cdots,\mathbf{c}_{n-M})\begin{bmatrix} a_1 & & & \\ b_1 & a_2 & & \\ & b_2 & \ddots & \\ & & \ddots & a_{n-M-1} & \\ & & & b_{n-M-1} & a_{n-M-1} \end{bmatrix}=\mathbf{I}$$

we have

$$a_{n-M}\mathbf{c}_{n-M}=\mathbf{e}_{n-M}, \qquad a_i\mathbf{c}_i=\mathbf{e}_i-b_i\mathbf{c}_{i+1}, \quad i=n-M-1,\cdots,1$$

where $\mathbf{e}_i$'s are unit vectors. Because $\mathbf{C}^{-T}$ is lower triangular, $\mathbf{c}_{i+1}$ is orthogonal to $\mathbf{e}_i$. Thus we have the recursive formula

$$\|\mathbf{c}_{n-M}\|^2=a_{n-M}^{-2},$$

$$\|\mathbf{c}_i\|^2=(1+b_i^2\|\mathbf{c}_{i+1}\|^2)a_i^{-2}, \qquad i=n-M-1,\cdots,1,$$

which can be calculated in $O(n)$ flops.

3.2. Comparison with other algorithms. Other possible algorithms applicable to our setting are the DCD approach, and the *eigenvalue decomposition* approach, which is the direct analogue of the SVD approach in the ridge regression setting.

In modern literature on numerical linear algebra, the recommended algorithm for the eigenvalue decomposition of a symmetric matrix is through Householder tridiagonalization followed by symmetric QR iterations (see [17, Chap. 8]). Similar to BD versus SVD, our *tridiagonalization* approach (TD henceforth) stops before getting into the iterative step, but takes in the extra complexity of evaluating the GCV and likelihood functions from the tridiagonal matrix instead of from the diagonal form. This can also be viewed as introducing the band structure through tridiagonalization, which makes the grid search based on Cholesky decomposition very efficient, but avoids further exploring the eigenstructure that is not necessary for the purpose of evaluating the GCV function.

To compare TD with DCD, we must distinguish between the likelihood function and the GCV function. For the likelihood function, the operational cost for each grid evaluation is mostly in the full matrix Cholesky decomposition, which is about $(1/6)n^3$ flops. See [13, Chap. 8]. On the other hand, the cost for TD is mostly in the Householder tridiagonalization, which takes $(2/3)n^3$ flops, see [17, p. 277]. Thus the TD approach is roughly equivalent to four grid evaluations of the likelihood function in computational cost. When DCD is used to evaluate the GCV function, however, we need to invert the Cholesky factor to compute the trace of the inverse matrix. This needs an extra $(1/6)n^3$ flops (see [13, Chaps. 3, 6]). Hence the TD approach takes twice the operations needed for one grid evaluation of the DCD approach for GCV computation. Remember that the Cholesky decomposition might be the least expensive approach for solving the system when λ is known, so it provides a reasonable baseline for comparison.

Our method is in the spirit of Elden's BD approach. It also suggests an alternative method for handling the bidiagonal form after the bidiagonalization in BD. In his original proposal, Elden [14] applies a sequence of Givens rotations on the bidiagonal matrix $\mathbf{B}$ to compute the numerator of GCV function for each λ grid. Instead, we can form a tridiagonal matrix $\mathbf{T} = \mathbf{B}^T\mathbf{B}$, and use our step c to evaluate the GCV function at various trial values of λ based on the tridiagonal form $\mathbf{T}$. Both Elden's method and our step c are of linear order, although our method can easily be implemented by calling the appropriate LINPACK routines.

4. Householder tridiagonalization and truncation. In this section, we will present a *distributed truncation* strategy in the implementation of the Householder tridiagonalization algorithm described in [17, pp. 276–277]. The truncation is controlled by the Wielandt–Hoffman theorem, through a lemma proved at the end of the section. To make the discussion self-contained, we first review the Householder tridiagonalization algorithm.

4.1. The Householder tridiagonalization. Given a nonzero vector $\mathbf{b}$ of dimension l, the Householder matrix

$$\mathbf{H} = \mathbf{I} - \frac{2}{\mathbf{v}^T\mathbf{v}}\mathbf{v}\mathbf{v}^T \tag{4.1}$$

where $\mathbf{v} = \mathbf{b} \pm \|\mathbf{b}\|\mathbf{e}_1$, is an orthogonal projection matrix projecting $\mathbf{b}$ onto the linear space spanned by $\mathbf{e}_1$. That is, $\mathbf{H}$ zeros all components of $\mathbf{b}$ except the first one. If $\mathbf{b}$ is of unit length, the Householder matrix (4.1) can be written as

$$\mathbf{H} = \mathbf{I} - \frac{\mathbf{v}\mathbf{v}^T}{\nu_1} \tag{4.2}$$

where $\mathbf{v} = \mathbf{e}_1 \pm \mathbf{b}$ and $\nu_1 = 1 \pm b_1$ (see also [13, Chap. 9]).

For any symmetric matrix $\mathbf{A}$ of size $n \times n$, let

$$\mathbf{b}_1 = (a_{12}, \cdots, a_{1n})^T.$$

If $\mathbf{b}_1 \neq \mathbf{0}$, forming the Householder matrix $\mathbf{H}_1$ from $\mathbf{b}_1$ through (4.1), we have

$$\operatorname{diag}(1, \mathbf{H}_1)\mathbf{A} \operatorname{diag}(1, \mathbf{H}_1) = \begin{bmatrix} a_{11} & \tilde{\mathbf{b}}_1^T \\ \tilde{\mathbf{b}}_1 & \mathbf{D}_1 \end{bmatrix}$$

where $\tilde{\mathbf{b}}_1$ has all components 0 except the first and $\mathbf{D}_1$ is a symmetric matrix of size $(n-1) \times (n-1)$. We can write $\mathbf{A}$ as $\mathbf{D}_0$, and it is easy to see that, by applying the procedure described above successively to $\mathbf{D}_k$'s, we will reduce the matrix to tridiagonal form. Actually, before step k, we have

$$\mathbf{A}_{k-1} = (\mathbf{P}_1 \cdots \mathbf{P}_{k-1})^T \mathbf{A} (\mathbf{P}_1 \cdots \mathbf{P}_{k-1}) = \begin{bmatrix} \mathbf{T}_{k-1} & & \mathbf{O} \\ & & \mathbf{b}_k^T \\ \mathbf{O} & \mathbf{b}_k & \mathbf{D}_{k-1} \end{bmatrix}$$

where $\mathbf{T}_{k-1}$ is $k \times k$ tridiagonal, $\mathbf{D}_k$ is an $(n-k) \times (n-k)$ general symmetric matrix, call it *tail*, and $\mathbf{b}_k$ is an $n-k$ vector. If $\mathbf{b}_k \neq \mathbf{0}$, we apply an $(n-k) \times (n-k)$ Householder matrix from both sides to set all but the first components of $\mathbf{b}_k$ to zero. If $\mathbf{b}_k = \mathbf{0}$, we call it a *diagonal separation*, and skip the step. The matrices $\mathbf{P}_k$ are generally in the form

$$\mathbf{P}_k = \operatorname{diag}(\mathbf{I}_k, \mathbf{H}_k)$$

where $\mathbf{I}_k$ is the identity matrix of size $k \times k$, and $\mathbf{H}_k$ is a Householder matrix of size $(n-k) \times (n-k)$. If step k is skipped due to a diagonal separation, then $\mathbf{P}_k = \mathbf{I}_n$. At an unskipped step k, the operational count is about $2(n-k)^2$ flops.

4.2. Implementation. In the implementation of the Householder tridiagonalization algorithm, our main task is to update the tail $\mathbf{D}_{k-1}$ by $\mathbf{D}_{k*} = \mathbf{H}_k \mathbf{D}_{k-1} \mathbf{H}_k$. We first standardize the $\mathbf{b}_k$'s at each step, and use (4.2) to compute the Householder matrix $\mathbf{H}_k$. We drop all subscripts since no confusion is caused, and assume $\mathbf{b}$ is prestandardized to have length one. It can be shown as in [17, p. 277] that

$$\mathbf{D}_* = \mathbf{HDH} = \mathbf{D} - \mathbf{v}\mathbf{w}^T - \mathbf{w}\mathbf{v}^T$$

where $\mathbf{v}$ is as in (4.2), and

$$\mathbf{w} = \mathbf{p} - \frac{\mathbf{p}^T \mathbf{v}}{2\nu_1} \mathbf{v}$$

where

$$\mathbf{p} = \frac{\mathbf{D}\mathbf{v}}{\nu_1}.$$

An important aspect of the implementation is the appropriate criterion for setting the diagonal separation, which is the main issue of this section. Diagonal separations at early steps may speed up the process considerably.

4.3. The Wielandt-Hoffman theorem. We include here the Wielandt-Hoffman theorem [17, p. 270] that leads to the truncation scheme discussed later.

THEOREM 1 (Wielandt-Hoffman). *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be* $n \times n$ *symmetric matrices, having ordered eigenvalues* $\{d_i\}$, $\{s_i\}$, *respectively. Then*

$$\sum_{i=1}^{n} (d_i - s_i)^2 \leqq \operatorname{tr}[(\mathbf{X}-\mathbf{Y})^T(\mathbf{X}-\mathbf{Y})] = \|\mathbf{X}-\mathbf{Y}\|_F^2.$$

The theorem implies that the difference of the eigenstructures of two matrices is controlled by the Frobenius-norm of their difference.

4.4. Distributed truncation during tridiagonalization. We now describe a distributed truncation strategy for speeding up the Householder tridiagonalization algorithm. The strategy skips the Householder transform whenever appropriate by truncating the small $\mathbf{b}$'s to zero. Remember that at step-k we need about $2(n-k)^2$ flops to perform the Householder transform. We propose the following algorithm.

ALGORITHM 2 (Distributed Truncation in Householder Tridiagonalization).

(1) *Initialization*: Given the tolerance ε, compute $u=\varepsilon/C$, where $C=\sum_{k=1}^{n-2}(n-k)^2=[n(n-1)(2n-1)/6]-1$. Set $\tau=0$.

(2) *Tridiagonalization*: For $k=1,\cdots,n-2$

  (a) Set $\tau=\tau+(n-k)^2u$.

  (b) If $2\,\|\mathbf{b}_k\|^2\leqq\tau$, then set $\tau=\tau-2\|\mathbf{b}_k\|^2$, set $\mathbf{b}_k=0$, skip the Householder transform. Otherwise, perform the Householder transform as usual.

As the justification of the truncation strategy, we have Lemma 1.

LEMMA 1. *For the above truncation strategy, denote the matrix* $\mathbf{A}_*$ *as the matrix restored from the final tridiagonal form by reversing the Householder transforms applied. Then* $\|\mathbf{A}-\mathbf{A}_*\|_F^2<\varepsilon$, *where* ε *is the prespecified tolerance.*

Usually the tolerance ε is made to be a small proportion of the total square norm, $\delta\ \|\mathbf{A}\|_F^2$, say, where δ is specified by the user. If great precision is desired, δ will be set to the square of the machine precision. The proof of the lemma is given at the end of the section. This lemma assures that the truncation described above is under our control in the sense of the Wielandt–Hoffman theorem. It is easy to see that we *will* have diagonal separation when the matrix $\mathbf{A}$ is rank-deficient, as assured by the following lemma.

LEMMA 2. *For any* $k\times k$ *tridiagonal matrix* $\mathbf{T}$, *if all of its off-diagonal elements are nonzero, then* $\operatorname{rank}(\mathbf{T})\geqq k-1$.

The validity of Lemma 2 can easily be seen by evaluating the determinant of the left-bottom corner $(k-1)\times(k-1)$ submatrix. Each application of the Householder transform adds one to the rank of the tridiagonal part $\mathbf{T}$. So we have Theorem 2.

THEOREM 2. *For an* $n\times n$ *symmetric matrix* $\mathbf{A}$ *with* $\operatorname{rank}(\mathbf{A})=k$, *the tridiagonalization algorithm will perform the Householder transformation at most k times. The operations involved are at most* $(2/3)[n^3-(n-k)^3]$ *flops.*

It is observed that we may have quite a large number of diagonal separations even when the matrix $\mathbf{A}$ is of computationally full rank. We should also truncate the tail when enough norm is accumulated at the upper-left corner, to make the strategy complete.

4.5. Proof of Lemma 1. First we assume we only truncate twice. Denote the product of the Householder matrices up to the first truncation as $\mathbf{U}_1$, and the product of Householder matrices between the first truncation and the second truncation as $\mathbf{U}_2$. Denote $\mathbf{A}_1$ as the matrix restored after the first truncation, and $\mathbf{A}_2$ as $\mathbf{A}_*$ defined in the lemma. We have

$$\|\mathbf{A}-\mathbf{A}_2\|_F^2=\|\mathbf{U}_1(\mathbf{A}-\mathbf{A}_1)\mathbf{U}_1^T+\mathbf{U}_1(\mathbf{A}_1-\mathbf{A}_2)\mathbf{U}_1^{\mathrm{T}}\|_F^2, \tag{4.3}$$

since

$$\mathbf{U}_1(\mathbf{A}-\mathbf{A}_1)\mathbf{U}_1^T=\left[\begin{array}{ccc} & & \mathbf{O} \\ & \mathbf{T}_1 & \mathbf{b}_1^T \\ \mathbf{O} & \mathbf{b}_1 & \mathbf{D}_1 \end{array}\right]-\begin{bmatrix}\mathbf{T}_1 & \mathbf{O}\\ \mathbf{O} & \mathbf{D}_1\end{bmatrix}=\left[\begin{array}{ccc} & & \mathbf{O} \\ & \mathbf{O} & \mathbf{b}_1^T \\ \mathbf{O} & \mathbf{b}_1 & \mathbf{O} \end{array}\right]$$

and

$$\mathbf{U}_1(\mathbf{A}_1-\mathbf{A}_2)\mathbf{U}_1^T=\begin{bmatrix}\mathbf{T}_1 & \mathbf{O}\\ \mathbf{O} & \mathbf{D}_1\end{bmatrix}-\begin{bmatrix}\mathbf{T}_1 & \mathbf{O}\\ \mathbf{O} & \mathbf{D}_{1*}\end{bmatrix}=\begin{bmatrix}\mathbf{O} & \mathbf{O}\\ \mathbf{O} & \mathbf{D}_1-\mathbf{D}_{1*}\end{bmatrix}$$

where $\mathbf{D}_{1*}$ is the submatrix restored after the second truncation at this corner. Thus the two matrices in the right-hand side of (4.3), treated as two big vectors, are orthogonal. The square Frobenius norm of part one is $2\|\mathbf{b}_1\|^2$. Note that $\mathbf{U}_2 = \operatorname{diag}(\mathbf{I}, \bar{\mathbf{U}}_2)$, where $\bar{\mathbf{U}}_2$ matches $\mathbf{D}_1$ in size. We can do the same process with $\mathbf{D}_1 - \mathbf{D}_{1*}$, as we do with $\mathbf{A} - \mathbf{A}_1$. So the square norm of the second part is $2\|\mathbf{b}_2\|^2$, where $\mathbf{b}_2$ is the second truncated $\mathbf{b}$. It is easy to continue this procedure by induction in the general case where we have more than two diagonal separations.

5. Application to interaction spline smoothing with multiple smoothing parameters.

5.1. Interaction splines. Let $H = H_0 \oplus H_1$ be a Hilbert space where H_0 is of dimension M and H_1 is the direct sum of p orthogonal subspaces $H^1, \cdots, H^p$,

$$H_1=\sum_{\beta=1}^{p}\oplus H^\beta \tag{5.1}$$

where H^β are orthogonal subspaces, and suppose we wish to find $f \in H$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-L_if)^2+\lambda\sum_{\beta=1}^{p}\theta_\beta^{-1}\|P^\beta f\|^2 \tag{5.2}$$

where $\theta_1 = 1$ and P^β is the orthogonal projector in H onto H^β. By replacing the square norm $\|P_1 f\|^2 = \sum_{\beta=1}^{p}\|P^\beta f\|^2$ in (1.1) with $\sum_{\beta=1}^{p}\theta_\beta^{-1}\|P^\beta f\|^2$ it can easily be seen that the solution to the problem (5.2) is of the form (1.5) with

$$\xi_i=\sum_\beta P^\beta\xi_i$$

replaced by

$$\xi_i^\theta=\sum_\beta \theta_\beta P^\beta\xi_i$$

and that $\tilde{\Sigma}$ in (1.6) is of the form

$$\tilde{\Sigma}=\tilde{\Sigma}_1+\theta_2\tilde{\Sigma}_2+\cdots+\theta_p\tilde{\Sigma}_p$$

where the ijth entry of $\tilde{\Sigma}_\beta$ is $\langle P^\beta\xi_i, P^\beta\xi_j\rangle_H$.

The additive spline models and their generalizations, the interaction spline models, fall into this framework. The additive spline models have become popular in the analysis of medical data (see [18], [7], and references cited therein). The interaction spline models have been discussed by Barry [2], [3], Wahba [37], and Chen [8]. These models, which in a sense generalize analysis of variance to function spaces, have strong potential for the empirical modeling of responses to economic and medical varibles, given large data sets of responses with several independent variables, and represent a major advance over the usual parametric (mostly linear) models. We have chosen a relatively simple special case of an interaction spline model, based on synthetic data, for the first test of the algorithm, because it has many of the features of the general case, and because numerical methods for efficient computation of interaction smoothing splines with irregular data have not, to our knowledge, been presented elsewhere.

We now describe these models. Let W_2^m be the Sobolev space

$$W_2^m=\{f\colon f,f',\cdots,f^{(m-1)}\text{ absolutely continuous}, f^{(m)}\in L_2[0,1]\}$$

with the squared norm

$$\|f\|^2_{W_2^m}=\sum_{\nu=0}^{m-1}(R_\nu f)^2+\int_0^1 (f^{(m)}(x))^2\,dx$$

where

$$R_\nu f=\int_0^1 f^{(\nu)}(x)\,dx, \qquad \nu=0,1,\cdots,m-1.$$

Let $k_l(x) = B_l(x)/l!$, where B_l is the lth Bernoulli polynomial, and we have $R_\nu B_l = \delta_{\nu-1}$ where $\delta_i = 1$, $i = 0$, and 0 otherwise. With this norm, W_2^m can be decomposed as the direct sum of m orthogonal one-dimensional subspaces $\{k_l\}$, $l = 0, 1, \cdots, m-1$, where $\{k_l\}$ is the one-dimensional subspace spanned by k_l, and H_*, which is the subspace (orthogonal to $\Sigma \oplus \{k_l\}$) satisfying $R_\nu f = 0$, $\nu = 0, 1, \cdots, m-1$, that is,

$$W_2^m=\{k_0\}\oplus\{k_1\}\oplus\cdots\oplus\{k_{m-1}\}\oplus H_*.$$

This construction can be found, e.g., in [11]. Letting $\overset{d}{\otimes}\, W_2^m$ be the tensor product of W_2^m with itself d times, we have

$$\overset{d}{\otimes}\, W_2^m=\overset{d}{\otimes}[\{k_0\}\oplus\cdots\oplus\{k_{m-1}\}\oplus H_*]$$

and $\overset{d}{\otimes}\, W_2^m$ may be decomposed into the direct sum of $(m+1)^d$ fundamental subspaces, each of the form

(5.3) $$[\quad]\otimes[\quad]\otimes\cdots\otimes[\quad] \qquad (d\,\text{boxes})$$

where each box ([ ]) is filled with either $\{k_l\}$ for some l, or H_*. Additive and interaction spline models are obtained by letting the H^β's of (5.2) be various of these $(m+1)^d$ fundamental subspaces. To obtain (purely) additive spline models, similar to those in [18], we retain only those subspaces of the form (5.3) above whose elements have a dependency on at most one variable; this means that (at most) one box is filled with an entry other than $\{k_0\} \equiv \{1\}$. We may construct spline models that are nonparametric in one variable and polynomial in all the others, nonparametric in two variables, and so forth.

The form of the induced norms on the various subspaces can be seen most easily by an example. Suppose $d = 4$ and consider, for example, the subspace

$$[\{k_l\}]\otimes[H_*]\otimes[H_*]\otimes[\{k_r\}],$$

which we assign the index $l{*}{*}r$. Then the square norm of the projection of f in $\overset{4}{\otimes}\, W_2^m$ onto this subspace is

$$\|P_{l**r}f\|^2=\int_0^1\int_0^1\left[\frac{\partial^{2m}}{\partial x_2^m\partial x_3^m}R_{l(x_1)}R_{r(x_4)}f(x_1,x_2,x_3,x_4)\right]^2 dx_2\,dx_3$$

where $R_{k(x_\alpha)}$ means R_k applied to what follows as a function of x_α. When we use the fact that the reproducing kernel (r.k.) for $\{k_l\}$ is $k_l(x)k_l(x')$ and the r.k. for H_* is $Q(x, x')$ given by

$$Q(x,x')=k_m(x)k_m(x')+(-1)^{m-1}k_{2m}([x-x'])$$

where $[u]$ is the fractional part of u (see [11]), it is easy to see that the r.k. for this subspace, call it $Q_{l**r}(x_1, x_2, x_3, x_4;\, x'_1, x'_2, x'_3, x'_4) = Q_{l**r}(\mathbf{x}; \mathbf{x}')$, is

$$Q_{l**r}(\mathbf{x};\mathbf{x}') = k_l(x_1)k_l(x'_1)Q(x_2,x'_2)Q(x_3,x'_3)k_r(x_4)k_r(x'_4).$$

For the properties of tensor products of r.k. spaces (see [1], [40]). If $L_i f = f(\mathbf{x}(i))$, where $\mathbf{x}(i)$ is the ith value of $\mathbf{x}$, then

$$(P_{l**r}\xi_i)(\mathbf{x}) = Q_{l**r}(\mathbf{x}(i),\mathbf{x})$$

and

$$\langle P_{l**r}\xi_i, P_{l**r}\xi_j\rangle = Q_{l**r}(\mathbf{x}(i),\mathbf{x}(j)).$$

In the purely additive model $f(x_1, \cdots, x_d)$ is of the form

$$f(x_1, \cdots, x_d) = \mu + \sum_{\alpha=1}^{d} g_\alpha(x_\alpha)$$

where $g_\alpha \in \{k_l\} \oplus \cdots \oplus \{k_{m-1}\} \oplus H_*$ and the penalty term in (5.2) can be taken as

$$\sum_{\alpha=1}^{d} \theta_\alpha^{-1} \int_0^1 \left(\frac{\partial^m g_\alpha}{\partial x_\alpha^m}\right)^2 dx_\alpha.$$

Although the popular purely additive model could be fitted with the methods described here, we do not do so because the present method does not use the special structure that is available for computing univariate polynomial splines (see, e.g., [19]).

5.2. Numerical examples. For simple yet nontrivial examples, we consider fitting interaction models with $d = 2$, $m = 2$. Let

$$W_2^2 \otimes W_2^2 = H_0 \oplus H_A \oplus H_I$$

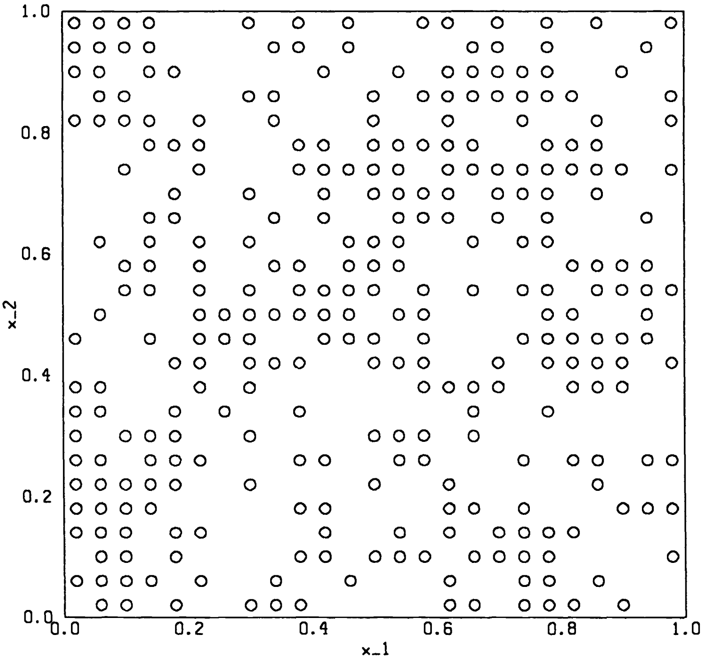


FIG. 1. *Sampling points for the numerical experiment.*

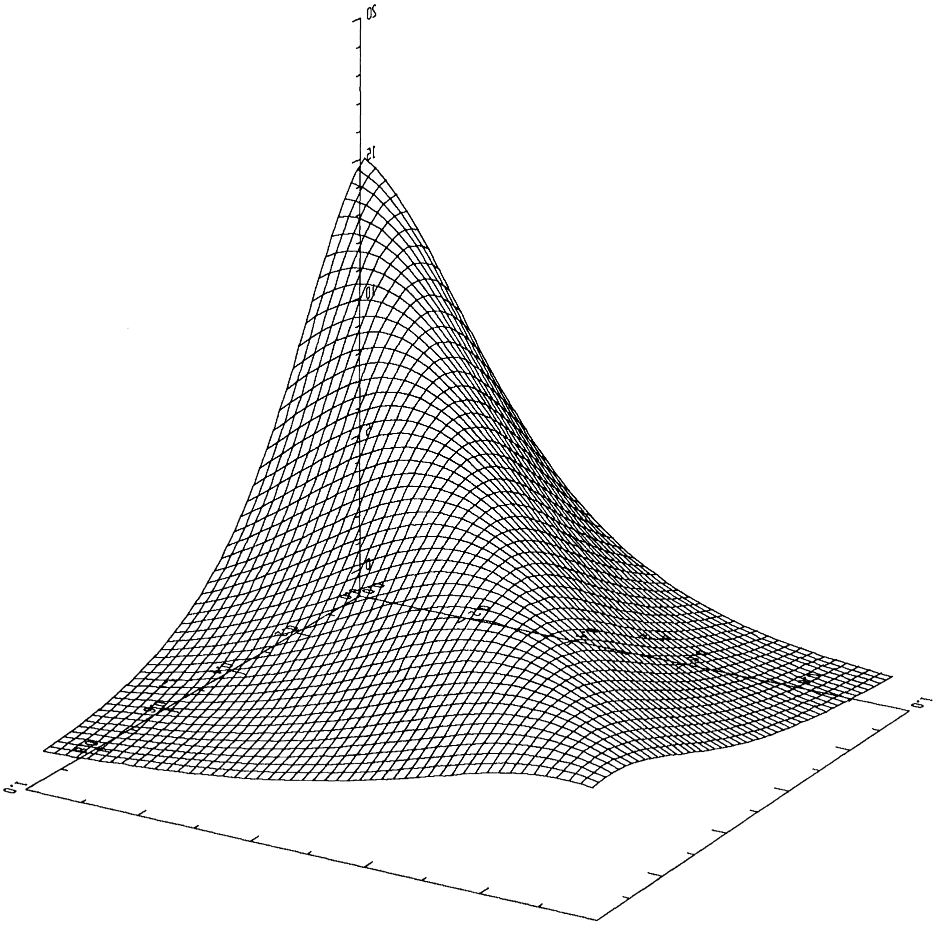


FIG. 2. *The "interactive" test function.*

where H_0, the null space of the penalty functional, is the $M = 4$ dimensional space

$$H_0 = \sum_{i,j=0}^{1} \oplus [\{k_i\} \otimes \{k_j\}].$$

We will denote by H_A the union of subspaces that are linear in one variable and "smooth" in the other. H_A is thus

$$H_A = [(\{k_0\} \oplus \{k_1\}) \otimes H_*] \oplus [H_* \otimes (\{k_0\} \oplus \{k_1\})]$$

and H_I is the nonlinear two-factor interaction subspace

$$[H_* \otimes H_*].$$

Then $f \in W_2^2 \otimes W_2^2$ has a (unique) decomposition of the form

$$f(x_1, x_2) = \{ \sum_{\nu,\mu=0}^{1} d_{\nu,\mu} k_\nu(x_1) k_\mu(x_2) \}$$
$$+ \{ f_{01}(x_1) + k_1(x_2) f_{11}(x_1) + f_{02}(x_2) + k_1(x_1) f_{12}(x_2) \} + \{ f_I(x_1, x_2) \}$$

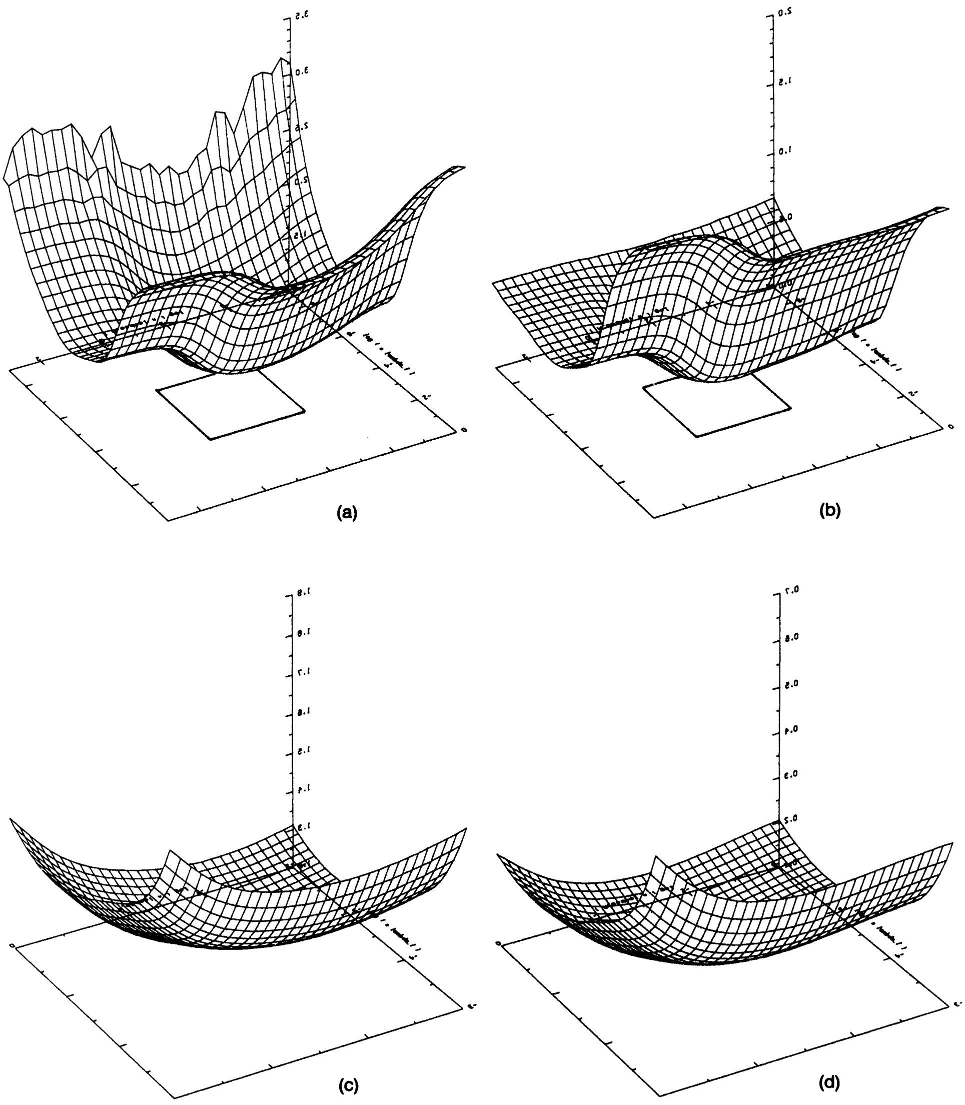


FIG. 3. *The* $V(\lambda)$ *and* MSE (λ) *for the "interactive" example.* (a) $V(\lambda)$ *in a broad area.* (b) MSE (λ) *in the same area as* (a). (c) *Enlargement of* (a) *in the boxed area.* (d) *Enlargement of* (b) *in the boxed area.*

where f_{01}, f_{02}, f_{11}, and f_{12} are in H_* and f_I is in $H_* \otimes H_*$, the components in { } are in H_0, H_A, and H_I, respectively. We consider the smoothing problem as follows. Find $f \in W_2^m \otimes W_2^m$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_1(i), x_2(i)))^2 + \lambda(\|P_A f\|^2 + \theta^{-1}\|P_I f\|^2), \tag{5.4}$$

which is actually a reparameterization, for numerical purposes, of the more natural form

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_1(i), x_2(i)))^2 + \lambda_A\|P_A f\|^2 + \lambda_I\|P_I f\|^2, \tag{5.5}$$

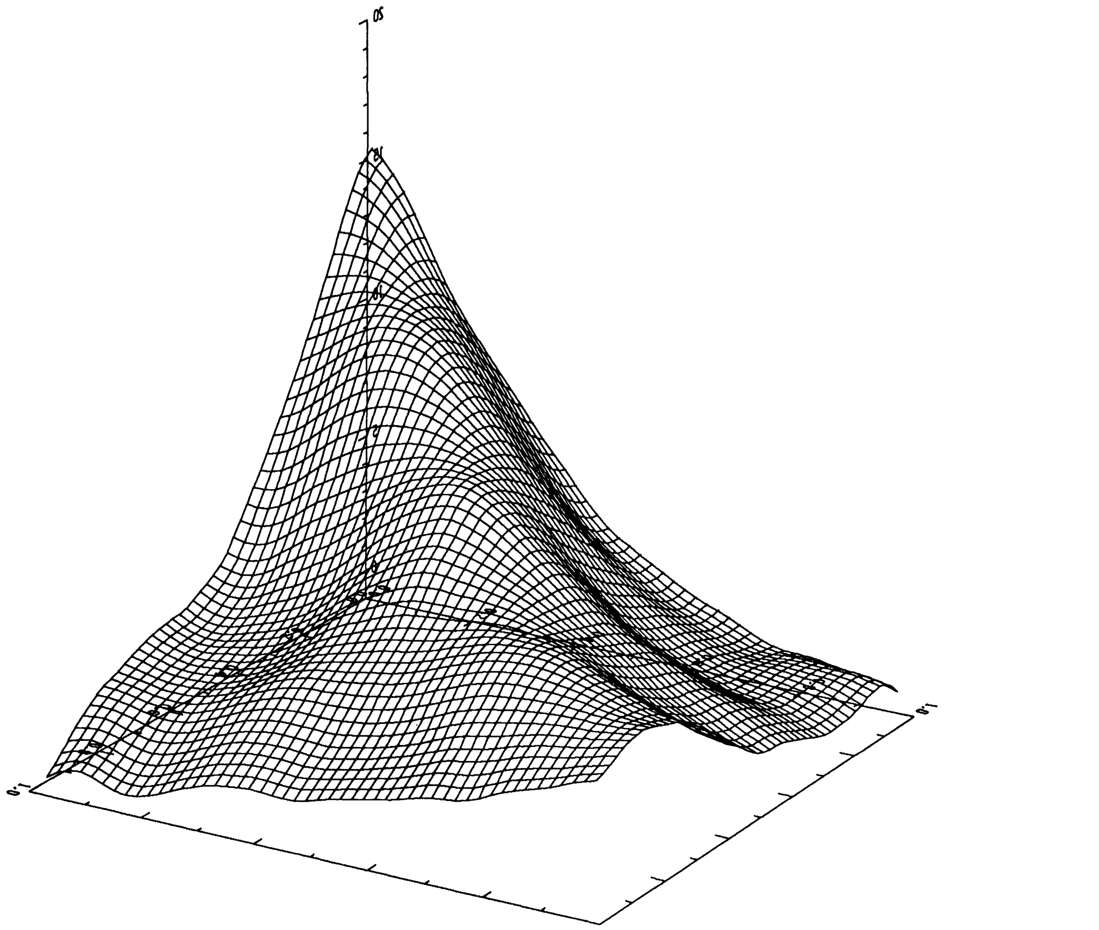


FIG. 4. *The estimated "interactive" function, with the* GCV *estimates of* λ_A *and* λ_I.

with $\theta = \lambda_A/\lambda_I$. (We note that we could have allowed two or more separate λ's for different components of H_A but we chose not to do that here.) In modeling response data, if a good estimate of λ_I turns out to be sufficiently large, then the interaction term will not be "significant," and the user may choose to delete it. If this term is not "significant," then consideration may be given to deleting the "cross" terms $k_1(x_2)f_{11}(x_1)$ and $k_1(x_1)f_{12}(x_2)$.

The r.k.'s Q_A for H_A and Q_I for H_I are given by

$$Q_A(x_1,x_2;x_1',x_2')$$

$$=Q(x_1,x_1')+Q(x_2,x_2')+k_1(x_1)k_1(x_1')Q(x_2,x_2')+Q(x_1,x_1')k_1(x_2)k_1(x_2')$$

and

$$Q_I(x_1,x_2;x_1',x_2')=Q(x_1,x_1')Q(x_2,x_2').$$

The ith row of the $n \times 4$ matrix S of (1.8) is

$$(1,k_1(x_1(i)),k_1(x_2(i)),k_1(x_1(i))k_1(x_2(i))).$$

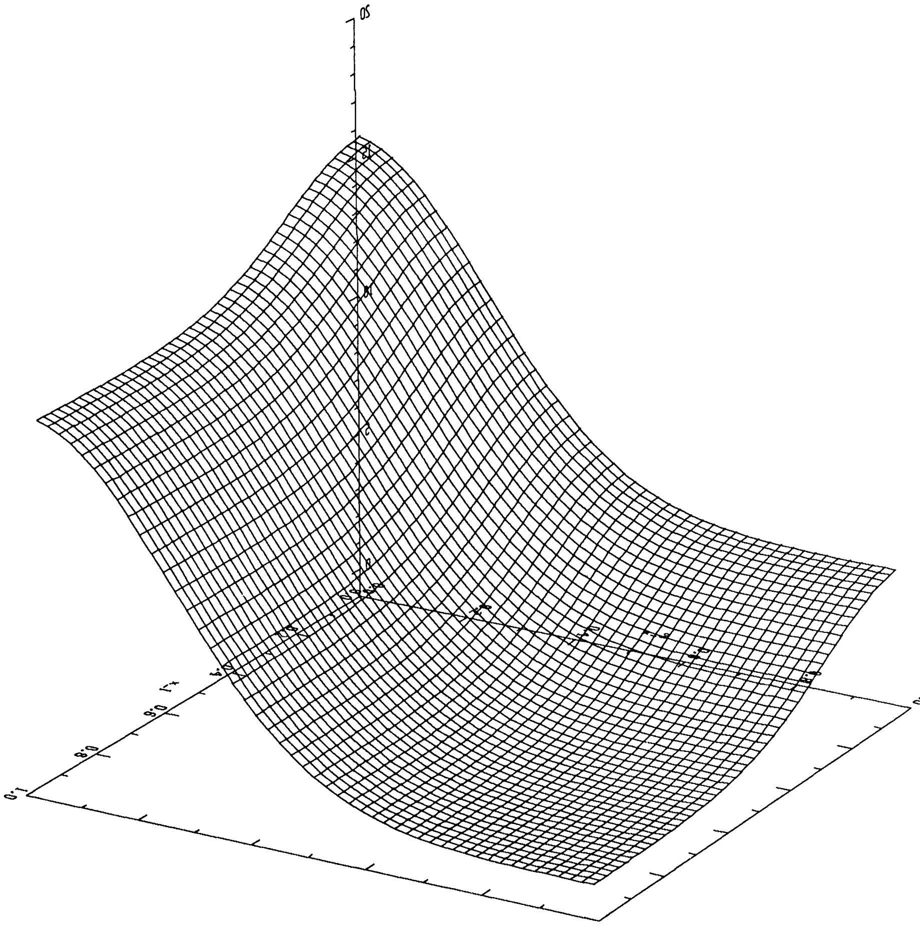


FIG. 5. *The "additive" test function.*

$\xi_i(\mathbf{x})$ in (1.5) is given by

$$\xi_i(\mathbf{x}) = Q_A(\mathbf{x}(i), \mathbf{x}) + \theta Q_I(\mathbf{x}(i), \mathbf{x})$$

and the ijth entry of $\tilde{\Sigma}$ of (1.7) is

$$Q_A(\mathbf{x}(i), \mathbf{x}(j)) + \theta Q_I(\mathbf{x}(i), \mathbf{x}(j)). \tag{5.6}$$

If the $\mathbf{x}(i)$'s form a regular pattern, then there will be some special structure that, in principle, could be exploited (see [31]). Our work is aimed at the case where no special structure can be assumed.

The sampling points we chose in the numerical experiment are shown in Fig. 1. We first drew a 25×25 regular mesh, with each side being .02(.04).98. Then we divided the mesh into 5×5 regular blocks, with each block containing 25 meshpoints. From 10 of the blocks randomly chosen out of the 25, we randomly deleted seven meshpoints per block, and from the remaining 15 blocks we randomly deleted 17 meshpoints per block. This procedure leaves us with 300 "patchier" data points as shown in the plot. These 300 data points $\mathbf{x}(i)$, $i = 1, \cdots, 300$, are ordered so that

$$i<j \Rightarrow x_1(i)<x_1(j) \quad \text{or} \quad x_1(i)=x_1(j), \quad x_2(i)<x_2(j).$$

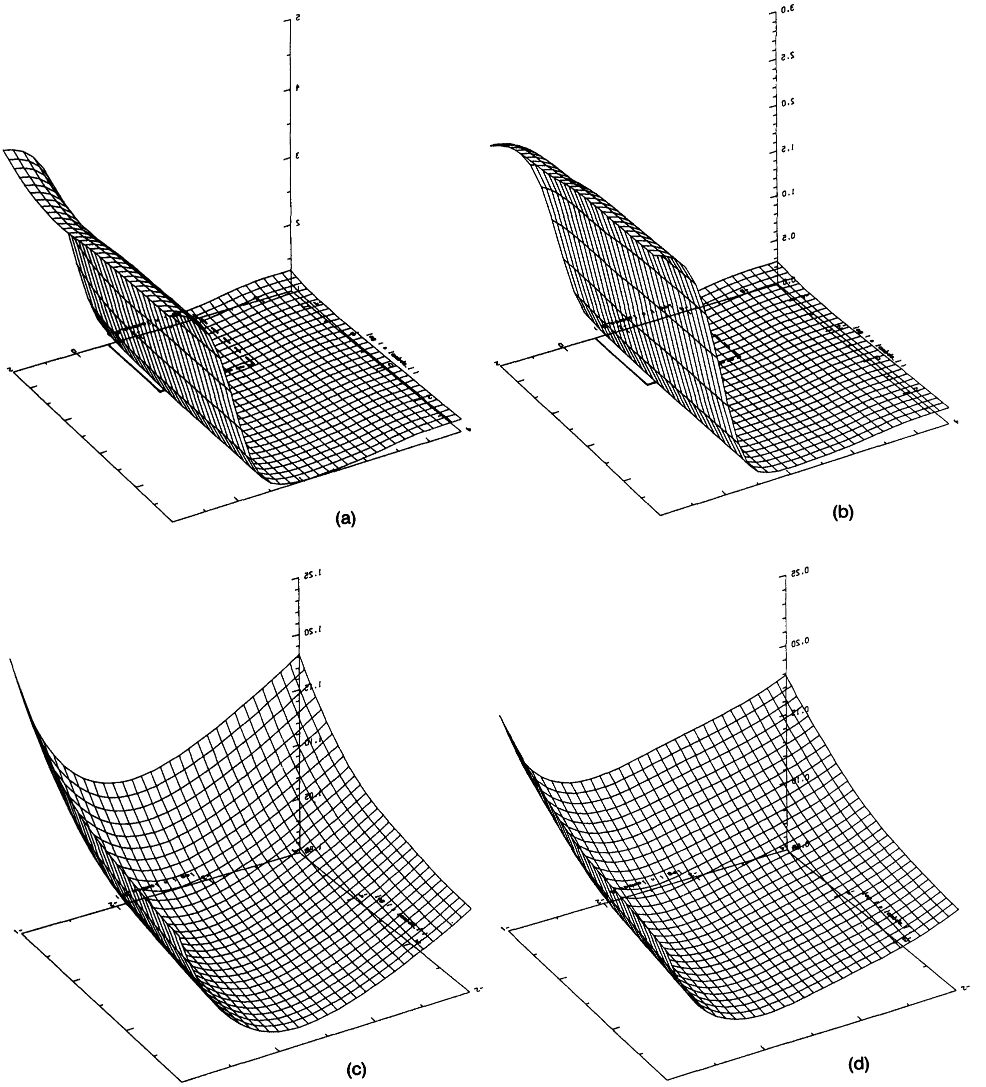


FIG. 6. *The* $V(\lambda)$ *and* MSE (λ) *for the "additive" example.* (a) $V(\lambda)$ *in a broad area.* (b) MSE (λ) *in the same area of* (a). (c) *Enlargement of* (a) *in the boxed area.* (d) *Enlargement of* (b) *in the boxed area.*

We use the random number generator *rnor* (which further calls *uni*) [22], as implemented in the Core Mathematics Library (CMLIB) from National Bureau of Standards, to generate 300 independent normal $N(0, 1)$ deviates ε_i to go with these data points, with the first dummy call using seed 4321, and variable *mdig* in routines *rnor* and *uni* fixed to 16. Thus the standard deviation of the ε_i's was equal to the distance between two successive hatch marks in the vertical scale in Figs. 2, 4, 5, and 7. We ran the experiment on the data $y_i = f(\mathbf{x}(i)) + \varepsilon_i$ for two different test functions. One has a strong interaction component, and the other is purely additive. We present them in sequence.

The "interactive" test function is shown in Fig. 2:

$$f(x_1, x_2) = \frac{40 \exp(8[(x_1-.5)^2+(x_2-.5)^2])}{\exp(8[(x_1-.2)^2+(x_2-.7)^2])+\exp(8[(x_1-.7)^2+(x_2-.2)^2])}.$$

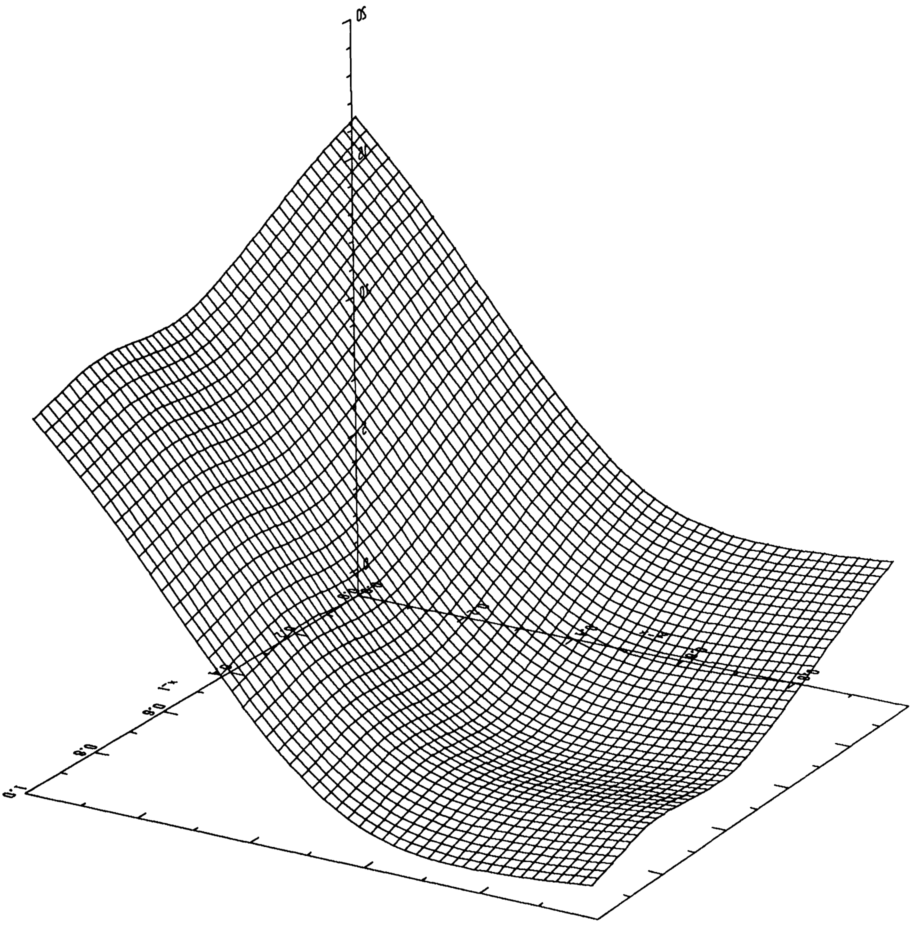


FIG. 7. *The estimated "additive" function, with the* GCV *estimates of* λ_A *and* λ_I.

Figure 3(a) shows $V(\lambda) = V(\lambda_A, \lambda_I)$ for a wide range of λ_A, λ_I, with a tensor product mesh of $\log_{10}(n\lambda_A) \in -6(.3)3$ and $\log_{10}(n\lambda_I) \in -9(.3)0$. Defining MSE$(\lambda)$ as

$$\text{MSE}(\lambda) = \frac{1}{n}\sum_i (L_i f - L_i f_\lambda)^2 = \frac{1}{n}\sum_i (f(\mathbf{x}(i)) - f_\lambda(\mathbf{x}(i)))^2,$$

Fig. 3(b) shows MSE(λ) over the same mesh as in Fig. 3(a). In theory the minimizer of $V(\lambda)$ is an estimate of the minimizer of MSE(λ), and so, in a numerical experiment such as this one, MSE(λ) can be inspected to see how good the estimates (minimizer of $V(\lambda)$) are. Figure 3(c) gives these surfaces over a narrower range of values of λ_A and λ_I containing their minima (boxed region of Figs. 3(a) and 3(b)), using mesh $\log_{10}(n\lambda_A) \in -4(.1){-1}$ times $\log_{10}(n\lambda_I) \in -7(.1){-4}$. The minimum GCV value 1.14694 on these meshpoints is obtained at $\log_{10}(n\lambda_A) = -2.5$ and $\log_{10}(n\lambda_I) = -6.0$. This combination gives the fitted surface shown in Fig. 4. This surface has an MSE of 0.10874 compared to the minimum possible MSE value of 0.09141 on the meshpoints. By setting one of the smoothing parameters to infinity, i.e., leaving out H_I or H_A, we can estimate the "marginally" optimal smoothing parameters. There are some theoretical results, to be presented elsewhere, that suggest the "marginally" optimal smoothing parameters can be good starting guesses for the simultaneously optimal smoothing parameters under certain conditions. For the current example, the "marginally" optimal smoothing parameters are $\log_{10}(n\lambda_A) = -2.568$ with GCV value 2.356, and $\log_{10}(n\lambda_I) = -5.928$ with GCV value 1.661.

The "additive" test function is shown in Fig. 5, and is

$$f(x_1, x_2) = 4(\exp(-8x_1^2) + 3\exp(-8x_2^2)).$$

The model $f = f_0 + f_A + f_I$, where $f_0 \in H_0$, $f_A \in H_A$, and $f_I \in H_I$ as before, was fitted, with $V(\lambda)$ and MSE(λ) searched on the mesh $\log_{10}(n\lambda_A) \in -7(.3)2$ times $\log_{10}(n\lambda_I) \in -5(.3)4$. The minimum GCV value of 1.01669 was obtained at $\log_{10}(n\lambda_A) = -2.2$ and $\log_{10}(n\lambda_I) = 4$, indicating that λ_I wants to go to infinity. The MSE at $(-2.2, 4)$ was .078457, and the minimum MSE on the mesh was .075363. The model $f = f_0 + f_A$ was then fitted, and $\log(n\lambda_A) = -2.13$ obtained by a golden-section search, with a minimum GCV of 1.01647, which suggests that $\lambda_I = \infty$ is, in fact, the minimizer. We also tried the modified model $f = f_0 + f_{A'} + f_{I'}$, where the primes are here intended to indicate that the "cross" terms, that is, the terms of the form $k_1(x_2)f_{11}(x_1)$ and $k_1(x_1)f_{12}(x_2)$, were removed from H_A and added to H_I. This gives new spaces $H_{A'}$ and $H_{I'}$ with corresponding smoothing parameters $\lambda_{A'}$ and $\lambda_{I'}$. The same mesh was searched (in $\log_{10}(n\lambda_{A'})$ times $\log_{10}(n\lambda_{I'})$) and the same minimizer was found, but with a minimum GCV value of 1.00940, which is less than 1.01647. $V(\lambda)$ and MSE(λ) for this modified model are plotted in Fig. 6. Figures 6(a) and 6(b) contain the edges of the high plateau in both GCV and MSE surfaces that appear at large $\lambda_{A'}$. Figures 6(c) and 6(d) give the surfaces in an area where $\lambda_{A'}$ is around the optimum and the decreasing trend in the $\lambda_{I'}$ direction can be visualized (boxed region of Figs. 6(a) and 6(b)). The purely additive model was tried (that is, $\lambda_{I'} = \infty$) and this model was fitted using $\log_{10}(n\lambda_{A'}) = -2.2$, which gave a GCV value of 1.00940 and an MSE of 0.073606. The minimum MSE on the meshpoints of Fig. 6 is 0.072730. The fitted purely additive function is shown in Fig. 7.

5.3. Numerical strategies and timing results. Now we briefly describe the basic numerical strategies in fitting the above presented models, as well as some timing results of interest.

It is clear from (5.6) that the $\tilde{\Sigma}$ matrix in (1.6) varies with the parameter θ. As discussed in § 3, the major numerical work of the algorithm rests on the tridiagonalization process, so the GCV evaluations on various grid points of λ for fixed θ (the notation of (5.4)) are essentially free given the tridiagonalization. Thus the two parameters θ and λ bear different numerical properties. θ is "cubic" and λ is "linear." For obtaining a "quick" solution, a double grid search can be done on (θ, λ), with the execution time being basically proportional to the number of θ values tried. However, to visualize the GCV and MSE surfaces, the parameterization in (5.4) is not appropriate. We need to draw the surfaces on the product mesh of λ_A and λ_I as parameterized in (5.5). When both λ_A and λ_I are in a log scale, as they should be, and when the stepsizes on both axes are the same, it is easy to see that the meshpoints on each diagonal line have the same θ parameter. Our strategy of obtaining those surfaces presented before is based on this observation. Hence the numerical cost of each surface is proportional to the sum of the grid numbers on both axes minus 1.

For our examples, the matrix Σ is of size 296 × 296. We set the truncation rate in the tridiagonalization to the square of the machine precision, so essentially no real mass will be truncated. The timing for one run of tridiagonalization of a full rank matrix Σ, on the Celerity at the Yale University Computer Sciences Department, is 86.0 seconds.

One evaluation of $V(\lambda)$ given the tridiagonal form takes 0.083 seconds. Obtaining the coefficients $\mathbf{c}$ and $\mathbf{d}$ for fixed λ takes a further 0.600 seconds. In fitting the purely additive model, the matrix Σ has a rank of no more than 50, due to the way we chose the sampling points. The above three numbers become 35.0, 0.083, and 0.233 in the additive case. It shows that the truncation strategy (most likely the tail truncation in this situation) in our implementation of the Householder tridiagonalization does work properly. It is noted that the first and third numbers reduce significantly, for both of them are affected by the number of Householder transforms needed to tridiagonalize the Σ matrix, although the second number is not affected.

As a matter of fact, the major block of our algorithm is just a GCV solver for the system (1.6). A famous one smoothing parameter example that satisfies (1.6) is the thin plate smoothing spline. Dr. Fred Reames has conducted a numerical experiment to compare the proposed algorithm against the SVD approach for the thin plate smoothing spline implemented in GCVPACK. On a data set of size $n = 496$, with $M = 3$, the proposed algorithm solves (1.6) in 2,342 seconds on a VaxStation, and the GCVPACK does this in 9,646 seconds. Remember that our tridiagonalization method is the symmetric version of Elden's bidiagonalization method. This in a sense verifies the conjecture of Bates et al. [6] that the bidiagonalization would speed up the numerical process significantly for large data sets. For details about the thin plate smoothing spline models, see Wahba and Wendelberger [34] and Bates et al. [6].

6. Some remarks on further topics.

6.1. Anisotropic thin plate smoothing splines. The usual thin plate spline penalty functional is

$$\|P_1 f\|^2 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{\alpha_1 + \cdots + \alpha_d = m} \frac{m!}{\alpha_1! \cdots \alpha_d!} \left(\frac{\partial^m f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \right)^2 dx_1 \cdots dx_d$$

(see, e.g., [34]). In some applications an "elliptical" penalty functional is desirable, for example, in modeling diffusion processes with two or three space variables and one time variable, it is desired to have a free scale factor in time. In this case $f(x_1, \cdots, x_d)$ is replaced by $f(x_1, \theta_1 \tilde{x}_2, \cdots, \theta_d \tilde{x}_d)$ where the $\theta_2, \cdots, \theta_d$ are scale factors to be chosen. Then Σ will depend on the θ's. The numerical formulation is similar to that of interaction smoothing spline models.

6.2. A problem from meteorology. Another type of problem that arises in meteorology, can be put in a form such that the present algorithm allows efficient solution. Suppose we have two information systems providing data on same function g (meteorological variables such as atmospheric wind and temperature), for example,

$$\begin{aligned} y_i^{(1)} &= g(x_i) + \varepsilon_i^{(1)}, \\ y_i^{(2)} &= g(x_i) + \varepsilon_i^{(2)}, \end{aligned} \qquad i = 1, \cdots, n$$

where $\varepsilon^{(1)}$ is distributed as $N(\mathbf{0}, \sigma^2\mathbf{I})$ and $\varepsilon^{(2)}$ is distributed as $N(\mathbf{0}, b\Sigma_\theta)$. Here Σ_θ is a correlation matrix sufficiently different from $\mathbf{I}$, and possibly depending on a parameter θ. σ^2, b, and θ are unknown, and it is desired to estimate the relative error $r = \sigma^2/b$ and (possibly) θ. Details of an important meteorological example and a family of practical generalizations are discussed by Wahba in [39].

Let $\mathbf{z} = \mathbf{y}^{(1)} - \mathbf{y}^{(2)}$, then $\mathbf{z}$ is distributed as $N(\mathbf{0}, b(\Sigma_\theta + r\mathbf{I}))$. Maximizing the likelihood with respect to b, r, and θ gives that the maximum likelihood estimate of (r, θ) is given by the minimizer of

$$M(r,\theta)=\frac{\mathbf{z}^T(\Sigma_\theta+r\mathbf{I})^{-1}\mathbf{z}}{[\det(\Sigma_\theta+r\mathbf{I})]^{-1/n}},$$

so that the algorithm for (1.10) applies.

6.3. The use of $M(\lambda)$ in hypothesis testing/model building problems. If H_1 possess a reproducing kernel, say $Q(s, t)$, s, $t \in S$, then there is a Bayesian model that leads to the variational problem (1.1). That is, suppose f is a stochastic process with

$$f(s)=\sum_{\nu=1}^{M}\delta_\nu\phi_\nu(s)+bX(s), \qquad s\in S$$

with $X(s)$, $s \in S$, a zero mean Gaussian stochastic process with $E[X(s)X(t)] = Q(s, t)$, $\delta = (\delta_1, \cdots, \delta_M)^T$ is a random Gaussian vector with $E[\delta\delta^T] = \eta I$, then

$$f_\lambda(s)=\lim_{\eta\to\infty} E(f(s)\,|\,y_1,\cdots,y_n), \qquad \lambda=\frac{\sigma^2}{nb}$$

(see [33]). We remark that if f is a sample function from a stochastic process and H_* is an infinite-dimensional space, then $f \notin H_*$ with probability one. This remark accounts for the differing range of applicability of V and M (see also [36]). In the Bayesian setting the null hypothesis $b = 0$ (equivalently $\gamma = 1/\lambda = nb/\sigma^2 = 0$) is equivalent to $f = \Sigma\delta_\nu\phi_\nu$ for some δ. The alternative, $\gamma > 0$, is equivalent to $f \in H$. An important example is the case of

$$H_{\text{null}}: f \text{ linear} \quad \text{vs.} \quad H_{\text{alternative}}: f\in W_2^2.$$

Some of these hypothesis tests are discussed in [42], [4], [10], and [38]. The likelihood ratio test statistic is

$$t_{\text{ML}}=\inf_\lambda\frac{M(\lambda)}{M(\infty)},$$

and the GCV test statistic is

$$t_{\text{GCV}}=\inf_\lambda\frac{V(\lambda)}{V(\infty)}.$$

For computational purposes it is convenient to change the parameter from λ to $1/\lambda = \gamma$.

If the tridiagonalization is carried one step further to diagonalization, then it is a relatively simple matter to generate the distribution of these statistics under the null hypothesis and specific alternative hypotheses. In the purely Bayesian model, t_{ML} is a standard choice. It is not known how the ML and GCV tests would compare against non-Bayesian alternatives. These tests can, in principle, be used to test hypotheses about whether θ_β's are zero in the interaction spline models.

6.4. Further numerical strategies. Some preliminary numerical experiments indicate that to get eigenvalues from the symmetric tridiagonal form is comparatively cheaper than to get singular values from the bidiagonal form, provided the orthogonal matrix involved is not accumulated. Also of interest is that poorer conditioning seems to speed up convergence in the symmetric QR iterations. Since the eigenstructure will provide more information about the adequacy of the model rather than simply a fitted model, and the usually negligible difference between the evaluations of GCV function from the tridiagonal form and from the diagonal form would become serious when resampling statistical inferences are applied to the problem, it is preferable to go further from tridiagonal to diagonal form, at least as an option for special purposes. To avoid accumulating the orthogonal matrix involved, it is possible to dynamically update the data $\mathbf{x}$ in (3.1), (3.2) during the reduction from tridiagonal $\mathbf{T}$ to the desired eigenstructure. This process needs further special coding as opposed to simply assembling the routines from standard libraries. We plan to implement and test this procedure in further study, thus providing an option beyond the tridiagonalization model fitting.

An alternative to the BD approach for ridge regression setting is as follows. Form $\mathbf{X}^T\mathbf{X}$ at the outset as in DCD, and then apply TD on $\mathbf{X}^T\mathbf{X}$ instead of BD on $\mathbf{X}$. The TD approach (including the formation of $\mathbf{X}^T\mathbf{X}$) is usually slightly faster than BD approach. This is because in TD we can take advantage of the symmetry of matrix $\mathbf{X}^T\mathbf{X}$, which makes the operations needed for tridiagonalization half of those needed for bidiagonalizing unsymmetric matrix of comparable size. More details of this approach to ridge regression will also be explored in later work.

One thing of interest here is the role of the matrix condition in the GCV computations. In the case where linear systems are solved directly as in the standard least square problems, great precision in the computation of small singular/eigenvalues are of course very crucial to the numerical stability of the solver. Thus the singular value decomposition of $\mathbf{X}$ is commonly preferred to the normal equations for standard least square problems (see [17, Chap. 6]). When we are applying regularization procedures to the ill-posed problems, however, the effect of very small singular/eigenvalues are supposed to be thresholded by the smoothing parameters. Hence the solution should be robust to the precision of the very small singular/eigenvalues. This is the reason that appropriate truncation can reduce the numerical burden without sacrificing the precision of the solution (see also [5], [6]). This phenomena makes the worse matrix condition a welcome event numerically in this specific situation. Also, this phenomena, together with the benefit offered by symmetry discussed earlier, might make the methods based on normal equations preferred over the singular value decomposition of $\mathbf{X}$ for the regularized solver of linear systems.

Acknowledgments. The computation of the examples was done on the Celerity computer at the Yale University Computer Sciences Department. We thank Prof. Martin Schultz (ONR grant N00014-86-K-0310) for providing access to this system. We thank Dr. Fred Reames for help in testing our algorithm against GCVPACK and Barry and Hartigan and Utreras and Schumaker for providing draft versions of their manuscripts.

REFERENCES

[1] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[2] D. Barry, *Nonparametric Bayesian regression*, Ph.D. thesis, Yale University, New Haven, CT, 1983.

[3] ———, *Nonparametric Bayesian regression*, Ann. Statist., 14 (1986), pp. 934–953.

[4] D. Barry and J. Hartigan, *An omnibus test for departures from constant mean*, manuscript, 1988.

[5] D. M. Bates and G. Wahba, *Computational methods for generalized cross-validation with large data sets*, in Treatment of Integral Equations by Numerical Methods, C. T. Baker and G. F. Miller, eds., Academic Press, London, 1982, pp. 283–296.

[6] D. M. Bates, M. J. Lindstrom, G. Wahba, and B. Yandell, GCVPACK—*Routines for generalized cross validation*, Commun. Statist. B, 16 (1987), pp. 263–297.

[7] A. BUJA, T. HASTIE, AND R. TIBSHIRANI, *Linear smoothers and additive models*, Ann. Statist., 17 (1989), to appear.

[8] Z. CHEN, *A stepwise approach for purely periodic interaction spline models*, Commun. Statist. A, 16 (1987), pp. 877–895.

[9] D. COX, *Multivariate smoothing spline functions*, SIAM J. Numer. Anal., 21 (1984), pp. 789-813.

[10] D. COX, E. KOH, G. WAHBA, AND B. S. YANDELL, *Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models*, Ann. Statist., 16 (1988), pp. 113–119.

[11] P. CRAVEN AND G. WAHBA, *Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation*, Numer. Math., 31 (1979), pp. 377–403.

[12] J. CRUMP AND J. SEINFELD, *A new algorithm for inversion of aerosol size distribution data*, Aerosol Sci. Tech., 1 (1982), pp. 15–34.

[13] J. DONGARRA, J. BUNCH, C. MOLER, AND G. W. STEWART, *Linpack Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.

[14] L. ELDEN, *A note on the computation of the generalized cross-validation function for ill-conditioned least squares problems*, BIT, 24 (1984), pp. 467–472.

[15] D. GIRARD, *Un algorithme simple et rapide pour la validation croissée generalisée sur des problèmes de grande taille*, RR 669-M, Informatique et Mathématiques Appliquées de Grenoble, Grenoble, France, May 1987.

[16] G. H. GOLUB, M. HEATH, AND G. WAHBA, *Generalised cross validation as a method for choosing a good ridge parameter*, Technometrics, 21 (1979), pp. 215–224.

[17] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computation*, The John Hopkins University Press, Baltimore, MD, 1983.

[18] T. HASTIE AND R. TIBSHIRANI, *Generalized additive models*, Statist. Sci., 1 (1986), pp. 297–318.

[19] M. F. HUTCHINSON AND F. R. DEHOOG, *Smoothing noisy data with spline functions*, Numer. Math., 47 (1985), pp. 99–106.

[20] G. KIMELDORF AND G. WAHBA, *Some results on Tchebycheffian spline functions*, J. Math. Anal. Appl., 33 (1971), pp. 82–95.

[21] K. LI, *Asymptotic optimality of C sub L and generalized cross-validation in ridge regression with application to spline smoothing*, Ann. Statist., 14 (1986), pp. 1101–1112.

[22] G. MARSAGLIA AND W. W. TSANG, *A fast, easily implemented method for sampling from decreasing or symmetric unimodal density functions*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 349–359.

[23] P. MERZ, *Determination of adsorption energy distribution by regularization and a characterization of certain adsorption isotherms*, J. Comput. Phys., 38 (1980), pp. 64–85.

[24] D. NYCHKA AND D. COX, *Convergence rates for regularized solutions of integral equations from discrete noisy data*, Tech. Report 752, Department of Statistics, University of Wisconsin, Madison, WI, 1984.

[25] D. NYCHKA, G. WAHBA, S. GOLDFARB, AND T. PUGH, *Cross-validated spline methods for the estimation of three dimensional tumor size distributions from observations on two dimensional cross sections*, J. Amer. Statist. Assoc., 79 (1984), pp. 832–846.

[26] F. O'SULLIVAN AND G. WAHBA, *A cross validated Bayesian retrieval algorithm for non-linear remote sensing experiments*, J. Comput. Phys., 59 (1985), pp. 441–455.

[27] F. O'SULLIVAN, B. YANDELL, AND W. RAYNOR, *Automatic smoothing of regression functions in generalized linear models*, J. Amer. Statist. Assoc., 81 (1986), pp. 96–103.

[28] F. O'SULLIVAN, *Fast computation of fully automated log-density and log-hazard estimators*, SIAM. J. Sci. Statist. Comput., 9 (1988), pp. 363–379.

[29] P. SPECKMAN, *Spline smoothing and optimal rates of convergence in nonparametric regression models*, Ann. Statist., 13 (1985), pp. 970–983.

[30] F. UTRERAS, *Optimal smoothing of noisy data using spline functions*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 349-362.

[31] F. UTRERAS AND L. SCHUMAKER, *On generalized cross-validation for tensor smoothing splines*, SIAM J. Sci. Statist. Comput., 11 (1990), to appear.

[32] G. WAHBA, *Practical approximate solutions to linear operator equations when the data are noisy*, SIAM J. Numer. Anal., 14 (1977), pp. 651-667.

[33] ———, *Improper priors, spline smoothing and the problem of guarding against model errors in regression*, J. Roy. Statist. Soc. Ser. B, 40 (1978), pp. 364-372.

[34] G. WAHBA AND J. WENDELBERGER, *Some new mathematical methods for variational objective analysis using splines and cross-validation*, Monthly Weather Rev., 108 (1980), pp. 1122–1145.

[35] G. WAHBA, *Ill-posed problems: Numerical and statistical methods for mildly, moderately, and severely ill-posed problems with noisy data*, Tech. Report 595, Department of Statistics, University of Wisconsin, Madison, WI, 1980; Proc. International Conference on Ill-Posed Problems, M. Z. Nashed, ed., to appear.

[36] ———, *A comparison of* GCV *and* GML *for choosing the smoothing parameter in the generalized spline smoothing problem*, Ann. Statist., 13 (1985), pp. 1378–1402.

[37] ———, *Partial and interaction splines for the semiparametric estimation of functions of several variables*, in Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface, T. J. Boardman, ed., American Statistical Association, Washington, DC, 1986, pp. 75–80.

[38] ———, *Spline and Partial Spline Models in Statistics*, CBMS Lecture Notes at Ohio State University, Columbus, OH, 1987, to appear.

[39] ———, *On the dynamic estimation of relative weights for observation and forecast in numerical weather prediction*, Tech. Report 818, Department of Statistics, University of Wisconsin, Madison, WI, 1988; Proc. International Workshop on Remote Sensing Retrieval Methods, A. Deepak, ed., to appear.

[40] H. L. WEINERT, *Reproducing Kernel Hilbert Spaces*, Hutchinson Ross, Stroudsburg, PA, 1982.

[41] H. J. WOLTRING, *On optimal smoothing and derivative estimation from noisy displacement data in biomechanics*, Human Movement Sci., 4 (1985), pp. 229–245.

[42] T. YANAGIMOTO AND M. YANAGIMOTO, *The use of marginal likelihood for a diagnostic test for the goodness of fit of the simple linear regression model*, Technometrics, 29 (1987), pp. 95–101.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 4, pp. 481–493, October 1989

005

MULTISPLITTING WITH DIFFERENT WEIGHTING SCHEMES*

R. E. WHITE†

Abstract. Parallel algorithms generated by multisplittings are considered. A parallel algorithm may be formed, first, by concurrently executing the iteration associated with each splitting, and second, by forming a weighted sum of these computations. However, it is not imperative that the weighting be done last. Convergence results are obtained for a variety of other weighting schemes. In particular, it is shown that preweighting is in some cases more desirable than the traditional postweighting. Furthermore, we indicate how one can use a symmetric weighting scheme to obtain a good multisplitting version of the SSOR preconditioner. These algorithms are illustrated by computations done on an Alliant FX/8.

Key words. multisplitting, parallel computation, preweighting

AMS(MOS) subject classifications. 65F10, 65N20

1. Introduction. In this paper we consider the linear algebraic system

$$Ax = d. \tag{1}$$

Iterative methods for approximating the solution of (1) are usually based on a single splitting $A = B - C$. In order to utilize multiprocessing computers, O'Leary and White [9] introduced multisplittings. A *multisplitting* of a matrix A is a sequence of splittings $A = B_k - C_k$ for $k = 1, \cdots, K$, where B_k are nonsingular. When coupled with *weighting matrices* D_k for $k = 1, \cdots, K$ with $D_k \geqq 0$, diagonal and $\sum_{k=1}^{K} D_k = I$, one can form a parallel algorithm.

Parallel algorithm with postweighting.

$$\begin{aligned} x^{m+1} &= Hx^m + Gd \text{ where} \\ H &= \sum_{k=1}^{K} D_k B_k^{-1} C_k \text{ and} \\ G &= \sum_{k=1}^{K} D_k B_k^{-1}. \end{aligned} \tag{2}$$

The terms in Hx^m and Gd may be computed concurrently.

Note $B_k^{-1}A = I - B_k^{-1}C_k$ and $(\sum_{k=1}^{K} D_k B_k^{-1})A = I - (\sum_{k=1}^{K} D_k B_k^{-1} C_k)$ or $GA = I - H$. So given G and A one can define $H = I - GA$ and define an algorithm by

$$\begin{aligned} x^{m+1} &= Hx^m + Gd \\ &= (I - GA)x^m + Gd \\ &= x^m + G(d - Ax^m) \\ &= x^m + Gr(x^m). \end{aligned} \tag{3}$$

In this paper we consider different forms of G given by a multisplitting with $0 \leqq \lambda \leqq 1$:

$$G_\lambda = \sum_{k=1}^{K} D_k^\lambda B_k^{-1} D_k^{1-\lambda}. \tag{4}$$

* Received by the editors December 12, 1988; accepted for publication (in revised form) February 9, 1989.

† Department of Mathematics, Box 8205, North Carolina State University, Raleigh, North Carolina 27695-8205 (white%crscfx40@ncsuvx.ncsu.edu).

When $\lambda = 1$, this is the *postweighting* used in (2). For $\lambda = 0$ we call this *preweighting*, and $\lambda = \frac{1}{2}$ is *symmetric weighting*. As we shall see, when the matrix A has dissection form and the multisplittings are associated with the SOR method, then $G_0 = (1/\omega(D - \omega L))^{-1}$ and the iterative method (3) with $G = G_o$ is the serial SOR method. When $\lambda = \frac{1}{2}$ and each B_k^{-1} is symmetric positive definite, then $G_{1/2}$ will be symmetric positive definite. Consequently, $G_{1/2}$ can be used as a parallel preconditioner for the conjugate gradient method; this means the preconditioned portions may be broken into parts corresponding to the terms in (4) and executed concurrently.

The convergence for (2) was first considered in O'Leary and White [9] for the case A is an M-matrix, and later in Neumann and Plemmons [8]. In White [15] an analysis is given when A is a symmetric positive definite matrix. Related overlapping blocks schemes have been considered by Ostrowski [11], Robert [13], Hayes [4], and McBryan and Van de Velde [7]. White [16], Neumann and Plemmons [8], and Elsner [2] consider comparison results for $\rho(H)$ in (2).

In the next section we give convergence results for the algorithm (3) for general G and for $G = G_\lambda$ in (4) (see Theorems 2 and 4). Both the M-matrix condition on A (Theorems 1, 2, and 4), and the symmetric positive definite condition (Theorem 6) on A will be considered. The third section contains a discussion of preconditioners for the conjugate gradient method. This will include the case where $\lambda = \frac{1}{2}$ in (4) and B_k^{-1} are from SSOR multisplittings (Theorem 7). The last section contains a number of numerical experiments done on the Alliant FX/8 multiprocessing vector computer at Argonne National Laboratory. These experiments will indicate speedups (relative to computations done with one CPU or CE and with no vectorization) of about ten for the SOR multisplitting method, and six for the SSOR multisplitting preconditioned conjugate gradient method.

2. Weighting schemes for the multisplitting algorithm. In order to simplify our discussion of (3) with (4), we initially restrict our consideration to A of the form

$$A = \begin{bmatrix} A_1 & -C_{12} & -C_{13} \\ -C_{21} & A_2 & -C_{23} \\ -C_{31} & -C_{32} & A_3 \end{bmatrix} \quad \text{where } A_k = M_k - N_k.$$

The nodes have been partitioned into P_1, P_2, and P_3 as given in Fig. 1. Consider two splittings of A given by

$$B_1 = \begin{bmatrix} M_1 & & \\ & M_2 & \\ -C_{31} & & M_3 \end{bmatrix} \quad \text{with } D_1 = \begin{bmatrix} I & & \\ & 0 & \\ & & d_{13}I \end{bmatrix}$$

$$B_2 = \begin{bmatrix} M_1 & & \\ & M_2 & \\ & -C_{32} & M_3 \end{bmatrix} \quad \text{with } D_2 = \begin{bmatrix} 0 & & \\ & I & \\ & & d_{23}I \end{bmatrix}.$$

Assume $d_{13} + d_{23} = 1$ so that $D_1 + D_2 = I$. The multisplitting form of the SOR algorithm is given by

$$M_k = \frac{1}{\omega}(D^{(k)} - \omega L^{(k)}) \quad \text{where}$$

$$D^{(k)} = \text{diag}(a_{ii}), \quad i \in P_k$$

$$L^{(k)} = (l_{ij}^k)$$

$$l_{ij}^k = \begin{cases} -a_{ij}, & j < i, i, j \in P_k \\ 0, & j \geqq i. \end{cases}$$

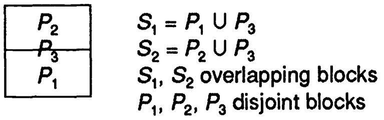

P_2
P_3
P_1
$S_1 = P_1 \cup P_3$
$S_2 = P_2 \cup P_3$
S_1, S_2 overlapping blocks
P_1, P_2, P_3 disjoint blocks

FIG. 1. *Blocks of nodes.*

Serial SOR. The SOR algorithm is given by a single splitting $A = B - C$, where

$$B = \begin{bmatrix} M_1 & & \\ -C_{21} & M_2 & \\ -C_{31} & -C_{32} & M_3 \end{bmatrix}.$$

Multisplitting (3) *and* (4) *with* $0 \leqq \lambda \leqq 1$. The above B_k matrices have block elementary form, and so,

$$B_1^{-1} = \begin{bmatrix} M_1 & & \\ & M_2 & \\ -C_{31} & & M_3 \end{bmatrix}^{-1} = \begin{bmatrix} M_1^{-1} & & \\ & M_2^{-1} & \\ M_3^{-1}C_{31}M_1^{-1} & & M_3^{-1} \end{bmatrix},$$

$$B_2^{-1} = \begin{bmatrix} M_1^{-1} & & \\ & M_2^{-1} & \\ & M_3^{-1}C_{32}M_2^{-1} & M_3^{-1} \end{bmatrix},$$

$$\begin{aligned} G_\lambda &= D_1^\lambda B_1^{-1} D_1^{1-\lambda} + D_2^\lambda B_2^{-1} D_2^{1-\lambda} \\ &= \begin{bmatrix} M_1^{-1} & & \\ & 0 & \\ d_{13}^\lambda M_3^{-1}C_{31}M_1^{-1} & 0 & d_{13}^\lambda M_3^{-1} d_{13}^{1-\lambda} \end{bmatrix} + \begin{bmatrix} 0 & & \\ & M_2^{-1} & \\ 0 & d_{23}^\lambda M_3^{-1}C_{32}M_2^{-1} & d_{23}^\lambda M_3^{-1} d_{23}^{1-\lambda} \end{bmatrix} \\ &= \begin{bmatrix} M_1^{-1} & & \\ & M_2^{-1} & \\ d_{13}^\lambda M_3^{-1}C_{31}M_1^{-1} & d_{23}^\lambda M_3^{-1}C_{32}M_2^{-1} & M_3^{-1} \end{bmatrix} \quad \text{as } d_{13} + d_{23} = 1. \end{aligned}$$

G_λ^{-1} exists and so (3) may be represented by a single splitting $A = M - N$, where

$$M = G_\lambda^{-1} = \begin{bmatrix} M_1 & & \\ & M_2 & \\ -d_{13}^\lambda C_{31} & -d_{23}^\lambda C_{32} & M_3 \end{bmatrix}.$$

Remark. If A has dissection form ($C_{21} = C_{12} = 0$), then $G_o = B$ is the serial SOR algorithm. If C_{21} or C_{12} are not zero, then we may consider these splittings as "incomplete" SOR splittings.

Several interesting questions will be answered in the following theorems. Under what conditions on A and the multisplittings will the scheme in (3) and (4) converge? How does the rate of convergence depend on the parameter λ in (4)? We first consider a general form of G and assume nonnegativity conditions as in a weak regular splitting.

THEOREM 1. *Let G and A be given with A nonsingular. Consider the algorithm* (3), *where H is defined by $H = I - GA$. Assume A^{-1}, H, $G \geqq 0$. Then the following are equivalent*:

(i) *The algorithm* (3) *converges and its limit is the solution of* (1).

(ii) $\rho(H) < 1$.

(iii) *Each row of G has at least one nonzero component.*

(iv) *G is nonsingular.*

Proof. The first four implications of the following proof are well known. In the literature [9] and [10], the condition (iii) implies (i) is implicitly stated in the proof of convergence for an algorithm associated with a single weak regular splitting. Condition (iii) is usually verified to determine convergence.

(iv) $\Rightarrow$ (iii): If (iii) is not true, then $\det(G) = 0$ and G is singular.

(ii) $\Rightarrow$ (iv): By Theorem 2.45 in [10] $I - H$ is nonsingular, and so, $G = (I - H)A^{-1}$ is nonsingular.

(iv) $\Rightarrow$ (ii): $GA = I - H$. Since both G, A are nonsingular, $I - H$ is nonsingular. By Theorem 2.45 in [10], $\rho(H) < 1$.

(i) $\Rightarrow$ (ii): By the definition of H, $x^{m+1} = Hx^m + Gd$ and $H \geqq 0$. Since x^{m+1} converges, $\rho(H) < 1$.

(iii) $\Rightarrow$ (i): If $I + \cdots + H^m$ is uniformly bounded, we are done. Since G and H are nonnegative,

$$\begin{aligned} 0 \leqq (I + \cdots + H^m)G &= (I + \cdots + H^m)(I - H)A^{-1} \\ &= (I - H^{m+1})A^{-1} \\ &\leqq A^{-1}. \end{aligned}$$

Let $I + \cdots + H^m = (s_{ij}^m)$, $G = (g_{ij})$ and $A^{-1} = (a_{ij})$. For all i and j

$$0 \leqq \sum_{\ell} s_{i\ell}^m g_{\ell j} \leqq a_{ij}.$$

Let ℓ be fixed and choose $j = j(\ell)$ such that $g_{\ell, j(\ell)} > 0$. Since $s_{i\ell}^m$, $g_{\ell j}$, $a_{ij} \geqq 0$,

$$s_{i\ell}^m g_{\ell, j(\ell)} \leqq a_{i, j(\ell)}$$

$$s_{i\ell}^m \leqq a_{i, j(\ell)} / g_{\ell, j(\ell)}.$$

Since the right side is independent of m, algorithm (3) converges. The convergence as given in (i) $\Rightarrow$ (ii) implies $\rho(H) < 1$ and, therefore, G is nonsingular. Thus, for $x^m \rightarrow x$ $x = Hx + Gd = (I - GA)x + Gd$. $GAx = Gd$, and consequently, $Ax = d$. □

The following three corollaries are proved by using condition (iii) of the above theorem.

COROLLARY 1. *Consider problem* (1) *with algorithm* (3) *and* $G = G_1$ *in* (4). *If* $A^{-1} \geqq 0$ *and* $A = B_k - C_k$ *are weak regular splittings, then* (3) *converges to the solution of* (1).

Proof. This is Theorem 1(a) in [9]. By assumption $B_k^{-1} \geqq 0$ and $B_k^{-1} C_k \geqq 0$. Thus, $G = G_1 = \sum D_k B_k^{-1} \geqq 0$ and $GA = \sum D_k B_k^{-1} A = \mathrm{I} - \sum D_k B_k^{-1} C_k$, and so, $H = \sum D_k B_k^{-1} C_k \geqq 0$. It remains to show that condition (iii) in Theorem 1 holds. Let $G = (g_{ij})$ and $\sum (d_i^k h_{ij}^k) = \sum D_k B_k^{-1}$, where $D_k = \operatorname{diag}(d_i^k)$ and $B_k^{-1} = (h_{ij}^k)$. Let i be fixed. Since $\sum D_k = I$, there is some $k = k(i)$ and that $d_i^{k(i)} > 0$. Thus, $g_{ij} = \sum d_i^k h_{ij}^k \geqq d_i^{k(i)} h_{ij}^{k(i)}$. Since each B_k^{-1} exists, every row of B_k^{-1} must have a nonzero component, that is, $h_{i,j(i)}^{k(i)} > 0$. Thus $g_{i,j(i)} > 0$. □

In the next corollary we assume $G_o, H_o \geqq 0$. Theorem 2 gives a class of multisplittings which satisfy $G_o, H_o \geqq 0$.

COROLLARY 2. *Consider problem* (1) *with algorithm* (3) *and* $G = G_o$ *in* (4). *If* A^{-1}, G_o, $H_o \geqq 0$ *and* $A = B_k - C_k$ *are weak regular splittings, then* (3) *converges to the solution of* (1).

Proof. We show condition (iii) in Theorem 1 holds. Let $G_o = (g_{ij}) = \sum (h_{ij}^k d_j^k) = \sum B_k^{-1} D_k$, where $D_k = \operatorname{diag}(d_j^k)$ and $B_k^{-1} = (h_{ij}^k)$. Let $k = k(j)$ such that $d_j^{k(j)} > 0$, fix i, and note $g_{ij} \geqq h_{ij}^{k(j)} d_j^{k(j)}$. Since B_k^{-1} exists, every row of $B_{k(j)}^{-1} = (h_{ij}^{k(j)})$ must have at least one nonzero component, that is, $h_{i,j(i)}^{k(j(i))} > 0$. Therefore, $g_{i,j(i)} \geqq h_{i,j(i)}^{k(j(i))} d_{j(i)}^{k(j(i))} > 0$. □

COROLLARY 3. *Consider problem* (1) *with algorithm* (3) *and* $G = G_\lambda$ *in* (4) *and* $0 \leqq \lambda \leqq 1$. *If* A^{-1}, G_λ, $H_\lambda \geqq 0$ *and* $A = B_k - C_k$ *are weak regular splittings, then* (3) *converges to the solution of* (1).

Proof. As above let $G_\lambda = (g_{ij}) = \sum (d_i^k)^\lambda h_{ij}^k (d_j^k)^{1-\lambda}$. Let $d_i^{k(i)} > 0$ and let $j = i$, $g_{ii} \geqq d_i^{k(i)} h_{ii}^{k(i)} > 0$. □

In order to apply Theorem 1 to algorithm (3) with $G = G_\lambda$ for $0 \leqq \lambda < 1$, we must make further restrictions on the multisplitting. We consider K splittings. This corresponds to K overlapping block similar to that illustrated for $K = 2$ in Fig. 1.

$$A = \begin{bmatrix} \ddots & & & & \\ & A_k & \cdots & -C_{kj} & \\ & \vdots & \ddots & \vdots & \\ & -C_{jk} & \cdots & A_j & \\ & & & & \ddots & \\ & & & & & A_{K+1} \end{bmatrix}, \quad 1 \leqq k, j \leqq K+1$$

$$A_k = M_k - N_k$$

$$(5) \qquad B_k = \begin{bmatrix} \ddots & & & \\ & M_k & & \\ & & \ddots & \\ & & & \ddots & \\ & -C_{K+1,k} & & & M_{K+1} \end{bmatrix}, \quad 1 \leqq k \leqq K$$

$$D_k = \begin{bmatrix} 0 & & & & \\ & \ddots & & & \\ & & I & & \\ & & & \ddots & \\ & & & & 0 & \\ & & & & & d_{k,K+1} I \end{bmatrix}, \quad I \text{ is in the } k\text{th block.}$$

Assume $A_{K+1} = M_{K+1} - N_{K+1}$ is block diagonal, for example, see [16]. Let $0 \leqq d_{k,K+1}$ be diagonal matrices that are constant on these blocks and $\sum_{k=1}^{K} d_{k,K+1} = I$. Then $M_{K+1} d_{k,K+1}^\lambda M_{K+1}^{-1} = d_{k,K+1}^\lambda$; this is used below in the derivation of line (6).

THEOREM 2. *Consider the multisplitting of A given by* (5). *Assume $A^{-1} \geqq 0$, $A_k = M_k - N_k$ are weak regular splittings and $C_{kj} \geqq 0$ for $1 \leqq k, j \leqq K + 1$. Then algorithm* (3) *with $G = G_\lambda$ in* (4) *for $0 \leqq \lambda \leqq 1$ converges to the solution of* (1).

Proof. Use Theorem 1 by showing $G_\lambda \geqq 0$, $H_\lambda \geqq 0$, where $G_\lambda A = I - H_\lambda$ and G_λ satisfies (iii) or (iv). The reader may find the notation for $K = 2$ or 3 easier to deal with than the general case as presented below.

$$G_\lambda = \sum_{k=1}^{K} D_k^\lambda B_k^{-1} D_k^{1-\lambda}$$

$$= \sum_{k=1}^{K} D_k^\lambda \begin{bmatrix} \ddots & & \\ & M_k^{-1} & \\ & & \ddots & \\ & M_{K+1}^{-1} C_{K+1,k} M_k^{-1} & & M_{K+1}^{-1} \end{bmatrix} D_k^{1-\lambda}$$

$$= \begin{bmatrix} \ddots & & & \\ & M_k^{-1} & & \\ & & \ddots & \\ \cdots & d_{k,K+1}^\lambda M_{K+1}^{-1} C_{K+1,k} M_k^{-1} & \cdots & M_{K+1}^{-1} \end{bmatrix} \quad \text{as } \sum_{k=1}^{K} d_{k,K+1} = I.$$

Since $M_k^{-1} \geqq 0$ and $C_{K+1,k} \geqq 0$, $G_\lambda \geqq 0$. Since M_k^{-1}, M_{K+1}^{-1} exist, G_λ^{-1} exists and equals

$$G_\lambda^{-1} = \begin{bmatrix} \ddots & & & & \\ & & M_k & & \\ & & & \ddots & \\ \cdots & -d^\lambda_{k,K+1}C_{K+1,k} & & \cdots & M_{K+1} \end{bmatrix}. \tag{6}$$

In order to show $H_\lambda \geqq 0$, compute $G_\lambda A = I - H_\lambda$.

$$G_\lambda A = \begin{bmatrix} \ddots & & & & & \\ & I-M_k^{-1}N_k & \cdots & -M_k^{-1}C_{kj} & & \\ & \vdots & \ddots & \vdots & & \\ & -M_j^{-1}C_{jk} & \cdots & I-M_j^{-1}N_j & & \\ & & & & \ddots & \\ \cdots & \gamma_k & \cdots & & & \gamma_{K+1} \end{bmatrix} \begin{matrix} 1\leqq j\leqq K+1 \\ 1\leqq k\leqq K \end{matrix} \quad \text{where}$$

$$\gamma_k = [\cdots d^\lambda_{k,K+1}M_{K+1}^{-1}C_{K+1,k}M_k^{-1}\cdots M_{K+1}^{-1}] \begin{bmatrix} -C_{1k} \\ \vdots \\ -C_{k-1,k} \\ M_k - N_k \\ -C_{k+1,k} \\ \vdots \\ -C_{K+1,k} \end{bmatrix}$$

$$= (\text{nonpositive terms}) + d^\lambda_{k,K+1}M_{K+1}^{-1}C_{K+1,k}(I-M_k^{-1}N_k) - M_{K+1}^{-1}C_{K+1,k}$$

$$= (\text{nonpositive terms}) + (I-d^\lambda_{k,K+1})M_{K+1}^{-1}C_{K+1,k} - d^\lambda_{k,K+1}M_{K+1}^{-1}C_{K+1,k}M_k^{-1}N_k$$

$$= \text{nonpositive, for } 1\leqq k\leqq K, M_k^{-1}N_k \geqq 0, M_k^{-1}\geqq 0, C_{kj}\geqq 0. \text{ Also,}$$

$$\gamma_{K+1} = [\cdots d^\lambda_{k,K+1}M_{K+1}^{-1}C_{K+1,k}M_k^{-1}\cdots M_{K+1}^{-1}] \begin{bmatrix} \vdots \\ -C_{k,K+1} \\ \vdots \\ M_{K+1}-N_{K+1} \end{bmatrix}$$

$$= -\sum_{k=1}^{K} d^\lambda_{k,K+1}M_{K+1}^{-1}C_{K+1,k}M_k^{-1}C_{k,K+1} + I - M_{K+1}^{-1}N_{K+1}$$

$$= (\text{nonpositive terms}) + I.$$

Thus, $G_\lambda A = I + (\text{nonpositive terms})$, and so, $H_\lambda \geqq 0$. □

The next theorem is a comparison result with respect to λ for the algorithm given in Theorem 2. We will use the comparison Theorem 3.15 given in Varga [14], which we state here as Theorem 3. In both Theorems 3 and 4 notice the stronger condition $A^{-1} > 0$ and regular splittings.

THEOREM 3. *Let $A = B - C = \bar{B} - \bar{C}$ be regular splittings of A and $A^{-1} > 0$. If $C \geqq \bar{C}$ with equality excluded, then*

$$\rho(\bar{B}^{-1}\bar{C}) < \rho(B^{-1}C) < 1.$$

THEOREM 4. *Consider the multisplitting of A given by (5). Assume $A^{-1} > 0$, $A_k = M_k - N_k$ are regular splittings of A_k, and $C_{kj} \geqq 0$ for $1 \leqq k, j \leqq K+1$. Also assume $C_{K+1,k} \neq 0$ for $1 \leqq k \leqq K$. Let H_λ be defined by the algorithm (3), where $G = G_\lambda$ in (4) and $G_\lambda A = I - H_\lambda$. If $\bar{\lambda} < \lambda$, then*

$$\rho(H_{\bar{\lambda}}) < \rho(H_\lambda) < 1.$$

Proof. By line (6) in the proof of Theorem 2, H_λ may be given by a single splitting of $A = B_\lambda - C_\lambda$, where $B_\lambda = G_\lambda^{-1}$ and $C_\lambda = B_\lambda - A$

$$(7)\qquad C_\lambda = \begin{bmatrix} \ddots & & & & \\ & N_k & \cdots & C_{kj} & \\ & \vdots & \ddots & \vdots & \\ & C_{jk} & \cdots & N_j & \\ & & & & \ddots \\ \cdots & \tilde{C}_{K+1,k} & \cdots & \tilde{C}_{K+1,j} & \cdots N_{K+1} \end{bmatrix}, \quad \tilde{C}_{K+1,k} \equiv (I - d_{k,K+1}^{\lambda})C_{K+1,k}.$$

Since $0 \leqq \bar{\lambda} < \lambda \leqq 1$ and $0 < d_{k,K+1} < I$ for some k, $I - d_{k,K+1}^{\lambda} > I - d_{k,K+1}^{\bar{\lambda}} > 0$. Since $C_{K+1,k} \geqq 0$, $C_\lambda \geqq C_{\bar{\lambda}}$ with equality excluded when $C_{K+1,k} \neq 0$. By Theorem 3

$$\rho(H_{\bar{\lambda}}) = \rho(B_{\bar{\lambda}}^{-1} C_{\bar{\lambda}}) < \rho(B_\lambda^{-1} C_\lambda) = \rho(H_\lambda) < 1. \qquad \square$$

The above theorem suggests that when forming multisplitting algorithms of the form (3) with $G = G_\lambda$, one ought to use preweighting, that is, $\lambda = 0$. Furthermore, the remark before Theorem 1 also holds for the more general multisplitting given by (5) if $A_k = M_k - N_k$ represents the SOR algorithm for the kth block. In particular, if A has dissection form ($C_{kj} = C_{jk} = 0$ for $1 \leqq k, j \leqq K$), then the algorithm for $\lambda = 0$ is exactly the serial SOR algorithm.

The next result considers the case when A is assumed to be symmetric and positive definite. We will utilize the Householder–John theorem [5], which we state here as Theorem 5.

THEOREM 5 (Householder–John). *Let $A = B - C$ be Hermitian and $B^* + C$ be positive definite. Then $\rho(B^{-1}C) < 1$ if and only if A is positive definite.*

THEOREM 6. *Consider the multisplitting of A given by (5) and the algorithm given by (3) and $G = G_\lambda$ in (4). If*

(i) *A is a real symmetric positive definite and*

(ii) *$B_\lambda^T + C_\lambda$ is positive definite where*

$$(8)\qquad B_\lambda^T + C_\lambda = \begin{bmatrix} \ddots & & & & \\ & M_k^T + N_k & \cdots & C_{kj} & \tilde{C}_{k,K+1} \\ & \vdots & \ddots & \vdots & \\ & C_{jk} & \cdots & M_j^T + N_j & \tilde{C}_{j,K+1} \\ & & & & \ddots \\ \cdots & \tilde{C}_{K+1,k} & \cdots & \tilde{C}_{K+1,j} & \cdots M_{K+1}^T N_{K+1} \end{bmatrix}$$

$$\tilde{C}_{k,K+1} = (I - d_{k,K+1}^{\lambda})C_{k,K+1}$$

$$\tilde{C}_{K+1,k} = (I - d_{k,K+1}^{\lambda})C_{K+1,k} = (I - d_{k,K+1}^{\lambda})C_{k,K+1}^T,$$

then the algorithm (3) converges to the solution of (1).

Proof. We need to show $\rho(B_\lambda^{-1}C_\lambda) < 1$, where $B_\lambda = G_\lambda^{-1}$ in line (6) and $C_\lambda = A - B_\lambda$ in line (7). An easy calculation gives $B_\lambda^T + C_\lambda$ in line (8). Thus by the Householder–John theorem $\rho(B_\lambda^{-1}C_\lambda) < 1$. $\square$

Remarks. (1) When

$$A_k = M_k - N_k = (1/\omega)(D^{(k)} - \omega L^{(k)}) - (1/\omega)((1-\omega)D^{(k)} + \omega L^{(k)T})$$

($D^{(k)}$, $L^{(k)}$ for $1 \leqq k \leqq K+1$ are defined as in the beginning of § 2) and A has dissection form ($C_{kj} = C_{jk} = 0$ for $1 \leqq j, k \leqq K$), then $B_o^T + C_o = (2-\omega)/\omega D$, where $A = D - L - L^T$. Thus, Theorem 5 may be considered as a generalization of classical convergence result for the SOR algorithm, $0 < \omega < 2$ and real symmetric positive definite matrices.

(2) As indicated in [16], $B_\lambda^T + C_\lambda$ being positive definite may be viewed as a further constraint on ω. In [16] it is shown that $0 < \omega \leqq \omega_o < 2$ implies, in many cases, $B_\lambda^T + C_\lambda$ is positive definite.

3. Multisplitting preconditioned conjugate gradient method. In this section we consider (1) with A symmetric positive definite. Preconditioners, M, are usually associated with single splitting of $A = M - N$ and are also symmetric and positive definite. The idea is to choose M so that $M^{-1}A$ approximates the identity matrix, and $Mz = r$ can "easily" be solved. This is more precisely stated in Golub and Van Loan [3] or Ortega and Rheinboldt [10]. Several preconditioners are now described.

SSOR *preconditioner*. The SSOR preconditioner is given by a forward and backward sweep of SOR. Let $A = D - L - U$. The SSOR iteration matrix is

$$G=\left[\frac{1}{\omega}(D-\omega U)\right]^{-1}\frac{1}{\omega}[(1-\omega)D+\omega L]\left[\frac{1}{\omega}(D-\omega L)\right]^{-1}\frac{1}{\omega}[(1-\omega)D+\omega U].$$

By using the identity $H(I-H)^{-1} = (I-H)^{-1} - I$ for invertible $I - H$ and using routine matrix algebra, we may calculate M^{-1} from $A = M - N$ and $M^{-1}N = G$, namely,

$$M^{-1}=\left[\frac{1}{\omega}(D-\omega U)\right]^{-1}\frac{2-\omega}{\omega}D\left[\frac{1}{\omega}(D-\omega L)\right]^{-1}. \tag{9}$$

This is the M in the algorithm because $Mz_{n-1} = r_{n-1} = d - Ax_{n-1}$ and $z_{n-1} = M^{-1}(b + Nx_{n-1}) - x_{n-1}$. This form of M^{-1} requires less computational effort. The importance of this preconditioner is that it significantly reduces the number of iterations needed for convergence and the overall computation time.

Vector preconditioners have been investigated by Poole and Ortega [12]. Jalby, Meier, and Sameh [6] have studied preconditioners that use multiprocessors systems. Adams and Ong [1] have introduced an additive preconditioner that can be viewed as a multisplitting.

Additive preconditioner. Let $A = B - C$ and A be symmetric. Then a second splitting of A is $A = A^T = B^T - C^T$. So, $K = 2$ with $D_1 = \frac{1}{2}I$ and $D_2 = \frac{1}{2}I$ gives $H = D_1B_1^{-1}C_1 + D_2B_2^{-1}C_2 = \frac{1}{2}B^{-1}C + \frac{1}{2}B^{-T}C^T$ and $G = D_1B_1^{-1} + D_2B_2^{-1} = \frac{1}{2}(B^{-1} + B^{-T})$. Note, G is symmetric. Adams and Ong [1] used this to construct an additive preconditioner where $B = (1/\omega)(D - \omega L)$. Then $M^{-1} = \frac{1}{2}(B^{-1} + B^{-T}) = \frac{1}{2}((1/\omega)(D - \omega L))^{-1} + \frac{1}{2}((1/\omega)(D - \omega L))^{-T}$.

Symmetric multisplitting preconditioner. Let $A = B_k - C_k$ and B_k be symmetric positive definite. The matrix $\sum_{k=1}^K D_kB_k^{-1}$ may not be symmetric. We can use the additive preconditioner with $B = (\sum_{k=1}^K D_kB_k^{-1})^{-1}$. This gives the preconditioner

$$\begin{aligned} M^{-1} &= \frac{1}{2}\left(\sum_{k=1}^K D_kB_k^{-1} + \left(\sum_{k=1}^K D_kB_k^{-1}\right)^T\right) \\ &= \frac{1}{2}\left(\sum_{k=1}^K D_kB_k^{-1} + \sum_{k=1}^K B_k^{-T}D_k\right) \qquad (10) \\ &= \frac{1}{2}\sum_{k=1}^K (D_kB_k^{-1} + B_k^{-1}D_k). \end{aligned}$$

This preconditioner requires two calculations with each B_k^{-1}. In the following, symmetric weights allow one to reduce the computation almost in half.

Symmetric weighted multisplitting preconditioner. Let $A = \mathbf{B}_k - \mathbf{C}_k$, where $\mathbf{B}_k$ is symmetric positive definite. Use the symmetric weight $G_{1/2}$ for the multisplitting algorithm in (3)

$$M^{-1} = \sum_{k=1}^{K} D_k^{1/2} \mathbf{B}_k^{-1} D_k^{1/2} \tag{11}$$

THEOREM 7. *Consider* (11), *where* $\mathbf{B}_k$ *are symmetric positive definite. Then* M^{-1} *is symmetric positive definite.*

Proof. The following calculation shows that M^{-1} is symmetric;

$$\begin{aligned}(M^{-1})^T &= \left(\sum_{k=1}^{K} D_k^{1/2} \mathbf{B}_k^{-1} D_k^{1/2} \right)^T \\ &= \sum_{k=1}^{K} D_k^{1/2} \mathbf{B}_k^{-T} D_k^{1/2} \\ &= \sum_{k=1}^{K} D_k^{1/2} \mathbf{B}_k^{-1} D_k^{1/2} \\ &= M^{-1}.\end{aligned}$$

Next we show that M^{-1} is positive definite. Let $x \neq 0$ and assume $D_k^{1/2} x \neq 0$ for some $k = k_o$ that

$$\begin{aligned}x^T M^{-1} x &= \sum_{k=1}^{K} x^T D_k^{1/2} \mathbf{B}_k^{-1} D_k^{1/2} x \\ &= (D_{k_o}^{1/2} x)^T \mathbf{B}_{k_o}^{-1} (D_{k_o}^{1/2} x) + \sum_{k \neq k_o} (D_k^{1/2} x)^T \mathbf{B}_k^{-1} (D_k^{1/2} x).\end{aligned}$$

Since $\mathbf{B}_k^{-1}$ is symmetric positive definite, the terms in $\sum_{k \neq k_o}$ are nonnegative and $(D_{k_o}^{1/2} x)^T \mathbf{B}_{k_o}^{-1} (D_{k_o}^{1/2} x) > 0$. Thus, $x^T M^{-1} x > 0$. □

A particular choice of $\mathbf{B}_k^{-1}$ in (11) is taken from the serial SSOR in (9) applied to the kth block. Let B_k^{-1} be as in (5) with $M_k = (1/\omega)(D^{(k)} - \omega L^{(k)})$ and $N_k = (1/\omega)((1-\omega)D^{(k)} + \omega L^{(k)T})$. Define $\mathbf{B}_k^{-1}$ as follows

$$\mathbf{B}_k^{-1} = B_k^{-T} \frac{2-\omega}{\omega} D B_k^{-1}. \tag{12}$$

The calculations given in the next section will indicate the relative effectiveness of the preconditioner given by (11) and (12).

4. Numerical experiments on the Alliant FX/8. In this section we consider the numerical solution of the algebraic problem obtained from the five-point finite-difference method applied to the elliptic partial differential equation

$$\begin{aligned}-\Delta u &= 10.0 \quad \text{on} \quad \Omega = (0,1) \times (0,1) \\ u &= 0.0 \quad \text{on} \quad \partial\Omega.\end{aligned} \tag{13}$$

The resulting algebraic equation is scaled by

$$(D^{-1/2} A D^{-1/2})(D^{1/2} u) = D^{1/2} d$$

where D is the diagonal of A. The multisplitting version of SOR and the multisplitting version of the SSOR preconditioned conjugate gradient methods will be illustrated.

The calculations were done on the Alliant FX/8 at Argonne National Laboratory. The Alliant FX/8 has eight processors or CEs (computational elements). Each CE has a vector pipeline and processes data in 32 64-bit groups. There is a high-speed local cache of 512 KB, and the global memory is 64 MB. The software has parallel extensions of C and FORTRAN and includes vector operations. The parallelism is loop based and can be turned on or off by either compiler options or simple directives before each loop. Concurrency is distributed to the outer loops and vectorization is done in the inner loops. We used the following compiler options:

$-0g$	optimized serial code (one CE),
$-0gv$	optimized serial code with vectorization,
$-0gc$	optimized code with eight CEs,
-0	optimized code with eight CEs and vectorization.

The multisplittings were formed by considering overlapping blocks of Ω. Fig. 2 indicates Ω with four blocks which overlap by one row.

In general, let $\{1, \cdots, N\}$ represent the unknowns and $\{1, \cdots, N\} = \cup_{k=1}^{K} S_k$, where S_k may overlap. Consider the component form of algorithm (3) with $G = G_\lambda$ and $D_k = \text{diag}\,(d_i^k)$ with $d_i^k = 0$ for $i \notin S_k$, $d_i^k \geqq 0$ and $\sum_{k=1}^{K} d_i^k = 1$.

SOR *version of Algorithm* (3) *with* $G = G_\lambda$ *in* (4). Let $A = (a_{ij})$ and $d = (d_i)$. Then

$$\left.\begin{aligned} r_i^m &= d_i - \sum_j a_{ij} u_j^m \\ r_i^{k,m+1} &= \omega\left((d_i^k)^{1-\lambda} r_i^m - \sum_{\substack{j<i \\ j \in S_k}} a_{ij} r_j^{k,m+1} \right) \Big/ a_{ii} \\ u_i^{m+1} &= u_i^m + \sum_{k=1}^{K} (d_i^k)^\lambda r_i^{k,m+1}. \end{aligned}\right\} \tag{14}$$

The following calculations were for $N = 57 \times 57 = 3{,}249$ and the blocks were as in Fig. 2, with only one overlapping row of unknowns. The stopping criteria was

$$|u_i^{m+1} - u_i^m| \leqq 0.0001 \quad \text{for all } i.$$

The values for the SOR parameter ω were optimized by numerical experimentation to within ±0.005. The nodes were ordered as in (5), where the overlapping nodes were listed last. Within the disjoint blocks the nodes were listed by rows, starting with the bottom row of each disjoint block. Table 1 indicates the effect of different compiler options. In all calculations the SOR parameter was $\omega = 1.900$. The numbers from those in White [16] are indicated in Table 1 by parenthesis. There are two possible reasons for time differences. First, in [16] red-black ordering was used within the disjoint blocks, and consequently, one expects to obtain faster vectorized code. Second, the local high-speed cache for the calculations in Table 1 was 512 KB. This is four times larger than that used in [16].

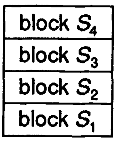


FIG. 2. Ω *with four blocks.*

TABLE 1
$K = 1$ and compiler options.

Option	Times	(Times in [16] with RB)
$-0g$	11.75	(15.36)
$-0gv$	5.54	(4.56)
$-0gc$	8.15	(10.08)
-0	4.23	(2.77)

In our discussion *speedup* will refer to the ratio of the time for the given method to the time for calculation with the $-0g$ option (one CE and no vectorization). The differences in the speedups in Table 1 are primarily due to the red-black ordering used in [16]. The red-black ordering cannot be used for algorithms of the form (3) with $G = G_\lambda$ in (4). This is because a permutation may destroy the property of a matrix being lower triangular, for example,

$$1,2,3 \rightarrow 1,3,2$$

$$\begin{bmatrix} a & & \\ & a & \\ b & c & a \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a & & \\ & a & \\ b & c & a \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} a & 0 & 0 \\ b & a & c \\ 0 & 0 & a \end{bmatrix}.$$

However, some permutations, such as the diagonal ordering, do not have this undesirable property. This will be described in the discussion of the preconditioned conjugate gradient method.

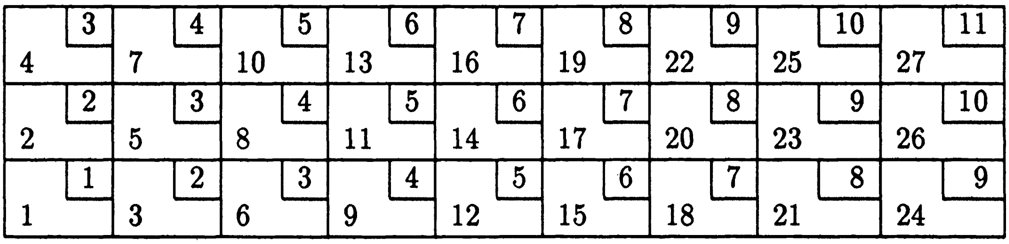
Table 2 contains different λ and different numbers of blocks K. The -0 option was used in all calculations. The decreasing times as λ decreases verifies the theoretical results contained in Theorem 4. The best speedup of about 10 was obtained by letting $\lambda = 0$ and $K = 8$.

The convergence rate for the SOR method can be very sensitive to the choice of SOR parameter ω. This is especially true when the number of unknowns is large. The conjugate gradient method is a good alternative, and the serial SSOR preconditioner serves as a good preconditioner when the choice of SOR parameter does not radically alter convergence rates. In §3 we mentioned three preconditioners. The serial SSOR (8) and the symmetric weighted multisplitting (11) preconditioners seem to be the best for our example (13). When the -0 option was used (see Table 4), the speedups ranged from 4.5 to 6.4.

In our numerical experiments we had $N = 191 \times 191 = 36{,}481$ unknowns. The stopping criteria was $r_n^T M^{-1} r_n < 10.0e - 6$, where $r_n = d - Ax^n$ and M is the preconditioner. The SOR parameter was found by experimentation and 64-bit reals were used. The following is the component form of the preconditioner in (11) with $\mathbf{B}_k^{-1}$ in (12).

TABLE 2
Multisplitting algorithm with $G = G_\lambda$.

K	λ	Iterations	ω	Times
2	1	184	1.925	3.64
2	$\frac{1}{2}$	145	1.920	3.29
2	0	128	1.900	2.90
4	0	126	1.900	1.66
8	0	125	1.900	1.20



4 $\boxed{3}$	7 $\boxed{4}$	10 $\boxed{5}$	13 $\boxed{6}$	16 $\boxed{7}$	19 $\boxed{8}$	22 $\boxed{9}$	25 $\boxed{10}$	27 $\boxed{11}$
2 $\boxed{2}$	5 $\boxed{3}$	8 $\boxed{4}$	11 $\boxed{5}$	14 $\boxed{6}$	17 $\boxed{7}$	20 $\boxed{8}$	23 $\boxed{9}$	26 $\boxed{10}$
1 $\boxed{1}$	3 $\boxed{2}$	6 $\boxed{3}$	9 $\boxed{4}$	12 $\boxed{5}$	15 $\boxed{6}$	18 $\boxed{7}$	21 $\boxed{8}$	24 $\boxed{9}$

n = node number $\boxed{\mathrm{m}}$ = vector number in pipe of length equal to three

FIG. 3. *Diagonal ordering.*

The weighting scheme in (11) is symmetric about $\mathbf{B}_k^{-1}$ and implies that the preconditioner is symmetric, as is established in Theorem 7.

Symmetric weighted SSOR *multisplitting preconditioner*. The computation of $M^{-1}r$, where M is the preconditioner:

$$(15)\qquad \left.\begin{array}{ll} a_{ii}u_i^{k,m+1} = \omega(d_i^k)^{1/2}r_i - \omega \displaystyle\sum_{\substack{j<i\\ j\in S_k}} a_{ij}u_j^{k,m+1}, & \begin{array}{c} j\in S_k\\ j \text{ increases}\end{array} \\ v_i^{k,m+1} = (2-\omega)a_{ii}u_i^{k,m+1} - \omega \displaystyle\sum_{\substack{j>i\\ j\in S_k}} a_{ij}v_j^{k,m+1} & \begin{array}{c} j\in S_k\\ j \text{ decreases}\end{array} \\ [M^{-1}r]_i = \displaystyle\sum_{k=1}^{K} (d_i^k)^{1/2} v_i^{k,m+1}. & \end{array}\right\}$$

Table 3 contains calculations with the serial SSOR preconditioner, different compiler options, and two types of orderings of the nodes. The row ordering is the classical ordering of nodes, starting with the bottom row and moving from left to right in each row. In order to take advantage of the vector pipelines in each CE, we have used diagonal ordering of the nodes as illustrated in Fig. 3. There are 27 unknowns with three rows and nine columns. The numbers in the big boxes indicate the diagonal numbering. The numbers in the small boxes indicate the nodes in a pipe that can be computed by vectorization. This ordering preserves the form of the existing lower triangular matrices. Note, the red-black ordering does not do this.

In Tables 3 and 4 the total time is broken into two parts, the cg time and the pg time. The cg time refers to the time required to compute the conjugate gradient portion (the preconditioner is the identity matrix), and the pg time is the time required to compute the preconditioner portion.

In all the calculations in Table 3 the SOR parameter was ω = 1.92, and it took 28 iterations to satisfy the stopping criteria. The effect of the diagonal ordering was significant,

TABLE 3
One block, K = 1.

K	Ordering	Option	cg time	pc time	Total time
1	diagonal	−0*g*	17.98	25.75	43.73
1	diagonal	−0*gv*	10.32	7.07	17.39
1	diagonal	−0*gc*	4.82	18.43	23.25
1	diagonal	−0	2.54	7.17	9.71
1	row	−0	2.54	16.10	18.64

TABLE 4
Variable blocks.

K	Overlap	Iterations	ω	cg time	pc time	Total time
1	—	28	1.92	2.54	7.17	9.71
2	1	38	1.90	3.35	4.99	8.34
4	1	45	1.86	4.10	3.94	8.04
8	1	48	1.85	4.29	3.23	7.52
8	3	44	1.83	4.02	3.15	7.17
8	5	41	1.86	3.71	3.10	6.81

and it allowed full use of the vector pipelines. The best speedup was given by the diagonal ordering and the −0 option (all eight CEs with vector operations) and was 4.5. In this calculation the conjugate gradient part of the code took 2.54 seconds and the preconditioned part took 7.17 seconds. The next table shows how one can reduce the time to compute the preconditioner (pc time) by using the multisplitting preconditioner.

Table 4 indicates that as the number of blocks increases, the number of iterations will increase to meet the stopping criteria. The fastest computation is given by $K = 8$ with an overlap of five rows between the blocks. The overall speedup is 6.4. In this case, the preconditioner time is reduced to 3.10 seconds, and the conjugate gradient time is 3.71. The speedup for the preconditioner was 8.3. The increased conjugate gradient time is a result of increasing the number of iterations from 28 to 41.

REFERENCES

[1] L. ADAMS AND E. G. ONG, *Additive polynomial preconditioners for parallel computers*, preprint.

[2] L. ELSNER, *Comparisons of weak regular splittings and multisplitting methods*, Third SIAM Conference on Applied Linear Algebra, Madison, WI, May 1988.

[3] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[4] L. J. HAYES, *A vectorized matrix-vector multiply and overlapping block iterative method*, in Supercomputer Applications, R. W. Numrich, ed., Plenum Press, New York, 1984, pp. 91–100.

[5] A. S. HOUSEHOLDER, *On the convergence of matrix iterations*, Oak Ridge National Laboratory Tech. Report 1883, 1953.

[6] W. JALBY, U. MEIER, AND A. SAMEH, *The behavior of conjugate gradient based algorithms on a multivector processor with a memory hierarchy*, Center for Supercomputer Research and Development, University of Illinois, Urbana, IL, 1986, preprint.

[7] O. A. MCBRYAN AND E. F. VAN DE VELDE, *Parallel algorithms for elliptic equations*, Commun. Pure and Appl. Math., 38 (1985), pp. 769–795.

[8] M. NEUMANN AND R. J. PLEMMONS, *Convergence of parallel multisplittings and iterative methods for M-matrices*, Linear Algebra Appl., 88/89 (1987), pp. 559–573.

[9] D. P. O'LEARY AND R. E. WHITE, *Multi-splittings of matrices and parallel solution of linear systems*, SIAM J. Algebraic Discrete Methods, 6 (1985), pp. 630–640.

[10] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[11] A. M. OSTROWSKI, *Iterative solution of linear systems of functional equations*, J. Math. Anal. Appl., 2 (1961), pp. 351–369.

[12] E. L. POOLE AND J. M. ORTEGA, *Multicoloring* ICCG *methods for vector computers*, SIAM J. Numer. Anal., 24 (1987), pp. 1394–1418.

[13] F. ROBERT, *Méthodes iterative série parallèle*, C.R. Acad. Sc. Paris, A-271 (1970), pp. 847–850.

[14] R. S. VARGA, *Matrix Iterative Analysis*, Prentice Hall, Englewood Cliffs, NJ, 1962.

[15] R. E. WHITE, *A nonlinear parallel algorithm with application to the Stefan problem*, SIAM J. Num. Anal., 23 (1986), pp. 639–652.

[16] ———, *Multisplittings of a symmetric positive definite matrix*, SIAM J. Matrix Anal. Appl., 11 (1990), to appear.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 4, pp. 494–499, October 1989

006

THE STRONG STABILITY OF ALGORITHMS FOR SOLVING SYMMETRIC LINEAR SYSTEMS*

JAMES R. BUNCH†, JAMES W. DEMMEL‡, AND CHARLES F. VAN LOAN§

Abstract. An algorithm for solving linear equations is stable on the class of nonsingular symmetric matrices or on the class of symmetric positive definite matrices if the computed solution solves a system that is near the original problem. Here it is shown that any stable algorithm is also strongly stable on the same matrix class if the computed solution solves a nearby problem that is also symmetric or symmetric positive definite.

Key words. stability, symmetric matrices

AMS(MOS) subject classifications. 65F05, 65F30, 65G05

1. Introduction. When applied to a linear system $Ax = b$, a stable algorithm for solving systems of linear equations produces a computed solution $\hat{x}$ that is the solution to a nearby system

$$\hat{A}\hat{x} = \hat{b},$$

where $\|\hat{A} - A\| / \|A\|$ is small and $\|\hat{b} - b\| / \|b\|$ is small, for some norm $\|\cdot\|$. How "small" is small enough depends on the accuracy desired in the solution (and on the condition number of A) [16, pp. 189–191]. A proof of the stability of an algorithm usually involves showing that $\|\hat{A} - A\| / \|A\|$ and $\|\hat{b} - b\| / \|b\|$ are bounded by $p(n)u$, where p is a low degree polynomial, n is the order of A, and u is the unit roundoff (machine precision). We would like $p(n)u << 1$.

In solving structured linear equations, it is often important that the perturbed matrix $\hat{A}$ have the same structure as A. For example, solving electrical network problems gives rise to symmetric systems of linear equations, $Ax = b$. If the computed solution $\hat{x}$ to $Ax = b$ satisfies $\hat{A}\hat{x} = \hat{b}$, but $\hat{A}$ is not symmetric, then the system $\hat{A}\hat{x} = \hat{b}$ could never have arisen from an electrical network problem. But if $\hat{A}$ is symmetric, then we hope that there is an electrical network near our original network that gives rise to the system $\hat{A}\hat{x} = \hat{b}$.

Another situation where it is important that the perturbed matrix remain symmetric is in the analysis of Algorithm 5 in [3]. That algorithm uses a variation of inverse iteration to find the eigenvectors of a certain class of symmetric matrices to high accuracy. The class includes all symmetric positive definite matrices that can be consistently ordered. The error analysis uses a new perturbation theorem about symmetric perturbations of symmetric matrices, and to apply it one needs to know that a nearby symmetric matrix exists which exactly satisfies the equations at each step of inverse iteration.

The term *strongly stable*, developed in [4], is used in this context. An algorithm for solving linear equations is *strongly stable* for a class of matrices **A** if for each A in **A**

* Received by the editors May 31, 1988; accepted for publication (in revised form) February 6, 1989.

† Department of Mathematics, C-012, University of California–San Diego, La Jolla, California 92093. The research of this author was supported by National Science Foundation grant DMS 83-18412.

‡ Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, New York, New York 10012. The research of this author was partially supported by a Presidential Young Investigator award.

§ Department of Computer Science, Cornell University, Ithaca, New York 14853. The research of this author was partially supported by Office of Naval Research contract N00014-83-K-0640 and National Science Foundation grant DCR 86-02310.

and for arbitrary b the computed solution $\hat{x}$ solves a nearby system $\hat{A}\hat{x} = \hat{b}$ with $\hat{A}$ in $\mathbf{A}$. Note that for stability we do not require $\hat{A}$ to be in $\mathbf{A}$, but for strong stability we do. (Other stability concepts were introduced in [2], [13], [14].)

In [4] it is shown that the following algorithms for solving linear equations are strongly stable for their respective classes of matrices:

(1) Gaussian elimination with partial or complete pivoting on $\mathbf{A}_1 = \{$nonsingular matrices$\}$ [16];

(2) Cholesky on $\mathbf{A}_2 = \{$symmetric positive definite matrices$\}$ [16];

(3) LDL^T (symmetric Gaussian elimination) on $\mathbf{A}_2$ [16];

(4) Symmetric indefinite algorithm (diagonal pivoting method [5], [6], [9]) on $\mathbf{A}_3 = \{$nonsingular symmetric matrices$\}$;

(5) LU decomposition (Gaussian elimination without pivoting) on $\mathbf{A}_4 = \{$strictly column diagonally dominant matrices ($|a_{ii}| > \sum_{j \neq i} |a_{ji}|$ for all i)$\}$ or $\mathbf{A}_5 = \{$strictly column diagonally dominant band matrices$\}$. (See Appendix.)

(6) Gaussian elimination with partial or complete pivoting followed by iterative refinement on $\mathbf{A}_6 = \{$nonsingular matrices with an arbitrary but fixed sparsity pattern and which are not too ill conditioned$\}$. (See [2], [13], [14] for discussion.)

In [4] it was noted that while Gaussian elimination with partial pivoting and Gaussian elimination with complete pivoting are stable on $\mathbf{A}_2$ and $\mathbf{A}_3$ and Aasen's method [1], [10] is stable on $\mathbf{A}_3$, it does not follow from their error analyses that these algorithms are strongly stable. Thus, the strong stability of these algorithms on $\mathbf{A}_2$ and $\mathbf{A}_3$, respectively, was left as an open question.

Here we will extend the list of strongly stable "situations" developed in [4]. In particular, we show that if a method is stable for the class of nonsingular symmetric matrices or the class of symmetric positive definite matrices, then it is strongly stable for the same class.

2. Constructing a symmetric perturbed system. If $A = A^T$, $(A + E)z = b$, $z \neq 0$, where E might be nonsymmetric, then we shall construct $F = F^T$ such that $(A + F)z = b$ and $\|F\|$ is within a small constant of $\|E\|$ for the 2-norm and the Frobenius norm. We shall do this in two different ways. The first will use the Powell-Symmetric-Broyden (PSB) update [12]; the second will use a construction via the QR decomposition; in either case we shall show that z is the exact solution of a symmetric perturbed system. We include both since the analyses are instructive in their own right.

The problem of nearby symmetric systems has already been addressed in the theory for quasi-Newton methods. For the first approach we shall use the following [7], [8, p. 196].

THEOREM 1. *If H_c is symmetric, $s_c \neq 0$, then the unique solution to*

$$\text{minimize}\,\{\|H - H_c\|_F : H = H^T, Hs_c = y_c\}$$

is given by the PSB-*update*:

$$H_+ = H_c + \frac{(y_c - H_c s_c)s_c^T + s_c(y_c - H_c s_c)^T}{s_c^T s_c} - \frac{\langle y_c - H_c s_c, s_c\rangle s_c s_c^T}{(s_c^T s_c)^2}.$$

Here, $\| \quad \|_F$ is the Frobenius norm and $\langle u, v\rangle = u^T v$. We will use this to prove the following theorem.

THEOREM 2. *If $A = A^T$, $(A + E)z = b$, $r \equiv b - Az$, $z \neq 0$, then*

$$\hat{F} = \frac{rz^T + zr^T}{z^T z} - \frac{(z^T r)}{(z^T z)^2} zz^T$$

satisfies $(A+\hat{F})z=b$, $\hat{F}=\hat{F}^T$, *and* $\|\hat{F}\|\leqq 3\|E\|$ *for the* 2*-norm and the Frobenius norm. Furthermore,* $\hat{F}$ *is the unique solution to*

$$\text{minimize}\,\{\,\|F\|_F\colon F=F^T,(A+F)z=b\}.$$

Proof. In Theorem 1, take $H_c=A$, $s_c=z$, $y_c=b$. Then $y_c-H_cs_c=b-Az=r$. Thus, the unique $\hat{F}$ minimizing $\{\,\|F\|_F\colon(A+F)z=b,F=F^T\}$ is the PSB update

$$\hat{F}=\frac{rz^T+zr^T}{z^Tz}-\frac{(z^Tr)}{(z^Tz)^2}zz^T.$$

Thus,

$$\|\hat{F}\|_2\leqq\|\hat{F}\|_F\leqq\frac{\|rz^T\|_F+\|zr^T\|_F}{z^Tz}+\frac{|z^Tr|}{(z^Tz)^2}\|zz^T\|_F.$$

But

$$\|uv^T\|_F=\|uv^T\|_2=\|u\|_2\|v\|_2\,[10,\text{p. }16].$$

Hence,

$$\|\hat{F}\|_2\leqq\|\hat{F}\|_F\leqq\frac{2\|r\|_2\|z\|_2}{\|z\|_2^2}+\frac{\|z\|_2^3\|r\|_2}{\|z\|_2^4}=3\frac{\|r\|_2}{\|z\|_2}.$$

However, $r\equiv b-Az=Ez$, so $\|r\|_2\leqq\|E\|_2\|z\|_2$. Thus,

$$\|\hat{F}\|_2\leqq\|\hat{F}\|_F\leqq 3\|E\|_2\leqq 3\|E\|_F. \qquad \square$$

Now, we shall construct a symmetric perturbed system by an approach via the QR decomposition which will give a slightly sharper bound. But first we need the following lemma.

LEMMA 1. *Given any two unit vectors u and v, there exists a symmetric orthogonal matrix P such that* $Pu=v$.

Proof. If u and v are parallel, then P is a multiple of the identity. If u and v are not parallel, P can be taken to be a Householder matrix that reflects in a plane containing $u+v$ and is orthogonal to the plane containing u and v. □

THEOREM 3. *If* $A=A^T$, $(A+E)z=b$, $z\neq 0$, *then there exists* $\tilde{F}=\tilde{F}^T$ *such that* $(A+\tilde{F})z=b$, $\|\tilde{F}\|_2\leqq\|E\|_2$ *and* $\|\tilde{F}\|_F\leqq\sqrt{2}\|E\|_F$. (*The bounds are sharp.*)

Proof. We need to determine $\tilde{F}$ so that

$$\tilde{F}^T=\tilde{F}\,\text{and}\,\tilde{F}z=r,$$

where $r\equiv b-Az=Ez$. If $r=0$, let $\tilde{F}\equiv 0$. Suppose $r\neq 0$.

$$\text{Let}\,X\equiv[z\,|\,r]=QR,\quad\text{where}\,R\equiv[\hat{z}\,|\,\hat{r}],\hat{z}\equiv\begin{bmatrix}\hat{z}_1\\0\\0\\\vdots\\0\end{bmatrix},\quad\text{and}\,\hat{r}\equiv\begin{bmatrix}\hat{r}_1\\\hat{r}_2\\0\\\vdots\\0\end{bmatrix},$$

be the QR decomposition of X. Note that, expressing $\tilde{F}=QFQ^T$, it is sufficient to determine F so that

$$(QFQ^T)^T=QFQ^T\quad\text{and}\quad QFQ^Tz=r,$$

or, more simply, so that

$$F^T = F \quad \text{and} \quad F\hat{z} = \hat{r}.$$

These can both be satisfied by choosing $F = \operatorname{diag}(F_{11}, 0)$ if F_{11} can be determined so that

$$F_{11}^T = F_{11} \quad \text{and} \quad F_{11}\begin{bmatrix} \hat{z}_1 \\ 0 \end{bmatrix} = \begin{bmatrix} \hat{r}_1 \\ \hat{r}_2 \end{bmatrix}.$$

Since $z \neq 0$ and $r \neq 0$, $\hat{z}_1 \neq 0$, and $\hat{r}_1 \neq 0$ or $\hat{r}_2 \neq 0$. Let

$$u = \begin{bmatrix} \hat{z}_1 \\ 0 \end{bmatrix} \Big/ |\hat{z}_1|$$

and

$$v = \begin{bmatrix} \hat{r}_1 \\ \hat{r}_2 \end{bmatrix} \Big/ \left\| \begin{bmatrix} \hat{r}_1 \\ \hat{r}_2 \end{bmatrix} \right\|_2.$$

By Lemma 1, there exists $P = P^T = P^{-1}$ such that $Pu = v$. Let $F_{11} \equiv \alpha P$, where $\alpha \equiv \|\hat{r}\|_2 / \|\hat{z}\|_2 = \|r\|_2 / \|z\|_2$. Then $Fz = r$ and $F^T = F$.

$$\|F\|_2 = \alpha \|P\|_2 = \alpha = \frac{\|r\|_2}{\|z\|_2} = \frac{\|Ez\|_2}{\|z\|_2} \leqq \|E\|_2.$$

If $\|Ez\|_2 = \|E\|_2 \|z\|_2$, then $\|E\|_2 = \|F\|_2$ and the bound is sharp.

Since F_{11} is a multiple of a 2×2 orthogonal matrix, $\|F_{11}\|_F = \sqrt{2}\|F_{11}\|_2$. Thus,

$$\|F\|_F = \sqrt{2}\|F\|_2 \leqq \sqrt{2}\|E\|_2 \leqq \sqrt{2}\|E\|_F.$$

Setting $\tilde{F} = QFQ^T$ gives us the result. □

However, the $\hat{F}$ constructed in Theorem 2 minimizes

$$\{\|F\|_F : F = F^T, (A+F)z = b\},$$

and, hence

$$\|\hat{F}\|_F \leqq \|\tilde{F}\|_F \leqq \sqrt{2}\|E\|_F.$$

Thus, Theorem 3 gives us the following Corollary.

COROLLARY. *The matrix $\hat{F}$ in Theorem 2 satisfies* $\|F\|_F \leqq \sqrt{2}\|E\|_F$.

(*Note*: In [11] Higham gives a result similar to this Corollary.)

3. Applications. Gaussian elimination with pivoting and Aasen's method are stable for symmetric systems [10]. But, while the computed solution $\hat{x}$ solves a nearby system

$$(A+E)\hat{x} = b,$$

it is *not* the case that the matrix E is symmetric, at least not from the traditional backwards error analyses. Our results show that there is a symmetric F with $\|F\|_2 \leqq \|E\|_2$ and $\|F\|_F \leqq \sqrt{2}\|E\|_F$ so that

$$(A+F)\hat{x} = b.$$

Thus, Gaussian elimination with pivoting and Aasen's method are strongly stable when applied to symmetric systems. In [4], only the diagonal pivoting method [5], [6], [9] was shown to be strongly stable on symmetric systems. More generally, we have Theorem 4.

THEOREM 4. *If a method is stable for (nonsingular) symmetric matrices, then it is strongly stable for (nonsingular) symmetric matrices.*

Finally we make some observations about strong stability of algorithms for symmetric *positive definite* systems. The BFGS update [8, p. 201] and the DFP update [8, p. 205] do not give an F near E in the symmetric positive definite case. However, we can make an existence argument as follows.

THEOREM 5. *If A is symmetric positive definite and*

$$(A+E)\hat{x}=b$$

with $\|E\|_2 < \lambda_{\min}(A)$, then there exists a symmetric F so that

$$(A+F)\hat{x}=b, \tag{1}$$

$$\|F\|_2 \leqq \|E\|_2, \tag{2}$$

and

$$\lambda_{\min}(A+F)>0. \tag{3}$$

Proof. Theorem 4 ensures that (1) and (2) hold. From [10, p. 269] or [16, pp. 101–102] we have that

$$\lambda_{\min}(A+F) \geqq \lambda_{\min}(A)+\lambda_{\min}(F) \geqq \lambda_{\min}(A)-\|F\|_2.$$

Since $\|E\|_2 < \lambda_{\min}(A)$ and $\|F\|_2 \leqq \|E\|_2$, we have $\lambda_{\min}(A+F) > 0$. □

If A is symmetric positive definite, then $\lambda_{\min}(A) = \|A\|_2$. Hence, Theorem 5 says that if $(A+E)\hat{x} = b$ with $\|E\|_2/\|A\|_2 < 1$, then there exists a symmetric F such that $A+F$ is positive definite, $(A+F)\hat{x}=b$, and $\|F\|_2/\|A\|_2 \leqq \|E\|_2/\|A\|_2$. □

We shall state this more formally in Theorem 6.

THEOREM 6. *If a method is stable for symmetric positive definite matrices, then it is strongly stable for symmetric positive definite matrices.*

4. Conclusions. We have shown that any algorithm for linear equations that is stable on $\mathbf{A}_2 = \{$symmetric positive definite matrices$\}$ or $\mathbf{A}_3 = \{$nonsingular symmetric matrices$\}$ will also be strongly stable on the same matrix class.

Appendix. A matrix A is *strictly column (row) diagonally dominant* if $|a_{ii}| > \sum_{j\neq i}|a_{ji}|$ for each i ($|a_{ii}| > \sum_{j\neq i}|a_{ij}|$ for each i). Let us perturb A to $\tilde{A} = A+E$. The following lemma shows that if the perturbation E is small enough then $A+E$ is still strictly column (row) diagonally dominant.

LEMMA 2. *If A is strictly column (row) diagonally dominant and if $\|E\|_1 < \delta$, where $\delta \equiv \min_i\{|a_{ii}| - \sum_{j\neq i}|a_{ji}|\}$ (if $\|E\|_\infty < \varepsilon \equiv \min_i\{|a_{ii}| - \sum_{j\neq i}|a_{ij}|\}$), then $A+E$ is strictly column (row) diagonally dominant.*

Proof. We shall prove it for column dominance; the proof for row dominance is similar. Since

$$\sum_j |e_{ij}| \leqq \|E\|_1 < \delta \quad \text{and} \quad \sum_{j\neq i}|a_{ji}| \leqq |a_{ii}|-\delta,$$

we have

$$\sum_{j\neq i}|a_{ji}+e_{ji}| \leqq \sum_{j\neq i}|a_{ji}| + \sum_{j\neq i}|e_{ji}| < |a_{ii}|-\delta+\delta-|e_{ii}|$$

$$\leqq |a_{ii}+e_{ii}| \quad \text{for each } i. \qquad \square$$

The following theorem shows that if the machine precision u is small enough then Gaussian elimination without pivoting (LU decomposition) is strongly stable for column strictly diagonally dominant matrices.

THEOREM 7. *Let A be a column strictly diagonally dominant; let z be the computed solution by Gaussian elimination without pivoting. Then there exists an E such that $(A + E)z = b$, where $\|E\|_1 \leqq p(n)ua$, $p(n)$ is a low degree polynomial in n, u is the machine precision, and $a \equiv \max_{i,j} |a_{ij}|$. If, also, $u < \delta/(p(n)a)$, where $\delta \equiv \max_i \{|a_{ii}| - \sum_{j\neq i} |a_{ji}|\}$, then $A + E$ is strictly column diagonally dominant.*

Proof. From [10], [15], [16], there is an E such that $(A + E)z = b$ with $\|E\|_1 < \frac{1}{2}p(n)u \max_{i,j,k} |a_{ij}^{(k)}|$, where p is a polynomial of degree 3 and $a_{ij}^{(k)}$ are the elements in the reduced matrices. From [15, Chap. 3], $\max_{i,j,k} |a_{ij}^{(k)}| \leqq 2a$. If $u < \delta/(p(n)a)$, then $\|E\|_1 < \delta$, and by Lemma 2, $A + E$ is strictly column diagonally dominant. □

REFERENCES

[1] J. O. AASEN, *On the reduction of a symmetric matrix to tridiagonal form*, BIT, 11 (1971), pp. 233–242.

[2] M. ARIOLI, J. DEMMEL, AND I. DUFF, *Solving sparse linear systems with sparse backward error*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 165–190.

[3] J. BARLOW AND J. DEMMEL, *Computing accurate eigensystems of scaled diagonally dominant matrices*, Computer Science Department Report 421, Courant Institute, New York University, New York, NY, 1988; SIAM J. Numer. Anal., 27 (1990), to appear.

[4] J. R. BUNCH, *The weak and strong stability of algorithms in numerical linear algebra*, Linear Algebra Appl., 88/89 (1987), pp. 49–66.

[5] J. R. BUNCH AND L. KAUFMAN, *Some stable methods for calculating inertia and solving symmetric linear systems*, Math. Comp., 31 (1977), pp. 162–179.

[6] 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.

[7] J. E. DENNIS, JR. AND J. J. MORÉ, *Quasi-Newton methods, motivations, and theory*, SIAM Rev., 19 (1977), pp. 46–89.

[8] J. E. DENNIS, JR. AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[9] 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.

[10] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.

[11] N. J. HIGHAM, *Matrix nearness problems and applications*, Numerical Analysis Report 161, University of Manchester, Manchester, United Kingdom, 1988.

[12] M. J. D. POWELL, *A new algorithm for unconstrained optimization*, in *Nonlinear Programming*, J. B. Rosen, D. L. Mangasarian, and K. Ritter, eds., Academic Press, New York, 1970, pp. 31–65.

[13] R. D. SKEEL, *Scaling for numerical stability in Gaussian elimination*, J. Assoc. Comput. Mach., 26 (1979), pp. 494–526.

[14] ———, *Iterative refinement implies numerical stability for Gaussian elimination*, Math. Comp. 35 (1980), pp. 817–832.

[15] G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[16] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 4, pp. 500–519, October 1989

007

MATRICES, DIGRAPHS, AND DETERMINANTS*

JOHN S. MAYBEE†, D. D. OLESKY‡, P. VAN DEN DRIESSCHE§, AND G. WIENER¶

Abstract. A detailed account of various determinantal formulas is presented in a graph-theoretic form involving paths and cycles in the digraph of the matrix. For cases in which the digraph has special local properties, for example, a cutpoint or a bridge, particular formulas are given that are more efficient for computing the determinant than simply using the matrix representation. Applications are also given to characteristic determinants, general minors, and cofactors.

Key words. digraph, cycle, determinant, minor, cofactor

AMS(MOS) subject classifications. 05C50, 15A15

1. Introduction. The connection between the digraph of a matrix and the determinant of a matrix has been pointed out by many authors during the past three decades (see, e.g., [2]–[5], [9], [10], [12], [15], [18], [19], [21]). These papers include applications to solving linear systems, finding spectra of graphs and solving qualitative problems; however, no systematic exposition of this subject has appeared. Our purpose here is to derive some fundamental formulas and give some new applications. The fundamental formulas are in § 2. The remaining sections are independent except that the cofactor formulas in § 7 depend on results on nonprincipal minors in § 6. Sections 3 and 4 contain formulas for digraphs with special local properties; applications are given in § 5. We conclude in § 8 with an example illustrating several of our formulas.

The motivation for this work comes from the fact that although the evaluation of a determinant may be very difficult using the matrix representation, the matrix digraph often indicates efficient means of evaluation. We give specific examples of this involving the graph-theoretic concepts of cutpoints, critical subdigraphs, and bridges.

We now introduce our notation. With an $n \times n$ matrix $A = [a_{ij}]$ we associate the digraph $D(A) = (V, \mathscr{A})$, having vertex set $V = \{1, 2, \cdots, n\}$ and arc set $\mathscr{A}$ containing the arc (i, j) if and only if $a_{ij} \neq 0$ for $i \neq j$. In addition we suppose that there is a subset $V_0 \subseteq V$ of distinguished vertices of $D(A)$. The vertex $i \in V_0$ if and only if $a_{ii} \neq 0$. (It should be noted that, for reasons motivated by applications, some authors prefer to put (i, j) in $\mathscr{A}$ if and only if $a_{ji} \neq 0$; see [1] for example. The development of determinant formulas is the same in either case, however.) In order to fix our terminology, we shall call a sequence $(i_1, i_2, \cdots, i_r)$ of distinct vertices a *path* in $D(A)$ if each of the arcs $(i_1, i_2), (i_2, i_3), \cdots, (i_{r-1}, i_r)$ belongs to $\mathscr{A}$. The length of such a path is $r - 1$. We sometimes use a notation like $p(i \rightarrow j)$ to denote a path in $D(A)$ from i to j. The length of p will then be denoted by $l(p)$ and the set of vertices belonging to the path will be denoted by $V[p]$. The set of vertices of $D(A)$ *not* belonging to p will be denoted by

* Received by the editors December 9, 1987; accepted for publication (in revised form) January 6, 1989.

† Department of Mathematics, University of Colorado, Boulder, Colorado 80309.

‡ Department of Computer Science, University of Victoria, Victoria, British Columbia, Canada V8W 2Y2 (dolesky@uvunix.bitnet). The research of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant A-8214, and the University of Victoria Committee on Faculty Research and Travel.

§ Department of Mathematics, University of Victoria, British Columbia, Canada V8W 2Y2 (pvdd@uvvm.bitnet). The research of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant A-8965, and the University of Victoria Committee on Faculty Research and Travel.

¶ National Center for Atmospheric Research, P.O. Box 3000, Boulder, Colorado 80307-3000.

$V(p)$. We call a sequence $(i_1, i_2, \cdots, i_r, i_1)$, where $i_1, i_2, \cdots, i_r$ are distinct vertices of $D(A)$ and each of the arcs $(i_1, i_2), \cdots, (i_r, i_1)$ belongs to $\mathscr{A}$, a *cycle* of $D(A)$. Its length is $r \geqq 2$. We also call the distinguished vertices of $D(A)$ cycles of length one or 1-cycles. A cycle of $D(A)$ will be denoted by c; the length of c is $l(c) \geqq 1$. $V[c]$ is the set of vertices in c and $V(c)$ the set of vertices of $D(A)$ not in c.

Suppose $I \subseteq V$. We use the notation $\langle I \rangle$ to denote the subdigraph of $D(A)$ generated (induced) by the vertices in I; that is, the arc set of this subdigraph is exactly the arcs of $\mathscr{A}$ joining vertices of I. See [11] or [20] for a discussion of this concept. Usually we regard subsets of V as being ordered sets since they are subsets of the set of the first n integers. If I is a subset of V, we denote by $A[I]$ the principal submatrix of A in the rows and columns defined by I. Similarly, we denote by $A(I)$ the complementary principal submatrix, i.e., the principal submatrix in the rows and columns defined by $V - I$. The determinants of these submatrices are denoted by $\det A[I]$ and $\det A(I)$, respectively. These are *principal* minors of the matrix A. If $I = \phi$, we set $\det A[I] = 1$, and thus $\det A(V) = 1$. Note that $\det A[V]$ is equal to the determinant of A, denoted by $\det A$. The relationship between principal submatrices of A and generated subdigraphs of $D(A)$ is given by $D(A[I]) = \langle I \rangle$.

If p is a path in $D(A)$, we let $A[p]$ denote the corresponding product of elements of A, which we call a *path* of A. Similarly, if c is a cycle of $D(A)$, then $A[c]$ denotes the corresponding product of elements of A, which we call a *cycle* of A. (Note that if $c = i$, a distinguished vertex of $D(A)$, then $A[c] = a_{ii}$.) When p is a path of $D(A)$ we denote by $\det A[V(p)] = \det A(V[p])$ the principal minor of A in the rows and columns defined by $V(p)$, i.e., the indices not on the path. Similarly, $\det A[V(c)]$ is defined for c a cycle of $D(A)$. We call $\det A[V(p)]$ the *cominor* of p and $\det A[V(c)]$ the *cominor* of c. Thus to each path and cycle of A there is associated a uniquely defined principal minor of A called the cominor of the path or cycle.

If $I \subseteq V$ and $J \subseteq V$ with $|I| = |J|$, we denote by $A[I, J]$ the submatrix of A in rows I and columns J and by $A(I, J)$ the complementary submatrix; note that $A[I, I] = A[I]$ and $A(I, I) = A(I)$. Then $\det A[I, J]$ and $\det A(I, J)$ denote the corresponding determinants.

2. Fundamental formulas.

By definition, for $A = [a_{ij}]$ an $n \times n$ matrix,

$$\det A = \sum_{\phi} (\operatorname{sgn} \phi) \prod_{i=1}^{n} a_{i,\phi(i)}$$

where ϕ is an arbitrary permutation of V and $\operatorname{sgn} \phi$ is the sign of the permutation ϕ. Any permutation ϕ can be uniquely factored (up to the order of factors) into a product of disjoint permutation cycles. Let $\phi = \phi_1, \phi_2, \cdots, \phi_t$ be the unique factorization of ϕ. Each of the ϕ_j, $j = 1, 2, \cdots, t$, is actually a sequence $\phi_j = (l_1, l_2, \cdots, l_{r_j})$ of distinct integers. Thus, provided each of the arcs $(l_1, l_2), (l_2, l_3), \cdots, (l_{r_j}, l_1)$ belongs to $\mathscr{A}$, ϕ_j defines a unique cycle c_j of $D(A)$, namely, $(l_1, l_2, \cdots, l_{r_j}, l_1)$. In this case ϕ_j also determines a unique cycle $A[c_j]$ of A. The sign of the permutation ϕ can be computed from

$$\operatorname{sgn} \phi = (\operatorname{sgn} \phi_1)(\operatorname{sgn} \phi_2) \cdots (\operatorname{sgn} \phi_t).$$

Consequently we have the formula

$$(\operatorname{sgn} \phi) \prod_{i=1}^{n} a_{i,\phi(i)} = (\operatorname{sgn} \phi_1) A[c_1] (\operatorname{sgn} \phi_2) A[c_2] \cdots (\operatorname{sgn} \phi_t) A[c_t].$$

This term is nonzero if and only if each of the permutation cycles $\phi_1, \cdots, \phi_t$ corresponds to a cycle of $D(A)$. Note that sgn ϕ_j is positive if ϕ_j is a cycle of odd length and negative if ϕ_j is a cycle of even length.

We now restrict ϕ to permutations in which each permutation cycle corresponds to a cycle present in $D(A)$. Observe that the set of cycles $f = \{c_1, c_2, \cdots, c_t\}$ of $D(A)$ defined by the factorization of ϕ consists of pairwise disjoint cycles and every vertex of $D(A)$ belongs to one of them. Such a set f is called a *factor of* $D(A)$. (Graph theorists define a 1-factor for a digraph D to be a spanning subdigraph for which each vertex has indegree and outdegree equal to one (see, e.g., [3]). Thus our use of the term factor coincides with the concept of a 1-factor as used in graph theory except that we use distinguished vertices in place of loops.) Thus there exists a one-to-one correspondence between the factors of $D(A)$ and the nonzero terms in the expansion of det A. With each factor f of $D(A)$ we associate the unique integer μ_f equal to the number of cycles of even length belonging to f. Also we set $A[f]$ equal to the product of the $A[c]$ over the cycles c in f.

From this discussion we can deduce the fundamental determinant formula in the following graph-theoretic form (see, e.g., [2], [4], [9], [19]).

THEOREM 1. *Let A be an $n \times n$ matrix with digraph $D(A)$. Suppose $D(A)$ has the factors $f_k = \{c_{k1}, c_{k2}, \cdots, c_{km_k}\}$, $k = 1, 2, \cdots, q$, and let μ_k be the number of cycles of even length in f_k. Then*

$$(1) \qquad \det A = \sum_{k=1}^{q} (-1)^{\mu_k} A[c_{k1}]A[c_{k2}]\cdots A[c_{km_k}] = \sum_{k=1}^{q} (-1)^{\mu_k} A[f_k].$$

We observe that (1) applied to $\langle I \rangle$ gives a formula for computing det $A[I]$. Formula (1) is simply a restatement of the formula for the determinant of a square matrix in graph-theoretic terms. It can have as many as $n!$ terms in the event that $D(A)$ is a complete digraph on n vertices and $V_0 = V$. Thus the utility of such a reformulation depends on whatever special structural properties the matrix A may have, as defined by its digraph $D(A)$. Note that even if $a_{ij} \neq 0$, the term a_{ij} occurs in det A if and only if i and j are in a cycle that is in a factor of $D(A)$. We present later some special cases for which (1) yields efficient formulas for det A.

The sign $(-1)^{\mu_k}$ appearing in (1) can also be written in another way. For $1 \leqq k \leqq q$ and $1 \leqq j \leqq m_k$, the sign contributed by the cycle c_{kj} is $(-1)^{l_{kj}+1}$, where l_{kj} is the length of c_{kj}. But $l_{k1} + l_{k2} + \cdots + l_{km_k} = n$ for each k. Consequently, as μ_k and $n + m_k$ are both even or odd, we can also write (1) in the form

$$(1') \qquad \det A = (-1)^n \sum_{k=1}^{q} (-1)^{m_k} A[c_{k1}]A[c_{k2}]\cdots A[c_{km_k}]$$

where m_k is the number of cycles in the factor f_k. This is the form derived in [3], where the formula is attributed to Coates [2] and historical remarks are also given. At this point we observe that when A is a (0, 1) matrix, our results can be stated in terms of the adjacency matrix of a digraph (or graph; see, e.g., [3]).

A classical tool in the theory of determinants is the expansion of det A by rows or columns and, more generally, by the Laplace expansion formula. These tools have led to many useful theoretical results. By using the concept of a cycle in the digraph $D(A)$, theoretically useful expansions of det A in terms of principal minors of A can be derived. We turn next to such expansions.

THEOREM 2 [18]. *Let A be an $n \times n$ matrix with directed graph $D(A)$. Let i be a fixed vertex in V, suppose the set of all cycles of $D(A)$ containing the vertex i is $\{c_1, c_2, \cdots, c_q\}$ and the length of c_k is l_k. Then*

$$\det A = \sum_{k=1}^{q} (-1)^{l_k+1} A[c_k] \det A[V(c_k)]. \tag{2}$$

Proof. We can partition the set of factors of $D(A)$ into subsets $F_1, \cdots, F_q$ according to which one of the cycles belongs to the factor. All factors containing c_k are placed in F_k. Each term in the expansion of det A corresponding to a factor $f \in F_k$ contains the product (sgn $c_k)A[c_k]$ which equals $(-1)^{l_k+1}A[c_k]$. The remaining cycles in the factor f generate a factor of $\langle V(c_k)\rangle$. Thus, when we sum over all factors belonging to the set F_k, we generate the product $(-1)^{l_k+1}A[c_k] \det A[V(c_k)]$. Formula (2) now follows from (1) by summation on k. □

Several applications of (2) are given in § 5. This formula can be looked on as an expansion of the determinant of A relative to a fixed diagonal element, namely the ith diagonal element, i.e., relative to a fixed vertex of $D(A)$. Here is a generalization.

Let I be a fixed subset of V. A set f_I of disjoint cycles in $D(A)$ *spans* I if every cycle in f_I contains at least one vertex of I and every vertex in I belongs to one of the cycles. Such a spanning set of cycles will be called *minimal* if the set of vertices in f_I is equal to I. Corresponding to each f_I we have a unique cominor det $A[V(f_I)]$, where $V(f_I)$ is the set of vertices not in f_I. If f_I is minimal, then det $A[V(f_I)]$ = det $A(I)$. If f_I is not minimal, then det $A[V(f_I)]$ is a principal minor of $A(I)$. Denote by E_I the sets of f_I that are spanning sets of cycles for I, and which are *not* minimal spanning sets of I.

THEOREM 3 [15]. *Let A be an $n \times n$ matrix with digraph $D(A)$. Then, in terms of the notation above, we have*

$$\det A = \det A[I] \det A(I) + \sum_{f_I \in E_I} (-1)^{\mu(f_I)} A[f_I] \det A[V(f_I)] \tag{3}$$

where $A[f_I]$ is the product of all $A[c]$ for $c \in f_I$ and $\mu(f_I)$ is the number of cycles of even length in f_I.

Proof. From (1), det $A = \sum_{f_I}(-1)^{\mu(f_I)}A[f_I]$ det $A[V(f_I)]$. Formula (3) follows by separating minimal sets from those that are not minimal. □

We may regard (3) as an expansion of det A relative to a fixed set of vertices of $D(A)$. When $I = \{i\}$, (3) coincides with (2).

3. The cutpoint and critical subdigraph formulas. Our ability to relate the expansion of det A to $D(A)$ in (1)–(3) can be used to obtain useful special results in the event that $D(A)$ has special local properties. Here and in § 4 we present some applications based on this idea.

Recall that the vertex i of $D(A)$ is called a *cutpoint* if the number of weak components of $D(A) - \{i\}$ is larger than the number of weak components of $D(A)$. (We remind the reader that $D(A) - \{i\}$ is obtained by removing the vertex i from $D(A)$ together with any arcs of $\mathscr{A}$ incident at i. Weak components are discussed in [11] and [20].) Suppose i is a cutpoint of $D(A)$ and the components of $D(A) - \{i\}$ are D_j, $1 \leqq j \leqq p(i)$. Set $I_j = V[D_j]$, the vertex set of D_j, and let $\bar{D}_j = \langle \bar{I}_j \rangle$ where $\bar{I}_j = I_j \cup \{i\}$, $1 \leqq j \leqq p(i)$. Thus $D_j = \bar{D}_j - \{i\}$.

THEOREM 4. *Let A be an $n \times n$ matrix and suppose $D(A)$ has a cutpoint i. Let $p(i)$, I_j and $\bar{I}_j$ be defined as above. Then*

$$\det A = \sum_{j=1}^{p(i)} \left[\det A[\bar{I}_j] \prod_{\substack{k=1 \\ k \neq j}}^{p(i)} \det A[I_k] \right] - (p(i)-1) a_{ii} \prod_{k=1}^{p(i)} \det A[I_k]. \tag{4}$$

Proof. Let ϕ be a permutation of V. Suppose first that $\phi(i) \neq i$. Then $\phi(i) \in I_j$ for some j. Factor ϕ into the disjoint cycles $\phi_1, \phi_2, \cdots, \phi_t$. There will be a unique cycle, say ϕ_m, that moves i, i.e., $\phi_m(i) \neq i$. Since i is a cutpoint of $D(A)$, the corresponding cycle c_m in $D(A)$ must lie entirely in $\bar{D}_j$ (because $c_m - \{i\}$ is a path it must lie entirely in some weak component). Any other cycle c_k determined by ϕ_k for $k \neq m$ must lie entirely in some D_l, $1 \leqq l \leqq p(i)$, $l \neq j$. Therefore the product

$$(\operatorname{sgn} \phi_1) A[c_1] (\operatorname{sgn} \phi_2) A[c_2] \cdots (\operatorname{sgn} \phi_t) A[c_t]$$

appears exactly once in the expansion of

$$\det A[\bar{I}_q] \prod_{\substack{k=1 \\ k \neq q}}^{p(i)} \det A[I_k]$$

when $q = j$, and does not appear in any of the terms when $q \neq j$, $1 \leqq q \leqq p(i)$. Next suppose $\phi(i) = i$. In this case there is a unique cycle, ϕ_m say, such that $\phi_m(i) = i$. But again because i is a cutpoint of $D(A)$, any cycle $c_k \in D(A)$ determined by ϕ_k, $k \neq m$, must lie entirely in some D_l, $1 \leqq l \leqq p(i)$. In this case we observe that the product must appear exactly once in each term

$$\det A[\bar{I}_j] \prod_{\substack{k=1 \\ k \neq j}}^{p(i)} \det A[I_k], \qquad 1 \leqq j \leqq p(i).$$

Since the permutation ϕ either moves i or fixes i, every nonzero term in the expansion of det A will appear at least once in

$$\sum_{j=1}^{p(i)} \det A[\bar{I}_j] \prod_{\substack{k=1 \\ k \neq j}}^{p(i)} \det A[I_k]. \tag{5}$$

Finally we note that the terms in det A falling under the first case will be counted exactly once in (5). The terms in det A falling under the second case will be counted $p(i)$ times in (5). Since the expression $a_{ii} \prod_{k=1}^{p(i)} \det A[I_k]$ comes from precisely those terms in the second case, (4) holds. □

We illustrate (4) for a cluster of cycles in § 5 and also in our example in § 8.

We remark that, in the case where i is not a distinguished vertex of $D(A)$, the expression in (5) equals det A. This can be viewed as a generalization of the formula for the determinant of the coalescence of two graphs without loops (see, e.g., [21]).

We now derive another form of the cutpoint formula (4). For each $j = 1, 2, \cdots, p(i)$ let $\{c_{j1}, \cdots, c_{jm_j}\}$ be the set of cycles of $D(A)$ incident at the cutpoint i and such that $V[c_{jk}] \cap I_j \neq \phi$. Also let l_{jk} be the length of c_{jk}, $k = 1, 2, \cdots, m_j$. Then we can apply the vertex expansion formula (2) at the vertex i to evaluate each of the determinants det $A[\bar{I}_j]$. To simplify the notation let us set $I_{jk} = \bar{I}_j - V[c_{jk}]$, so that $I_{jk} \subseteq \bar{I}_j$. Then we have

$$\det A[\bar{I}_j] = \sum_{k=1}^{m_j} (-1)^{l_{jk}+1} A[c_{jk}] \det A[I_{jk}] + a_{ii} \det A[I_j].$$

Substituting this into (4) yields

$$\det A = \sum_{j=1}^{p(i)} \left[\sum_{k=1}^{m_j} (-1)^{l_{jk}+1} A[c_{jk}] \det A[I_{jk}] + a_{ii} \det A[I_j] \right] \prod_{\substack{k=1\\ k\neq j}}^{p(i)} \det A[I_k]$$

$$-(p(i)-1)a_{ii} \prod_{k=1}^{p(i)} \det A[I_k],$$

which simplifies to

$$(4') \quad \det A = a_{ii} \prod_{k=1}^{p(i)} \det A[I_k] + \sum_{j=1}^{p(i)} \left[\sum_{k=1}^{m_j} (-1)^{l_{jk}+1} A[c_{jk}] \det A[I_{jk}] \right] \prod_{\substack{k=1\\ k\neq j}}^{p(i)} \det A[I_k].$$

Here is a special case of (4′). Suppose for all $j = 1, 2, \cdots, p(i)$, there is a unique cycle c_j of length l_j incident at i such that $V[c_j] \cap I_j \neq \phi$. We then obtain the expansion

$$(6) \quad \det A = a_{ii} \prod_{k=1}^{p(i)} \det A[I_k] + \sum_{j=1}^{p(i)} (-1)^{l_j+1} A[c_j] \det A[\bar{I}_j - V[c_j]] \prod_{\substack{k=1\\ k\neq j}}^{p(i)} \det A[I_k].$$

The key property of a cutpoint that permits us to prove (4) and (4′) is that each cycle in $D(A)$ must be contained entirely within one of the sets $\bar{D}_j$, and hence each factor of $D(A)$ consists of factors of $\bar{D}_j$ for some j and of D_k for $k \neq j$. If this property can be generalized to some larger subdigraph of $D(A)$, we say that $D(A)$ has a critical subdigraph. More precisely we use the following concept.

Let D be a digraph. The subdigraph D_0 will be called a *critical subdigraph* of D if:

(a) $D_0 = \langle I_0 \rangle$ for some $I_0 \subset V$;

(b) $\langle V - I_0 \rangle \equiv D - D_0$ has more weak components than D; and

(c) If $D_j = \langle I_j \rangle, j = 1, 2, \cdots, p$ $(p \geqq 2)$ are the weak components of $D - D_0$, then every factor of D consists of a factor of $\langle I_0 \cup I_j \rangle$ for some fixed j together with factors of $\langle I_k \rangle$ for $k = 1, 2, \cdots, p$ $(k \neq j)$.

We can now prove the following result.

THEOREM 5. *Suppose the digraph $D(A)$ of the matrix A has a critical subdigraph D_0. Then in the above notation*

$$(7) \quad \det A = \sum_{j=1}^{p} \det A[I_0 \cup I_j] \prod_{\substack{k=1\\ k\neq j}}^{p} \det A[I_k] - (p-1) \det A[I_0] \prod_{k=1}^{p} \det A[I_k].$$

Proof. By property (c), the sum on the right above contains every term in the expansion of det A. However the term det $A[I_0] \prod_{k=1}^{p}$ det $A[I_k]$ occurs p times in the summation, hence it must be subtracted off $(p-1)$ times, giving (7). □

We shall see, by way of some examples, that Theorem 5 offers a substantial generalization of the cutpoint formulas (4) and (4′). On the other hand, it is certainly not clear even from the graph-theoretic point of view how to characterize a critical subdigraph. Here, however, is a sufficient condition that a subdigraph be critical.

LEMMA 1. *Let $D_0 = \langle I_0 \rangle$ be a subdigraph of D satisfying* (a) *and* (b). *If there exist vertices v_{in} and v_{out} of D_0 such that every cycle c with $V[c] \cap I_0 \neq \phi$ and $V[c] \cap (V - I_0) \neq \phi$ enters D_0 at v_{in} and leaves D_0 at v_{out}, then D_0 is a critical subdigraph of D.*

Proof. If c is any cycle of D, then either $V[c] \subset I_0$, $V[c] \subset I_j$ for some fixed $j \in \{1, 2, \cdots, p\}$ where $D_j = \langle I_j \rangle$ is the jth weak component of $D - D_0$, or $V[c]$ satisfies the condition of the lemma. But in the last case it is clear that $V[c] \cap (V - I_0) \subset I_j$ for some fixed j. It follows that every factor of D consists of a factor of $\langle I_0 \cup I_j \rangle$ for some fixed j together with factors of $\langle I_k \rangle$ for $k \neq j$. □

We now give two applications of the critical subdigraph formula.

Let $D = (V, \mathscr{A})$ be a digraph. We call D a *ladder digraph* if $V = V_1 \cup V_2 \cup \cdots \cup V_k$, where $V_i \cap V_j = \phi$, $i \neq j$, $k \geqq 3$, and every $(x, y) \in \mathscr{A}$ is such that $x \in V_i$, $y \in V_j$ with $|i-j| \leqq 1$. The subdigraphs $\langle V_i \rangle$, $i = 1, 2, \cdots, k$, are called the *rungs* of D and, for $i = 2, 3, \cdots, k-1$, the *interior rungs* of D.

Let A be a block tridiagonal matrix, i.e.,

$$A = \begin{bmatrix} A_1 & B_1 & 0 & \cdots & 0 & 0 \\ C_1 & A_2 & B_2 & \cdots & 0 & 0 \\ 0 & C_2 & A_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A_{p-1} & B_{p-1} \\ 0 & 0 & 0 & \cdots & C_{p-1} & A_p \end{bmatrix}$$

where $p \geqq 3$, and A_j is an $r_j \times r_j$ block, $j = 1, 2, \cdots, p$, $\sum_1^p r_j = n$. Then we can write $D(A) = (V, \mathscr{A})$ where $V = V_1 \cup V_2 \cup \cdots \cup V_p$, and setting $r_i^* = \sum_{j=1}^{i-1} r_j$, $r_1^* = 0$, $V_i = (r_i^* + 1, \cdots, r_i^* + r_i)$, $i = 1, 2, \cdots, p$. Note that $V_i \cap V_j = \phi$ if $i \neq j$. Also we have $(i, j) \in \mathscr{A}$ if and only if i and j belong to V_k for some $k = 1, 2, \cdots, p$, or $i \in V_k$, $j \in V_{k+1}$, $k = 1, 2, \cdots, p-1$, or $i \in V_k$, $j \in V_{k-1}$, $k = 2, 3, \cdots, p$. Thus, if A is block tridiagonal, $D(A)$ is a ladder digraph. Conversely, if $D(A)$ is a ladder digraph, there exists a permutation matrix P such that P^TAP is a block tridiagonal matrix.

We call the block tridiagonal matrix *critical* if $D(A)$ is a ladder digraph and each interior rung of D is a critical subdigraph. The concepts are illustrated in Fig. 1, where ⊗ denotes a distinguished vertex. By Lemma 1, the subdigraphs $\langle 3, 4 \rangle$, $\langle 5, 6 \rangle$, $\langle 7, 8 \rangle$, and $\langle 9, 10 \rangle$ are all critical.

Now consider the interior rung $\langle V_2 \rangle$. Applying Theorem 5 we get

$$\begin{aligned} \det A = {} & \det A[V_1 \cup V_2] \det A[V_3 \cup \cdots \cup V_k] + \det A[V_1] \det A[V_2 \cup \cdots \cup V_k] \\ & - \det A[V_1] \det A[V_2] \det A[V_3 \cup \cdots \cup V_k]. \end{aligned}$$

But we can apply Theorem 3 to $\det A[V_1 \cup V_2]$. In fact, let $E_{1,2}$ be the set of all nonminimal sets of cycles which span V_1 in $\langle V_1 \cup V_2 \rangle$. Then

$$\det A[V_1 \cup V_2] = \det A[V_1] \det A[V_2] + \sum_{f \in E_{1,2}} (-1)^{\mu(f)} A[f] \det A[V(f)].$$

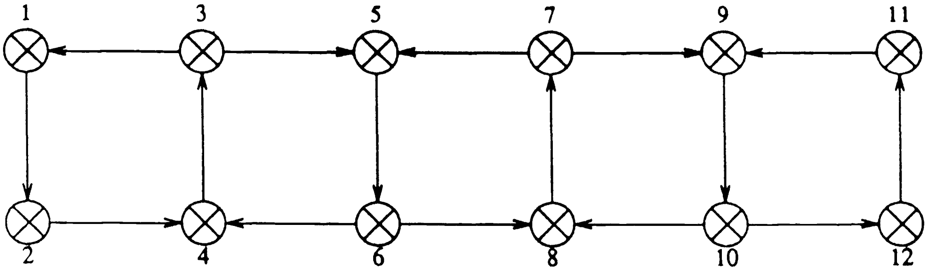


FIG. 1. *A ladder digraph.*

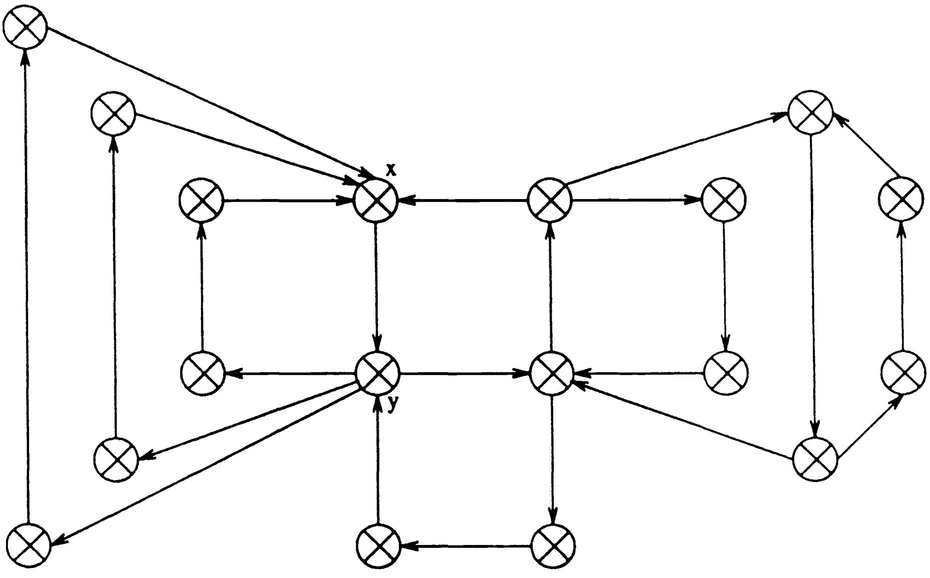


FIG. 2. *A directed* 4-*cockade*.

Here each $\det A[V(f)]$ is a principal minor of $\det A[V_2]$. Thus in this case formula (7) becomes

$$\det A = \det A[V_1]\det A[V_2\cup\cdots\cup V_k] + \sum_{f\in E_{1,2}} (-1)^{\mu(f)}A[f]\det A[V(f)]\det A[V_3\cup\cdots\cup V_k].$$

Observe the analogy between this formula and the recurrence formula for the ordinary tridiagonal case. Also observe that for $k > 3$ we can apply the same reasoning to $\det A[V_2\cup\cdots\cup V_k]$ using the interior rung $\langle V_3\rangle$, etc. In this way we can associate a generalized recurrence relation with a critical ladder digraph.

As a second application, consider the following class of digraphs (see [17]). If D is a digraph and (x, y) is an arc of D, a 3-*path operation* adds two new vertices z_1, z_2 and three new arcs (y, z_1), (z_1, z_2), (z_2, x) to D. We call D a *directed* 4-*cockade* if it can be obtained from a 4-cycle by a finite sequence of 3-path operations. This is introduced for undirected graphs in [22]. It is easy to see that every directed 4-cockade is strongly connected and that every cycle has length four. We illustrate in Fig. 2 a directed 4-cockade with all vertices distinguished. Observe that for $D_0 = \langle x, y\rangle$, $D - D_0$ has four weak components, but there do not exist vertices v_{in} and v_{out} in $\langle x, y\rangle$. This shows that the condition of Lemma 1 is not necessary for a subdigraph to be critical. However, we can use our formula (7) in this example as D_0 is a critical subdigraph.

More generally, if (x, y) is an arc of a directed 4-cockade D with all vertices distinguished, then $D - \langle x, y\rangle$ has more weak components than D if and only if the arc (x, y) belongs to at least two 4-cycles. Note that, as pointed out by a referee, such a digraph $\langle x, y\rangle$ is not necessarily a critical subdigraph. However, we can prove that such an $\langle x, y\rangle$ is not critical if and only if there is a 4-cycle in $D - \{x\}$ that is not in $D - \{y\}$, and a 4-cycle in $D - \{y\}$ that is not in $D - \{x\}$.

4. The bridge formulas. Another type of local behavior lending itself to a simple determinantal formula is the following. The arcs (i, j) and (j, i) constitute a *bridge* of the digraph D if their removal increases the number of weak components of D. Suppose D is a weakly connected digraph and that (i, j) and (j, i) constitute a bridge of D. Then i and j are in different weak components and there are exactly two weak components. Let $\bar{D}_i$, $\bar{D}_j$ be the weak component of $D(A) - \{(i,j), (j,i)\}$ containing i, j, respectively. Set $D_i = \bar{D}_i - \{i\}$, $D_j = \bar{D}_j - \{j\}$ and define $\bar{I} = V[\bar{D}_i]$, $\bar{J} = V[\bar{D}_j]$, $I = V[D_i]$, $J = V[D_j]$.

THEOREM 6. *Let A be an $n \times n$ matrix with a weakly connected digraph $D(A)$. If the arcs (i, j) and (j, i) constitute a bridge of $D(A)$ and the subsets I, J, $\bar{I}$, $\bar{J}$ of V are defined as above, then*

$$\det A = \det A[\bar{I}] \det A[\bar{J}] - a_{ij}a_{ji} \det A[I] \det A[J]. \tag{8}$$

Proof. Any nonzero term $(\operatorname{sgn} \phi) \prod_{i=1}^{n} a_{i,\phi(i)}$ in the expansion of det A is uniquely determined by its representation as a factor of $D(A)$. If f is a factor, then f falls into one of the following mutually exclusive classes:

(a) Every cycle of f lies in $\bar{D}_i$ or in $\bar{D}_j$;

(b) f contains the cycle (i, j, i) and every other cycle of f lies in either D_i or D_j.

The factors in (a) are uniquely determined by all the terms in det $A[\bar{I}]$ det $A[\bar{J}]$, while those in (b) are uniquely determined by all the terms in $a_{ij}a_{ji}$ det $A[I]$ det $A[J]$. Formula (8) follows on taking account of the sign of the 2-cycle $a_{ij}a_{ji}$. □

When A is a $(0, 1)$ symmetric matrix, (8) can be viewed as a formula for computing the determinant of a graph from the determinants of subgraphs (see [3]).

The *bridge formula* of Theorem 6 can be looked on as a special case of the following situation. Let c be a cycle of length $l \geqq 2$ of a weakly connected digraph $D(A)$, and assume that $D(A) - \{\text{arcs of } c\}$ consists of a set of isolated points and a set $\bar{D}_1, \cdots, \bar{D}_p$ of disjoint subdigraphs each containing two or more points and exactly one point of c. Here $0 \leqq p \leqq l$. Let $V[c] \cap V(\bar{D}_j) = \{x_j\}, j = 1, 2, \cdots, p$, and $I_0 = V[c] - \{x_1, \cdots, x_p\}$. We can assume that $p \geqq 1$ since $p = 0$ implies that $D(A) = c$. Setting $\bar{I}_j = V[\bar{D}_j]$ and $I_j = V[D_j] - \{x_j\}, j = 1, 2, \cdots, p$, we have

$$\det A = \prod_{j=1}^{p} \det A[\bar{I}_j] \prod_{\sigma \in I_0} a_{\sigma\sigma} + (-1)^{l+1} A[c] \prod_{j=1}^{p} \det A[I_j]. \tag{9}$$

This *generalized bridge formula* is an obvious extension of (8) and we omit the proof. When $l = p = 2$ it reduces precisely to (8). Also, in the special case where $p = l$ we have

$$\det A = \prod_{j=1}^{l} \det A[\bar{I}_j] + (-1)^{l+1} A[c] \prod_{j=1}^{l} \det A[I_j].$$

Next we present an application of this generalized bridge formula. Let $c = (1, 2, \cdots, l, 1)$ be a cycle of length $l \geqq 2$ and suppose there is at most one cycle $c_j, j = 1, 2, \cdots, l$, of length $l_j \geqq 2$ such that $V[c_j] \cap V[c] = \{j\}$. Setting $\bar{I}_j = V[c_j]$ and $I_j = V[c_j] - \{j\}$, if in addition c_j is the only cycle of length $\geqq 2$ in $\bar{I}_j$, then we have

$$\det A[\bar{I}_j] = \prod_{\sigma \in I_j} a_{\sigma\sigma} + (-1)^{l_j+1} A[c_j],$$

and det $A[I_j] = \prod_{\sigma \in I_j} a_{\sigma\sigma}$. We then obtain from (9) the following result:

$$\det A = \prod_{j=1}^{l} \left[\prod_{\sigma \in I_j} a_{\sigma\sigma} + (-1)^{l_j+1} A[c_j] \right] + (-1)^{l+1} A[c] \prod_{\sigma = l+1}^{n} a_{\sigma\sigma}. \tag{10}$$

Observe also in connection with (10) that $A[c] = a_{12}a_{23}\cdots a_{l-1,l}a_{l1}$. In the particular case where $l_j = 2$, $j = 1, 2, \cdots, l$, we can write $A[c_j] = a_{j,l+j}a_{l+j,j}$, $j = 1, 2, \cdots, l$, and thus

$$\det A = \prod_{j=1}^{l} \{a_{jj}a_{l+j,l+j} - a_{j,l+j}a_{l+j,j}\} + (-1)^{l+1}a_{12}\cdots a_{l1} \prod_{j=l+1}^{2l} a_{jj}.$$

5. Applications. Let us begin with two applications of the expansion formula (2) relative to a vertex. First consider an $n \times n$ matrix $A = [a_{ij}]$ with $a_{ij} \neq 0$ if and only if $-2 \leqq i - j \leqq 1$; this is a special case of an upper Hessenberg matrix. The digraph of A is shown in Fig. 3.

Let us denote the leading principal minor of A of order r by $\det A_r$, $r = 0, 1, 2, \cdots, n$, where $\det A_0 = 1$, and $\det A_n = \det A$. There are three cycles incident at vertex n, namely, the 1-cycle at n, the 2-cycle $(n, n-1, n)$, and the 3-cycle $(n, n-1, n-2, n)$. The corresponding cycles of A are a_{nn}, $a_{n,n-1}a_{n-1,n}$ and $a_{n,n-1}a_{n-1,n-2}a_{n-2,n}$ with cominors $\det A_{n-1}$, $\det A_{n-2}$, and $\det A_{n-3}$, respectively. Consequently, we derive from (2) that

$$\det A = a_{nn}\det A_{n-1} - a_{n,n-1}a_{n-1,n}\det A_{n-2} + a_{n,n-1}a_{n-1,n-2}a_{n-2,n}\det A_{n-3}.$$

This formula expresses the determinant of A in terms of the determinants of three successive principal minors of A. Obviously the same reasoning can be applied to any of the generated subdigraphs $\langle 1, 2, \cdots, r\rangle$ for $r \geqq 3$. In this way we obtain the recurrence formulas

$$(11) \quad \det A_r = a_{rr}\det A_{r-1} - a_{r,r-1}a_{r-1,r}\det A_{r-2} + a_{r,r-1}a_{r-1,r-2}a_{r-2,r}\det A_{r-3}$$

for $r \geqq 3$ with $\det A_0 = 1$, $\det A_1 = a_{11}$, $\det A_2 = a_{22}\det A_1 - a_{12}a_{21}\det A_0$.

The recurrence formulas (11) may also be readily applied to the characteristic matrix $A - \lambda I = A(\lambda)$ to yield

$$(12) \quad \begin{aligned} \det A_r(\lambda) &= (a_{rr} - \lambda)\det A_{r-1}(\lambda) - a_{r,r-1}a_{r-1,r}\det A_{r-2}(\lambda) \\ &\quad + a_{r,r-1}a_{r-1,r-2}a_{r-2,r}\det A_{r-3}(\lambda) \end{aligned}$$

for $r \geqq 3$ with initial conditions $\det A_0(\lambda) = 1$, $\det A_1(\lambda) = a_{11} - \lambda$, $\det A_2(\lambda) = (a_{22} - \lambda)\det A_1(\lambda) - a_{12}a_{21}\det A_0(\lambda)$. Note that when $a_{r-2,r} = 0$ this reduces to the recurrence formulas for a tridiagonal matrix (see, e.g., [7]). These relations could be used to investigate the spectral properties of such an upper Hessenberg matrix A (see, e.g., [17]).

A second example of the application of (2) arises from a modification of an econometric model currently used by the U.S. Department of Energy [16]. Suppose an $n \times n$ matrix $A = [a_{ij}]$ is such that $a_{ij} \neq 0$ if and only if $i = 1, j = 1, i = j$, or $i = n$. The digraph of A is shown in Fig. 4.

Here again each vertex of $D(A)$ is distinguished. In this example every cycle of length $\geqq 2$ is incident at vertex 1. This means that the cominor of each such cycle, as

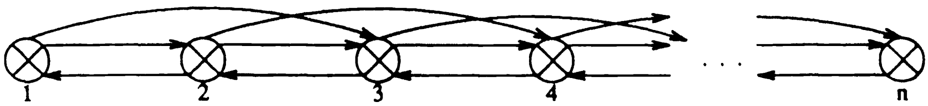


FIG. 3. *The digraph* $D(A)$ *for* $A = [a_{ij}]$ *with* $a_{ij} \neq 0$ *if and only if* $-2 \leqq i - j \leqq 1$.

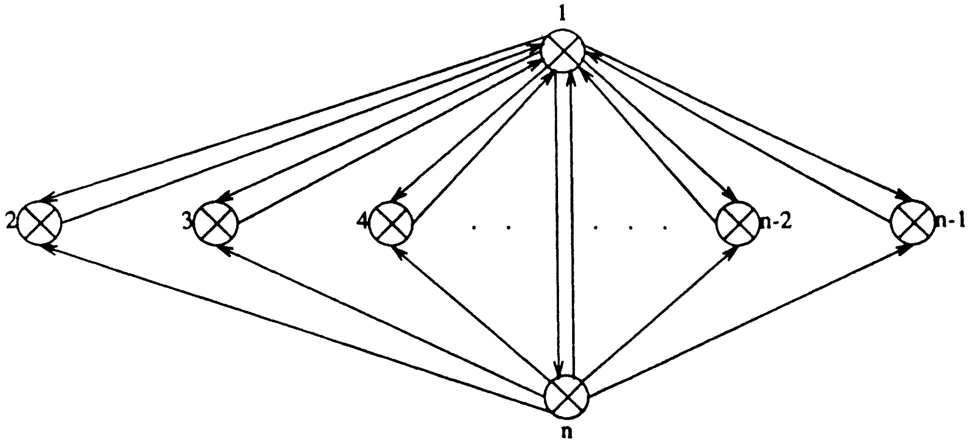


FIG. 4. *The digraph $D(A)$ for $A = [a_{ij}]$ with $a_{ij} \neq 0$ if and only if $i = 1, j = 1, i = j$, or $i = n$.*

well as the cominor of the 1-cycle at vertex 1 itself, is simply a product of elements of A on the principal diagonal. Choosing $i = 1$ in (2), we obtain

$$\det A = a_{11} \prod_{i=2}^{n} a_{ii} - \sum_{i=2}^{n} a_{1i}a_{i1} \prod_{\substack{k=2 \\ k\neq i}}^{n} a_{kk} + \sum_{i=2}^{n-1} a_{1n}a_{ni}a_{i1} \prod_{\substack{k=2 \\ k\neq i}}^{n-1} a_{kk}, \tag{13}$$

an explicit formula that can be readily evaluated.

Again we may apply our result to the matrix $A(\lambda) = A - \lambda I$ to obtain the following formula for the characteristic determinant of A:

(14)

$$\det A(\lambda) = (a_{11} - \lambda) \prod_{i=2}^{n} (a_{ii} - \lambda) - \sum_{i=2}^{n} a_{1i}a_{i1} \prod_{\substack{k=2 \\ k\neq i}}^{n} (a_{kk} - \lambda) + \sum_{i=2}^{n-1} a_{1n}a_{ni}a_{i1} \prod_{\substack{k=2 \\ k\neq i}}^{n-1} (a_{kk} - \lambda).$$

We note that this formula can be used to give information about the spectrum of A under various special hypotheses about the signs of the nonzero elements of A. In the formulas (13) and (14) we have separated off the factor involving a_{11} in the first term on the right-hand side because this is the only place at which it occurs. All other elements along the principal diagonal appear in at least $n - 1$ terms.

We now derive an eigenvalue property from the critical subdigraph expansion (7). Suppose the matrix A has a critical subdigraph and, as above, set $A(\lambda) = A - \lambda I$. Then the expansion (7) becomes

(15)

$$\det A(\lambda) = \sum_{j=1}^{p} \det A[I_0 \cup I_j; \lambda] \prod_{\substack{k=1 \\ k\neq j}}^{p} \det A[I_k; \lambda] - (p-1) \det A[I_0; \lambda] \prod_{k=1}^{p} \det A[I_k; \lambda]$$

where we have used the more compact notation $A[I; \lambda]$ instead of $A(\lambda)[I]$.

THEOREM 7. *Suppose the digraph of the matrix A has a critical subdigraph D_0 and that λ_0 is an eigenvalue of r of the submatrices $A[I_k]$, $2 \leqq r \leqq p$. Then λ_0 is an eigenvalue of A of multiplicity at least $r - 1$.*

Proof. By hypothesis $\det A[I_k; \lambda_0] = 0$ for $r \geqq 2$ values of k. It follows that λ_0 is a zero of each of the terms in the sum in formula (15) $(r-1)$-times. It is also a zero of the last term r-times. □

Next we turn to an application of (3). Let A be such that $D(A)$ has a pair of pendant vertices k and l each joined to another vertex of $D(A)$ by symmetric arcs. To be more specific, let i, j, k, and l be distinct vertices of $D(A)$ such that (i, k) and (k, i) are the only arcs of $D(A)$ incident at k, and (j, l) and (l, j) the only arcs of $D(A)$ incident at l. Applying (3) with $I = \{k, l\}$, and denoting $A(\{k, l\})$ by $A(k, l)$, we obtain

$$\text{(16)} \quad \begin{aligned} \det A = {} & a_{kk}a_{ll}\det A(k,l) - a_{kk}a_{jl}a_{lj}\det A(j,k,l) - a_{ll}a_{ik}a_{ki}\det A(i,k,l) \\ & + a_{ik}a_{ki}a_{jl}a_{lj}\det A(i,j,k,l). \end{aligned}$$

Observe that in (16) the minors $\det A(j, k, l)$, $\det A(i, k, l)$, and $\det A(i, j, k, l)$ are all principal minors of $A(k, l)$.

As an application of (4) consider the following. Let the matrix A be such that $D(A)$ has all vertices distinguished and consists of $m \geqq 2$ cycles $c_1, \cdots, c_m$ of lengths $l_1, \cdots, l_m$, respectively, all of which intersect at a single vertex and are otherwise disjoint. We call such a matrix a *cluster of cycles*. Without loss of generality, we label this unique cutpoint of $D(A)$ vertex 1. Letting I_k, $k = 1, 2, \cdots, m$, be the set of noncutpoints of c_k and $\bar{I}_k = I_k \cup \{1\}$, we have

$$\det A[I_k] = \prod_{\sigma \in I_k} a_{\sigma\sigma} \quad \text{and} \quad \det A[\bar{I}_k] = \prod_{\sigma \in \bar{I}_k} a_{\sigma\sigma} + (-1)^{l_k+1}A[c_k],$$

$k = 1, 2, \cdots, m$. From (4) we then obtain

$$\det A = \sum_{j=1}^{m}\left[\prod_{\sigma \in \bar{I}_j} a_{\sigma\sigma} + (-1)^{l_j+1}A[c_j]\right]\prod_{\substack{k=1 \\ k \neq j}}^{m}\prod_{\sigma \in I_k} a_{\sigma\sigma} - (m-1)a_{11}\prod_{k=2}^{n} a_{kk},$$

whence

$$\text{(17)} \qquad \det A = a_{11}\prod_{k=2}^{n} a_{kk} + \sum_{j=1}^{m}(-1)^{l_j+1}A[c_j]\prod_{\sigma \notin \bar{I}_j} a_{\sigma\sigma}.$$

Setting $A - \lambda I = A(\lambda)$ as before, we deduce from (17) the formula for the characteristic determinant of a cluster of cycles (with cutpoint at vertex 1), namely,

$$\text{(18)} \qquad \det A(\lambda) = (a_{11} - \lambda)\prod_{k=2}^{n}(a_{kk} - \lambda) + \sum_{j=1}^{m}(-1)^{l_j+1}A[c_j]\prod_{\sigma \notin \bar{I}_j}(a_{\sigma\sigma} - \lambda).$$

6. Nonprincipal minors. The expansion formulas derived in § 2 can be applied as in [14] to yield graph-theoretic insights into the expansion of an arbitrary minor of the matrix A. In particular this leads to a theoretically valuable formula for computing the matrix of cofactors of A, cof A, and for computing A^{-1} whenever it exists.

As before, let $D(A) = (V, \mathscr{A})$ with $V_0 \subseteq V$ the set of distinguished vertices of $D(A)$. Suppose $I \subseteq V$, $J \subseteq V$ with $|I| = |J|$, $I \neq J$. Let $L = I \cup J$, $s = |L|$ and $d(I, J) = |L| - |I|$. Following [13], we call $d(I, J)$ the *dispersion* of the pair of sets I and J. Note that $d(I, J) + |I| \leqq n$ and $d(I, J) \geqq 1$. Let $K = I \cap J$ (possibly empty). Then there exist nonempty sets I_0, J_0 such that $I = K \cup I_0$ and $J = K \cup J_0$, where $|I_0| = |J_0| = d(I, J)$ and $I_0 \cap J = J_0 \cap I = \phi$. Note that $L = K \cup I_0 \cup J_0$.

We define the $\langle I, J\rangle$-generated subdigraph of $D(A)$ as follows. Start with $\langle L\rangle$ and delete all arcs of $\langle L\rangle$ incident *to* a vertex of I_0 and all arcs of $\langle L\rangle$ incident *from* a vertex of J_0. In order to relate this subdigraph of $D(A)$ to a submatrix of A, consider the principal submatrix $A[L]$. In this submatrix set to zero all elements in the rows corresponding to J_0 and all elements in the columns corresponding to I_0. Call the resulting submatrix $A[L;\ I,\ J]$. Observe that the nonzero elements of $A[L;\ I,\ J]$ all appear in the rows $K \cup I_0$ and in the columns $K \cup J_0$, and they are precisely the same as the elements in the submatrix $A[I,\ J]$ of A. Note that $\det A[L;\ I,\ J] = 0$ because it has $d(I,\ J)$ rows of zeros and $d(I,\ J)$ columns of zeros.

The next step in our construction follows that in [14]. Set $d(I,\ J) = r$, $I_0 = \{i_1, i_2, \cdots, i_r\}$ and $J_0 = \{j_1, j_2, \cdots, j_r\}$, where $i_1 < i_2 < \cdots < i_r$ and $j_1 < j_2 < \cdots < j_r$. In the matrix $A[L;\ I,\ J]$ replace each of the zeros in the positions $(j_\sigma,\ i_\sigma)$, $\sigma = 1, 2, \cdots, r$, with a one. Call the resulting matrix $\hat{A}[L;\ I,\ J]$.

LEMMA 2. *For the matrix $\hat{A}[L;\ I,\ J]$ defined as above,*

$$\det \hat{A}[L; I, J] = (-1)^{\mu(I,J)} \det A[I, J]$$

where $\mu(I,\ J) = \sum_{\sigma=1}^{r} (\tau(i_\sigma) + \tau(j_\sigma))$ and $\tau(i_\sigma)$, $\tau(j_\sigma)$ are the relative positions of i_σ, j_σ, respectively, in the ordered set L.

Proof. Denoting L by $\{l_1, l_2, \cdots, l_s\}$ with $l_1 < l_2 < \cdots < l_s$, let τ denote the function such that $\tau(l_k) = k$, $1 \leqq k \leqq s$. Since $I_0, J_0 \subseteq L$, $\tau(i_\sigma)$ and $\tau(j_\sigma)$ denote, respectively, the positions of i_σ and j_σ $(1 \leqq \sigma \leqq r)$ in the ordered set L. The result now follows from the structure of $\hat{A}[L;\ I,\ J]$. □

In order to obtain a graph-theoretic understanding of this lemma, observe that inserting the ones in the matrix $A[L; I,\ J]$ as was done above can be interpreted in terms of the digraph $\langle I,\ J\rangle$ as adding the arcs $(j_\sigma,\ i_\sigma)$, $\sigma = 1, 2, \cdots, r$. Denoting the resulting digraph by $\langle \overline{I, J} \rangle$, observe that $\langle \overline{I, J} \rangle$ is *not* in general a subdigraph of $D(A)$ but that $D(\hat{A}[L;\ I,\ J]) = \langle \overline{I, J} \rangle$. We can interpret the calculation of $\det \hat{A}[L; I,\ J]$ graph theoretically with the help of $\langle \overline{I, J} \rangle$. Corresponding to a factor (a set of cycles) of $\langle \overline{I, J} \rangle$ is a "factor" that is a set of cycles and paths of $\langle I,\ J\rangle$. The extension of this notion of factor is used below in the context of nonprincipal minors.

THEOREM 8. *Let A be an $n \times n$ matrix with I, J, L, s, $\mu(I,\ J)$, and $\langle \overline{I, J} \rangle$ defined as above. A nonprincipal minor of A is given by*

$$\det A[I, J] = (-1)^{\mu(I,J)}(-1)^{s} \sum_{k} (-1)^{\nu_k} A[p_{k1}] \cdots A[p_{kr}] A[c_{k1}] \cdots A[c_{km_k}] \tag{19}$$

where the sum is taken over all factors f_k of $\langle \overline{I, J} \rangle$, and ν_k is the number of cycles of f_k.

Proof. Clearly the terms in the expansion of $\det \hat{A}[L; I,\ J]$ correspond to the factors of $\langle \overline{I, J} \rangle$. The distinguished vertices of $\langle \overline{I, J} \rangle$ are found in $V_0 \cap K$. Therefore every cycle of $\langle \overline{I, J} \rangle$ containing either of the vertices j_σ or i_σ, $\sigma = 1, 2, \cdots, r$, must contain the arc $(j_\sigma,\ i_\sigma)$ because j_σ is a sink vertex of $\langle I,\ J\rangle$ and i_σ is a source vertex of $\langle I,\ J\rangle$. It follows that every factor of $\langle \overline{I, J} \rangle$ contains all of the arcs $(j_\sigma,\ i_\sigma)$, $\sigma = 1, 2, \cdots, r$. But this implies that each factor of $\langle \overline{I, J} \rangle$ contains a set of paths $p_1, \cdots, p_r$ in $\langle I,\ J\rangle$ having the following properties (see [14]):

(a) $p_1, \cdots, p_r$ are disjoint;

(b) Each p_σ starts at a vertex of I_0 and ends at a vertex of J_0; and

(c) Each p_σ contains no other vertices in either I_0 or J_0.

Set ν_k equal to the number of cycles in the factor f_k of $\langle \overline{I, J} \rangle$. Then the sum of the lengths of the cycles in f_k is equal to s and the sign they contribute to the factor f_k is $(-1)^{s+\nu_k}$. Corresponding to the factor f_k of $\langle \overline{I, J} \rangle$ is a "factor" of $\langle I,\ J\rangle$ that we may write in the form $\hat{f}_k = \{p_{k1}, \cdots, p_{kr}, c_{k1}, \cdots, c_{km_k}\}$ as the element of A corresponding to each $(j_\sigma,\ i_\sigma)$ has the value 1, $\sigma = 1, 2, \cdots, r$. The r paths in $\hat{f}_k$ come from the r cycles of $\langle \overline{I, J} \rangle$ containing the arcs $(j_\sigma,\ i_\sigma)$, $\sigma = 1, 2, \cdots, r$. We associate with the factor $\hat{f}_k$

the sign $(-1)^{s+\nu_k}$, and thus using (1′) and Lemma 2 we have (19) for the nonprincipal minor $\det A[I, J]$. □

Since $\hat{A}[L; I, J]$ plays only an auxiliary role for the purpose of computing $\det A[I, J]$, we can interpret the computation of the determinant graph theoretically in terms of $D(A[I, J]) = \langle I, J\rangle$. Observe that each of the cycles $c_{k1}, \cdots, c_{km_k}$ in (19) belongs to the subdigraph $\langle K\rangle$. Therefore let us partition the factors $\hat{f}_k$ according to the set of r paths in the factor; thus two factors having the same set of paths are put into the same class. Denote by $V(k; L)$ the complementary set of indices in L to the set contained in the union of the paths $p_{k1}, p_{k2}, \cdots, p_{kr}$. Note that $V(k; L)$ is uniquely defined. We can now modify (19) to the form

$$(19') \qquad \det A[I,J] = (-1)^{\mu(I,J)} \sum_k (-1)^{\mu_k} A[p_{k1}] \cdots A[p_{kr}] \det A[V(k;L)]$$

where the sum is over all distinct sets of paths in $\langle I, J\rangle$ and μ_k equals the number of cycles of even length generated by the paths $p_{k1}, \cdots, p_{kr}$ in $\langle \overline{I, J}\rangle$. This result is used in [14] to show that if A is an M-matrix with its graph having no simple cycle of length greater than three, then the sign of any minor depends only on this graph (and not on the magnitudes of the matrix entries). The formulas of this section illustrate the fact that the expansions of nonprincipal minors involve paths in $D(A)$, whereas the expansions of principal minors involve only cycles (see [18], [19]).

We give now two special cases of Theorem 8. First consider a minor with maximum possible dispersion, i.e., the case $I \cap J = \phi$.

Corollary 8.1. *Let A be an $n \times n$ matrix and $d(I, J) = |I| = r$. Then a minor of maximum dispersion of A is given by*

$$(20) \qquad \det A[I,J] = (-1)^r \sum_k (-1)^{\nu_k} \prod_{\sigma=1}^{r} a_{i_\sigma, f_k(i_\sigma)}$$

where the sum is taken over all factors f_k of $\langle \overline{I, J}\rangle$, ν_k is the number of cycles in f_k, and $f_k(i_\sigma)$ is the element of J that is the endpoint of the arc with initial point i_σ for given f_k.

Proof. In this case the graph $\langle I, J\rangle$ is a directed bipartite graph, i.e., every arc of $\langle I, J\rangle$ has the form (i_σ, j_σ) where $i_\sigma \in I$ and $j_\sigma \in J$. For each factor f_k of $\langle \overline{I, J}\rangle$ every cycle has length $2l$ for some l, since $\langle \overline{I, J}\rangle$ is a bipartite digraph. Thus each cycle of f_k contributes a negative sign. Letting $I = \{i_1, i_2, \cdots, i_r\}$ and $J = \{j_1, j_2, \cdots, j_r\}$, we consider the mapping τ on $I \cup J$ for which $\tau(I \cup J) = \{1, 2, \cdots, 2r\}$. Now $\tau(i_\sigma) + \tau(j_\sigma)$ is even if both $\tau(i_\sigma)$ and $\tau(j_\sigma)$ are even or if they are both odd, and $\tau(i_\sigma) + \tau(j_\sigma)$ is odd if one of $\tau(i_\sigma)$, $\tau(j_\sigma)$ is odd and the other even. But, if r is odd, there must be an odd number of differences with one even and one odd and, if r is even, an even number of such differences. It follows that $\mu(I, J) = \sum_{\sigma=1}^{r} (\tau(i_\sigma) + \tau(j_\sigma))$ has the same parity as r. As s is even, we obtain the expansion formula (20) in the maximum dispersion case. □

Now consider the case of a minor of dispersion one. Such minors are sometimes called almost principal minors (see [13]). There is, however, some confusion in the literature concerning this term. Apparently Gantmacher and Krein [7] had a narrower concept in mind when they introduced almost principal minors. We use the term in the broad sense here.

Corollary 8.2. *Let A be an $n \times n$ matrix. Then, with $I = K \cup \{i_0\}$ and $J = K \cup \{j_0\}$, a minor of dispersion one of A is given by*

$$(21) \qquad \det A[I,J] = (-1)^{\tau(i_0)+\tau(j_0)} \sum_{k=1}^{m} (-1)^{l_k} A[p_k(i_0 \rightarrow j_0)] \det A[V(p_k,K)]$$

where the sum is over all distinct paths from i_0 to j_0 in $\langle I, J\rangle$; $\tau(i_0)$, $\tau(j_0)$ is the position of i_0, j_0, respectively, in the ordered set $I \cup J$; l_k is the length of path p_k from i_0 to j_0; and $V(p_k, K)$ is the set of vertices of $\langle I, J\rangle$ not belonging to p_k.

Proof. When $I = K \cup \{i_0\}$, $J = K \cup \{j_0\}$ each "factor" of $\langle I, J\rangle$ consists of a product of cycles of $\langle K\rangle$ and a path $p(i_0 \rightarrow j_0)$. The path contributes sign $(-1)^l$ where l is the length of the path. We can partition the factors of $\langle I, J\rangle$ into equivalence classes by holding the path $p(i_0 \rightarrow j_0)$ fixed and permitting the cycles of $\langle K\rangle$ to vary. Let $p_1(i_0 \rightarrow j_0), \cdots, p_m(i_0 \rightarrow j_0)$ be the distinct paths from i_0 to j_0 in $\langle I, J\rangle$. Let $V(p_k, K)$ be the set of vertices of $\langle I, J\rangle$ not belonging to p_k; these vertices are all in $\langle K\rangle$, and hence the notation. Then we have (21) as a general formula for an almost principal minor. □

Note that when $|I| = 1$ the almost principal minor is the nondiagonal element $a_{i_0j_0}$ of A. In this case $L = \{i_0, j_0\}$ and we have $\tau(\min\{i_0, j_0\}) = 1$, $\tau(\max\{i_0, j_0\}) = 2$, and $l_k = 1$. Thus, $(-1)^{\tau(i_0)+\tau(j_0)} = (-1)^{l_k} = -1$ so that (21) yields $\det A[i_0, j_0] = a_{i_0j_0}$, as it must. Note that if A is an M-matrix, then each term in the summation of (21) is nonnegative, so that $\det A[I, J] \det A[J, I] \geqq 0$ for any I, J with $d(I, J) = 1$. When this inequality holds, A is called weakly sign symmetric.

7. Cofactor formulas. We now use Corollary 8.2 to prove the following results, where the cofactor of a_{ij} is denoted by A_{ij}, and the matrix $\operatorname{cof} A = [A_{ij}]$.

THEOREM 9 [15]. *Let A be an $n \times n$ matrix with digraph $D(A)$. Let a_{ij} with $i \neq j$ be an arbitrary nondiagonal element of A. Then the cofactor of a_{ij} is given by*

$$A_{ij} = \sum_k (-1)^{l_k} A[p_k(j \rightarrow i)] \det A[V(p_k)] \tag{22}$$

where the sum is taken over all paths in $D(A)$ from j to i, and l_k is the length of path p_k.

Proof. Let us apply (21) to the almost principal minor $\det A(i,j)$, i.e., to the almost principal minor $\det A[I, J]$ where $I = V - \{i\}$, $J = V - \{j\}$. We then have $L = V$, $i_0 = j$ and $j_0 = i$, $\tau(j) = j$, $\tau(i) = i$. Note that the set of all paths from j to i in $\langle I, J\rangle$ is the same as the set of all paths from j to i in $D(A)$. Therefore we obtain the result

$$\det A(i,j) = (-1)^{i+j} \sum_k (-1)^{l_k} A[p_k(j \rightarrow i)] \det A[V(p_k)].$$

But the cofactor of a_{ij} is $A_{ij} = (-1)^{i+j} \det A(i,j)$, so the formula (22) follows at once. □

COROLLARY 9.1. *Let A be an $n \times n$ nonsingular matrix with digraph $D(A)$, and let $A^{-1} = [\alpha_{ij}]$. Then we have*

$$\alpha_{ii} = \det A(i) / \det A, \tag{23a}$$

and

$$\alpha_{ij} = \frac{1}{\det A} \sum_k (-1)^{l_k} A[p_k(i \rightarrow j)] \det A[V(p_k)], \qquad i \neq j, \tag{23b}$$

where the sum is taken over all paths in $D(A)$ from i to j, and l_k is the length of path p_k.

Proof. Formulas (23) follow at once from (22) and the fact that $(\det A)A^{-1} = (\operatorname{cof} A)^T$. □

In the following corollary we use these formulas to prove that if a matrix is nonsingular and irreducible and every vertex is distinguished, then, if cancellations are ignored, its inverse matrix is full. Other proofs of this have been given recently [6], [8] in the context of sparse matrices.

COROLLARY 9.2. *Let $A = [a_{ij}]$ be an $n \times n$ irreducible matrix with $a_{ii} \neq 0$ for all $i \in V$, and suppose $A^{-1} = [\alpha_{ij}]$ exists. Then, ignoring cancellations, $\alpha_{ij} \neq 0$ for all $i, j \in V$.*

Proof. Suppose $\alpha_{ii} = 0$. Then from (23a), $\det A(i) = 0$. As cancellations are ignored, this implies that at least one of $a_{pp} = 0$, $p \in V - \{i\}$, which is a contradiction.

Suppose $\alpha_{ij} = 0$, $i \neq j$. Then from (23b), $A[p_k(i \to j)] = 0$ for each path p_k in $D(A)$ from i to j. (Note that as each a_{ii} is assumed nonzero and cancellations are ignored, $\det A[V(p_k)]$ is nonzero.) Thus there is no path from i to j in $D(A)$. So A is reducible, which is a contradiction. □

Note that it is possible for every vertex not to be distinguished, but the inverse matrix to be full.

The basic cofactor formula is presented above as equation (22) of Theorem 9. We now elaborate on this result and indicate some applications.

Since $A(\operatorname{cof} A)^T = (\det A)I$,

$$\sum_{k=1}^{n} a_{ik}A_{jk} = \begin{cases} 0 & \text{if } i \neq j, \\ \det A & \text{if } i = j. \end{cases}$$

For $i = j$, we have

$$\begin{aligned} \det A &= \sum_{k=1}^{n} a_{ik}A_{ik} = a_{ii}A_{ii} + \sum_{k \neq i} a_{ik}A_{ik} \\ &= a_{ii}\det A(i) + \sum_{k \neq i} a_{ik} \sum_{m} (-1)^{l_{mk}} A[p_{mk}(k \to i)] \det A[V(p_{mk})] \end{aligned}$$

where m is taken over all paths in $D(A)$ from k to i. Clearly for each k such that both $a_{ik} \neq 0$ and there exists at least one path $p(k \to i)$ in $D(A)$, the product $a_{ik}A[p_{mk}(k \to i)]$ is a cycle containing the index i. The sign attached is $(-1)^{l+1}$ where l is the length of the cycle. So we have rederived Theorem 2 using the cofactor formula.

Now for $i \neq j$, we have

$$0 = \sum_{k=1}^{n} a_{ik}A_{jk} = a_{ii}A_{ji} + a_{ij}A_{jj} + \sum_{k \neq i,j} a_{ik}A_{jk}.$$

Thus,

$$\begin{aligned} 0 = {} & a_{ii} \sum_{k} (-1)^{l_k} A[p_k(i \to j)] \det A[V(p_k)] + a_{ij} \det A(j) \\ & + \sum_{k \neq i,j} a_{ik} \sum_{m} (-1)^{l_{mk}} A[p_{mk}(k \to j)] \det A[V(p_{mk})]. \end{aligned}$$

Now the sum in the first term is over all paths in $D(A)$ from i to j, and a given path from i to j in $D(A)$ appears exactly once in the third term. Observe that, since $p_{mk}(k \to j)$ does not contain the vertex i, the set $V(p_{mk})$ does. On the other hand, $V(p_k)$ appearing in the first term does not include the vertex i. Thus we have the following identity:

$$0 = a_{ij} \det A(j) + \sum_{k} (-1)^{l_k} A[p_k(i \to j)] \{ a_{ii} \det A[V(p_k)] - \det A(V(p_k) \cup \{i\}] \}.$$

In the remainder of this section, we consider particular cases of the cofactor formulas when $D(A)$ has special local properties.

Consider first the case where $D(A)$ has the cutpoint i; see § 3 for notation. There are four cases to consider in evaluating cof A.

Case (i). $\sigma \in I_j$, $\tau \in I_k$ for $j \neq k$. We have

$$A_{\sigma\tau} = A[\bar{I}_j]_{\sigma i} A[\bar{I}_k]_{i\tau} \prod_{m \neq j,k} \det A[I_m].$$

Case (ii). $\sigma = i$, $\tau \in I_j$ ($\tau = i$, $\sigma \in I_j$ is analogous). We have

$$A_{\sigma\tau} = A[\bar{I}_j]_{\sigma\tau} \prod_{k \neq j} \det A[I_k].$$

Case (iii). $\sigma, \tau \in I_j$, $\sigma \neq \tau$. We now obtain

$$A_{\sigma\tau} = A[\bar{I}_j]_{\sigma\tau} \prod_{k \neq j} \det A[I_k] + \sum_{k \neq j} A[I_j]_{\sigma\tau} \det A[\bar{I}_k] \prod_{m \neq j,k} \det A[I_m]$$
$$-(p(i)-1) a_{ii} A[I_j]_{\sigma\tau} \prod_{k \neq j} \det A[I_k].$$

Case (iv). $\sigma = \tau$. If $\sigma = i$, then $A_{ii} = \prod_{k=1}^{p(i)} \det A[I_k]$. If $\sigma \in I_j$, we obtain

$$A_{\sigma\sigma} = \det A[\bar{I}_j - \{\sigma\}] \prod_{k \neq j} \det A[I_k] + \sum_{k \neq j} \det A[\bar{I}_k] \det A[I_j - \{\sigma\}] \prod_{m \neq j,k} \det A[I_m]$$
$$-(p(i)-1) a_{ii} \det A[I_j - \{\sigma\}] \prod_{k \neq j} \det A[I_k].$$

Next consider the cofactors of A when $D(A)$ contains the bridge consisting of the arcs (i, j) and (j, i); see § 4 for notation. There are now three cases to consider in evaluating the $A_{\sigma\tau}$.

Case (i). $\sigma \in \bar{I}$, $\tau \in \bar{J}$ ($\sigma \in \bar{J}$, $\tau \in \bar{I}$ is done analogously). Let p be an arbitrary path in $D(A)$ from τ to σ. We can write $p(\tau \rightarrow \sigma) = p'(\tau \rightarrow j)(j, i)p''(i \rightarrow \sigma)$. Then $p'(\tau \rightarrow j)$ is contained in $\bar{D}_j$ and $p''(i \rightarrow \sigma)$ in $\bar{D}_i$. Therefore from (22) we may write

$$A_{\sigma\tau} = \sum_k (-1)^{l_k} A[p_k(\tau \rightarrow \sigma)] \det A[V(p_k)]$$
$$= \sum_k (-1)^{l(p'_k)+1+l(p''_k)}$$
$$\times A[p'_k(\tau \rightarrow j)] a_{ij} A[p''_k(i \rightarrow \sigma)] \det A[\bar{J} - V[p']] \det A[\bar{I} - V[p'']],$$

which is equivalent to $A_{\sigma\tau} = -a_{ji} A[\bar{I}]_{\sigma i} A[\bar{J}]_{j\tau}$. This expansion means that for $\sigma \in \bar{I}$ and $\tau \in \bar{J}$, the cofactor $A_{\sigma\tau}$ can be written as the product of $-a_{ji}$ and certain cofactors of the smaller matrices $A[\bar{I}]$ and $A[\bar{J}]$.

Case (ii). $\sigma, \tau \in \bar{I}$, $\sigma \neq \tau$ ($\sigma, \tau \in \bar{J}$, $\sigma \neq \tau$ can be done analogously). If σ and τ both differ from i, we obtain

$$A_{\sigma\tau} = A[\bar{I}]_{\sigma\tau} \det A[\bar{J}] - a_{ij} a_{ji} A[I]_{\sigma\tau} \det A[J].$$

For $\sigma = i$ ($\tau = i$ is done analogously), this reduces to

$$A_{\sigma\tau} = A[\bar{I}]_{\sigma\tau} \det A[\bar{J}].$$

Case (iii). $\sigma = \tau$. If $\sigma \neq i$, $\sigma \neq j$, $\sigma \in I$ ($\sigma \in J$ is analogous), then

$$A_{\sigma\sigma} = A[\bar{I}]_{\sigma\sigma} \det A[\bar{J}] - a_{ij} a_{ji} A[I]_{\sigma\sigma} \det A[J].$$

For $\sigma = i$ ($\sigma = j$ is done analogously), then $A_{\sigma\sigma} = \det A[I] \det A[\bar{J}]$.

From these three cases, we see that, when $D(A)$ has the bridge $\{(i, j), (j, i)\}$, the matrix $\operatorname{cof} A$ is completely determined by the matrices $\operatorname{cof} A[\bar{I}]$, $\operatorname{cof} A[I]$, $\operatorname{cof} A[\bar{J}]$, $\operatorname{cof} A[J]$, the principal minors $\det A[\bar{I}]$, $\det A[I]$, $\det A[\bar{J}]$, $\det A[J]$ and the elements a_{ij}, a_{ji}.

As a final application of our methods we mention the following double bridge formula. Given a matrix A, let i, j, k, l be distinct vertices in $D(A)$. A *double bridge* is a subset B of arcs of $D(A)$ such that $B \subseteq \{(i, k), (k, i), (j, l), (l, j)\}$, B contains at least one arc from $\{(i, k), (k, i)\}$ and at least one arc from $\{(j, l), (l, j)\}$, and $D(A) - B$ has more weak components than $D(A)$ with $\{i, j\}$ in one weak component and $\{k, l\}$ in another.

Let $D(A)$ be weakly connected and have a double bridge. Suppose $\langle I_1 \cup \{i, j\}\rangle$ and $\langle I_2 \cup \{k, l\}\rangle$ are the subdigraphs of $D(A) - B$ containing $\{i, j\}$ and $\{k, l\}$, respectively. When we let $\bar{I}_1 = I_1 \cup \{i, j\}$ and $\bar{I}_2 = I_2 \cup \{k, l\}$, the cycles of a factor f of $D(A)$ may be categorized as follows:

(i) f contains cycles lying entirely in $\langle \bar{I}_1\rangle$ or in $\langle \bar{I}_2\rangle$;

(ii) f contains the two 2-cycles (i, k, i), (j, l, j) and cycles lying entirely in $\langle I_1\rangle$ or in $\langle I_2\rangle$;

(iii) f contains the 2-cycle (j, l, j) and cycles lying either in $\langle I_1 \cup \{i\}\rangle$ or in $\langle I_2 \cup \{k\}\rangle$;

(iv) f contains the 2-cycle (i, k, i) and cycles lying either in $\langle I_1 \cup \{j\}\rangle$ or in $\langle I_2 \cup \{l\}\rangle$;

(v) f contains cycles that lie partly in $\langle \bar{I}_1\rangle$ and partly in $\langle \bar{I}_2\rangle$.

Thus,

$$\begin{aligned}\det A = {} & \det A[\bar{I}_1]\det A[\bar{I}_2] + a_{ik}a_{ki}a_{jl}a_{lj}\det A[I_1]\det A[I_2] \\ & - a_{jl}a_{lj}\det A[I_1 \cup \{i\}]\det A[I_2 \cup \{k\}] - a_{ik}a_{ki}\det A[I_1 \cup \{j\}]\det A[I_2 \cup \{l\}] \\ & - a_{jl}a_{ki}A[\bar{I}_1]_{ji}A[\bar{I}_2]_{kl} - a_{ik}a_{lj}A[\bar{I}_1]_{ij}A[\bar{I}_2]_{lk},\end{aligned}$$

where the first four terms correspond, respectively, to categories (i)–(iv) and the last two terms correspond to the two types of factors in (v).

8. An example. We conclude with a 7×7 example that illustrates several of the formulas given previously. Figure 5 displays the strongly connected digraph $D(A)$ for our example; all the vertices are distinguished. We use $A[i, j]$ to denote $A[\{i, j\}]$ and similar notation for other principal minors.

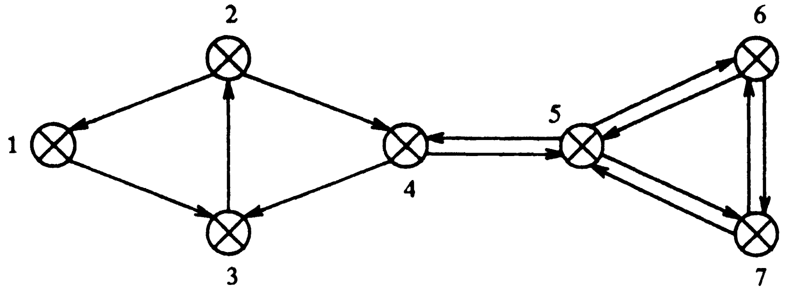


FIG. 5. *The digraph for our* 7×7 *example.*

Vertex 5 of $D(A)$ is a cutpoint and the number of weak components in $D(A) - \{5\}$ is two. When we take $I_1 = \{1, 2, 3, 4\}$ and $I_2 = \{6, 7\}$, our cutpoint formula (4) gives

$$\det A = \det A[1,2,3,4,5] \det A[6,7] + \det A[5,6,7] \det A[1,2,3,4] - a_{55} \det A[1,2,3,4] \det A[6,7].$$

When we work with cycles, our formula (4′) gives

$$\det A = \{a_{55} \det A[1,2,3,4] - a_{45}a_{54} \det A[1,2,3]\} \det A[6,7] + \{a_{57}a_{76}a_{65} + a_{56}a_{67}a_{75} - a_{65}a_{56}a_{77} - a_{75}a_{57}a_{66}\} \det A[1,2,3,4].$$

We can also regard $D_0 = \langle 2, 3 \rangle$ as a critical subdigraph of $D(A)$. Then, when we take $I_1 = \{1\}$ and $I_2 = \{4, 5, 6, 7\}$, our critical subdigraph formula (7) gives

$$\det A = \det A[1,2,3] \det A[4,5,6,7] + a_{11} \det A[2,3,4,5,6,7] - a_{11} \det A[2,3] \det A[4,5,6,7].$$

The arcs (4, 5) and (5, 4) constitute a bridge of $D(A)$, with subsets $I = \{1, 2, 3\}$ and $J = \{6, 7\}$. Formula (8) then yields

$$\det A = \det A[1,2,3,4] \det A[5,6,7] - a_{45}a_{54} \det A[1,2,3] \det A[6,7].$$

Expanding about row 4 of A, we have, by the usual cofactor expansion, $\det A = a_{43}A_{43} + a_{44}A_{44} + a_{45}A_{45}$. Clearly, $A_{44} = \det A[1, 2, 3] \det A[5, 6, 7]$, and from our cofactor formula (22) $A_{43} = a_{32}a_{24}a_{11} \det A[5, 6, 7]$, and

$$A_{45} = -a_{54} \det A[1,2,3] \det A[6,7].$$

REFERENCES

[1] T. BONE, C. JEFFRIES, AND V. KLEE, *A qualitative analysis of $\dot{x} = Ax + b$*, Discrete Appl. Math., 20 (1988), pp. 9–30.

[2] C. L. COATES, *Flow-graph solutions of linear algebraic equations*, IRE Trans. Circuit Theory, 6 (1959), pp. 170–187.

[3] C. CVETKOVIC, M. DOOB, AND H. SACKS, *Spectra of Graphs*, Academic Press, New York, 1980.

[4] C. A. DESOER, *Optimal formula for the gain of a flow graph or a simple derivation of Coates' formula*, Proc. Inst. Rad. Engrg., 48 (1960), pp. 883–889.

[5] M. DOOB, *Applications of graph theory in linear algebra*, Math. Mag., 57 (1984), pp. 67–76.

[6] I. S. DUFF, A. M. ERISMAN, C. W. GEAR, AND J. K. REID, *Some remarks on inverses of sparse matrices*, Tech. Memo. 51, Argonne National Laboratory, Argonne, IL, 1985.

[7] F. GANTMACHER AND M. KREIN, *Oscillation Matrices, Oscillation Kernels and Small Vibrations of Mechanical Systems*, Moscow, Leningrad, 1950. (English translation, U.S. Atomic Energy Commission, Washington, DC, 1961.)

[8] J. R. GILBERT, *Predicting structure in sparse matrix computations*, Tech. Report TR86-750, Cornell University, Ithaca, NY, 1986.

[9] K. GOLDBERG, *Random notes on matrices*, J. Res. Nat. Bur. Standards, 60 (1958), pp. 321–326.

[10] F. HARARY, *The determinant of the adjacency matrix of a graph*, SIAM Rev., 4 (1962), pp. 202–210.

[11] F. HARARY, R. Z. NORMAN, AND D. CARTWRIGHT, *Structural Models: An Introduction to the Theory of Directed Graphs*, John Wiley, New York, 1965.

[12] J. Z. HERON, *Theorems on linear systems*, Ann. N.Y. Acad. Sciences, 108 (1963), pp. 36–68.

[13] D. HERSHKOWITZ, V. MEHRMANN, AND H. SCHNEIDER, *Matrices with sign symmetric diagonal shifts or scalar shifts*, SIAM J. Algebraic Discrete Meth., 8 (1987), pp. 108–122.

[14] C. R. JOHNSON, D. D. OLESKY, BRIAN ROBERTSON, AND P. VAN DEN DRIESSCHE, *Sign determinancy of M-matrix minors*, Linear Algebra Appl., 91 (1987), pp. 133–141.

[15] J. S. MAYBEE, *Sign solvability*, in Proc. Symposium on Computer Assisted Analysis and Model Simplification, H. Greenberg and J. S. Maybee, eds., Academic Press, New York, 1981.

[16] ———, *Mathematical structure of the* OMS *model*, Office of Statistical Standards, Energy Information Administration, Washington, DC, 1986.

[17] ———, *Some possible new directions for combinatorial matrix analysis*, Linear Algebra Appl., 107 (1988), pp. 23–40.

[18] J. S. MAYBEE AND J. QUIRK, *Qualitative problems in matrix theory*, SIAM Rev., 11 (1969), pp. 30–51.

[19] J. PONSTEIN, *Self-avoiding paths and the adjacency matrix of a graph*, SIAM J. Appl. Math., 14 (1966), pp. 600–609.

[20] F. ROBERTS, *Discrete Mathematical Models*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[21] A. J. SCHWENK, *Computing the characteristic polynomial of a graph*, in Graphs and Combinatorics, Lecture Notes in Mathematics 406, Springer-Verlag, Berlin, New York, 1974, pp. 158–172.

[22] C. THOMASSEN, *Sign-nonsingular matrices and even cycles in directed graphs*, Linear Algebra and Appl., 75 (1986), pp. 27–41.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 4, pp. 520–532, October 1989

008

AN EIGENVECTOR TEST FOR INFLATION MATRICES AND ZME-MATRICES*

JEFFREY L. STUART†

Abstract. It is shown that a matrix A is of the form $A = B \times\times U + \rho G(V)$, where U is an inflator and $\times\times$ is the inflation product, if and only if A has a row and a column eigenvector for some eigenvalue such that the eigenvectors satisfy a simple restriction on their supports. This test is extended to recover the inflation sequence for a ZME-matrix. These results imply that the maximal eigenvalue (and hence spectral radius) of a ZME-matrix is the maximum of the maximal eigenvalues of its 2×2 principal submatrices. Additionally, it is shown that for every inflation sequence, there exists an equivalent normalized inflation sequence.

Key words. inflation, Z-matrix, ZME-matrix, spectral radius

AMS(MOS) subject classifications. 15A18, 15A48

1. Introduction. The inflation product introduced by Friedland, Hershkowitz, and Schneider in [2] has been studied in a number of recent papers [1], [2], [5]–[8]. This paper addresses two questions associated with inflation that have arisen in or been suggested by the previous papers. The first is that of recognizing when a matrix arises by inflation. In § 8, we develop a necessary condition based on the existence of a pair of eigenvectors with a particular support structure. In § 12, we extend that test to show how an inflation sequence for a ZME-matrix can be recovered. In § 13, we use these results to derive a simple computational method for determining the maximal eigenvalue (and hence spectral radius) of an $n \times n$ ZME-matrix based on finding the roots of $(n^2 - n)/2$ quadratic equations.

The second question addressed is whether the existence of an inflation sequence guarantees the existence of a normalized inflation sequence. This arises from the fact that several results in [5] require the existence of normalized inflation sequences, whereas certain constructions for producing inflation sequences (such as in [7]) do not necessarily produce normalized inflation sequences. In § 7, we show that the answer is affirmative, and we present an algorithm for transforming an inflation sequence into an equivalent normalized inflation sequence.

2. Basic definitions and notation. Throughout this paper, $\mathcal{M}_n(\mathcal{F})$ will be the set of all $n \times n$ matrices over the set $\mathcal{F}$ where $\mathcal{F}$ is either $\mathbb{R}$ or $\mathbb{C}$. The set of $1 \times n$ matrices over $\mathcal{F}$ will be denoted as $\mathcal{F}^n$, and the term "vector" will always mean row vector. The symbol $\mathcal{O}$ will always denote the zero vector or zero matrix, as determined by context. If v is in $\mathcal{F}^n$, the *support of* v, denoted supp (v), is the set defined by supp $(v) = \{i: v_i \neq 0\}$. A *strictly nonzero matrix* (*strictly nonzero vector*) will be a matrix (vector) each of whose entries is nonzero. A *strictly positive matrix* (*strictly positive vector*) will be a real matrix (vector) each of whose entries is positive. A *positive diagonal matrix* is a diagonal matrix each of whose diagonal entries is positive.

Let A be in $\mathcal{M}_n(\mathbb{C})$. The spectral radius of A will be denoted by $\rho(A)$. If the spectrum of A is real, $\lambda_{\max}(\lambda)$ will denote the maximum eigenvalue. If ω is a nonempty subset of $\{1, 2, \cdots, n\}$, then $A[\omega]$ will be the principal submatrix of A with entries indexed by the elements of ω.

* Received by the editors May 9, 1988; accepted for publication (in revised form) February 6, 1989.

† Department of Mathematics, University of Southern Mississippi, Hattiesburg, Mississippi 39406.

Let A be in $\mathcal{M}_n(\mathbb{C})$. The matrix A is *reducible* if there is an $n \times n$ permutation matrix P such that

$$PAP^t = \begin{bmatrix} B_1 & B_2 \\ 0 & B_3 \end{bmatrix}$$

where the matrices B_1 and B_3 are square matrices. If no such permutation matrix P exists, then A is *irreducible*.

A *Z-matrix* is a real, square matrix with all of its off-diagonal entries nonpositive. A ZME-*matrix* is a matrix all of whose positive integer powers are Z-matrices, and all of whose positive, odd powers are irreducible. The properties of the class of ZME-matrices and certain of its subclasses have been extensively studied (see [1]-[3], [5], [9]).

3. Inflation, inflators, and the matrix $G(U)$. Let m and n be positive integers with $m \leqq n$. An *m-partition of n* is a partition of the set $\{1, 2, \cdots, n\}$ into an ordered collection of m nonempty, disjoint sets such that the elements within each set are arranged in ascending order.

Throughout this paper, the following conventions are assumed: First, m and n are positive integers with $m \leqq n$; and second, the set Π is an m-partition of n given by $B_1, B_2, \cdots, B_m$.

Let U be in $\mathcal{M}_n(\mathbb{C})$. The partition Π induces a block partitioning of the matrix U as follows. For $1 \leqq i, j \leqq m$, the i, j block of U consists of all entries $U_{\alpha\beta}$ such that α is in B_i and β is in B_j. Denote the i, j block of U by $U_{\langle i,j \rangle}$.

Let x be in $\mathbb{C}^n$. Then Π partitions x into m subvectors such that the ith subvector has entries x_α where $\alpha \in B_i$. Denote the ith subvector by $x_{\langle i \rangle}$.

Let A be in $\mathcal{M}_m(\mathbb{C})$. Let u be in $\mathcal{M}_n(\mathbb{C})$. The *inflation matrix of A by U with respect to the partition* Π is the $n \times n$ matrix denoted by $A \times\times U$ defined as follows. For each α and β in $\{1, 2, \cdots, n\}$, there exist unique indices r and s such that $\alpha \in B_r$ and $\beta \in B_s$; let $(A \times\times U)_{\alpha\beta} = a_{rs}U_{\alpha\beta}$. Equivalently, in the block partition induced by the partition Π, $(A \times\times U)_{\langle r,s \rangle} = a_{rs}U_{\langle r,s \rangle}$ for each r and s. When the partitions are clear, $A \times\times U$ will be called an *inflation matrix*. (This is the definition of inflation given in Definition 4.1 of [2].)

Let U be in $\mathcal{M}_n(\mathbb{C})$. The matrix U is called an *inflator* (*with respect to* Π) if there exist vectors u and $\hat{u}$ in $\mathbb{C}^n$ that are partitioned by Π such that the following conditions hold:

(i) u and $\hat{u}$ are strictly nonzero vectors;
(ii) For $1 \leqq i, j \leqq m$, $U_{\langle i,j \rangle} = [u_{\langle i \rangle}]^t[\hat{u}_{\langle j \rangle}]$;
(iii) For $1 \leqq i \leqq m$, $u_{\langle i \rangle}[\hat{u}_{\langle i \rangle}]^t = 1$.

The pair of vectors u and $\hat{u}$ is called a *generating pair for the inflator U*. The matrix U is called a *normalized inflator* if u and $\hat{u}$ can be chosen so that they also satisfy a fourth condition:

(iv) For $1 \leqq i \leqq m$, $u_{\langle i \rangle}[u_{\langle i \rangle}]^* = \hat{u}_{\langle i \rangle}[\hat{u}_{\langle i \rangle}]^*$.

Observe that $U = u^t[\hat{u}]$. (These conditions are equivalent to Definition 4.3 of [2].)

Let U be an inflator associated with the m-partition Π of n. The matrix $G(U)$ is defined by $G(U) = I_n - (I_m \times\times U)$. Thus $G(U)$ can be expressed as the (internal) direct sum

$$G(U) = I_n - \left[\bigoplus_{i=1}^{m} U_{\langle i,i \rangle}\right] = \bigoplus_{i=1}^{m} [I - U_{\langle i,i \rangle}].$$

Thus $G(U)$ is permutation similar to a block-diagonal matrix. By convention, if $U = [0]$ is the 1×1 zero matrix, $G(U) = I_1$.

4. Inflation sequences and inflation-generated projectors. Let $n_1, n_2, \cdots, n_k$ be a sequence of integers such that $1 = n_1 < n_2 < \cdots < n_k = n$. For $1 < i \leqq k$, let $P_{i-1,i}$ be an n_{i-1}-partition of n_i. Let $U_1 = [0]$, the 1×1 zero matrix. For $1 < i \leqq k$, let U_i be an inflator associated with $P_{i-1,i}$. The sequence $\{U_i\}_{i=1}^{k}$ is called an *inflation sequence*. If each of the inflators U_i is normalized for $1 < i \leqq k$, then the sequence is called a *normalized inflation sequence*.

If $\{U_i\}_{i=1}^{k}$ is an inflation sequence, we will adopt the convention that $G(U_i) \times\times U_{i+1} \times\times \cdots \times\times U_k = G(U_k)$ when $i = k$. For $1 \leqq i \leqq k$, let $E_i = G(U_i) \times\times U_{i+1} \times\times \cdots \times\times U_k$. Let ε denote the set $\{E_i\colon 1 \leqq i \leqq k\}$. The set ε is called a *complete set of inflation-generated projectors*.

LEMMA 4.1. *Let $\{U_i\}_{i=1}^{k}$ be an inflation sequence. Let ε be the corresponding complete set of inflation-generated projectors. For $1 \leqq i \leqq k$, the $n \times n$ matrix E_i is an idempotent matrix of rank $(n_i - n_{i-1})$. Furthermore, $E_iE_j = \mathcal{O}$ when $i \neq j$, and*

$$\sum_{i=1}^{k} E_i = I_n.$$

Proof. See [2, § 6]. □

5. Inflation-generated matrices. The matrix A is called an *inflation-generated matrix* if there exist k pairwise distinct complex numbers $\alpha_1, \alpha_2, \cdots, \alpha_k$, and there exists an inflation sequence $\{U_i\}_{i=1}^{k}$ such that

$$A = \sum_{i=1}^{k} \alpha_i E_i$$

where for each i, the matrix E_i corresponds to U_i. (That is, the matrices E_i comprise the complete set of inflation-generated projectors corresponding to $\{U_i\}_{i=1}^{k}$.)

If the requirement that the α_i be distinct is relaxed, it is not clear that the resultant matrix is inflation-generated. (For one set of conditions under which the restriction can be relaxed, see [5].) It shall be seen that this will interfere with the iterated application of Theorem 8.1.

6. The normalization and its construction. The relationship between inflators and normalized inflators is partially revealed by the following result of Friedland, Hershkowitz, and Schneider [2, Lem. 4.16].

LEMMA 6.1. *Let U be an inflator with respect to Π. Then there exists a unique normalized inflator V with respect to Π such that $G(U) = G(V)$.*

The unique matrix V in Lemma 6.1 is called the *normalization of* U. In the remainder of this section, we show how V is obtained from U, and how V can be substituted for U in inflation products.

First, however, we present two results concerning inflation and diagonal matrices. The first follows directly from the definition of inflation.

LEMMA 6.2. *Let D be an $m \times m$ diagonal matrix. Let Π be an m-partition of n. Then $D \times\times I_n$ is a diagonal matrix. If D is a positive diagonal matrix, then $D \times\times I_n$ is a positive diagonal matrix. If D is nonsingular, then $D \times\times I_n$ is nonsingular with*

$$[D \times\times I_n]^{-1} = [D^{-1}] \times\times I_n.$$

LEMMA 6.3. *Let U be an inflator with respect to Π. Let U have generating pair u and $\hat{u}$ such that $U = u^t\hat{u}$. Let D be a nonsingular, $n \times n$ diagonal matrix. Then*

(i) *$\tilde{U} = DUD^{-1}$ is an inflator with respect to Π with generating pair uD and $\hat{u}D^{-1}$;*

(ii) $A \times\times \tilde{U} = D[A \times\times U]D^{-1}$ *for all* A *in* $\mathcal{M}_m(\mathbb{C})$;

(iii) $G(\tilde{U}) = DG(U)D^{-1}$.

Proof of (i). Since D is a nonsingular, diagonal matrix, it follows that uD and $\hat{u}D^{-1}$ are both strictly nonzero. Note that

$$\begin{aligned}\tilde{U}_{\langle i,j\rangle} &= [DUD^{-1}]_{\langle i,j\rangle} = D_{\langle i,i\rangle}U_{\langle i,j\rangle}[D^{-1}]_{\langle j,j\rangle}\\ &= D_{\langle i,i\rangle}[u_{\langle i\rangle}]^t\hat{u}_{\langle j\rangle}[D^{-1}]_{\langle j,j\rangle} = [[uD]_{\langle i\rangle}]^t[\hat{u}D^{-1}]_{\langle j\rangle}.\end{aligned}$$

Finally,

$$\begin{aligned}[uD]_{\langle i\rangle}[[\hat{u}D^{-1}]_{\langle i\rangle}]^t &= u_{\langle i\rangle}D_{\langle i,i\rangle}[D^{-1}]_{\langle i,i\rangle}[\hat{u}_{\langle i\rangle}]^t\\ &= u_{\langle i\rangle}[DD^{-1}]_{\langle i,i\rangle}[\hat{u}_{\langle i\rangle}]^t\\ &= u_{\langle i\rangle}[\hat{u}_{\langle i\rangle}]^t = 1.\end{aligned}$$

Proof of (ii). Let α, β be such that $i \in B_\alpha$ and $j \in B_\beta$. Then

$$\begin{aligned}[D(A \times\times U)D^{-1}]_{ij} &= D_{ii}(A \times\times U)_{ij}[D^{-1}]_{jj} = D_{ii}a_{\alpha\beta}U_{ij}[D^{-1}]_{jj}\\ &= a_{\alpha\beta}D_{ii}U_{ij}[D^{-1}]_{jj} = a_{\alpha\beta}[DUD^{-1}]_{ij} = [A \times\times \tilde{U}]_{ij}.\end{aligned}$$

Proof of (iii). From the definition of $G(U)$,

$$DG(U)D^{-1} = D[I_n - I_m \times\times U]D^{-1} = I_n - D[I_m \times\times U]D^{-1}.$$

Using part (ii), $D[I_m \times\times U]D^{-1} = I_m \times\times \tilde{U}$. The result follows from the definition of $G(\tilde{U})$. □

Let U be an inflator with respect to Π with generating pair u and $\hat{u}$. Since u and $\hat{u}$ are strictly nonzero vectors, there exist m unique, positive numbers λ_i that satisfy

$$u_{\langle i\rangle}[u_{\langle i\rangle}]^* = (\lambda_i)^2\hat{u}_{\langle i\rangle}[\hat{u}_{\langle i\rangle}]^*$$

for $1 \leqq i \leqq m$. Define $D(U)$ to be the diagonal matrix

$$D(U) = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_m).$$

The positive diagonal matrix $D(U)$ is called the *normalizer of* U.

The following lemma shows how the normalization of U is constructed using the normalizer of U.

LEMMA 6.4. *Let* U *be an inflator with respect to* Π *with generating pair* u *and* $\hat{u}$. *Let* $V = [D(U) \times\times I_n]^{-1}U[D(U) \times\times I_n]$. *Then* V *is the normalization of* U. *Furthermore,* V *has generating pair* $v = u[D(U) \times\times I_n]^{-1}$ *and* $\hat{v} = \hat{u}[D(U) \times\times I_n]$. *Equivalently,* $v_{\langle i\rangle} = (\lambda_i)^{-1}u_{\langle i\rangle}$ *and* $\hat{v}_{\langle i\rangle} = \lambda_i\hat{u}_{\langle i\rangle}$ *for each* i.

Proof. Let $D = [D(U) \times\times I_n]^{-1}$. By Lemma 6.2, D is a nonsingular, diagonal matrix. By Lemma 6.3, V is an inflator with respect to Π with generating pair v and $\hat{v}$. By the choice of the λ_i, V is a normalized inflator. Finally, by Lemma 6.3, $G(V) = DG(U)D^{-1} = I_n - D(I_m \times\times U)D^{-1}$. Thus $[G(V)]_{\langle i,j\rangle} = 0$ if $i \neq j$, and

$$\begin{aligned}[G(V)]_{\langle i,i\rangle} &= I - D_{\langle i,i\rangle}U_{\langle i,i\rangle}[D^{-1}]_{\langle i,i\rangle}\\ &= I - [(\lambda_i)^{-1}I]U_{\langle i,i\rangle}[\lambda_i I]\\ &= I - U_{\langle i,i\rangle} = [G(U)]_{\langle i,i\rangle}.\end{aligned}$$

Thus $G(V) = G(U)$. That is, V is the normalization of U. □

LEMMA 6.5. *Let* U *be an inflator with respect to* Π. *Let* V *be the normalization of* U. *Let* A *be in* $\mathcal{M}_m(\mathbb{C})$. *Then* $\tilde{A} = [D(U)]A[D(U)]^{-1}$ *is the unique matrix satisfying* $A \times\times U = \tilde{A} \times\times V$.

Proof. For each i and j,

$$V_{\langle i,j\rangle} = [[D(U)\times\times I_n]^{-1}]_{\langle i,i\rangle} U_{\langle i,j\rangle}[D(U)\times\times I_n]_{\langle j,j\rangle}$$
$$= [(\lambda_i)^{-1}I]U_{\langle i,j\rangle}[\lambda_j I] = [(\lambda_i)^{-1}\lambda_j]U_{\langle i,j\rangle}.$$

Hence for all A in $\mathcal{M}_m(\mathbb{C})$,

$$[A\times\times U]_{\langle i,j\rangle} = A_{ij}U_{\langle i,j\rangle} = A_{ij}[\lambda_i(\lambda_j)^{-1}]V_{\langle i,j\rangle}$$
$$= [[D(U)]A[D(U)]^{-1}]_{ij}V_{\langle i,j\rangle}$$
$$= [[[D(U)]A[D(U)]^{-1}]\times\times V]_{\langle i,j\rangle}.$$

Thus $A \times\times U = \tilde{A} \times\times V$ if and only if

$$\tilde{A} = [D(U)]A[D(U)]^{-1}. \qquad \square$$

Remark. By Lemma 6.4, the inflator U and its normalization V are positive diagonally similar. The matrices A and $\tilde{A}$ in the previous lemma are positive diagonally similar. In both cases, the normalizer of U is or generates the positive diagonal matrix needed for similarity.

7. The normalization theorem. The principal result in this section is actually Theorem 2.4.4 of [5]. (In [5], it appears without a proof.)

THEOREM 7.1 (The Normalization Theorem). *Let $\{U_i\}_{i=1}^k$ be an inflation sequence. Then there exists an inflation sequence $\{V_i\}_{i=1}^k$ such that for each i, V_i is a normalized inflator with respect to the same partition as that of U_i, and*

$$G(V_i)\times\times V_{i+1}\times\times \cdots \times\times V_k = G(U_i)\times\times U_{i+1}\times\times\cdots\times\times U_k.$$

Furthermore, the following algorithm constructs the sequence $\{V_i\}_{i=1}^k$ from the sequence $\{U_i\}_{i=1}^k$.

ALGORITHM.

(i) *Let $D^{(k)}$ be the normalizer of U_k.*
(ii) *Let V_k be the normalization of U_k.*
(iii) *For $i = k-1, k-2, \cdots, 2$, let $D^{(i)}$ be the normalizer of the matrix $D^{(i+1)}U_i[D^{(i+1)}]^{-1}$.*
(iv) *For $i = k-1, k-2, \cdots, 2$, let V_i be the normalization of the matrix $D^{(i+1)}U_i[D^{(i+1)}]^{-1}$.*
(v) *Let $V_1 = U_1$, the 1×1 zero matrix.*

Proof. Observe that $D^{(k)}$ is a positive diagonal matrix, and that V_k is the unique normalized inflator with respect to the partition of U_k such that $G(U_k) = G(V_k)$. Suppose that the matrices $V_k, V_{k-1}, V_{k-2}, \cdots, V_{i+1}$ have been constructed and that they are normalized inflators with respect to the appropriate partitions. Then $D^{(i+1)}$ is a normalizer, and hence a positive diagonal matrix. If $i > 1$, let $\hat{U} = D^{(i+1)}U_i[D^{(i+1)}]^{-1}$. By Lemma 6.3, $\hat{U}$ is an inflator with respect to the same partition as U_i. Thus $D^{(i)} = D(\hat{U})$ is well defined. Let V_i be the normalization of $\hat{U}$. Then V_i is a normalized inflator with respect to the same partition as that of U_i.

Let $i < k$. Let $E = G(U_i) \times\times U_{i+1} \times\times \cdots \times\times U_k$. By Lemma 6.4, $D^{(k)}$ normalizes U_k. Then by Lemma 6.5,

$$E = [D^{(k)}[G(U_i)\times\times U_{i+1}\times\times\cdots\times\times U_{k-1}][D^{(k)}]^{-1}]\times\times V_k.$$

By Lemma 6.3,

$$E = [[G(U_i)\times\times U_{i+1}\times\times\cdots\times\times U_{k-2}]\times\times[D^{(k)}U_{k-1}[D^{(k)}]^{-1}]]\times\times V_k.$$

By its definition, $D^{(k-1)}$ is the normalizer for $D^{(k)}U_{k-1}[D^{(k)}]^{-1}$, transforming it into V_{k-1}. Then by Lemma 6.5,

$$E=[[D^{(k-1)}[G(U_i)\times\times U_{i+1}\times\times\cdots\times\times U_{k-2}][D^{(k-1)}]^{-1}]\times\times V_{k-1}]\times\times V_k.$$

Iterating this process $(k-i)$ times yields

$$E=[D^{(i+1)}G(U_i)[D^{(i+1)}]^{-1}]\times\times V_{i+1}\times\times V_{i+2}\times\times\cdots\times\times V_k.$$

If $i=1$, then $G(U_1)=I_1$, so $D^{(2)}G(U_1)[D^{(2)}]^{-1}=I_1=G(V_1)$. If $i>1$, then by Lemma 6.3,

$$D^{(i+1)}G(U_i)[D^{(i+1)}]^{-1}=G([D^{(i+1)}U_i[D^{(i+1)}]^{-1}]).$$

By Lemma 6.4, the right-hand side of the preceding expression is equal to $G(V_i)$. Thus $E=G(V_i)\times\times V_{i+1}\times\times\cdots\times\times V_k$. □

COROLLARY 7.2. *Let $\{V_i\}_{i=1}^{k}$ be the inflation sequence constructed from $\{U_i\}_{i=1}^{k}$ by the algorithm in Theorem 7.1. Then for each i, U_i and V_i are positive diagonally similar. For $i=1$, $U_1=V_1=[0]$, and any positive 1×1 matrix will produce the transformation. For $1<i<k$, the positive diagonal matrix $[D^{(i)}\times\times I_{n_i}]^{-1}D^{(i+1)}$ will produce the transformation. For $i=k$, the matrix $[D^{(k)}\times\times I_n]$ will produce the transformation.*

Proof. This follows from the definitions, Lemma 6.2, and Theorem 7.1. □

8. An eigenvector test for inflation matrices. The following theorem, which contains the eigenvector test as one of its equivalent conditions, is proven in § 9.

THEOREM 8.1. *Let A be in $\mathcal{M}_n(\mathbb{C})$. Let ρ be in $\mathbb{C}$. The following are equivalent:*

(i) *$A=B\times\times U+\rho G(U)$ for some inflator U;*

(ii) *$A=B\times\times U+\rho G(U)$ for some normalized inflator U;*

(iii) *$A=C\times\times V+\rho G(V)$ where V is an inflator corresponding to an $(n-1)$-partition of n;*

(iv) *$A=C\times\times V+\rho G(V)$ where V is a normalized inflator corresponding to an $(n-1)$-partition of n;*

(v) *A has a row eigenvector x and a column eigenvector y^t both corresponding to the eigenvalue ρ such that $xy^t\neq 0$, such that* supp (x) = supp (y), *and such that* $|\text{supp}\,(x)|=2$.

Remark. In the proof of (v) implies (iii), it is shown how to explicitly construct the matrices C and V in (iii) from x and y in (v). Using the results on normalization, the matrices B and U in (ii) can be constructed from the matrices B and U in (i), and the matrices C and V in (iv) can be constructed from the matrices C and V in (iii). For a construction of x and y in (v), given the inflator U or V from (i)–(iv), see Theorem 5.1 of [6].

Remark. Recall that an inflation matrix is merely a matrix of the form $A=B\times\times U$ for some inflator U. That is, $A=B\times\times U+\rho G(U)$ where $\rho=0$. Consequently, Theorem 8.1 yields an obvious corollary relating the existence of eigenvectors for the zero eigenvalue of a matrix with that matrix being an inflation matrix.

As the following example demonstrates, the condition $xy^t\neq 0$ cannot be relaxed if x and y in (v) are to be eigenvectors for $G(V)$ where V is as in (iii) or (iv).

Example. Let $u=(i\ \ 1\ \ 1)$. Let $U=u^tu$. Then U is an inflator corresponding to the unique 1-partition of 3: $\{1, 2, 3\}$. Let $A=G(U)$, then A has row and column eigenvectors x and y^t, respectively, where $x=y=(i\ \ 1\ \ 0)$. Note that $xy^t=0$. Suppose that there eixsts an inflator V such that $A=B\times\times V+G(V)$ where $G(V)$ is rank one, $xG(V)=x$, and $[G(V)]y^t=y^t$. Let Π be the partition corresponding to V. Let v and $\hat{v}$ be a generating pair for V. By Theorem 5.1 of [6], $v_{\langle i\rangle}[x_{\langle i\rangle}]^t=0$ for each i.

Thus v has the form $v = (\alpha \ \ \beta | 1)$ for some nonzero complex numbers α and β. Then $v_{\langle 1\rangle}[x_{\langle 1\rangle}]^t = 0$; that is, $\alpha i + \beta = 0$, forcing $v_{\langle 1\rangle} = cx_{\langle 1\rangle}$ for some $c \neq 0$. Similarly, $\hat{v}_{\langle 1\rangle} = \hat{c}y_{\langle 1\rangle}$ for some $\hat{c} \neq 0$. Note that $\hat{v}_{\langle 1\rangle}[v_{\langle 1\rangle}]^t = c\hat{c}x_{\langle 1\rangle}[y_{\langle 1\rangle}]^t = 0$, contradicting the fact that v and $\hat{v}$ are a generating pair. Note that $G(U)$ does have a pair of eigenvectors that satisfy condition (v) of the theorem: $x = y = (0 \ \ 1 \ \ -1)$. For this pair, $A = G(W) \times\times V + G(V)$ where W is the inflator $W = ww^t$ and V is the inflator $V = v^tv$, with $w = (\sqrt{2} \ \ i)$ and $v = \frac{1}{2}(2 | \sqrt{2} \ \ \sqrt{2})$.

The condition $|\text{supp}\,(x)| = 2$ cannot be replaced by a simple, weaker condition. The essential problem is that if a vector x has exactly two nonzero entries, if x and u have the same zero pattern, and if $xu^t = 0$, then u is uniquely determined (up to a scalar multiple) by x. If, however, x has three or more nonzero entries, then the space of solutions to $xu^t = 0$ is at least two-dimensional. Consider the following example.

Example. Let A be the matrix

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

The matrix A has spectrum $\{1, 1, 4\}$. Clearly, each of the vectors $(1 \ \ -1 \ \ 0)$ and $(0 \ \ 1 \ \ -1)$ is a row eigenvector and the transpose of a column eigenvector for A. Since each of these vectors satisfies condition (v) of Theorem 8.1, A must be inflation-generated. Since the eigenspace for the eigenvalue one is two-dimensional, it is easy to construct an orthonormal basis of eigenvectors for A such that each eigenvector has three nonzero entries. For example, let $a = (\sqrt{42})^{-1}(-5, 4, 1)$, $b = (\sqrt{14})^{-1}(1, 2, -3)$, and $c = (\sqrt{3})^{-1}(1, 1, 1)$. Let $P = a^ta$, $Q = b^tb$, and $R = c^tc$. Then $A = 1(P + Q) + 4R$. The projector for the eigenvalue one is $P + Q = G(U)$ where $U = c^tc$. Let B be the matrix $B = 1P + \lambda Q + 4R$. Suppose that $\lambda \in \mathbb{C} \setminus \{1, 4\}$. If B were of the form $B = C \times\times V + \rho G(V)$ for some inflator V, then one of P, Q, and R would be the matrix $G(V)$. Since each of P, Q, and R has no nonzero entries, each corresponds to a one-partition of three. Thus if one of these matrices is to be of the form $G(V)$, it must be rank 2, which is clearly false. Thus we have a matrix that has row and column eigenvector pairs x, y^t such that $x = y$, $xy^t \neq 0$, $|\text{supp}\,(x)| = 3$, but such that $B \neq C \times\times V + \rho G(V)$.

9. The proof of Theorem 8.1. The following equivalences are proved: (i) ⇔ (ii), (i) ⇔ (iii) ⇔ (v), and (iii) ⇔ (iv). Furthermore, (iv) ⇒ (iii) ⇒ (i), and (ii) ⇒ (i) are obvious.

Proof of (i) ⇒ (ii). If U is normalized, then the result holds. If U is not normalized, construct a normalized inflator $\hat{U}$ using Lemma 6.4. Let $\hat{B}$ be constructed using Lemma 6.5. Then

$$B \times\times U + \rho G(U) = \hat{B} \times\times \hat{U} + \rho G(\hat{U}).$$

Proof of (iii) ⇒ (iv). Apply the proof of (i) ⇒ (ii); substitute V for U, and C for B.

Proof of (i) ⇒ (iii). Let Π be the partition corresponding to the inflator U. If Π is an $(n - 1)$-partition of n, the result is clear. If Π is an m-partition of n where $m < (n - 1)$, then by Theorem 4.2 of [7], there is a sequence of inflators $U_1, U_2, \cdots, U_{n-m}$ corresponding to a sequence of partitions such that the partition for U_k is a $(m + k - 1)$-partition of $(m + k)$, and such that $U = U_1 \times\times U_2 \times\times \cdots \times\times U_{n-m}$. By Theorem 2.6 of [7], $W = U_1 \times\times U_2 \times\times \cdots \times\times U_{n-m-1}$ is an inflator corresponding to an m-partition of $(n - 1)$. By Theorem 2.6 of [7], $G(U) = [G(W)] \times\times U_{n-m} + G(U_{n-m})$. Let $B = C \times\times W + \rho G(W)$. Then $A = B \times\times V + \rho G(V)$ where V is the inflator U_{n-m} corresponding to an $(n - 1)$-partition of n.

Proof of (iii) $\Rightarrow$ (v). Let Π be the $(n-1)$-partition of n. Then there is a unique subset in Π that is not a singleton, and without loss of generality, it can be assumed that the subsets of Π have been labeled so that this set is B_1. Then $B_1 = \{\alpha, \beta\}$ for some α and β with $\alpha \neq \beta$ and $1 \leqq \alpha, \beta \leqq n$. Let v and $\hat{v}$ be a generating pair for U. Then $v_{\langle i\rangle} = \hat{v}_{\langle i\rangle} = [1] \in \mathbb{R}^1$ for $1 < i \leqq (n-1)$, and both of $v_{\langle 1\rangle}$ and $\hat{v}_{\langle 1\rangle}$ are strictly nonzero vectors in $\mathbb{C}^2$. By Theorem 5.1 of [6], there exist nonzero vectors w and $\hat{w}$ in $\mathbb{C}^n$ that are partitioned by Π and that satisfy the following conditions:

(1) $w_{\langle 1\rangle}[v_{\langle 1\rangle}]^t = 0$ and $\hat{w}_{\langle 1\rangle}[\hat{v}_{\langle 1\rangle}]^t = 0$;

(2) $w_{\langle i\rangle} = \hat{w}_{\langle i\rangle} = [0]$ for $1 < i \leqq (n-1)$;

(3) $[G(U)]\hat{w}^t = \hat{w}^t$ and $w[G(U)] = w$;

(4) $w[B \times\times U] = \mathcal{O}$ and $[B \times\times U]\hat{w}^t = \mathcal{O}^t$ where $\mathcal{O}$ denotes the zero vector in $\mathbb{C}^n$.

That is, there are row and column eigenvectors for A corresponding to the eigenvalue ρ such that their supports are B_1. Since V is an inflator, $v_{\langle 1\rangle}[\hat{v}_{\langle 1\rangle}]^t = 1$. Since $v_{\langle 1\rangle}$, $\hat{v}_{\langle 1\rangle}$, $w_{\langle 1\rangle}$, and $\hat{w}_{\langle 1\rangle}$ all lie in $\mathbb{C}^2$, condition (1) implies $w_{\langle 1\rangle}[\hat{w}_{\langle 1\rangle}]^t \neq 0$. Hence $w\hat{w}^t \neq 0$.

Proof of (v) $\Rightarrow$ (iii). Without loss of generality, it may be assumed that x and y have been normalized so that $xy^t = 1$. Let $B_1 = \operatorname{supp}(x)$; and for $1 < i \leqq (n-1)$, let B_i be the ith element of $\{1, 2, \cdots, n\} \backslash \operatorname{supp}(x)$. Let Π be the $(n-1)$-partition of n formed by the sets $B_1, \cdots, B_{n-1}$. Let v and $\hat{v}$ be strictly nonzero vectors in $\mathbb{C}^n$ partitioned by Π such that $v_{\langle i\rangle} = \hat{v}_{\langle i\rangle} = [1] \in \mathbb{R}^1$ for $1 < i \leqq (n-1)$, and such that

$$v_{\langle 1\rangle}[x_{\langle 1\rangle}]^t = \hat{v}_{\langle 1\rangle}[y_{\langle 1\rangle}]^t = 0.$$

Since $1 = xy^t = x_{\langle 1\rangle}[y_{\langle 1\rangle}]^t$, this implies $v_{\langle 1\rangle}[\hat{v}_{\langle 1\rangle}]^t \neq 0$. Finally, by scaling the subvectors $v_{\langle 1\rangle}$ and $\hat{v}_{\langle 1\rangle}$ if necessary, it may be assumed that $v_{\langle 1\rangle}[\hat{v}_{\langle 1\rangle}]^t = 1$. Then v and $\hat{v}$ satisfy the definition of a generating pair for an inflator V corresponding to Π. It remains to show that $A = C \times\times V + \rho G(V)$.

First, it is shown that x and y are, respectively, row and column eigenvectors for $G(V)$ for the eigenvalue one. Computing, we have $xG(V) = x[I - I \times\times V] = x - x[I \times\times V]$. For $1 \leqq i \leqq (n-1)$,

$$[x[I \times\times V]]_{\langle i\rangle} = \sum_{j=1}^{n-1} x_{\langle j\rangle}[I \times\times V]_{\langle j,i\rangle} = x_{\langle i\rangle}[I \times\times V]_{\langle i,i\rangle}$$

$$= x_{\langle i\rangle}[v_{\langle i\rangle}]^t\hat{v}_{\langle i\rangle} = 0\hat{v}_{\langle i\rangle} = \mathcal{O}.$$

Thus $x[I \times\times V] = \mathcal{O}$, and hence $x[G(V)] = x$. Similarly, $[G(V)]y^t = y^t$.

It is known (see [2, § 4]) that $G(V)$ is a rank one, idempotent matrix; hence it is diagonalizable with eigenvalues one (multiplicity one) and zero (multiplicity $n-1$). Choose a basis for $\mathbb{C}^n$ of the form $\{x, \beta_2, \cdots, \beta_n\}$ where the β_i are row eigenvectors for $G(V)$ corresponding to zero. Now let $H = G(V)[A - \rho G(V)]$. Then

$$xH = xG(V)[A - \rho G(V)] = x[A - \rho G(V)] = \rho x - \rho x = \mathcal{O}.$$

For each i,

$$\beta_i H = \mathcal{O}[A - \rho G(V)] = \mathcal{O}.$$

Thus H annihilates a basis for $\mathbb{C}^n$. Thus $G(V)[A - \rho G(V)] = \mathcal{O}$, the $n \times n$ zero matrix. Similarly, $[A - \rho G(V)]G(V) = \mathcal{O}$. By [2, Lemma 4.23], $A - \rho G(V) = C \times\times V$ for some C in $\mathcal{M}_{n-1}(\mathbb{C})$. □

10. Recovering inflation sequences. Suppose that the matrix A is an inflation-generated matrix as in § 5. Then it is apparent that A can be expressed as $A = \hat{A} \times\times U_k +$

$\alpha_k G(U_k)$, where $\hat{A}$ is an inflation-generated matrix corresponding to the numbers α_1, $\alpha_2, \cdots, \alpha_{k-1}$, and the inflation sequence $\{U_i\}_{i=1}^{k-1}$. Thus Theorem 8.1 may be applied to $\hat{A}$. Care must be exercised, however, for it is not a priori apparent that given eigenvectors x and y for A that satisfy condition (v) of Theorem 8.1, that the eigenvalue ρ will be α_k. Consequently, when A is written as $A = B \times\times V + \rho G(V)$, it is not clear that B is inflation-generated. *Thus even when A is known to have an inflation sequence, it is not clear that that sequence (or any other sequence) can be recovered by repeated applications of Theorem* 8.1.

The example in the following section demonstrates the difficulties that may arise. We shall construct a complete set of inflation-generated projectors $\{E_1, E_2, E_3, E_4\}$ that is generated in that sequence, such that $E_3 = F_1 + F_2$ where F_1 and F_2 are a decomposition of E_3 into a pair of lower rank, orthogonal projectors. It shall be demonstrated that E_3 and E_4 both have eigenvectors that satisfy condition (v) of Theorem 8.1, and that for E_3, the eigenvectors actually correspond to F_1. Let A be the inflation-generated matrix

$$A = \alpha_1 E_1 + \alpha_2 E_2 + \alpha_3 E_3 + \alpha_4 E_4$$

where the α_i are distinct. Applying Theorem 8.1 to A using the eigenvectors corresponding to F_1, we obtain

$$A = B \times\times V_5 + \alpha_3 G(V_5)$$

where $G(V_5) = F_1$, and

$$B \times\times V_5 = \alpha_1 E_1 + \alpha_2 E_2 + \alpha_3 F_2 + \alpha_4 E_4.$$

The matrix B has only one pair of eigenvectors satisfying condition (v) of Theorem 8.1, the pair corresponding to H_4 where $H_4 \times\times V_5 = E_4$. Thus the inflation sequence, $\{V_i\}_{i=1}^{5}$ that we construct by repeated applications of Theorem 8.1 generates the projectors in the sequence E_1, E_2, F_2, E_4, F_1. Since α_3 is assigned to the projectors corresponding to both V_3 and V_5, *A is not an inflation-generated matrix with respect to this inflation sequence.*

11. A cautionary example. Let $U_1 = [0]$. Let $\omega = (\sqrt{2})^{-1}$. For $2 \leqq i \leqq 4$, let $U_i = [u^{(i)}]^t[u_i^{(i)}]$ where $u^{(2)} = (\omega\ \ \omega)$, $u^{(3)} = (\omega\ \ \omega | \omega\ \ \omega)$, and $u^{(4)} = (1|1|1|\omega\ \ \omega)$. Then $\{U_i\}_{i=1}^{4}$ is a normalized inflation sequence. Let $\{E_1, E_2, E_3, E_4\}$ be the corresponding complete set of inflation-generated projectors, with E_i corresponding to U_i for each i:

$$E_1 = \frac{1}{8}\begin{bmatrix} 2 & 2 & 2 & \sqrt{2} & \sqrt{2} \\ 2 & 2 & 2 & \sqrt{2} & \sqrt{2} \\ 2 & 2 & 2 & \sqrt{2} & \sqrt{2} \\ \sqrt{2} & \sqrt{2} & \sqrt{2} & 1 & 1 \\ \sqrt{2} & \sqrt{2} & \sqrt{2} & 1 & 1 \end{bmatrix},$$

$$E_2 = \frac{1}{8}\begin{bmatrix} 2 & 2 & -2 & -\sqrt{2} & -\sqrt{2} \\ 2 & 2 & -2 & -\sqrt{2} & -\sqrt{2} \\ -2 & -2 & 2 & \sqrt{2} & \sqrt{2} \\ -\sqrt{2} & -\sqrt{2} & \sqrt{2} & 1 & 1 \\ -\sqrt{2} & -\sqrt{2} & \sqrt{2} & 1 & 1 \end{bmatrix},$$

$$E_3 = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \oplus \frac{1}{4}\begin{bmatrix} 2 & -\sqrt{2} & -\sqrt{2} \\ -\sqrt{2} & 1 & 1 \\ -\sqrt{2} & 1 & 1 \end{bmatrix},$$

$$E_4 = [0] \oplus [0] \oplus [0] \oplus \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

Observe that $E_3 = F_1 + F_2$ where

$$F_1 = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \oplus [0] \oplus [0] \oplus [0],$$

$$F_2 = [0] \oplus [0] \oplus \frac{1}{4}\begin{bmatrix} 2 & -\sqrt{2} & -\sqrt{2} \\ -\sqrt{2} & 1 & 1 \\ -\sqrt{2} & 1 & 1 \end{bmatrix}.$$

Since all of the matrices are symmetric, row eigenvectors are also column eigenvectors. Both of E_3 and F_1 have eigenvector $[1 \ \ -1 \ \ 0 \ \ 0 \ \ 0]$ that corresponds to the eigenvalue one. For E_4, the vector $[0 \ \ 0 \ \ 0 \ \ 1 \ \ -1]$ is an eigenvector for eigenvalue one.

If A is the inflation-generated matrix

$$A = \alpha_1 E_1 + \alpha_2 E_2 + \alpha_3 E_3 + \alpha_4 E_4$$

where the α_i are distinct, then there are two distinct eigenvector pairs that satisfy condition (v) of Theorem 8.1. If the vectors for E_4 are chosen, then the inflation sequence $\{U_i\}_{i=1}^{4}$ is recovered by repeated applications of Theorem 8.1. If, however, $x = y = [1 \ \ -1 \ \ 0 \ \ 0 \ \ 0]$ is chosen, then the result is

$$A = B \times\times V + \alpha_3 G(V)$$

where $V = v^t v$ for $v = (\omega \ \ \omega \,|\, 1 \,|\, 1 \,|\, 1)$, where $G(V) = F_1$, and where

$$B \times\times V = \alpha_1 E_1 + \alpha_2 E_2 + \alpha_3 F_2 + \alpha_4 E_4.$$

It can be shown by direct computation that there are pairwise orthogonal, idempotent matrices H_1, H_2, H_3, and H_4 such that $E_1 = H_1 \times\times V$, $E_2 = H_2 \times\times V$, $F_2 = H_3 \times\times V$, and $E_4 = H \times\times V$. Furthermore, $H_i = G(W)$ for some normalized inflator W only if $i = 4$. In this case, $W = w^t w$ where $w = (1 \,|\, 1 \,|\, \omega \ \ \omega)$. Thus the matrix B has a unique pair of vectors satisfying condition (v) of Theorem 8.1, $x = y = [0 \ \ 0 \ \ 1 \ \ -1]$. By repeatedly applying Theorem 8.1, an inflation sequence $\{V_i\}_{i=1}^{5}$ is constructed such that E_1 corresponds to V_1, such that E_2 corresponds to V_2, such that F_2 corresponds to V_3, such that E_4 corresponds to $V_4 = W$, and such that F_1 corresponds to $V_5 = V$. Note that the eigenvalue α_3 of A is associated with two (nonconsecutive) inflators. Thus A is not inflation-generated with respect to the constructed sequence $\{V_i\}_{i=1}^{5}$, even though A is known to be inflation-generated.

12. Recovering inflation sequences for ZME-matrices. The difficulty of knowing a priori that we have chosen a correct pair of eigenvectors satisfying condition (v) of Theorem 8.1 can be alleviated under certain conditions, most notably when the original matrix is a ZME-matrix. Note that if A is a ZME-matrix, then by Lemma 3.1 of [2], the spectrum of A is real.

THEOREM 12.1. *Let A be a* ZME-*matrix. Let $\rho = \lambda_{\max}(A)$. Then there exist row and column eigenvectors of A corresponding to ρ that satisfy condition* (v) *of Theorem* 8.1. *Furthermore, if V is the inflator constructed using the eigenvectors, and B is the unique matrix satisfying $B \times\times V = A - \rho G(V)$, then B is a* ZME-*matrix. Thus repeated applications of Theorem* 8.1 *will produce an inflation sequence with respect to which A is an inflation-generated matrix.*

Before proceeding with the proof of Theorem 12.1, we prove the following lemma.

LEMMA 12.2. *Let U be an inflator corresponding to an m-partition of n where $m < (n-1)$. Let x and y^t be row and column eigenvectors, respectively, for $G(U)$ corresponding to the eigenvalue one. Suppose that $xy^t \neq 0$, that* $\operatorname{supp}(x) = \operatorname{supp}(y)$, *and that* $|\operatorname{supp}(x)| = 2$. *Then there is a pair of inflators W and V satisfying the following properties*:

(i) $U = W \times\times V$;

(ii) $G(U) = G(W) \times\times V + G(V)$;

(iii) *V corresponds to an $(n-1)$-partition of n;*

(iv) *x and y^t are row and column eigenvectors, respectively, for $G(V)$ corresponding to the eigenvalue one;*

(v) *if U is strictly positive, then W and V may be chosen to be strictly positive.*

Proof. The proof will consist of constructing an inflator V from U such that (iii) and (iv) hold. The construction of V also generates W satisfying (i). Finally, if (i) is shown to hold, then (ii) holds by § 4 of [7].

Let Π be the partition corresponding to U. By Theorem 5.1 of [6], $xG(U) = x$ if and only if $x_{\langle i\rangle}U_{\langle i,i\rangle} = \mathcal{O}$ for each i. Since U is strictly positive, this implies that the support of x (and hence of y) must be contained in one of the partition sets of Π. Without loss of generality, this is the set B_1.

Let u and $\hat{u}$ be a generating pair for U. There are two cases: $|B_1| = 2$; $|B_1| > 2$. In the first case, let w and $\hat{w}$ be vectors in $\mathbb{C}^{n-1}$ with blocks defined by

$$w_{\langle i\rangle} = \begin{cases} [1] & \text{if } i = 1, \\ u_{\langle i\rangle} & \text{if } i > 1, \end{cases}$$

$$\hat{w}_{\langle i\rangle} = \begin{cases} [1] & \text{if } i = 1, \\ \hat{u}_{\langle i\rangle} & \text{if } i > 1. \end{cases}$$

Clearly, w and $\hat{w}$ are a generating pair for an inflator W corresponding to an m-partition of $(n-1)$. Furthermore, if U is strictly positive, then so is W. Let v and $\hat{v}$ be the vectors in $\mathbb{C}^n$ with blocks defined by

$$v_{\langle i\rangle} = \begin{cases} u_{\langle 1\rangle} & \text{if } i = 1, \\ [1] & \text{if } 2 \leqq i \leqq n-1, \end{cases}$$

$$\hat{v}_{\langle i\rangle} = \begin{cases} \hat{u}_{\langle 1\rangle} & \text{if } i = 1, \\ [1] & \text{if } 2 \leqq i \leqq n-1. \end{cases}$$

Clearly, v and $\hat{v}$ are a generating pair for an inflator V corresponding to an $(n-1)$-partition of n. Since $x_{\langle 1\rangle}V_{\langle 1,1\rangle} = x_{\langle 1\rangle}U_{\langle 1,1\rangle} = \mathcal{O}$, it follows that $xG(V) = x$ by Theorem 5.1 of [6]. Similarly, $[G(V)]y^t = y^t$. It is a direct computation that $U = W \times\times V$. Finally, if U is strictly positive, then u and $\hat{u}$ are strictly positive, and clearly so are W and V.

In the second case, $|B_1| \geqq 3$. Let $k = |B_1|$. Let α and β be the indices in supp (x). For convenience, let $a = u_\alpha$, $\hat{a} = \hat{u}_\alpha$, $b = u_\beta$, and $\hat{b} = \hat{u}_\beta$. Since x is an eigenvector for $G(U)$, it follows from Theorem 5.1 of [6] that $x_{\langle 1\rangle}[u_{\langle 1\rangle}]^t = 0$. That is, $x_\alpha a + x_\beta b = 0$. Since u is a strictly nonzero vector, $(a \ \ b) = c(x_\beta \ \ -x_\alpha)$ for some nonzero complex number c. Similarly, $(\hat{a} \ \ \hat{b}) = \hat{c}(y_\beta \ \ -y_\alpha)$ for some $\hat{c} \neq 0$. Then $xy^t \neq 0$ implies $a\hat{a} + b\hat{b} \neq 0$. Let λ be a complex number (chosen positive if possible) such that

$$\lambda = [a\hat{a} + b\hat{b}]^{-1/2}.$$

Let v and $\hat{v}$ be the vectors in $\mathbb{C}^n$ with blocks defined by

$$v_{\langle i\rangle} = \begin{cases} [\lambda a \ \ \lambda b] & \text{if } i = 1, \\ [1] & \text{if } 2 \leqq i \leqq n-2, \end{cases}$$

$$\hat{v}_{\langle i\rangle} = \begin{cases} [\lambda\hat{a} \ \ \lambda\hat{b}] & \text{if } i = 1, \\ [1] & \text{if } 2 \leqq i \leqq n-2. \end{cases}$$

Clearly, v and $\hat{v}$ are a generating pair for an inflator V such that V corresponds to the $(n-1)$-partition Ω of n given by $\{\alpha, \beta\}$ and $(n-1)$ singletons. Furthermore, if U is strictly positive, then so are u and $\hat{u}$, and hence so is V. Since $x_{\langle 1\rangle}U_{\langle 1,1\rangle} = \mathcal{O}(\langle\ \ ,\ \ \rangle$ with respect to Π), and since $\operatorname{supp}(x) = \{\alpha, \beta\}$, it follows that $x_{\langle 1\rangle}V_{\langle 1,1\rangle} = \mathcal{O}(\langle\ \ ,\ \ \rangle$ with respect to Ω). Thus $xG(V) = x$ by Theorem 5.1 of [6]. Similarly, $[G(V)]y^t = y^t$.

Let z be the vector in $\mathbb{C}^{k-1}$ with entries defined by $z_1 = \lambda^{-1}$, and $z_i = [u_{\langle 1\rangle}]_{i+1}$ for $2 \leqq i < k$. Then z is strictly nonzero. If u and $\hat{u}$ are strictly positive, then so is z. Similarly define $\hat{z}$ in terms of λ and $\hat{u}_{\langle 1\rangle}$. Computing, $z\hat{z}^t = u_{\langle 1\rangle}[\hat{u}_{\langle 1\rangle}]^t = 1$ since $\lambda^{-2} = a\hat{a} + b\hat{b}$. Now let w and $\hat{w}$ be the vectors in $\mathbb{C}^{n-1}$ defined by

$$w_{\langle i\rangle} = \begin{cases} z & \text{if } i = 1, \\ u_{\langle i\rangle} & \text{if } 2 \leqq i \leqq m, \end{cases}$$

$$\hat{w}_{\langle i\rangle} = \begin{cases} z & \text{if } i = 1, \\ \hat{u}_{\langle i\rangle} & \text{if } 2 \leqq i \leqq m. \end{cases}$$

Then w and $\hat{w}$ are a generating pair for an inflator corresponding to an m-partition of $(n-1)$. Again, if U is strictly positive, then so is W. By direct computation, $U = W \times\times V$. □

Proof of Theorem 12.1. Let A be an $n \times n$ ZME-matrix. By Theorem 6.18 of [2], there is a normalized inflation sequence $\{U_i\}_{i=1}^{k}$ and a sequence of real numbers α_i satisfying $-\alpha_2 \leqq \alpha_1 < \cdots < \alpha_k$ such that

$$A = \sum_{i=1}^{k} \alpha_i E_i$$

where the E_i are in the complete set of inflation-generated projectors corresponding to $\{U_i\}_{i=1}^{k}$, and where E_i corresponds to U_i for each i. The eigenvalue of maximum modulus is α_k, and $E_k = G(U_k)$.

There are two cases to consider: (i) $G(U_k)$ has rank one, and (ii) $G(U_k)$ has rank at least two. In case (i), the row and column eigenvectors for A corresponding to α_k are unique up to scalar multiplication, hence they must be multiples of x and y, the pair of eigenvectors satisfying condition (v) of Theorem 8.1. Let V be the normalized, rank one inflator obtained from x and y. Then $G(V) = G(U_k)$; and since V and U_k are both normalized, $V = U_k$ by Lemma 4.16 of [2]. Then $A = B \times\times V + \alpha_k G(V)$ where B has inflation sequence $\{U_i\}_{i=1}^{k-1}$ and eigenvalues α_i satisfying $-\alpha_2 \leqq \alpha_1 < \cdots < \alpha_{k-1}$. By Theorem 6.18 of [2], B is a ZME-matrix.

In case (ii), let x and y^t be the eigenvectors satisfying condition (v) of Theorem 8.1. Then x and y^t are eigenvectors for $G(U_k)$ satisfying the hypotheses of Lemma 12.2. Then it follows that $U_k = \tilde{W} \times\times \tilde{U}$ where $\tilde{W}$ and $\tilde{U}$ are strictly positive inflators. By Lemmas 6.4 and 6.5, $\tilde{W}$ and $\tilde{U}$ may be chosen so that they are also normalized. Since $G(U_k) = G(\tilde{W}) \times\times \tilde{U} + G(\tilde{U})$, it follows that $A = B \times\times \tilde{U} + \alpha_k G(\tilde{U})$ where B has inflation sequence $\{U_1, U_2, \cdots, U_{k-1}, \tilde{W}\}$ and eigenvalues $-\alpha_2 \leqq \alpha_1 < \cdots < \alpha_k$, and hence B is a ZME-matrix. By Theorem 6.18 of [2], B is a ZME-matrix. Finally, if V is the normalized, rank one inflator constructed from x and y, then since $G(\tilde{U})$ is also rank one with eigenvectors x and y, $G(V) = G(\tilde{U})$. Hence by Lemma 4.16 of [2], $V = \tilde{U}$. □

13. The spectral radius of a ZME-matrix. In this section, it is shown that the spectral radius of an $n \times n$ ZME-matrix is the maximum of the maximal eigenvalues of the 2×2 principal submatrices. Since the maximal eigenvalue of a 2×2 matrix is just a root of a quadratic equation based on its entries, this provides a simple computational tool for finding the spectral radius of a ZME-matrix.

THEOREM 13.1. *Let A be in $\mathcal{M}_n(\mathbb{R})$ for $n \geqq 2$. If A is a* ZME-*matrix, then*

$$\rho(A)=\lambda_{\max}(A)=\max_{1\leqq i<j\leqq n}\rho(A[\{i,j\}]) \tag{13.2}$$

$$=\max_{1\leqq i<j\leqq n}\frac{1}{2}[a_{ii}+a_{jj}+[(a_{ii}-a_{jj})^2+4a_{ij}a_{ji}]^{1/2}]. \tag{13.3}$$

Proof. Suppose that A is a ZME-matrix. By Lemma 3.1 of [2], A has a real spectrum and $\rho(A) = \lambda_{\max}(A)$. By Theorem 9.6 of [2], there exist a diagonal matrix D with all its diagonal entries positive and a symmetric ZME-matrix $\hat{A}$ such that $A = D\hat{A}D^{-1}$. For all nonempty ω in $\{1, 2, \cdots, n\}$, $A[\omega] = D[\omega]\hat{A}[\omega](D[\omega])^{-1}$. Since such similarity transformations preserve spectra, it suffices to prove the theorem for symmetric A.

Suppose that A is a symmetric ZME-matrix. Then A is Hermitian, and the Cauchy eigenvalue interlacing inequalities hold [4, Result II.4.4.7, p. 119]. In particular, $\lambda_{\max}(A[\omega]) \leqq \lambda_{\max}(A)$ for all ω in $\{1, 2, \cdots, n\}$. Since A is an $n \times n$ ZME-matrix with $n \geqq 2$, Theorem 12.1 asserts that there is an eigenvector v for A corresponding to $\lambda_{\max}(A)$ such that $|\text{supp}(v)| = 2$. Let $\hat{\omega} = \text{supp}(v)$. Then the subvector of v consisting of the two nonzero entries is an eigenvector for $A[\hat{\omega}]$ for the eigenvalue $\lambda_{\max}(A)$. Thus $\lambda_{\max}(A) \leqq \lambda_{\max}(A[\hat{\omega}])$. Noting that $\hat{\omega} = \{i, j\}$ for some i and j with $1 \leqq i < j \leqq n$, (13.2) clearly holds. Since $A[\hat{\omega}]$ is a Hermitian matrix, $\lambda_{\max}(A[\hat{\omega}])$ is the larger root of the quadratic equation

$$\lambda^2-[\text{tr}(A[\hat{\omega}])]\lambda+\det(A[\hat{\omega}])=0.$$

By applying the quadratic formula and simplifying, (13.3) holds. □

Remark. Given an $n \times n$ ZME-matrix A, we can simplify the task of determining an inflator V for which $A = C \times\times V + \rho(A)G(V)$. The unique, nontrivial 2×2 block of V will correspond to one of the 2×2 submatrices that yields $\lambda_{\max}(A)$. Given a 2×2 submatrix that yields $\lambda_{\max}(A)$, we need only construct a row and a column eigenvector for $\lambda_{\max}(A)$ for this submatrix. These vectors are then extended n-vectors by adjoining $(n-2)$ zeros so that the nonzero entries have indices corresponding to the submatrix. If the resultant n-vectors are eigenvectors for A, then they give rise to V as in Theorem 8.1. Furthermore, at least one pair of vectors will extend to a pair of eigenvectors for A that give rise to V.

REFERENCES

[1] M. FIEDLER, *Characterizations of* MMA-*matrices*, Linear Algebra Appl., to appear.

[2] S. FRIEDLAND, D. HERSHKOWITZ, AND H. SCHNEIDER, *Matrices whose powers are Z-matrices or M-matrices*, Trans. Amer. Math. Soc., 300 (1987), pp. 343–366.

[3] D. HERSHKOWITZ AND H. SCHNEIDER, *Matrices with a sequence of accretive powers*, Israel Math. J., 55 (1986), pp. 344–372.

[4] M. MARCUS AND H. MINC, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, MA, 1964.

[5] H. SCHNEIDER AND J. STUART, *Allowable spectral perturbations for* ZME-*matrices*, Linear Algebra Appl., 111 (1988), pp. 63–118.

[6] ———, *Eigenvectors for inflation matrices and inflation-generated matrices*, Linear and Multilinear Algebra, 22 (1987), pp. 249–265.

[7] ———, *The decomposition of idempotents associated with inflators*, Linear Algebra Appl., 97 (1987), pp. 171–184.

[8] ———, *Inflation matrices and* ZME-*matrices that commute with a permutation matrix*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 408–418.

[9] ———, *The solutions to an infinite family of matrix inequalities with* ZME-*matrix coefficients*, Linear Algebra Appl., 108 (1988), pp. 141–156.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 4, pp. 533-541, October 1989

009

A NOTE ON LOCAL BEHAVIOR OF MULTIPLE EIGENVALUES*

JI-GUANG SUN†

Abstract. This note is a continuation of the work in [*J. Comput. Math.*, 6 (1988), pp. 28-38]. The directional derivatives of multiple eigenvalues of a symmetric eigenproblem analytically dependent on several parameters are given. The result can be used to define the sensitivity of multiple eigenvalues, and it is useful for investigating structural vibration design and control system design.

Key words. symmetric matrices, multiple eigenvalues, directional derivatives, sensitivity of eigenvalues

AMS(MOS) subject classifications. 15A18, 65F15

1. Introduction. Although the investigation of the sensitivity analysis of eigenvalues has a long history [4]-[6], [8]-[11], the case of multiple eigenvalues dependent on several parameters is rarely treated in the literature. The local behavior of multiple eigenvalues dependent on several parameters is quite different from the case of one parameter [8]-[10].

In this paper, which is a continuation of the work in [10], we investigate the local behavior of multiple eigenvalues of a symmetric eigenproblem analytically dependent on several parameters. The directional derivatives of the eigenvalues are discussed. The results of this note are useful for investigating structural vibration design and control system design.

Notation. The symbol $\mathcal{R}^{m\times n}$ denotes the set of real $m \times n$ matrices. We set $\mathcal{R}^n = \mathcal{R}^{n\times 1}$, $\mathcal{R} = \mathcal{R}^1$,

$$\mathcal{R}_r^{m\times n} = \{A \in \mathcal{R}^{m\times n}: \operatorname{rank}(A) = r\}$$

and

$$\mathcal{S}\mathcal{R}^{n\times n} = \{A \in \mathcal{R}^{n\times n}: A^T = A\}.$$

The matrix $I^{(n)}$ is the $n \times n$ identity and zero is the null matrix. The matrix $A > 0$ means that A is positive definite. We use $\rho(\ )$ for the spectral radius and $\|\ \|_2$ for the spectral norm and the usual Euclidean vector norm. The set of eigenvalues of an eigenproblem $Ax = \lambda x$ is denoted by $\lambda(A)$, and the set of eigenvalues of an eigenproblem $Ax = \lambda Bx$ is denoted by $\lambda(A, B)$. The eigenvalues of an $n \times n$ matrix A will be denoted by $\lambda_1(A), \cdots, \lambda_n(A)$.

Let $p = (p_1, \cdots, p_N)^T \in \mathcal{R}^N$. In this paper we consider the eigenproblem

$$A(p)x(p) = \lambda(p)B(p)x(p), \quad \lambda(p) \in \mathcal{R}, \quad x(p) \in \mathcal{R}^n, \quad p \in \mathcal{S}, \tag{1.1}$$

where $\mathcal{S}$ is an open subset of $\mathcal{R}^N$, and the matrices $A(p)$, $B(p) \in \mathcal{S}\mathcal{R}^{n\times n}$ are real analytic functions in $\mathcal{S}$ and $B(p) > 0$, for all $p \in \mathcal{S}$. Without loss of generality we may assume that the set $\mathcal{S}$ contains the origin, and throughout this paper we shall use the symbols $S_j(\eta)$ defined by

$$S_j(\eta) = \left(\frac{\partial A(p)}{\partial p_j}\right)_{p=0} - \eta\left(\frac{\partial B(p)}{\partial p_j}\right)_{p=0}, \qquad j = 1, \cdots, N,$$

where $A(p)$, $B(p)$ are the matrices appearing in the eigenproblem (1.1), and $\eta \in \mathcal{R}$.

* Received by the editors December 7, 1987; accepted for publication (in revised form) October 31, 1988. This project was supported by the National Natural Science Foundation of China and the Institute for Advanced Computer Studies of the University of Maryland.

† Computing Center, Academia Sinica, Beijing, China 100080, and Institute for Advanced Computer Studies, University of Maryland, College Park, Maryland 20742.

Let $\lambda(p)$ be a function defined in $\mathcal{S}$. The directional derivative of $\lambda(p)$ at $p^* \in \mathcal{S}$ in the direction ν, denoted by $D_\nu\lambda(p^*)$, is defined as follows:

$$D_\nu\lambda(p^*) \equiv \lim_{\tau \to +0} \frac{\lambda(p^*+\tau\nu)-\lambda(p^*)}{\tau}, \tag{1.2}$$

where $\nu \in \mathcal{R}^N$ with $\|\nu\|_2 = 1$ and τ is a positive scalar.

In § 2 we shall prove that every eigenvalue $\lambda(p)$ of the eigenproblem (1.1) has directional derivatives at each point of $\mathcal{S}$, and give expressions of the directional derivatives. On these grounds we can define the sensitivity of multiple eigenvalues useful for investigating structural vibration design and control system design.

2. Directional derivatives. Without loss of generality, we may investigate the directional derivatives of the eigenvalues at the origin of $\mathcal{R}^N$. The main result of this section is the following theorem.

THEOREM 2.1. *Let $p = (p_1, \cdots, p_N)^T \in \mathcal{R}^N$, and let $A(p), B(p) \in \mathcal{S}\mathcal{R}^{n\times n}$ be real analytic functions of p in some neighborhood $\mathcal{B}(0)$ of the origin of $\mathcal{R}^N$, where $B(p) > 0$, for all $p \in \mathcal{B}(0)$. Suppose that there is a matrix $X = (X_1, X_2) \in \mathcal{R}_n^{n\times n}$ with $X_1 \in \mathcal{R}^{n\times r}$ such that*

$$X^TA(0)X=\begin{pmatrix} \lambda_1 I^{(r)} & 0 \\ 0 & A_2 \end{pmatrix}, \quad X^TB(0)X=I, \quad \lambda_1 \notin \lambda(A_2). \tag{2.1}$$

Then there exist r continuous functions $\lambda_1(p), \cdots, \lambda_r(p)$ that are the eigenvalues of the eigenproblem (1.1) *satisfying*

$$\lambda_s(0)=\lambda_1, \qquad s=1, \cdots, r,$$

and for any direction $\nu = (\nu_1, \cdots, \nu_N)^T \in \mathcal{R}^N$ with $\|\nu\|_2 = 1$ there is a permutation π of $\{1, \cdots, r\}$ dependent on ν such that

$$D_\nu\lambda_s(0)=\lambda_{\pi(s)}\left(\sum_{j=1}^{N} \nu_j X_1^T S_j(\lambda_1) X_1 \right), \qquad s=1, \cdots, r. \tag{2.2}$$

Before we give the proof, we cite the Implicit Function Theorem [3, p. 277] and the Rellich theorem [9, pp. 42–45].

IMPLICIT FUNCTION THEOREM. *If the real-valued functions*

$$f_i(\xi_1, \cdots, \xi_s; \eta_1, \cdots, \eta_t), \qquad i=1, \cdots, s$$

are real analytic functions of $s + t$ real variables in some neighborhood of the origin of $\mathcal{R}^{s+t}$, if $f_i(0; 0) = 0$, $i = 1, \cdots, s$, and if

$$\det \frac{\partial(f_1, \cdots, f_s)}{\partial(\xi_1, \cdots, \xi_s)} \neq 0 \quad \text{for } \xi_1 = \cdots = \xi_s = \eta_1 = \cdots = \eta_t = 0,$$

then the equations

$$f_i(\xi_1, \cdots, \xi_s; \eta_1, \cdots, \eta_t)=0, \qquad i=1, \cdots, s$$

have a unique solution

$$\xi_i=g_i(\eta_1, \cdots, \eta_t), \qquad i=1, \cdots, s$$

vanishing for $\eta_1 = \cdots = \eta_t = 0$ and being real analytic in some neighborhood of the origin of $\mathcal{R}^t$.

RELLICH'S THEOREM. *Let $A(\zeta) \in \mathscr{S}\mathscr{R}^{n\times n}$ be an analytic function of a single real variable ζ in a neighborhood of the origin, and let λ_1 be an eigenvalue of $A(0)$ with multiplicity r. Then there exist r real analytic functions $\lambda_1(\zeta), \cdots, \lambda_r(\zeta)$ in a neighborhood of the origin, such that $\lambda_1(\zeta), \cdots, \lambda_r(\zeta)$ are eigenvalues of $A(\zeta)$ and*

$$\lambda_s(0)=\lambda_1, \qquad s=1, \cdots, r.$$

Proof of Theorem 2.1. The proof consists of the following three steps.

(1) Let

$$\tilde{A}(p)=X^TA(p)X=\begin{pmatrix} \tilde{A}_{11}(p) & \tilde{A}_{21}(p)^T \\ \tilde{A}_{21}(p) & \tilde{A}_{22}(p) \end{pmatrix} \tag{2.3}$$

and

$$\tilde{B}(p)=X^TB(p)X=\begin{pmatrix} \tilde{B}_{11}(p) & \tilde{B}_{21}(p)^T \\ \tilde{B}_{21}(p) & \tilde{B}_{22}(p) \end{pmatrix}, \tag{2.4}$$

where $\tilde{A}_{11}(p), \tilde{B}_{11}(p) \in \mathscr{S}\mathscr{R}^{r\times r}$. We introduce matrix-valued functions

$$F(Z, W,p)=\tilde{A}_{21}(p)+\tilde{A}_{22}(p)Z+W\tilde{A}_{11}(p)+W\tilde{A}_{21}(p)^TZ$$

and

$$G(Z, W,p)=\tilde{B}_{21}(p)+\tilde{B}_{22}(p)Z+W\tilde{B}_{11}(p)+W\tilde{B}_{21}(p)^TZ,$$

where

$$Z=(\zeta_{ij}), W=(\omega_{ij})\in \mathscr{R}^{(n-r)\times r}, \qquad p=(p_1, \cdots, p_N)^T \in \mathscr{R}^N$$

and

$$F(Z, W,p)=(f_{ij}(Z, W,p)), G(Z, W,p)=(g_{ij}(Z, W,p))\in \mathscr{R}^{(n-r)\times r}.$$

Observe that the functions $F(Z, W, p)$ and $G(Z, W, p)$ are analytic for Z, $W \in \mathscr{R}^{(n-r)\times r}$ and $p \in \mathscr{B}(0)$,

$$f_{ij}(0,0,0)=0, \quad g_{ij}(0,0,0)=0, \quad i=1, \cdots, n-r, \quad j=1, \cdots, r$$

and

$$\begin{aligned}\left(\det \frac{\partial(f,g)}{\partial(Z,w)}\right)_{Z=0,w=0,p=0} &= \det\begin{pmatrix} I^{(r)}\otimes \tilde{A}_{22}(0) & I^{(r)}\otimes \tilde{B}_{22}(0) \\ \tilde{A}_{11}(0)\otimes I^{(n-r)} & \tilde{B}_{11}(0)^T\otimes I^{(n-r)} \end{pmatrix} \\ &= \det\begin{pmatrix} I^{(r)}\otimes A_2 & I^{(r)}\otimes I^{(n-r)} \\ \lambda_1 I^{(r)}\otimes I^{(n-r)} & I^{(r)}\otimes I^{(n-r)} \end{pmatrix} \\ &= \left[\det\begin{pmatrix} A_2 & I^{(n-r)} \\ \lambda_1 I^{(n-r)} & I^{(n-r)} \end{pmatrix}\right]^r = \det(A_2-\lambda_1 I)^r \neq 0,\end{aligned}$$

where

$$f=(f_{11}, \cdots, f_{1r}, \cdots, f_{n-r,1}, \cdots, f_{n-r,r}),$$

$$g=(g_{11}, \cdots, g_{1r}, \cdots, g_{n-r,1}, \cdots, g_{n-r,r}),$$

$$Z=(\zeta_{11}, \cdots, \zeta_{1r}, \cdots, \zeta_{n-r,1}, \cdots, \zeta_{n-r,r}),$$

$$W=(\omega_{11}, \cdots, \omega_{1r}, \cdots, \omega_{n-r,1}, \cdots, \omega_{n-r,r}),$$

and $\otimes$ denotes the Kronecker product symbol [7, pp. 8–9]. Hence, by the Implicit Function Theorem the equation

$$F(Z,W,p)=0, \qquad G(Z,W,p)=0$$

has a unique real analytic solution

$$Z=Z(p), \qquad W=W(p)$$

in some neighborhood $\mathscr{B}_0$ $(\in\mathscr{B}(0))$ of the origin of $\mathscr{R}^N$ with $Z(0)=0$ and $W(0)=0$, and

$$\det(I^{(n-r)}-W(p)Z(p)^T)\neq 0 \quad \forall p\in\mathscr{B}_0. \tag{2.5}$$

From (2.5) the matrix

$$\begin{pmatrix} I & W(p)^T \\ Z(p) & I \end{pmatrix}$$

is nonsingular for $p\in\mathscr{B}_0$. Therefore we have

$$\begin{pmatrix} I & W(p)^T \\ Z(p) & I \end{pmatrix}^T \tilde{A}(p) \begin{pmatrix} I & W(p)^T \\ Z(p) & I \end{pmatrix} = \begin{pmatrix} A_1(p) & 0 \\ 0 & A_2(p) \end{pmatrix} \tag{2.6}$$

and

$$\begin{pmatrix} I & W(p)^T \\ Z(p) & I \end{pmatrix}^T \tilde{B}(p) \begin{pmatrix} I & W(p)^T \\ Z(p) & I \end{pmatrix} = \begin{pmatrix} B_1(p) & 0 \\ 0 & B_2(p) \end{pmatrix} > 0, \tag{2.7}$$

in which $A_1(p), B_1(p)\in\mathscr{S}\mathscr{R}^{r\times r}$,

$$A_1(p)=\tilde{A}_{11}(p)+Z(p)^T\tilde{A}_{21}(p)+\tilde{A}_{21}(p)^TZ(p)+Z(p)^T\tilde{A}_{22}(p)Z(p)$$

and

$$B_1(p)=\tilde{B}_{11}(p)+Z(p)^T\tilde{B}_{21}(p)+\tilde{B}_{21}(p)^TZ(p)+Z(p)^T\tilde{B}_{22}(p)Z(p).$$

From (2.6) and (2.7)

$$\tilde{A}(p)\begin{pmatrix} I \\ Z(p) \end{pmatrix} = \tilde{B}(p)\begin{pmatrix} I \\ Z(p) \end{pmatrix} B_1(p)^{-1}A_1(p).$$

Combining with (2.3), (2.4) and writing

$$X_1(p)=X\begin{pmatrix} I \\ Z(p) \end{pmatrix}, \tag{2.8}$$

we get

$$A(p)X_1(p)=B(p)X_1(p)B_1(p)^{-1}A_1(p) \tag{2.9}$$

and

$$A_1(0)=\lambda_1 I^{(r)}, \quad B_1(0)=I^{(r)}, \quad X_1(0)=X_1. \tag{2.10}$$

From (2.9), we have

$$B_1(p)^{-1}A_1(p)=[X_1(p)^TB(p)X_1(p)]^{-1}[X_1(p)^TA(p)X_1(p)]. \tag{2.11}$$

Let

$$\lambda(B_1(p)^{-1}A_1(p))=\{\lambda_s(p)\}_{s=1}^r, \qquad p\in\mathscr{B}_0.$$

Then the relation (2.1) and (2.4)–(2.7) show that

$$\lambda_s(p)\in\lambda(A(p),B(p)), \quad \lambda_s(0)=\lambda_1, \quad s=1,\cdots,r$$

and $\lambda_1(p),\cdots,\lambda_r(p)$ are near λ_1, provided that $\mathscr{B}_0$ is sufficiently small.

(2) Let $\nu\in\mathscr{R}^N$ be any fixed direction and $\|\nu\|_2=1$. Take $p=\tau\nu$ in which $\tau\in[-\beta,\beta]$ and β is a small positive scalar such that $\tau\nu\in\mathscr{B}_0$ for $\tau\in[-\beta,\beta]$. Let

$$\mu_s(\tau)=\lambda_s(\tau\nu), \qquad s=1,\cdots,r \tag{2.12}$$

and

$$H_1(p)=B_1(p)^{-1/2}A_1(p)B_1(p)^{-1/2}, \qquad \hat{H}_1(\tau)=H_1(\tau\nu). \tag{2.13}$$

Then clearly

$$\lambda(\hat{H}_1(\tau))=\{\mu_s(\tau)\}_{s=1}^r, \quad \tau\in[-\beta,\beta], \quad \mu_s(0)=\lambda_1 \quad \forall s.$$

But, on the other hand, since $\hat{H}_1(\tau)\in\mathscr{S}\mathscr{R}^{r\times r}$ is real analytic on $[-\beta,\beta]$ and $\hat{H}_1(0)=\lambda_1 I^{(r)}$, by Rellich's theorem there is a positive scalar $\beta_1\leqq\beta$ and real analytic functions $\hat{\lambda}_1(\tau),\cdots,\hat{\lambda}_r(\tau)$ on $[-\beta_1,\beta_1]$, such that

$$\lambda(\hat{H}_1(\tau))=\{\hat{\lambda}_t(\tau)\}_{t=1}^r, \quad \tau\in[-\beta_1,\beta_1], \quad \hat{\lambda}_t(0)=\lambda_1 \quad \forall t.$$

Observe the following facts:

(i) Since the zeros of a real analytic function of one real variable are isolated [1, p. 41], we have

$$\hat{\lambda}_i(\tau)\neq\hat{\lambda}_j(\tau) \quad \forall\tau\in(0,\beta_1], \quad i\neq j$$

provided that $\hat{\lambda}_i(\tau)\not\equiv\hat{\lambda}_j(\tau)$ for $\tau\in(0,\beta_1)$ and the positive scalar β_1 is sufficiently small.

(ii) The functions $\mu_1(\tau),\cdots,\mu_r(\tau)$ are continuous on $[0,\beta_1]$.

(iii) The sets $\{\mu_s(\tau)\}_{s=1}^r$ and $\{\hat{\lambda}_t(\tau)\}_{t=1}^r$ are just the same for any point $\tau\in[0,\beta_1]$, and there is a one-to-one correspondence between the elements of the two sets. Therefore there is a permutation π of $\{1,\cdots,r\}$ dependent on ν such that

$$\mu_s(\tau)=\hat{\lambda}_{\pi(s)}(\tau) \quad \forall s, \quad \tau\in[0,\beta_1]. \tag{2.14}$$

Consequently, from (1.2), (2.12), and (2.14), we get

$$\begin{aligned} D_\nu\lambda_s(0)&=\lim_{\tau\to+0}\frac{\lambda_s(\tau\nu)-\lambda_s(0)}{\tau}=\lim_{\tau\to+0}\frac{\mu_s(\tau)-\mu_s(0)}{\tau}\\ &=\lim_{\tau\to0}\frac{\hat{\lambda}_{\pi(s)}(\tau)-\hat{\lambda}_{\pi(s)}(0)}{\tau}=\left(\frac{d\hat{\lambda}_{\pi(s)}(\tau)}{d\tau}\right)_{\tau=0}, \qquad s=1,\cdots,r. \end{aligned} \tag{2.15}$$

(3) Let

$$G_1(p)=B_1(p)^{-1}A_1(p), \qquad \hat{G}_1(\tau)=G_1(\tau\nu). \tag{2.16}$$

Obviously,

$$\lambda(\hat{G}_1(\tau))=\lambda(\hat{H}_1(\tau)) \quad \forall\tau\in[0,\beta].$$

From (2.16), (2.8), and (2.11), we have

$$\begin{aligned}\left(\frac{d\hat{G}_1(\tau)}{d\tau}\right)_{\tau=0} &= \left(\frac{dG_1(\tau\nu)}{d\tau}\right)_{\tau=0}\\ &= \sum_{j=1}^{N} \nu_j\left(\frac{\partial G_1(p)}{\partial p_j}\right)_{p=0} = \sum_{j=1}^{N} \nu_j X_1^T S_j(\lambda_1) X_1.\end{aligned} \tag{2.17}$$

The relation (2.17) shows that $(d\hat{G}_1(\tau)/d\tau)_{\tau=0} \in \mathscr{S}\mathscr{R}^{r\times r}$, and hence there is a real orthogonal matrix $W_1 \in \mathscr{R}^{r\times r}$ such that

$$W_1^T\left(\frac{d\hat{G}_1(\tau)}{d\tau}\right)_{\tau=0} W_1 = \operatorname{diag}(\delta_1, \cdots, \delta_r), \qquad \delta_1 \leqq \cdots \leqq \delta_r. \tag{2.18}$$

Now we write

$$W_1^T\hat{G}_1(\tau)W_1 = (\gamma_{kl}(\tau))_{1\leq k,l\leq r},$$

in which the functions $\gamma_{kl}(\tau)$ are real analytic and so may be written as the following convergent power series:

$$\gamma_{kl}(\tau) = \gamma_{kl}^{(0)} + \gamma_{kl}^{(1)}\tau + \gamma_{kl}^{(2)}\tau^2 + \cdots, \qquad k, l = 1, \cdots, r.$$

From

$$(W_1^T\hat{G}_1(\tau)W_1)_{\tau=0} = \lambda_1 I^{(r)},$$

and

$$\left[\frac{d(W_1^T\hat{G}_1(\tau)W_1)}{d\tau}\right]_{\tau=0} = W_1^T\left(\frac{d\hat{G}_1(\tau)}{d\tau}\right)_{\tau=0} W_1$$

as well as (2.18), it follows that

$$\gamma_{kl}^{(0)} = \begin{cases}\lambda_1 & \text{if } k = l,\\ 0 & \text{otherwise,}\end{cases} \qquad \gamma_{kl}^{(1)} = \begin{cases}\delta_k & \text{if } k = l,\\ 0 & \text{otherwise.}\end{cases}$$

Therefore

$$\gamma_{kl}(\tau) = \begin{cases}\lambda_1 + \delta_k\tau + \gamma_{kk}^{(2)}\tau^2 + \gamma_{kk}^{(3)}\tau^3 + \cdots & \text{if } k = l,\\ \gamma_{kl}^{(2)}\tau^2 + \gamma_{kl}^{(3)}\tau^3 + \cdots & \text{otherwise.}\end{cases} \tag{2.19}$$

Assume that

$$\begin{aligned}&\delta_1 = \cdots = \delta_{r_1} < \delta_{r_1+1} = \cdots = \delta_{r_1+r_2}\\ &\quad < \cdots < \delta_{r_1+\cdots+r_{q-1}+1} = \cdots = \delta_{r_1+\cdots+r_{q-1}+r_q}, \qquad r_1 + \cdots + r_q = r\end{aligned} \tag{2.20}$$

and write

$$\delta_{r_1} = d_1, \quad \delta_{r_1+r_2} = d_2, \cdots, \delta_{r_1+\cdots+r_q} = d_q; \tag{2.21}$$

then by the Gerschgorin theorem from (2.19)–(2.21) we know that there are precisely q circular discs $\mathscr{D}_1, \cdots, \mathscr{D}_q$ with centers $\lambda_1 + d_1\tau, \cdots, \lambda_1 + d_q\tau$ and with radii of magnitude $O(\tau^2)$, respectively, and the union $\bigcup_{j=1}^{q} \mathscr{D}_j$ contains all of the eigenvalues $\hat{\lambda}_1(\tau), \cdots, \hat{\lambda}_r(\tau)$. Besides, the discs $\mathscr{D}_1, \cdots, \mathscr{D}_q$ are mutually disjoint provided that τ

belong to a sufficiently small segment $[-\beta_1, \beta_1]$, and in such a case every disk $\mathscr{D}_j$ contains exactly r_j eigenvalues which may be written as the following convergent power series:

$$\lambda_1 + d_j\tau + g^{(2)}_{r_1+\cdots+r_{j-1}+k}\tau^2 + g^{(3)}_{r_1+\cdots+r_{j-1}+k}\tau^3 + \cdots, \qquad k = 1, \cdots, r_j, \tag{2.22}$$

where $\tau \in [-\beta_1, \beta_1]$, $j = 1, \cdots, q$ and $r_0 = 0$.

Combining (2.22) with (2.20) and (2.21), we may rewrite the expressions of (2.22) as

$$\hat{\lambda}_t(\tau) = \lambda_1 + \delta_t\tau + g_t^{(2)}\tau^2 + g_t^{(3)}\tau^3 + \cdots, \qquad t = 1, \cdots, r.$$

Consequently, we obtain

$$\left(\frac{d\hat{\lambda}_t(\tau)}{d\tau}\right)_{\tau=0} = \delta_t, \qquad t = 1, \cdots, r. \tag{2.23}$$

Combining (2.17), (2.18), (2.23) with (2.15), we get the relations (2.2). □

Let e_j denote the jth column vector of the identity $I^{(N)}$. From Theorem 2.1 we get the following corollary.

COROLLARY 2.2. *Under the hypotheses of Theorem* 2.1, *there are permutations* π *and* π' *of* $\{1, \cdots, r\}$ *such that the relations*

$$D_{e_j}\lambda_s(0) = \lambda_{\pi(s)}(X_1^T S_j(\lambda_1) X_1) \tag{2.24}$$

and

$$D_{-e_j}\lambda_s(0) = \lambda_{\pi'(s)}(X_1^T S_j(\lambda_1) X_1), \qquad s = 1, \cdots, r \tag{2.25}$$

are validforj $= 1, \cdots, N$. *Where thefunctions* $\lambda_1(p), \cdots, \lambda_r(p)$ *are described in Theorem* 2.1. *Especially, if* $r = 1$ *then the eigenvalue* $\lambda_1(p)$ *has the partial derivatives with respect to* p_j *at the origin*

$$\left(\frac{\partial\lambda_1(p)}{\partial p_j}\right)_{p=0} = x_1^T S_j x_1, \qquad j = 1, \cdots, N, \tag{2.26}$$

where x_1 *is the associated eigenvector with* λ_1 *satisfying* (2.1).

We note that the relations (2.24) and (2.25) have been proved in [10] in a slightly different way, and the relations (2.26) have been obtained by Fox and Kapoor [4].

According to Theorem 2.1 we may introduce the following definition.

DEFINITION 2.3. Let $A(p)$, $B(p)$, X and λ_1 be as in Theorem 2.1. Then the quantity

$$s_p^{(\nu)}(\lambda_1) \equiv \rho\left(\sum_{j=1}^{N} \nu_j X_1^T S_j(\lambda_1) X_1\right) \tag{2.27}$$

is called the sensitivity of the multiple eigenvalue λ_1 in the direction $\nu = (\nu_1, \cdots, \nu_N)^T \in \mathscr{R}^N$ with $\|\nu\|_2 = 1$, the quantity

$$s_{p_j}(\lambda_1) \equiv \rho(X_1^T S_j(\lambda_1) X_1) \tag{2.28}$$

is called the sensitivity of the multiple eigenvalue λ_1 with respect to the parameter p_j, and the quantity

$$s_p(\lambda_1) \equiv \sqrt{\sum_{j=1}^{N} s_{p_j}^2(\lambda_1)} \tag{2.29}$$

is called the sensitivity of the multiple eigenvalue λ_1.

From (2.27)–(2.29) we get

$$\frac{1}{\sqrt{N}} s_p(\lambda_1) \leqq \max_{\substack{\nu \in \mathscr{R}^N \\ \|\nu\|_2 = 1}} s_p^{(\nu)}(\lambda_1) \leqq s_p(\lambda_1).$$

Example 2.4 (see [10]). Consider the eigenvalue problem

$$A(p)x(p) = \lambda(p)x(p), \quad \lambda(p) \in \mathscr{R}, \quad x(p) \in \mathscr{R}^2, \quad p = (p_1, p_2)^T \in \mathscr{R}^2$$

with

$$A(p) = \begin{pmatrix} 1 + 2p_1 + 2p_2 & p_2 \\ p_2 & 1 + 2p_2 \end{pmatrix}.$$

Obviously, the matrix $A(p)$ is a real analytic function of $p \in \mathscr{R}^2$, $A(0)$ has eigenvalue $\lambda_1 = 1$ with multiplicity 2, and the eigenvalues of $A(p)$ are

$$\lambda_1(p) = 1 + p_1 + 2p_2 + \sqrt{p_1^2 + p_2^2}, \qquad \lambda_2(p) = 1 + p_1 + 2p_2 - \sqrt{p_1^2 + p_2^2}.$$

It is well known that the functions $\lambda_1(p)$ and $\lambda_2(p)$ are not differentiable at $p = 0$. Straightforward calculations show that, for any direction $\nu = (\cos\theta, \sin\theta)^T \in \mathscr{R}^2$ with $\theta \in [0, 2\pi)$, the functions $\lambda_1(p)$ and $\lambda_2(p)$ have directional derivatives

$$D_\nu\lambda_1(0) = \cos\theta + 2\sin\theta + 1, \qquad D_\nu\lambda_2(0) = \cos\theta + 2\sin\theta - 1. \tag{2.30}$$

On the other hand, applying Theorem 2.1 we have

$$\{D_\nu\lambda_s(0)\}_{s=1}^2 = \left\{\lambda : \lambda \in \lambda\left(\cos\theta\begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} + \sin\theta\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\right)\right\}$$
$$= \{\cos\theta + 2\sin\theta + 1, \cos\theta + 2\sin\theta - 1\}.$$

This coincides with (2.30). Moreover, by Definition 2.3 we have

$$s_p^{(\nu)}(\lambda_1) = \max\{|\cos\theta + 2\sin\theta + 1|, |\cos\theta + 2\sin\theta - 1|\}$$

and

$$s_{p_1}(\lambda_1) = 2, \qquad s_{p_2}(\lambda_1) = 3,$$

where $\nu = (\cos\theta, \sin\theta)^T \in \mathscr{R}^2$ and $\lambda_1 = 1$. Further, we have

$$\max_{\substack{\nu \in \mathscr{R}^2 \\ \|\nu\|_2 = 1}} s_p^{(\nu)}(\lambda_1) = \sqrt{5} + 1 \approx 3.23607$$

and

$$s_p(\lambda_1) = \sqrt{13} \approx 3.60555.$$

Acknowledgments. I am grateful to Professors G. W. Stewart, Gene H. Golub, and P. Lancaster and to the referees for valuable comments and suggestions. I thank the referee who pointed out that the analyses of eigenvalues of symmetric, positive-definite matrices have been made by Polak and Wardi (1980, 1982; see [8]), and these analyses can be carried over to the cases of symmetric matrices and bilinear forms (e.g., the local Lipschitz continuity of the eigenvalues and the characterization for the generalized gradient introduced by Clarke [2] of multiple eigenvalues).

REFERENCES

[1] H. CARTAN, *Elementary Theory of Analytic Functions of One or Several Complex Variables*, Hermann, Paris, 1963.

[2] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[3] J. DIEUDONNÉ, *Éléments d'Analyse*, 1, *Fondements de l'Analyse Moderne*, Gauthier-Villars, Paris, 1968.

[4] R. L. FOX AND M. P. KAPOOR, *Rates of change of eigenvalues and eigenvectors*, AIAA J., 6 (1968), pp. 2426–2429.

[5] T. KATO, *A Short Introduction to Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, New York, 1982.

[6] P. LANCASTER, *On eigenvalues of matrices dependent on a parameter*, Numer. Math., 6 (1964), pp. 377–387.

[7] M. MARCUS AND H. MINC, *A Survey of Matrix Theory and Matrix Inequalities*, Allyn and Bacon, Boston, 1964.

[8] E. POLAK AND Y. WARDI, *Nondifferentiable optimization algorithm for designing control systems having singular value inequalities*, Automatica, 18 (1982), pp. 267–283.

[9] F. RELLICH, *Perturbation Theory of Eigenvalue Problems*, Lecture Notes, Mathematics Department, New York University, New York, 1953.

[10] JI-GUANG SUN, *Sensitivity analysis of multiple eigenvalues*. I, J. Comput. Math., 6 (1988), pp. 28–38.

[11] J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 4, pp. 542-550, October 1989

010

CIRCULANT PRECONDITIONERS FOR HERMITIAN TOEPLITZ SYSTEMS*

RAYMOND H. CHAN†

Abstract. The solutions of Hermitian positive definite Toeplitz systems $Ax = b$ by the preconditioned conjugate gradient method for three families of circulant preconditioners C is studied. The convergence rates of these iterative methods depend on the spectrum of $C^{-1}A$. For a Toeplitz matrix A with entries that are Fourier coefficients of a positive function f in the Wiener class, the invertibility of C is established, as well as that the spectrum of the preconditioned matrix $C^{-1}A$ clusters around one. It is proved that if f is $(l+1)$-times differentiable, with $l > 0$, then the error after $2q$ conjugate gradient steps will decrease like $((q-1)!)^{-2l}$. It is also shown that if C copies the central diagonals of A, then C minimizes $\|C-A\|_1$ and $\|C-A\|_\infty$.

Key words. Toeplitz matrix, circulant matrix, preconditioned conjugate gradient method

AMS(MOS) subject classifications. 65F10, 65F15

1. Introduction. In this paper we discuss the solutions to a class of Hermitian positive definite Toeplitz systems $Ax = b$ by the preconditioned conjugate gradient method. Direct methods that are based on the Levinson recursion formula are in constant use; see, for instance, Levinson [10] and Trench [12]. For an n-by-n Toeplitz matrix A_n, these methods require $O(n^2)$ operations. Faster algorithms that require $O(n \log^2 n)$ operations have been developed; see Bitmead and Anderson [1] and Brent, Gustavson, and Yun [2]. The stability properties of these direct methods for symmetric positive definite matrices are discussed in Bunch [3].

In [11], Strang proposed using preconditioned conjugate gradient method with circulant preconditioners for solving symmetric positive definite Toeplitz systems. The number of operations per iteration is of order $O(n \log n)$, as circulant systems can be solved efficiently by the Fast Fourier Transform. Chan and Strang [4] then considered using a circulant preconditioner S_n, obtained by copying the central diagonals of A_n and bringing them around to complete the circulant. In that paper, we proved that if the underlying generating function f, the Fourier coefficients of which give the entries of A_n, is a positive function in the Wiener class, then for n sufficiently large, S_n and S_n^{-1} are uniformly bounded in the l_2 norm and the eigenvalues of the preconditioned matrix $S_n^{-1}A_n$ cluster around 1. We note that f is an even function since the matrices A_n are symmetric.

In this paper, we extend these results to Hermitian positive definite Toeplitz systems. More precisely, we show in § 2 that if the generating function f is a real-valued positive function in the Wiener class, then the spectrum of $S_n^{-1}A_n$ is clustered around 1. We remark that the proof given in Chan and Strang [4] cannot be readily generalized to cover this case. In fact, for Hermitian A_n, the Hankel matrices $H_{n/2}$ used in the proof in [4] are not Hermitian, and the circulant-Toeplitz eigenvalue problem cannot be split into two similar Toeplitz–Hankel eigenvalue problems. In § 3, we establish the superlinear convergence rate of the conjugate gradient method when applied to these preconditioned systems. In particular, we show that if f is $(l+1)$-times differentiable, with $l > 0$, then the error after $2q$ conjugate gradient steps will decrease like $((q-1)!)^{-2l}$.

* Received by the editors September 28, 1988; accepted for publication (in revised form) January 25, 1989. This work was partially supported by National Science Foundation grants DCR86-02563 and CCR87-03768.

† Department of Mathematics, University of Hong Kong, Pokfulam Road, Hong Kong (HKUCS!HKUCC!HRSMCHR@UUNET.UU.NET).

In § 4, we discuss other viable preconditioners for the same problem. We show that the preconditioned systems for these preconditioners also have clustered spectra around 1 for large n and that they all have the same asymptotic convergence rate. In § 5, we show that the preconditioner that copies the central diagonals of A_n is optimal in the sense that it minimizes $\|C_n - A_n\|_1 = \|C_n - A_n\|_\infty$ over all Hermitian circulant matrices C_n. Finally, numerical results are given in § 6.

2. The spectrum of the preconditioned matrix. Let us first assume that the Hermitian Toeplitz matrices A_n are finite sections of a fixed singly infinite positive definite matrix A_∞; see Chan and Strang [4]. Thus the (i, j)th entries of A_n and A_∞ are a_{i-j}, with $a_k = \bar{a}_{-k}$ for all k. We associate with A_∞ the real-valued generating function

$$f(\theta) = \sum_{-\infty}^{\infty} a_k e^{-ik\theta},$$

defined on $[0, 2\pi)$. We will assume that f is a positive function and is in the Wiener class, i.e., the sequence $\{a_k\}_{k=-\infty}^{\infty}$ is in l_1. It then easily follows that the A_n are Hermitian positive definite matrices for all n; see for instance, Grenander and Szegö [8]. Moreover, if

$$0 < f_{\min} < f < f_{\max} < \infty,$$

then the spectrum $\sigma(A_n)$ of A_n satisfies

$$\sigma(A_n) \subseteq [f_{\min}, f_{\max}]. \tag{1}$$

Let S_n be the Hermitian circulant preconditioner that copies the central diagonals of A_n. More precisely, the entries $s_{ij} = s_{i-j}$ of S_n are given by

$$s_k = \begin{cases} a_k & 0 \leqq k \leqq m, \\ a_{k-n} & m < k < n, \\ \bar{s}_{-k} & 0 < -k < n. \end{cases} \tag{2}$$

For simplicity, we are assuming here and in the following equations that $n = 2m + 1$. The case where $n = 2m$ can be treated similarly, and in that case, we define $s_m = (a_m + a_{-m})/2$; see (17) below.

We will show that $S_n^{-1}A_n$ has a clustered spectrum. We first note the following theorem.

THEOREM 1. *Suppose f is positive and is in the Wiener class. Then for large n, the circulants S_n and S_n^{-1} are uniformly bounded in the l_2 norm. In fact, for large n, the spectrum $\sigma(S_n)$ of S_n satisfies*

$$\sigma(S_n) \subseteq [f_{\min}, f_{\max}]. \tag{3}$$

The proof of this theorem is similar to the proof of Theorem 1 of Chan and Strang [4], and we therefore omit it.

Next we show that $A_n - S_n$ has a clustered spectrum.

THEOREM 2. *Let f be a positive function in the Wiener class, then for all $\varepsilon > 0$, there exist M and $N > 0$ such that for all $n > N$, at most M eigenvalues of $S_n - A_n$ have absolute values exceeding ε.*

Proof. Clearly $B_n = S_n - A_n$ is a Hermitian Toeplitz matrix with entries $b_{ij} = b_{i-j}$ given by

$$(4) \qquad b_k = \begin{cases} 0 & 0 \leqq k \leqq m, \\ a_{k-n} - a_k & m < k < n, \\ \bar{b}_{-k} & 0 < -k < n. \end{cases}$$

Since f is in the Wiener class, for all given $\varepsilon > 0$, there exists an $N > 0$, such that $\sum_{k=N+1}^{\infty} |a_k| < \varepsilon$. Let $U_n^{(N)}$ be the n-by-n matrix obtained from B_n by replacing the $(n-N)$-by-$(n-N)$ leading principal submatrix of B_n by the zero matrix. Then rank $(U_n^{(N)}) \leqq 2N$. Let $W_n^{(N)} \equiv B_n - U_n^{(N)}$. The leading $(n-N)$-by-$(n-N)$ block of $W_n^{(N)}$ is the leading $(n-N)$-by-$(n-N)$ principal submatrix of B_n; hence this block is a Toeplitz matrix, and it is easy to see that the maximum absolute column sum of $W_n^{(N)}$ is attained at the first column (or the $(n-N-1)$th column). Thus

$$(5) \qquad \|W_n^{(N)}\|_1 = \sum_{k=m+1}^{n-N-1} |b_k| = \sum_{k=m+1}^{n-N-1} |a_{k-n} - a_k| \leqq \sum_{k=N+1}^{n-N-1} |a_k| < \varepsilon.$$

Since $W_n^{(N)}$ is Hermitian, we have $\|W_n^{(N)}\|_\infty = \|W_n^{(N)}\|_1$. Thus

$$\|W_n^{(N)}\|_2 \leqq (\|W_n^{(N)}\|_1 \cdot \|W_n^{(N)}\|_\infty)^{1/2} < \varepsilon.$$

Hence the spectrum of $W_n^{(N)}$ lies in $(-\varepsilon, \varepsilon)$. By Cauchy Interlace theorem, see Wilkinson [13], we see that at most $2N$ eigenvalues of $B_n = S_n - A_n$ have absolute values exceeding ε. □

Combining Theorems 1 and 2, and using the fact that

$$S_n^{-1} A_n = I_n + S_n^{-1}(A_n - S_n),$$

we have the following corollary.

COROLLARY. *Let f be a positive function in the Wiener class, then for all $\varepsilon > 0$, there exist N and $M > 0$, such that for all $n > M$, at most N eigenvalues of $S_n^{-1} A_n - I_n$ have absolute values larger than ε.*

Thus the spectrum of $S_n^{-1} A_n$ is clustered around one for large n.

3. Superlinear convergence rate. It follows easily from the Corollary of the last section that the conjugate gradient method, when applied to the preconditioned system $S_n^{-1} A_n$, converges superlinearly. More precisely, for all $\varepsilon > 0$, there exists a constant $C(\varepsilon) > 0$ such that the error vector e_q at the qth iteration satisfies

$$(6) \qquad \|e_q\| \leqq C(\varepsilon) \varepsilon^q \|e_0\|,$$

where $\|x\|^2 \equiv x^* S_n^{-1/2} A S_n^{-1/2} x$; see Chan and Strang [4] for a proof. Thus the number of iterations to achieve a fixed accuracy remains bounded as the matrix order n is increased. Since each iteration requires $O(n \log n)$ operations using the Fast Fourier Transform, see Strang [11], the work of solving the equation $A_n x = b$ to a given accuracy δ is $c(f, \delta) n \log n$, where $c(f, \delta)$ is a constant that depends on f and δ only.

We note that if extra smoothness conditions are imposed on f, we can get a more precise bound on the convergence rate.

THEOREM 3. *Let f be a $(l+1)$-times differentiable function with its $(l+1)$*th *derivative of f in $L^1[0, 2\pi)$, $l > 0$. Then for large n,*

$$\|e_{2q}\| \leqq \frac{c^q}{((q-1)!)^{2l}}\|e_0\|, \tag{7}$$

for some constant c that depends on f and l only.

Proof. We remark that from the standard error analysis of the conjugate gradient method, we have

$$\|e_q\| \leqq [\min_{P_q} \max_{\lambda} |P_q(\lambda)|]\|e_0\|, \tag{8}$$

where the minimum is taken over polynomials of degree q with constant term 1 and the maximum is taken over the spectrum of $S_n^{-1}A_n$, or equivalently, the spectrum of $S_n^{-1/2}A_nS_n^{-1/2}$; see for instance, Golub and Van Loan [7]. In the following, we will try to estimate that minimum.

We first note that the assumptions on f imply that

$$|a_j| \leqq \frac{\hat{c}}{|j|^{l+1}} \quad \forall j,$$

where $\hat{c} = \|f^{(l+1)}\|_{L^1}$; see, for instance, Katznelson [9]. Hence

$$\sum_{j=k+1}^{n-k-1} |a_j| \leqq \hat{c} \sum_{j=k+1}^{n-k-1} \frac{1}{|j|^{l+1}} \leqq \hat{c}\int_k^\infty \frac{dx}{x^{l+1}} \leqq \frac{\hat{c}}{k^l}, \quad \forall k \geqq 1. \tag{9}$$

As in Theorem 2, we write

$$B_n = W_n^{(k)} + U_n^{(k)}, \quad \forall k \geqq 1,$$

where $U_n^{(k)}$ is the matrix obtained from B_n by replacing its $(n-k)$-by-$(n-k)$ principal submatrix of B_n by a zero matrix. Using the arguments in Theorem 2, cf. (5) and (9), we see that rank $(U_n^{(k)}) \leqq 2k$ and $\|W_n^{(k)}\|_2 \leqq \hat{c}/k^l$, for all $k \geqq 1$. Now consider

$$S_n^{-1/2}B_nS_n^{-1/2} = S_n^{-1/2}W_n^{(k)}S_n^{-1/2} + S_n^{-1/2}U_n^{(k)}S_n^{-1/2} \equiv \tilde{W}_n^{(k)} + \tilde{U}_n^{(k)}.$$

By Theorem 1, we have, for large n, rank $(\tilde{U}_n^{(k)}) \leqq 2k$ and

$$\|\tilde{W}_n^{(k)}\|_2 \leqq \|S_n^{-1}\|_2\|W_n^{(k)}\|_2 \leqq \frac{\tilde{c}}{k^l}, \quad \forall k \geqq 1, \tag{10}$$

with $\tilde{c} = \hat{c}/f_{\min}$.

Next we note that $W_n^{(k)} - W_n^{(k+1)}$ can be written as the sum of two rank one matrices of the following form:

$$W_n^{(k)} - W_n^{(k+1)} = u_kv_k^* + v_ku_k^* = \tfrac{1}{2}(w_k^+w_k^{+*} - w_k^-w_k^{-*}), \quad \forall k \geqq 0.$$

Here u_k is the $(n-k)$th unit vector, $v_k = (b_{n-k-1}, \cdots, b_1, b_0/2, 0, \cdots, 0)$, with b_j given by (4), and $w_k^\pm = u_k \pm v_k$. Hence by letting $z_k^\pm = S_n^{-1/2}w_k^\pm$ for $k \geqq 0$, we have

$$\begin{aligned} S_n^{-1/2}B_nS_n^{-1/2} = \tilde{W}_n^{(0)} &= \tilde{W}_n^{(k)} + \frac{1}{2}\sum_{j=0}^{k-1}(z_j^+z_j^{+*} - z_j^-z_j^{-*}), \\ &= \tilde{W}_n^{(k)} + V_k^+ - V_k^-, \quad \forall k \geqq 1, \end{aligned} \tag{11}$$

where $V_k^\pm \equiv \frac{1}{2} \sum_{j=0}^{k-1} z_j^\pm z_j^{\pm *}$ are positive semidefinite matrices of rank k. Let us order the eigenvalues of $\tilde{W}_n^{(0)}$ as

$$\mu_0^- \leqq \mu_1^- \leqq \cdots \leqq \mu_1^+ \leqq \mu_0^+.$$

By applying the Cauchy Interlace Theorem to (11) and using the bound of $\|\tilde{W}_n^{(k)}\|_2$ in (10), we see that for all $k \geqq 1$, there are at most k eigenvalues of $\tilde{W}_n^{(0)}$ lying to the right of $\tilde{c}/k^l$, and there are at most k of them lying to the left of $-\tilde{c}/k^l$. More precisely, we have

$$|\mu_k^\pm| \leqq \|\tilde{W}_n^{(k)}\|_2 \leqq \frac{\tilde{c}}{k^l}, \quad \forall k \geqq 1.$$

Using the identity

$$S_n^{-1/2} A_n S_n^{-1/2} = I_n + S_n^{-1/2} B_n S_n^{-1/2} = I_n + \tilde{W}_n^{(0)},$$

we see that if we order the eigenvalues of $S_n^{-1/2} A_n S_n^{-1/2}$ as

$$\lambda_0^- \leqq \lambda_1^- \leqq \cdots \leqq \lambda_1^+ \leqq \lambda_0^+,$$

then $\lambda_k^\pm = 1 + \mu_k^\pm$ for all $k \geqq 0$ with

$$(12) \qquad 1 - \frac{\tilde{c}}{k^l} \leqq \lambda_k^- \leqq \lambda_k^+ \leqq 1 + \frac{\tilde{c}}{k^l}, \quad \forall k \geqq 1.$$

For $\lambda_0^\pm$, the bounds are obtained from (1) and (3):

$$(13) \qquad \frac{f_{\min}}{f_{\max}} \leqq \lambda_0^- \leqq \lambda_0^+ \leqq \frac{f_{\max}}{f_{\min}}.$$

Having obtained the bounds for $\lambda_k^\pm$, we can now construct the polynomial that will give us a bound for (8). Our idea is to choose P_{2q} that annihilates the q extreme pairs of eigenvalues. Thus consider

$$p_k(x) = \left(1 - \frac{x}{\lambda_k^+}\right)\left(1 - \frac{x}{\lambda_k^-}\right), \quad \forall k \geqq 1.$$

Between those roots $\lambda_k^\pm$, the maximum of $|p_k(x)|$ is attained at the average $x = \frac{1}{2}(\lambda_k^+ + \lambda_k^-)$, where by (12), we have

$$\max_{x \in [\lambda_k^-, \lambda_k^+]} |p_k(x)| = \frac{(\lambda_k^+ - \lambda_k^-)^2}{4\lambda_k^+ \lambda_k^-} \leqq \left(\frac{2\tilde{c}}{k^l}\right)^2 \cdot \left(\frac{f_{\max}}{2f_{\min}}\right)^2 = \left(\frac{\tilde{c} f_{\max}}{f_{\min}}\right)^2 \cdot \frac{1}{k^{2l}}, \quad \forall k \geqq 1.$$

Similarly, for $k = 0$, we have, by using (13),

$$\max_{x \in [\lambda_0^-, \lambda_0^+]} |p_0(x)| = \frac{(\lambda_0^+ - \lambda_0^-)^2}{4\lambda_0^+ \lambda_0^-} \leqq \frac{(f_{\max}^2 - f_{\min}^2)^2}{4 f_{\min}^4}.$$

Hence the polynomial $P_{2q} = p_0 p_1 \cdots p_{q-1}$, which annihilates the q extreme pairs of eigenvalues, satisfies

$$(14) \qquad |P_{2q}(x)| \leqq \frac{c^q}{((q-1)!)^{2l}},$$

for some constant c that depends only on f and l. This holds for all $\lambda_k^\pm$ in the inner interval between λ_{q-1}^- and λ_{q-1}^+, where the remaining eigenvalues are. Equation (7) now follows directly from (8) and (14). □

4. Other circulant preconditioners. The proof of Theorem 2 suggests that there are many other viable preconditioners that can give us the same asymptotic convergence rate. One example is given by the circulant matrix T_n proposed by Chan [6]. It is obtained by averaging the corresponding diagonals of A_n, with the diagonals of A_n being extended to length n by a wraparound. More precisely, the entries $t_{ij} = t_{i-j}$ of T_n are given by

$$t_k = \begin{cases} \frac{1}{n}\{ka_{k-n}+(n-k)a_k\} & 0 \leqq k < n, \\ \bar{t}_{-k} & 0 < -k < n, \end{cases}$$

where a_n is taken to be 0. He proved that such T_n minimizes the Frobenius norm $\|T_n - A_n\|_F$ over all possible circulant matrices T_n. The entries $b_{ij} = b_{i-j}$ of $T_n - A_n$ are given by

$$b_k = \begin{cases} \frac{k}{n}(a_{k-n} - a_k) & 0 \leqq k < n, \\ \bar{b}_{-k} & 0 < -k < n. \end{cases}$$

As in Theorem 2, we let $W_n^{(N)}$ be the matrix obtained from $T_n - A_n$ by replacing the last N rows and N columns of $T_n - A_n$ by zero vectors. We see that

$$\|W_n^{(N)}\|_1 \leqq 2 \sum_{k=0}^{n-N-1} |b_k| \leqq 2 \sum_{k=0}^{N} \frac{k}{n}|a_k| + 4 \sum_{k=N+1}^{n} |a_k|. \tag{15}$$

Now let $M > N$ be such that $(1/M)\sum_{k=0}^{N} k|a_k| < \varepsilon$. Then for all $n > M$, we have $\|W_n^{(N)}\|_1 < 6\varepsilon$. Hence the eigenvalues of $T_n - A_n$ are clustered around zero, except for at most $2N$ of them. We remark that by using results in Chan [5], we can show that $\lim_{n\to\infty} \|S_n - T_n\|_2 = 0$ and that the convergence rate of $S_n^{-1}A_n$ and $T_n^{-1}A_n$ are the same for large n. In particular, both will converge superlinearly.

As another example, let us consider the circulant matrix R_n with entries $r_{ij} = r_{i-j}$ given by

$$r_k = \begin{cases} a_{k-n} + a_k & 0 \leqq k < n, \\ \bar{r}_{-k} & 0 < -k < n, \end{cases}$$

where a_n is again taken to be 0. The entries $b_{ij} = b_{i-j}$ of $R_n - A_n$ are given by

$$b_k = \begin{cases} a_{k-n} & 0 \leqq k < n, \\ \bar{b}_{-k} & 0 < -k < n. \end{cases}$$

It is easily seen that the conclusion of Theorem 2 holds for this preconditioner, too; cf. (5) and (15). As was displayed in the similar case of T_n, we can show that $\lim_{n\to\infty} \|S_n - R_n\|_2 = 0$ and that the convergence rate of $S_n^{-1}A_n$ and $R_n^{-1}A_n$ are the same for large n; see Chan [5]. Numerical results in § 6 indeed show that the three preconditioners R_n, S_n, and T_n behave almost the same for large n.

5. The optimality of S_n. From the discussion in §§ 2 and 4, we know that it is interesting to obtain the Hermitian circulant matrix C_n that minimizes the norm $\|C_n - A_n\|_1 = \|C_n - A_n\|_\infty$. The minimum is attained by S_n.

THEOREM 4. *The circulant matrix S_n, whose entries are given by* (2), *minimizes $\|C_n - A_n\|_1 = \|C_n - A_n\|_\infty$ over all possible Hermitian circulant matrices C_n.*

Proof. Let us construct the circulant matrix C_n that minimizes the absolute column sums of $C_n - A_n$. Let the (i, j)th entries of C_n be $c_{ij} = c_{i-j}$. Since C_n is Hermitian and circulant, we have $c_k = \bar{c}_{n-k}$ for $k = 1, \cdots, m$, where $m = (n-1)/2$. Hence C_n is determined by $\{c_k\}_{k=0}^{m}$. For $j = 0, \cdots, n-1$, the jth absolute column sum u_j of $C_n - A_n$ is given by

$$u_j = \sum_{k=0}^{n-1-j} |a_k - c_k| + \sum_{k=1}^{j} |\bar{a}_k - \bar{c}_k|. \tag{16}$$

We note that $u_{n-1-j} = u_j$ for $0 \leqq j < n$. Hence it suffices to consider u_j for $0 \leqq j \leqq m$. The term involving c_0 in (16) is $|a_0 - c_0|$, which has its minimum at $c_0 = a_0$. For $k = 1, \cdots, m$, the terms involving c_k in (16) are either of the form

(a) $|a_k - c_k| + |\bar{a}_k - \bar{c}_k| = 2|a_k - c_k|$, or

(b) $|a_k - c_k| + |a_{n-k} - c_{n-k}| = |a_k - c_k| + |\bar{a}_{n-k} - c_k|$.

In case (a), the minimum is at $c_k = a_k$. In case (b), the minimum occurs at any c_k lying on the line segment joining a_k and $\bar{a}_{n-k}$. In particular (a) and (b) attain their minima at $c_k = a_k$. Thus C_n so constructed is the same as the S_n given by (2).

Now for any other Hermitian circulant matrix H_n, the jth absolute column sum v_j of $H_n - A_n$ will satisfy $u_j \leqq v_j$, for $j = 0, \cdots, n-1$. Hence,

$$\|S_n - A_n\|_1 = \max_j u_j \leqq \max_j v_j = \|H_n - A_n\|_1. \qquad \Box$$

Remark. When $n = 2m$ is even, c_m is real, since C_n is both Hermitian and circulant. The term involving c_m in u_j takes the form $|a_m - c_m|$ or $|\bar{a}_m - c_m|$. Since $u_j = u_{n-1-j}$ for $j = 0, \cdots, n-1$, we see that c_m should be chosen such that both terms are minimized, i.e.,

$$c_m = \tfrac{1}{2}(a_m + \bar{a}_m). \tag{17}$$

6. Numerical results. To test the convergence rates of the preconditioners, we have applied the preconditioned conjugate gradient method to $A_n x = b$ with

$$a_k = \begin{cases} \dfrac{1+\sqrt{-1}}{(1+k)^{1.1}} & k > 0, \\ 2 & k = 0, \\ \bar{a}_{-k} & k < 0. \end{cases}$$

The underlying generating function f is given by

$$f(\theta) = 2 \sum_{k=0}^{\infty} \frac{\sin(k\theta) + \cos(k\theta)}{(1+k)^{1.1}}.$$

Clearly, f is in the Wiener class. The spectra of A_n, $R_n^{-1}A_n$, $S_n^{-1}A_n$, and $T_n^{-1}A_n$ for $n = 32$ are represented in Fig. 1. Table 1 shows the number of iterations required to make $\|r_q\|_2 / \|r_0\|_2 < 10^{-7}$, where r_q is the residual vector after q iterations. The right-hand side b is the vector of all ones, and the zero vector is our initial guess. We see that as n increases, the number of iterations increases like $O(\log n)$ for the original matrix A_n, while it stays almost the same for the preconditioned matrices. Moreover, all preconditioned systems converge at the same rate for large n.

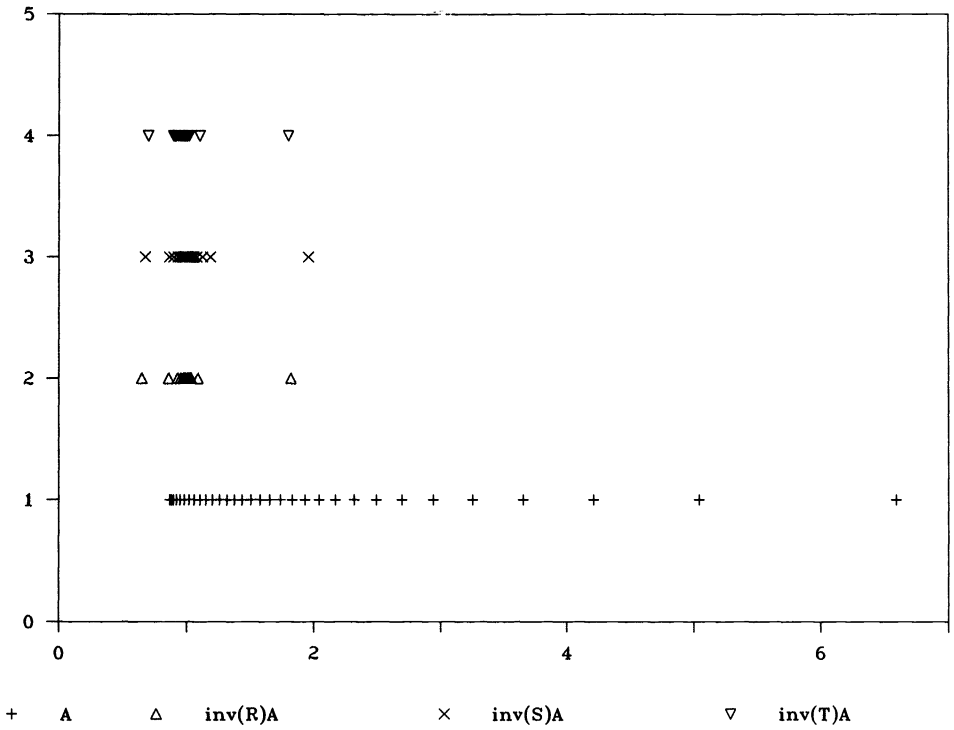


FIG. 1. *Spectra of the preconditioned systems.*

TABLE 1
Number of iterations for different systems.

n	A_n	$R_n^{-1}A_n$	$S_n^{-1}A_n$	$T_n^{-1}A_n$
16	13	7	8	7
32	15	6	7	6
64	18	7	7	7
128	19	7	7	7
256	21	7	7	7

Acknowledgment. The author acknowledges the help of Professor Olof Widlund for his help in the preparation of this paper and the hospitality of Professor Petter Bjørstad during his visit at the Institutt for Informatikk, University of Bergen, Norway in the summer of 1988.

REFERENCES

[1] R. BITMEAD AND B. ANDERSON, *Asymptotically fast solution of Toeplitz and related systems of equations*, Linear Algebra Appl., 34 (1980), pp. 103–116.

[2] R. BRENT, F. GUSTAVSON, AND D. YUN, *Fast solution of Toeplitz systems of equations and computations of Padé approximations*, J. Algorithms, 1 (1980), pp. 259–295.

[3] J. BUNCH, *Stability of methods for solving Toeplitz systems of equations*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 349–364.

[4] R. CHAN AND G. STRANG, *Toeplitz equations by conjugate gradients with circulant preconditioner*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 104–119.

[5] R. CHAN, *The spectrum of a family of circulant preconditioned Toeplitz systems*, SIAM J. Numer. Anal., 26 (1989), pp. 503–506.

[6] T. CHAN, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 766–771.

[7] G. H. GOLUB AND C. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[8] U. GRENANDER AND G. SZEGÖ, *Toeplitz Forms and Their Applications*, Second edition, Chelsea, New York, 1984.

[9] Y. KATZNELSON, *An Introduction to Harmonic Analysis*, Second edition, Dover, New York, 1976.

[10] N. LEVINSON, *The Wiener rms (root-mean-square) error criterion in filter design and prediction*, J. Math. Phys., 25 (1947), pp. 261–278.

[11] G. STRANG, *A proposal for Toeplitz matrix calculations*, Stud. Appl. Math., 74 (1986), pp. 171–176.

[12] W. TRENCH, *An algorithm for the inversion of finite Toeplitz matrices*, SIAM J. Appl. Math., 12 (1964), pp. 515–522.

[13] J. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 4, pp. 551-556, October 1989

011

LEVERRIER'S ALGORITHM: A NEW PROOF AND EXTENSIONS*

STEPHEN BARNETT†

Abstract. A new derivation is given of the Leverrier–Fadeev algorithm for simultaneous determination of the adjoint and determinant of the $n \times n$ characteristic matrix $\lambda I_n - A$. The proof uses an appropriate companion matrix and is of some interest in its own right. The method is extended to produce a corresponding scheme for the inverse of the polynomial matrix $\lambda^2 I_n - \lambda A_1 - A_2$, and indeed can be generalized for a regular polynomial matrix of arbitrary degree. The results have application to linear control systems theory.

Key words. characteristic matrix, characteristic polynomial, Leverrier–Fadeev algorithm, polynomial matrices

AMS(MOS) subject classifications. 15A18, 65F30

1. Introduction. Consider an $n \times n$ matrix A having characteristic polynomial

$$a(\lambda) = \det(\lambda I_n - A) \tag{1.1}$$

$$= \lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n \tag{1.2}$$

where I_n denotes the unit matrix of order n, and let

$$(\lambda I_n - A)^{-1} = (\lambda^{n-1}I_n + \lambda^{n-2}B_1 + \cdots + \lambda B_{n-2} + B_{n-1})/a(\lambda). \tag{1.3}$$

A well-known algorithm attributed to Leverrier, Fadeev, and others permits simultaneous determination of the coefficients a_k and the matrices B_k by means of the formulae

$$a_1 = -\mathrm{tr}\,(A), \qquad a_k = -\frac{1}{k}\mathrm{tr}\,(AB_{k-1}), \qquad k = 2, 3, \cdots, n, \tag{1.4}$$

$$B_1 = A + a_1 I_n, \qquad B_k = AB_{k-1} + a_k I_n, \qquad k = 2, 3, \cdots, n-1 \tag{1.5}$$

where tr (A) denotes the trace of A. The scheme is useful for theoretical, if not computational purposes, and finds application in linear control theory [3] and elsewhere. The usual proof relies on Newton's formulae that relate sums of powers of the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$ of A to the coefficients in (1.2). Details can be found, for example, in [4] and [8]. The purpose of this paper is first, to give in § 2 a simple alternative derivation of the algorithm using a companion matrix associated with $a(\lambda)$; and second, to present an extension in § 3 for the polynomial matrix $\lambda^2 I - \lambda A_1 - A_2$, this being readily generalized for a regular polynomial matrix of arbitrary degree.

2. Proof of the algorithm. From (1.2) and (1.3) we obtain the identity

$$(\lambda I_n - A)(\lambda^{n-1}I_n + \lambda^{n-2}B_1 + \cdots + B_{n-1}) = (\lambda^n + a_1\lambda^{n-1} + \cdots + a_n)I_n. \tag{2.1}$$

Equating coefficients of $\lambda^{n-1}, \lambda^{n-2}, \cdots, \lambda$ in (2.1) immediately produces (1.5)—note also that a check on the calculations is provided by $AB_{n-1} + a_n I_n = 0$.

Now define the companion matrix

$$C = \begin{bmatrix} 0 & I_{n-1} \\ -a_n \cdots & -a_2 - a_1 \end{bmatrix} \tag{2.2}$$

[*] Received by the editors April 20, 1988; accepted for publication (in revised form) January 9, 1989. This paper was presented at the Third SIAM Conference on Applied Linear Algebra, Madison, Wisconsin, May 23–26, 1988.

[†] Department of Mathematics, University of Bradford, Bradford BD7 1DP, West Yorkshire, United Kingdom.

whose characteristic polynomial is $a(\lambda)$ in (1.1). The derivation of (1.4) relies on the fact that since A and C have the same eigenvalues (although of course they need not be similar), then

$$\text{(2.3)} \qquad \operatorname{tr}(A^k) = \operatorname{tr}(C^k) = \sum_{i=1}^{n} \lambda_i^k.$$

Substituting (1.5) into (1.4) gives

$$\text{(2.4)} \quad a_1 = -\operatorname{tr}(A), \quad a_k = -\frac{1}{k}\operatorname{tr}(A^k + a_1A^{k-1} + \cdots + a_{k-1}A), \quad k = 2, \cdots, n.$$

In view of (2.3), it follows that the desired formulae (1.4) will be established if we can prove that

$$\text{(2.5)} \qquad a_k = -\frac{1}{k}\operatorname{tr}(C^k + a_1C^{k-1} + \cdots + a_{k-1}C), \qquad k = 2, \cdots, n,$$

since it is obvious from (2.2) that $a_1 = -\operatorname{tr}(C)$. Two preliminary results are needed, which are of some interest in their own right.

LEMMA 1. *The rows $x_1, x_2, \cdots, x_n$ of the matrix*

$$\text{(2.6)} \qquad \alpha_0C^m + \alpha_1C^{m-1} + \cdots + \alpha_mI_n, \qquad m < n$$

are given by

$$\text{(2.7)} \qquad x_1 = [\alpha_m, \alpha_{m-1}, \cdots, \alpha_0, 0, \cdots, 0], \qquad x_{i+1} = x_iC, \quad i \geqq 1.$$

A proof can be found in [1] and relies on the simple observation that

$$\text{(2.8)} \qquad e_iC^k = e_{i+k}, \qquad k \geqq 1$$

where e_i denotes the ith row of I_n.

Another proof of (1.4) due to Frame [2] is based on showing that

$$\operatorname{tr}(A^k + a_1A^{k-1} + \cdots + a_kI_n) = (n-k)a_k.$$

Although this is all that is required, we can in fact be more specific when A is the companion matrix C in (2.2). In this case we have Lemma 2.

LEMMA 2. *The principal diagonal of*

$$\text{(2.9)} \qquad E_k = C^k + a_1C^{k-1} + \cdots + a_{k-1}C + a_kI_n, \qquad k = 1, 2, \cdots, n-1$$

is $a_k, a_k, \cdots, a_k, 0, \cdots, 0$, where there are k zeros.

Proof. For $i = 1, 2, \cdots, n-k$ the i, i element of E_k is $e_iE_ke_i^T$, where superscript T denotes transpose. Then

$$\begin{aligned} e_iE_ke_i^T &= (e_iC^k + a_1e_iC^{k-1} + \cdots + a_ke_i)e_i^T \\ &= (e_{i+k} + a_1e_{i+k-1} + \cdots + a_ke_i)e_i^T \\ &= a_k \end{aligned}$$

using (2.8) and the property $e_je_i^T = \delta_{ji}$, the Kronecker delta.

To determine the remaining diagonal elements of E_k, for $i = n-k+1$, $n-k+2, \cdots, n$, consider the Cayley–Hamilton identity

$$C^n + a_1C^{n-1} + \cdots + a_nI_n = 0,$$

which can be rewritten as

$$(2.10)\qquad C^{n-k}E_k=-(a_{k+1}C^{n-k-1}+\cdots+a_{n-1}C+a_nI_n)$$

where E_k is defined in (2.9). Let the matrix within parentheses on the right in (2.10) have rows $y_1, y_2, \cdots, y_n$. Then by (2.7)

$$y_1=[a_n,a_{n-1},\cdots,a_{k+1},0,\cdots,0]$$

and

$$\begin{aligned} y_i&=y_1C^{i-1}, \qquad i\leqq k\\ (2.11)\qquad &=(a_ne_1+a_{n-1}e_2+\cdots+a_{k+1}e_{n-k})C^{i-1}\\ &=a_ne_i+a_{n-1}e_{i+1}+\cdots+a_{k+1}e_{n-k+i-1} \quad \text{by (2.8)}. \end{aligned}$$

Now denote the rows of E_k in (2.9) by $z_1, z_2, \cdots, z_n$. By a further application of (2.8) it follows that the ith row on the left side of (2.10) is

$$\begin{aligned} e_iC^{n-k}E_k&=e_{n-k+i}E_k\\ &=z_{n-k+i}, \qquad i=1,2,\cdots,k. \end{aligned}$$

Equating the first k rows on either side of (2.10) therefore produces

$$(2.12)\qquad z_{n-k+i}=-y_i, \qquad i=1,2,\cdots,k.$$

Comparing (2.12) with (2.11) shows that the coefficient of e_{n-k+i} in Z_{n-k+i} is zero for $i=1,\cdots,k$. In other words, the entries on the principal diagonal of rows $n-k+1$, $\cdots$, n of E_k are all zero, and this completes the proof of Lemma 2.

To conclude the derivation of (2.5), we have from (2.9)

$$\begin{aligned} \operatorname{tr}(C^k+a_1C^{k-1}+\cdots+a_kC)&=\operatorname{tr}(E_k)-\operatorname{tr}(a_kI_n)\\ &=(n-k)a_k-na_k \quad \text{by Lemma 2}\\ &=-ka_k, \end{aligned}$$

which is the required result.

3. Extension to polynomial matrices. We now give the algorithm corresponding to (1.4) and (1.5) for the polynomial matrix $\lambda^2I_n-\lambda A_1-A_2$.

THEOREM. *If*

$$(3.1)\qquad (\lambda^2I_n-\lambda A_1-A_2)^{-1}=(\lambda^{2n-2}\mathrm{I}_n+\lambda^{2n-3}\beta_1+\cdots+\beta_{2n-2})/d(\lambda)$$

where

$$(3.2)\qquad d(\lambda)=\lambda^{2n}+d_1\lambda^{2n-1}+\cdots+d_{2n-1}\lambda+d_{2n},$$

then the coefficients in (3.1) *and* (3.2) *can be determined sequentially by*

$$(3.3)\qquad d_1=-\operatorname{tr}(A_1), \qquad \beta_1=A_1+d_1I_n,$$

$$(3.4)\qquad d_k=-\frac{1}{k}\operatorname{tr}(A_1\beta_{k-1}+2A_2\beta_{k-2}), \qquad k=2,3,\cdots,2n-2,$$

$$(3.5)\qquad \beta_k=A_1\beta_{k-1}+A_2\beta_{k-2}+d_kI_n, \qquad k=2,3,\cdots,2n-2$$

where $\beta_0=I_n$; *and* d_{2n-1}, d_{2n} *are determined by*

$$(3.6)\qquad A_1\beta_{2n-2}+A_2\beta_{2n-3}=-d_{2n-1}I_n, \qquad A_2\beta_{2n-2}=-d_{2n}I_n.$$

Proof. By equating coefficients of $\lambda^{2n-1}, \lambda^{2n-2}, \cdots, \lambda^2$ in the identity

$$(3.7) \qquad (\lambda^2 I_n - \lambda A_1 - A_2)(\lambda^{2n-2}\mathrm{I}_n + \lambda^{2n-3}\beta_1 + \cdots + \beta_{2n-2}) = d(\lambda) I_n$$

the expressions for $\beta_1, \beta_2, \cdots, \beta_{2n-2}$ in (3.3) and (3.5) are obtained immediately. Consider the block companion matrix

$$(3.8) \qquad D = \begin{bmatrix} 0 & I_n \\ A_2 & A_1 \end{bmatrix}$$

for which

$$\det(\lambda I_{2n} - D) = \det(\lambda^2 I_n - \lambda A_1 - A_2) = d(\lambda).$$

The coefficients of $d(\lambda)$ in (3.2) can therefore be obtained by applying (1.4) and (1.5) to D. Hence $d_1 = -\mathrm{tr}\,(D) = -\mathrm{tr}\,(A_1)$, which is (3.3), and

$$(3.9) \qquad \begin{aligned} d_k &= -\frac{1}{k}\,\mathrm{tr}\,(D^k + d_1 D^{k-1} + \cdots + d_{k-1} D), \qquad k \geqq 2 \\ &= -\frac{1}{k}\,\mathrm{tr}\,(F), \quad \text{say.} \end{aligned}$$

The expressions (3.4) are obtained from (3.9) by establishing that

$$(3.10) \qquad F = \begin{bmatrix} \beta_{k-2}A_2 & \beta_{k-1} \\ \beta_{k-1}A_2 & (A_2\beta_{k-2} + A_1\beta_{k-1}) \end{bmatrix}.$$

This is easily done by induction as follows. It is routine to verify that (3.10) holds for $k = 2$. Using (3.8) and (3.10) produces

$$\begin{aligned} &D^{k+1} + d_1 D^k + \cdots + d_k D \\ &\quad = DF + d_k D \\ &\quad = \begin{bmatrix} \beta_{k-1}A_2 & (A_2\beta_{k-2} + A_1\beta_{k-1} + d_k I_n) \\ (A_2\beta_{k-2}A_2 + A_1\beta_{k-1}A_2 + d_k A_2) & (A_2\beta_{k-1} + A_1 A_2 \beta_{k-2} + A_1^2\beta_{k-1} + d_k A_1) \end{bmatrix} \\ &\quad = \begin{bmatrix} \beta_{k-1}A_2 & \beta_k \\ \beta_k A_2 & A_2\beta_{k-1} + A_1\beta_k \end{bmatrix} \end{aligned}$$

using the expression for β_k in (3.5). This verifies the induction hypothesis. Combining equations (3.9) and (3.10) then produces the required formula (3.4), on recalling that $\mathrm{tr}\,(A_2\beta_{k-2}) = \mathrm{tr}\,(\beta_{k-2}A_2)$. The expressions (3.6) follow at once by equating the coefficients of λ and λ^0 in (3.7). □

Remarks. (1) It follows at once from (3.6) that provided $d_{2n} \neq 0$, then $A_2^{-1} = -\beta_{2n-2}/d_{2n}$.

(2) It is also clear from (3.6) that only one element on the principal diagonal of each of $(A_1\beta_{2n-2} + A_2\beta_{2n-3})$ and $A_2\beta_{2n-2}$ is needed to determine d_{2n-1} and d_{2n}, respectively. However, evaluating these matrices in full can serve as a check on the calculations.

(3) Two alternative expressions for d_{2n-1} can be obtained after realizing that the argument used to derive (3.4) still holds for $k = 2n - 1$, so that

$$(3.11) \qquad d_{2n-1} = -\frac{1}{2n-1}\,\mathrm{tr}\,(A\beta_{2n-2} + 2A_2\beta_{2n-3}).$$

Equating the trace of each side of the first equation in (3.6) gives

$$d_{2n-1} = -\frac{1}{n}\operatorname{tr}(A_1\beta_{2n-2} + A_2\beta_{2n-3}). \tag{3.12}$$

Comparing (3.11) and (3.12) reveals that

$$\operatorname{tr}(A_2\beta_{2n-3}) = (n-1)\operatorname{tr}(A_1\beta_{2n-2}) \tag{3.13}$$

and substituting (3.13) into (3.11) produces Lemma 3.

LEMMA 3.

$$d_{2n-1} = -\operatorname{tr}(A_1\beta_{2n-2}) \tag{3.14}$$

$$= -\frac{1}{n-1}\operatorname{tr}(A_2\beta_{2n-3}). \tag{3.15}$$

It is clear that the theorem can be generalized to any regular polynomial matrix in the form

$$\lambda^N I_n - \lambda^{N-1}A_1 \cdots - \lambda A_{N-1} - A_N \tag{3.16}$$

by applying (1.4) and (1.5) to the block companion matrix

$$\begin{bmatrix} 0 & I_n & & 0 & \\ 0 & 0 & I_n & & \\ & & & \ddots & \\ & & & & I_n \\ A_N & \cdot & \cdot & \cdot & A_1 \end{bmatrix}.$$

It is surprising that extension of the method for $N > 1$ appears to be available only in a little-known Russian paper [5], which uses a different method of proof from that given above.

Example. Consider

$$A(\lambda) = \begin{bmatrix} \lambda^2 - 2\lambda - 4 & -3\lambda + 5 & \lambda - 3 \\ 4\lambda + 3 & \lambda^2 - \lambda & 2\lambda - 2 \\ 11 & -5\lambda - 2 & \lambda^2 - 9\lambda + 1 \end{bmatrix}$$

so that

$$A_1 = \begin{bmatrix} 2 & 3 & -1 \\ -4 & 1 & -2 \\ 0 & 5 & 9 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 4 & -5 & 3 \\ -3 & 0 & 2 \\ -11 & 2 & -1 \end{bmatrix}.$$

From (3.3) we have

$$d_1 = -\operatorname{tr}(A_1) = -12,$$

$$\beta_1 = A_1 - 12I$$

$$= \begin{bmatrix} -10 & 3 & -1 \\ -4 & -11 & -2 \\ 0 & 5 & -3 \end{bmatrix}.$$

Equation (3.4) gives

$$d_2 = -\tfrac{1}{2}\operatorname{tr}(A_1\beta_1 + 2A_2) = 48$$

and from (3.5)

$$\beta_2 = A_1\beta_1 + A_2 + 48I$$
$$= \begin{bmatrix} 20 & -37 & -2 \\ 33 & 15 & 10 \\ -31 & -8 & 10 \end{bmatrix}.$$

Continuing to apply (3.4) and (3.5) alternately, we obtain

$$d_3 = -\tfrac{1}{3}\operatorname{tr}(A_1\beta_2 + 2A_2\beta_1) = -157,$$

$$\beta_3 = A_1\beta_2 + A_2\beta_1 - 157I$$
$$= \begin{bmatrix} -7 & 61 & 13 \\ 45 & 23 & -5 \\ -12 & -57 & -7 \end{bmatrix},$$

$$d_4 = -\tfrac{1}{4}\operatorname{tr}(A_1\beta_3 + 2A_2\beta_2) = 41,$$

$$\beta_4 = A_1\beta_3 + A_2\beta_2 + 41I$$
$$= \begin{bmatrix} -4 & 1 & -10 \\ -25 & 29 & -17 \\ -6 & 47 & -15 \end{bmatrix}.$$

Finally, the expressions in (3.6) give

$$A_1\beta_4 + A_2\beta_3 = -366I, \qquad A_2\beta_4 = 91I$$

so that $d_5 = 366$, $d_6 = -91$. Alternatively, it is easy to verify that in (3.14) and (3.15), $-\operatorname{tr}(A_1\beta_4) = -\frac{1}{2}\operatorname{tr}(A_2\beta_3) = 366$. We therefore have

$$\det A(\lambda) = \lambda^6 - 12\lambda^5 + 48\lambda^4 - 157\lambda^3 + 41\lambda^2 + 366\lambda - 91,$$

adj $A(\lambda)$

$$= \begin{bmatrix} (\lambda^4 - 10\lambda^3 + 20\lambda^2 - 7\lambda - 4) & (3\lambda^3 - 37\lambda^2 + 61\lambda + 1) & (-\lambda^3 - 2\lambda^2 + 13\lambda - 10) \\ (-4\lambda^3 + 33\lambda^2 + 45\lambda - 25) & (\lambda^4 - 11\lambda^3 + 15\lambda^2 + 23\lambda + 29) & (-2\lambda^3 + 10\lambda^2 - 5\lambda - 17) \\ (-31\lambda^2 - 12\lambda - 6) & (5\lambda^3 - 8\lambda^2 - 57\lambda + 47) & (\lambda^4 - 3\lambda^3 + 10\lambda^2 - 7\lambda - 15) \end{bmatrix}.$$

Different extensions of Leverrier's algorithm arising from singular linear control systems have appeared in the literature recently [6], [7]. The problem is to expand $(\lambda J - A)^{-1}$ where J is singular but $\det(\lambda J - A) \not\equiv 0$, and an interesting question for future research is whether a scheme can be found to deal with (3.16) when I is replaced by J.

REFERENCES

[1] S. Barnett, *Polynomials and Linear Control Systems*, Marcel Dekker, New York, 1983, pp. 14–15.

[2] J. S. Frame, *Matrix functions and applications, Part* IV—*Matrix functions and constituent matrices*, IEEE Spectrum, 1 (1964), pp. 123–131.

[3] M. Clique and J.-C. Gille, *On companion matrices and state variable feedback*, Podstawy Sterowania, 15 (1985), pp. 367–376.

[4] F. R. Gantmacher, *The Theory of Matrices*, Vol. 1, Chelsea, New York, 1960, pp. 87–88.

[5] V. I. Gugnina, *Extension of D.K. Fadeev's method to polynomial matrices*, Dokl. Akad. Nauk UzSSR, 1 (1958), pp. 5–10. (In Russian.)

[6] F. L. Lewis, *Further remarks on the Cayley-Hamilton theorem and Leverrier's method for the matrix pencil* ($sE - A$), IEEE Trans. Automat. Control, 31 (1986), pp. 869–870.

[7] B. G. Mertzios, *Leverrier's algorithm for singular systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 652–653.

[8] H. H. Rosenbrock, *State-Space and Multivariable Theory*, Nelson, London, 1970, p. 13.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 4, pp. 557-592, October 1989

012

CHOLESKY FACTOR UPDATING TECHNIQUES FOR RANK 2 MATRIX MODIFICATIONS*

RICHARD BARTELS † AND LINDA KAUFMAN ‡

Abstract. Gill, Golub, Murray, and Saunders have described five methods by which the Cholesky factors of a positive definite matrix may be updated when the matrix is subjected to a symmetric rank 1 modification. For a negative rank 1 update, a modification of one of their methods was given by Lawson and Hanson and analyzed by Bojanczyk, Brent, van Dooren, and de Hoog. In many minimization algorithms, symmetric rank 2 modifications are found.

This paper shows how each of the rank 1 methods gives rise to a single-application rank 2 method. For some of the methods, this involves a new Householder transformation technique designed to eliminate elements of two vectors at once using a rank 1 correction of the identity matrix.

The authors' experiments on scalar, vector, and shared-memory multiple-instructions multiple-data machines show that it is more economical to perform rank 2 updates rather than two rank 1 updates. In their comparison, the authors do not consider pipelining two applications of the rank 1 algorithms, which in certain instances is possible.

Key words. Cholesky, updating, parallel

AMS(MOS) subject classifications. 65F05, 65W05

1. Introduction. In [5] Gill, Golub, Murray, and Saunders have described a number of methods by which the Cholesky factors of a real, $n \times n$, positive definite matrix A may be updated when the matrix is subjected to a symmetric rank 1 modification:

$$\bar{A} = A + b\mathbf{z}\mathbf{z}^t \ , \tag{1.1}$$

and the resulting matrix $\bar{A}$ is also positive definite. In minimization and root-finding algorithms based upon quasi-Newton techniques, symmetric rank 2 updates are used:

$$\begin{aligned} \bar{A} &= A + b_{11}\mathbf{z}_1\mathbf{z}_1^T + b_{12}(\mathbf{z}_1\mathbf{z}_2^T + \mathbf{z}_2\mathbf{z}_1^T) + b_{22}\mathbf{z}_2\mathbf{z}_2^T \\ &= A + [\mathbf{z}_1\mathbf{z}_2] \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} \mathbf{z}_1^T \\ \mathbf{z}_2^T \end{bmatrix} \\ &= A + ZBZ^T. \end{aligned} \tag{1.2}$$

The problem of multirank updates also appears in the QR factorization for Toeplitz matrices of Bojanczyk, Brent, and de Hoog [3] . The Gill, Golub, Murray, and Saunders techniques are applicable to rank k updates if they are expressed as k successive rank 1 changes.

With the exception of the first method given in [5], C1, which was based upon work by Bennet [1] for arbitrary rank modifications, all of the updating methods for Cholesky factors given by Gill, Golub, Murray, and Saunders are specific to the rank 1 case. Further, all save Bennet's method make use of Givens or Householder transformations.

*Received by the editors May 31, 1988; accepted for publication (in revised form) March 2, 1989. This paper was typeset at Bell Laboratories, Murray Hill, New Jersey, using the ***troff*** program running under the UNIX™ operating system. Final copy was produced on July 12, 1989.

† Computer Science Department, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 (rhbartels@waterloo.edu).

‡ Bell Laboratories, Murray Hill, New Jersey 07974 (att.com!lck).

In this work we use modifications of Householder transformations that enable us to translate the rank 1 methods C2-C5 of [5] into methods for updating Cholesky factors that are directly applicable to rank 2 modifications. We also give an algorithm that generalizes the modification of C3 that was given in Lawson and Hanson [8] and proved stable by Bojanczyk, Brent, van Dooren, and de Hoog [2].

We give operation counts to show that the rank 2 methods that we propose are no less efficient than two applications of their rank 1 counterparts, and we indicate where there might be an advantage in using the rank 2 methods.

Goldfarb [6] has also looked at methods for updating the Cholesky factorization of matrices that arise using variable metric methods for minimizing functions. In particular, he considers methods for updating the factorization of the matrix

$$A'=(I+\mathbf{v}\mathbf{u}^T)A(I+\mathbf{u}\mathbf{v}^T), \tag{1.3}$$

which is slightly less general than (1.2). In (1.3) one eigenvalue of the rank 2 change is nonnegative and one is nonpositive, while in (1.2) the signs of the eigenvalues of ZBZ^T are not restricted. Let L be the given lower triangular matrix such that $LL^T=A$, i.e., the Cholesky factorization of A. Goldfarb shows that (1.3) can be written as

$$A'=L(I+\mathbf{z}\mathbf{w}^T)(I+\mathbf{w}\mathbf{z}^T)L^T, \tag{1.4}$$

where $L\mathbf{z}=\mathbf{v}$ and $L\mathbf{w}=A\mathbf{u}$. He then gives two methods for finding an orthogonal matrix Q such that

$$(I+\mathbf{z}\mathbf{w}^{T)}Q=\tilde{L}, \tag{1.5}$$

where $\tilde{L}$ is a lower triangular matrix whose elements below the diagonal are essentially given by

$$\tilde{l}_{ij}=\mathbf{f}_i^T\mathbf{g}_j, \tag{1.6}$$

where $\mathbf{f}_i$ and $\mathbf{g}_j$ are vectors of length 2. Obtaining $\tilde{L}$ requires $O(n)$ operations, and multiplying L by $\tilde{L}$ to obtain the Cholesky factorization of A' requires $2n^2+O(n)$ multiplications.

Our generalization of methods C2, C4, and C5 of [5] in some sense parallels the ideas of Goldfarb. Rather than (1.4), we use a lemma proved in § 2 that there exists a matrix C such that

$$\bar{A}=L(I+VCV^T)(I+VC^TV^T)L^T, \tag{1.7}$$

where L plays the same role as in (1.4) and $LV=Z$. Our various generalizations of the methods in [5] give different methods for computing Q such that

$$(I+VCV^T)Q=\tilde{L}, \tag{1.8}$$

where again $\tilde{L}$ is lower triangular and has the same essential form as (1.6). As in Goldfarb's case, constructing $\tilde{L}$ requires $O(n)$ operations while computing $L\tilde{L}$ to form the new Cholesky factorization of A', which requires $2n^2+O(n)$ operations. The major difference is that because (1.5) is slightly simpler than (1.8), the coefficient of the $O(n)$ term in the construction of $\tilde{L}$ is smaller. Thus the advantage of Goldfarb's approach for the more limiting, although most prevalent problem, is a decrease in the coefficient of the lower-order terms in the theoretical complexity count.

When we began our examination of rank 2 updates, computational work was done on scalar machines. Theoretically, on these machines the computation saved using our rank 2 algorithms rather than two rank 1 updating algorithms was rather small. However, when we implemented our algorithms we discovered that, especially on vector and parallel machines, the results were more encouraging. The main reason for the difference between theory and practice is the structure of the algorithms.

The time-consuming sections in all the programs fall into three categories: triangular solves, sequences of planar transformations, and nested double updates, which we will explain later.

Algorithms C2, C4, C3 with a negative update, and C5 of [5] all begin with a triangular solve using the current L, which is updated later in the program. The generalizations to rank 2 algorithms start with a triangular solve involving two right-hand sides. Computing two solutions simultaneously rather than sequentially saves loop overhead. On a sophisticated machine without parallelization, some address calculation and vector loads and stores may be saved when the same triangular matrix is used. On machines with parallel processing the solution for each right-hand side with the rank 2 algorithms may be done concurrently. Thus, although the operation count is the same, two instantiations of a triangular solve with one right-hand side will probably take more time than a triangular solve with the same matrix but two right-hand sides.

Algorithms C3 and C4 of [5] both involve sequences of 2 plane transformations that are preceded by a triangular solve in C4 and in C3 with a negative update. In the generalizations for the rank 2 case, these 2 plane transformations are replaced by 3 plane transformations. Thus code of the form

Algorithm A

```
For $i=1, \cdots ,n$
    Compute numbers $\alpha$, $\beta$, and $\gamma$.
    For $j=i,i+1, \cdots ,n$
        Set $t=l_{j,i}+\alpha \times y_j$.
        Set $l_{j,i}=l_{j,i}+\beta \times t$.
        Set $y_j=y_j+t \times \gamma$.
```

for the most part was replaced by code of the form

Algorithm B

```
For $i=1, \cdots ,n$
    Compute numbers $\alpha$, $\beta$ , $\gamma$, $\delta$, and $\sigma$.
    For $j=i, \cdots ,n$
        Set $t=l_{j,i}+\alpha \times y_j+\delta \times w_j$.
        Set $l_{j,i}=l_{j,i}+\beta \times t$.
        Set $y_j=y_j+t \times \gamma$.
        Set $w_j=w_j+t \times \sigma$.
```

Obviously, two applications of Algorithm A would take more time than one application on Algorithm B. Not only are fewer arithmetic operations involved, but there are also fewer vector memory references in Algorithm B than in two applications of Algorithm A. In C3 with a positive update, one could pipeline two rank 1 corrections and begin the computation of the ith column of the new L for the second application as soon as the ith column of L is known for the first application. In fact, when the matrix B in (1.2) is indefinite, Algorithm B is essentially a pipelining of C4 for positive and negative updates. Algorithm C4 begins with a triangular solve and ends with a sequence of planar transformations. For two rank 1 corrections with that algorithm, one could pipeline the ending sequence of transformations for the first update with the triangular solve for the second update.

Double updating appears in C1, C2, and C5 and is essentially code of the form

Algorithm C

```
For $i=1, \cdots ,n$
    Compute numbers $a$ and $b$.
    For $j=i, \cdots ,n$
        Set $y_j=y_j+a\times l_{j,i}$.
        Set $l_{j,i}=l_{j,i}+b\times y_j$.
```

These loops for the rank 2 updating algorithms were changed to

Algorithm D

```
For $i=1,2, \cdots ,n$
    Compute numbers $a$, $b$, $c$, and $d$.
    For $j=i, \cdots ,n$
        Set $y_j=y_j+a\times l_{j,i}$.
        Set $w_j=w_j+c\times l_{j,i}$.
        Set $l_{j,i}=l_{j,i}+b\times y_j+d\times w_j$.
```

Although two applications of Algorithm C have the same floating point operation count as Algorithm D, the execution time of two instances of Algorithm C might be considerably more than one instance of Algorithm D. For a scalar machine, Algorithm D requires less loop control than two applications of Algorithm C. On a vector machine like the Cray-1, in which the number of vector memory references is the best measure of algorithm performance for algorithms like those given above, two applications of Algorithm C might require 1.2 as much time as one application of Algorithm D.

Algorithms C2 and C5 of [5] begin with a forward solve with the current L before Algorithm C is begun. For two rank 1 corrections with these algorithms, one could pipeline Algorithm C for the first rank 1 correction with a triangular solve for the second rank 1 correction. However, for C1, the least stable of the algorithms, one could pipeline Algorithm C itself as either

Algorithm E

```
For $i=1, \cdots ,n$
  Compute numbers $a$, $b$, $c$, and $d$.
  For $j=i, \cdots ,n$
      Set $y_j=y_j+a\times l_{j,i}$.
      Set $l_{j,i}=l_{j,i}+b\times y_j$.
      Set $w_j=w_j+c\times l_{j,i}$.
      Set $l_{j,i}=l_{j,i}+d\times w_j$.
```

or

Algorithm F

```
For $j=1, \cdots ,n$
    Set $y_j=y_j+a\times l_{j,1}$.
    Set $l_{j,1}=l_{j,1}+b\times y_j$.
For $i=2, \cdots ,n$
    Compute numbers $a$, $b$, $c$, and $d$.
    Set $w_{i-1}=w_{i-1}+c\times l_{i-1,i-1}$.
    Set $l_{i-1,i-1}=l_{i-1,i-1}+d\times w_{i-1}$.
    For $j=i, \cdots ,n$
        Set $y_j=y_j+a\times l_{j,i}$.
        Set $w_j=w_j+c\times l_{j,i-1}$.
        Set $l_{j,i}=l_{j,i}+b\times y_j$.
        Set $l_{j,i-1}=l_{j,i-1}+d\times w_j$.
Set $w_{n,n}=w_{n,n}+c\times l_{n,n}$.
Set $l_{n,n}=l_{n,n}+d\times w_n$.
```

On a machine with concurrency, Algorithm F could be faster than Algorithm E, because in the inner loop **w** and **y** could be handled simultaneously.

TABLE 1.1
Comparison of 2×*Algorithm C and Algorithms D, E, and F.*

	n	2× C	D	E	F	(2×C)/D
Sequent – no multiprocessing	100	.461	.407	.423	.418	1.13
	200	1.82	1.60	1.68	1.66	1.14
Sequent – with multiprocessing	100	.213	.135	.136	.144	1.57
	200	.539	.379	.380	.395	1.42
Vax 750 with floating point acceleration	100	.570	.394	.531	.555	1.45
	400	8.97	6.23	8.37	8.98	1.44
Cray XMP with vectorization but no multiprocessing	100	.000575	.000427	.000466	.000482	1.35
	800	.0209	.0180	.0197	.0187	1.16
Convex	100	.00349	.00341	.00293	.00348	1.02
	400	.0438	.0438	.0357	.0425	1.00
Alliant – with vectorization and concurrency	100	.00718	.00504	.00542	.00625	1.42
	1600	1.25	.927	.990	1.041	1.35

Table 1.1 gives the times for two applications of Algorithm C and one of Algorithms D, E, and F for several machines for $n=100$ and for some large values of n which were chosen to be considered a good large size for the machine in question. We also show the ratio for two applications of Algorithm C over one for Algorithm D. No computing was done to replace the phrase "Compute numbers" in the above algorithms. We have included Algorithms E and F to indicate what might happen if one knew that two rank 1

corrections were needed and changed the programs accordingly, without trying to use any of the theory developed in this paper. Certainly that would be the preferable mode on the Convex, but on the other machines there is the advantage of using Algorithm D versus Algorithms E or F, although not much on the Sequent. The generic speedups given in Table 1.1 will be reflected in the specific generalizations of C1 through C5 of [5] given in the later sections of this paper. A good indication of the performance in two different environments is the Sequent machine. To obtain multiprocessing, the user has to state explicitly around each "DO" loop in FORTRAN that the loop should use more processors and which variables should be shared and which are local. Thus the user has control of the situation. We see little speedup when multiprocessing is turned off, but when it is enabled, the speedup is large. The user should keep in mind that under the traditional measure of counting floating point operations, there should be no speedup.

Throughout this paper we will indicate opportunities for parallel and vector operations and show the performance on various machines. We will give speedup ratios for two applications of a rank 1 algorithm versus our rank 2 algorithm. Speedups of about 1.4 will be common. Since one rank 2 update cannot possibly cost less than one rank 1 update on a given machine, our speedups are bounded by 2. In our comparisons we do not consider pipelining two applications of the rank 1 algorithms, which, as we pointed out, could have been done in certain instances.

In § 2 of this paper we prove a lemma that is the foundation of our generalizations of algorithms C1-C5. We also give a modification of the traditional Householder transformations that permits the annihilation of elements of two vectors at once. In §§ 3 through 7 we present generalizations of Algorithms C1 through C5 of [5]. In § 5 we also generalize row removal method 3 of [8]. The user's particular problem, the mode in which the matrix factors are stored, the file structures used if the problem is large, the sparsity of the matrices being considered, and the fill-in properties of the method used all may play a role in the choice of an updating technique. As with Gill, Golub, Murray, and Saunders, it is not our purpose to recommend the use of a particular method where there exist more than one. Section 8 summarizes our computational results on various classes of machines.

2. Preliminaries.

Certain operations on pairs of vectors or on symmetric 2×2 matrices are basic to what will come later. We begin by establishing these operations.

2.1. Factoring $I+VBV^T$ in the positive definite rank 2 case.

In some of the updating methods to be discussed, the problem of finding the Cholesky factors of $\tilde{A}$, given those of A (as in (1.2)), involves the problem of finding the Cholesky factors of a positive definite matrix of the following form:

$$I+VBV^T \ , \tag{2.1.1}$$

where B is 2×2 and symmetric. As a first step toward producing the desired factors, we prove the following lemma.

LEMMA 1. *Assume $I+VBV^T$ is a symmetric positive definite matrix, where B is a 2×2 matrix and V is an n×2 matrix. Then there exists a symmetric matrix C such that*

$$I+VBV^T = (I+VCV^T)(I+VCV^T) \ . \tag{2.1.2}$$

Proof. Clearly (2.1.2) will be satisfied if

$$CXC + 2C - B = 0 \ , \tag{2.1.3}$$

where $X=V^TV$. It is easily verified that

$$C = X^{-1/2} \ [-I \pm (I+X^{1/2}BX^{1/2})^{1/2}]X^{-1/2} \tag{2.1.4}$$

will satisfy (2.1.3) formally, and we need only be assured that the square roots of X and $I+X^{1/2}BX^{1/2}$ exist.

Note that X is at least positive semidefinite, so it has a square root. As for the second square root, consider

$$V^T(I+VBV^T)V = X+XBX \ . \tag{2.1.5}$$

Since $I+VBV^T$ is assumed to be positive definite, this product must be at least positive semidefinite. That is,

$$X+XBX = X^{1/2}(I+X^{1/2}BX^{1/2})X^{1/2} \tag{2.1.6}$$

is positive semidefinite.

Now, consider any vector $\mathbf{y} \in E^2$. Then $\mathbf{y}=\mathbf{r}+\mathbf{s}$, where $\mathbf{r}$ is in the nullspace of $X^{1/2}$ and $\mathbf{s} = X^{1/2}\mathbf{t}$ for some $\mathbf{t}$. Consequently

$$\mathbf{y}^T(I+X^{1/2}BX^{1/2})\mathbf{y}= \mathbf{r}^T(I+X^{1/2}BX^{1/2})\mathbf{r}+\mathbf{r}^T(I+X^{1/2}BX^{1/2})\mathbf{s} + \mathbf{s}^T(I+X^{1/2}BX^{1/2})\mathbf{r} + \mathbf{t}^TX^{1/2}(I+X^{1/2}BX^{1/2})X^{1/2}\mathbf{t}. \tag{2.1.7}$$

The second and third terms are zero, the first term reduces to $\mathbf{r}^T\mathbf{r}$, which is nonnegative, and the last term is nonnegative because of our observation (2.1.6) above. Consequently, $I+X^{1/2}BX^{1/2}$ is positive semidefinite and has a square root. □

To compute the square root of a 2×2 matrix X is rather simple. Assume one can obtain the eigendecomposition of X in the form

$$X=QDQ^T, \tag{2.1.8}$$

where D is diagonal and Q is orthogonal. Then $X^{1/2}$ is given by

$$X^{1/2}=QD^{1/2}Q^T. \tag{2.1.9}$$

For a 2×2 matrix, Moler and Stewart [9] give a simple algorithm for determining the decomposition (2.1.8).

Fortunately, there is another lemma, similar to Lemma 1, that involves fewer operations to compute a factorization.

LEMMA 2. *Assume V is an n×2 full rank matrix and let $X=V^TV$. Assume B is a 2×2 diagonal matrix with $b_{11}>b_{22}$ if B is indefinite. If $I+VBV^T$ is a symmetric positive definite matrix, then there exists a lower triangular matrix C such that*

$$I+VBV^T = (I+VCV^T)(I+VC^TV^T) \ , \tag{2.1.10}$$

where

$$c_{11}=(-1+(1+x_{11}b_{11})^{1/2})/x_{11}, \tag{2.1.11}$$

c_{22} is the root of the equation

$$c_{22}^2y+2c_{22}-b_{22}=0, \tag{2.1.12}$$

where y in (2.1.12) *is given by*

$$y=(x_{22}(1+x_{11}b_{11})-x_{21}^2b_{11})/(1+x_{11}b_{11}), \tag{2.1.13}$$

and

$$c_{21}=-x_{21}c_{11}c_{22}/(1+x_{11}c_{11}). \tag{2.1.14}$$

Proof. If (2.1.11) is true, then c_{11} is the root of the equation

$$c_{11}^2 x_{11} + 2c_{11} - b_{11} = 0. \tag{2.1.15}$$

Using simple algebra, it is easy to show that if c_{11}, c_{22}, and c_{21} can be found satisfying (2.1.15), (2.1.12), and (2.1.14), then

$$B = C + C^T + CV^TVC^T,$$

which implies

$$I + VBV^T = V(C + C^T + CV^TVC^T)V^T.$$

Thus we must show that real roots of (2.1.15) and (2.1.12) can be computed.

Since X is positive definite, x_{11} is positive. If b_{11} is positive, then obviously $1 + x_{11}b_{11}$ is positive and hence c_{11} in (2.1.11) can be computed. Since $I + VBV^T$ is positive definite, $X + XBX$ must also be positive definite, which implies that

$$\mathbf{e}_1^T (X + XBX)\mathbf{e}_1 = x_{11}(1 + x_{11}b_{11}) + x_{21}^2 b_{22} > 0. \tag{2.1.16}$$

From the hypothesis of Lemma 2, if b_{11} is negative, so is b_{22}, which from (2.1.16) implies $1 + x_{11}b_{11} > 0$ so in all cases c_{11} can be computed.

The condition that c_{22} can be computed as a real root of (2.1.12) is equivalent to

$$(1 + x_{11}b_{11})(1 + b_{22}x_{22}) - b_{22}b_{11}x_{21}^2 \geq 0. \tag{2.1.17}$$

The left-hand side of (2.1.17) is just $det(I + BX)$. Since $X + XBX$ is positive definite, $det(X + XBX) = det(X)det(I + BX) > 0$, which implies $det(I + BX) > 0$, which implies (2.1.17), can be satisfied and c_{22} can be computed. □

When B does not satisfy the hypothesis of Lemma 2, but is symmetric, there exists an orthogonal matrix Q such that

$$B = Q\hat{B}Q^T,$$

where $\hat{B}$ is diagonal and $\hat{b}_{11} > \hat{b}_{22}$ if B is indefinite. If one lets $\hat{V} = VQ$, then one can find a lower triangular matrix $\hat{C}$ from Lemma 2 such that

$$I + \hat{V}\hat{B}\hat{V}^T = (I + \hat{V}\hat{C}\hat{V}^T)(I + \hat{V}\hat{C}^T\hat{V}^T)$$

and then the matrix $C = Q\hat{C}Q^T$ would satisfy (2.1.10).

2.2. A modified Householder transformation. Householder transformations, or elementary reflectors, as they are sometimes called, are symmetric and orthogonal matrices of the form

$$P = I - \frac{1}{\tau}\mathbf{u}\mathbf{u}^T, \tag{2.2.1}$$

where $\mathbf{u} \neq 0$ and $\tau = (\mathbf{u}^T\mathbf{u})/2$. They have the property that if $\mathbf{z}$ and $\mathbf{y}$ are two vectors of the same Euclidean length, the $\mathbf{u}$ can be chosen in (2.2.1) such that $P\mathbf{z} = \mathbf{y}$. The most frequent use of this property is made in transforming a given vector into another that has specified components equal to zero.

There are several features of Householder transformations that benefited [5], and we will take advantage of them. In the first place, if $\mathbf{z}$ can be partitioned as

$$\mathbf{z} = \begin{bmatrix} z_1 \\ \mathbf{z}_2 \end{bmatrix}$$

and we wish $P\mathbf{z}$ to have the form $(y_1, 0, \cdots, 0)^T$, then one can partition $\mathbf{u}$ as

$$\mathbf{u} = \begin{bmatrix} u_1 \\ \mathbf{z}_2 \end{bmatrix};$$

that is, the bottom $n-1$ elements of $\mathbf{z}$ and $\mathbf{u}$ will be the same.

Secondly, there are two notable cases of the matrices P. In the first case only the ith and jth components of $\mathbf{z}$ $(i \neq j)$ are transformed to produce a zero in the jth position . We will denote these matrices by P_j^i. In the second case the ith, jth, and kth components of $\mathbf{z}$ are transformed to produce a zero in the jth and kth positions. We will denote these matrices by P_{jk}^i.

When transformations of the form P_j^i and P_{jk}^i are applied to many vectors, it is more economical to rewrite (2.2.1) as

$$P = I - \mathbf{v}\mathbf{w}^T, \tag{2.2.2}$$

where $\mathbf{w}^T = (1, u_2/u_1, \cdots, u_n/u_1)^T$ and

$$\mathbf{v} = \frac{u_1}{\tau}\mathbf{u}.$$

For a transformation in two planes applied to k vectors, (2.2.2) requires $3k$ multiplications and for a transformation in three planes applied to k vectors, (2.2.2) requires $5k$ multiplications. Throughout this paper, when we give computational complexity counts involving standard Householder transformations in two and three planes, we will assume form (2.2.2) is being used.

For the purposes of our updating proposals, we would like to extend the notion of a Householder transformation in three planes somewhat. Let two vectors $\mathbf{z} = (\zeta_1, \cdots, \zeta_n)^T$ and $\mathbf{w} = (\omega_1, \cdots, \omega_n)^T$ be given $(n > 2$). We shall define an orthogonal matrix $P_{i,j,k}$ that when applied to both $\mathbf{z}$ and $\mathbf{w}$ will change the ith, jth, and kth entries and leave the kth entry zero $(i,j,k$ distinct) . Such a matrix will have the form

$$P_{i,j,k} = \begin{bmatrix} \alpha & \beta & \gamma \\ \beta & \delta & \eta \\ \gamma & \eta & \mu \end{bmatrix}. \tag{2.2.3}$$

Only rows and columns (i,j,k) have been indicated, and all components not explicitly given are equal to the components of the identity.

For the construction it is sufficient to consider the case

$$P_{1,2,3} = I - \frac{1}{\lambda}\mathbf{p}\mathbf{p}^T, \tag{2.2.4}$$

where $\mathbf{p} = [\pi_1, \pi_2, \pi_3]^T$ and $\lambda = \mathbf{p}^T\mathbf{p}/2$.

Any P of this form is easily seen to be orthogonal. The condition requiring that $P\mathbf{z}$ and $P\mathbf{w}$ both be vectors with a zero third component is

$$\zeta_3 - \frac{1}{\lambda}\pi_3(\pi_1\zeta_1 + \pi_2\zeta_2 + \pi_3\zeta_3) = \omega_3 - \frac{1}{\lambda}\pi_3(\pi_1\omega_1 + \pi_2\omega_2 + \pi_3\omega_3) = 0. \tag{2.2.5}$$

This condition can be satisfied by first finding π_1 and π_2 so that

$$\begin{aligned} \zeta_1\pi_1 + \zeta_2\pi_2 &= \zeta_3\kappa, \\ \omega_1\pi_1 + \omega_2\pi_2 &= \omega_3\kappa \end{aligned} \tag{2.2.6}$$

for some quantity κ, whose choice will be given in (2.2.8). Then (2.2.5) becomes

$$\zeta_3[1 - \frac{1}{\lambda}\pi_3(\kappa + \pi_3)] = \omega_3[1 - \frac{1}{\lambda}\pi_3(\kappa + \pi_3)] = 0.$$

This will be satisfied if we choose

$$\pi_3 = -\kappa - sgn(\kappa)\sqrt{\kappa^2+\pi_1^2+\pi_2^2}\,. \tag{2.2.7}$$

Condition (2.2.6) is most readily satisfied by letting

$$\pi_1 = \det\begin{bmatrix}\zeta_3\omega_3\\ \zeta_2\omega_2\end{bmatrix}, \tag{2.2.8}$$

$$\pi_2 = \det\begin{bmatrix}\zeta_1\omega_1\\ \zeta_3\omega_3\end{bmatrix},$$

$$\kappa = \det\begin{bmatrix}\zeta_1\omega_1\\ \zeta_2\omega_2\end{bmatrix} \quad .$$

This specification for $P_{1,2,3}$ could fail if $\pi_1=\pi_2=\pi_3=0$. From the definition of π_3, this implies that $\kappa=0$ as well. But in turn from (2.2.8) we find that this results in $\zeta_1=\zeta_2=\zeta_3=\omega_1=\omega_2=\omega_3=0$, in which case $P_{1,2,3}$ may be taken as the identity. The specification could also fail if **z** and **w** were collinear.

To draw the connection between the expressions given for $P_{i,j,k}$ in (2.2.3) and in (2.2.4), we note here that

$$\alpha = (\pi_2^2+\pi_3^2-\pi_1^2)/(\pi_2^2+\pi_2^2+\pi_3^2) = 1 - \frac{1}{\lambda}\pi_1^2, \tag{2.2.9}$$

$$\delta = (\pi_1^2+\pi_3^2-\pi_2^2)/(\pi_1^2+\pi_2^2+\pi_3^2) = 1 - \frac{1}{\lambda}\pi_2^2,$$

$$\mu = (\pi_1^2+\pi_2^2-\pi_3^2)/(\pi_1^2+\pi_2^2+\pi_3^2) = 1 - \frac{1}{\lambda}\pi_3^2,$$

$$\beta = -2\pi_1\pi_2/(\pi_1^2+\pi_2^2+\pi_3^2) = -\frac{1}{\lambda}\pi_1\pi_2,$$

$$\eta = -2\pi_2\pi_3/(\pi_1^2+\pi_2^2+\pi_3^2) = \frac{-1}{\lambda}\pi_2\pi_3,$$

$$\gamma = -2\pi_1\pi_3/(\pi_1^2+\pi_2^2+\pi_3^2) = -\frac{1}{\lambda}\pi_1\pi_3 \quad .$$

One should note that (2.2.4) is a rank 1 modification of the identity matrix that annihilates elements in two vectors and should not be confused with the generalized Householder transformations given in [4] , [10], and [7], which annihilate elements in k vectors with matrices that are rank k corrections to the identity matrix. Ironically, the work in [7] was stimulated by the current effort because it was thought necessary to generalize the algorithms in [5]. Although three planar versions of the Generalized Householder transformations of [7] and the block reflectors of [10] may be used instead of (2.2.2), they are a bit more costly.

2.3. Products of the $P_{i,j,k}$. In later sections we shall be employing matrices $P_{i,i+1,i+2}$, which are identity matrices, except for the 3x3 submatrices consisting of rows i, $i+1$, and $i+2$ in columns i, $i+1$, and $i+2$, where they have the form

$$\begin{bmatrix} \alpha_i & \beta_i & \gamma_i \\ \beta_i & \delta_i & \eta_i \\ \gamma_i & \eta_i & \mu_i \end{bmatrix}. \tag{2.3.1}$$

They will be applied to vectors $\mathbf{v}$ in the order

$$P_{1,2,3}P_{2,3,4} \cdots P_{n-2,n-1,n}\mathbf{v}. \tag{2.3.2}$$

We note that the product $P_{1,2,3}P_{2,3,4} \cdots P_{n-2,n-1,n}$ has the form

$$\begin{bmatrix}
\alpha_1 & f_1^T g_1 & f_1^T g_2 & \cdots & f_1^T g_{n-3} & f_1^T g_{n-2} & f_1^T g_{n-1} \\
\beta_1 & f_2^T g_1 & f_2^T g_2 & \cdots & f_2^T g_{n-3} & f_2^T g_{n-2} & f_2^T g_{n-1} \\
\gamma_1 & f_3^T g_1 & f_3^T g_2 & \cdots & f_3^T g_{n-3} & f_3^T g_{n-2} & f_3^T g_{n-1} \\
 & \gamma_2 & f_4^T g_2 & \cdots & f_4^T g_{n-3} & f_4^T g_{n-2} & f_4^T g_{n-1} \\
 & & \gamma_3 & \cdots & \cdot & \cdot & \cdot \\
 & & & \cdots & \cdot & \cdot & \cdot \\
 & & & \cdot & \cdot & \cdot & \cdot \\
 & & & \gamma_{n-3} & f_{n-1}^T g_{n-3} & f_{n-1}^T g_{n-2} & f_{n-1}^T g_{n-1} \\
 & & & & \gamma_{n-2} & \eta_{n-2} & \mu_{n-2}
\end{bmatrix}, \tag{2.3.3}$$

where each f_j and g_j satisfy each of the following recursions:

Forward Recursion.

Set $a_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $b_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $f_1 = \begin{bmatrix} \beta_1 \\ \gamma_1 \end{bmatrix}$, $f_2 = \begin{bmatrix} \delta_1 \\ \eta_1 \end{bmatrix}$.

For $j = 1, 2, \cdots, n-3$:

Set $H_j = \begin{bmatrix} a_{j-1} & b_{j-1} \end{bmatrix}$,

$$f_{j+2} = H_j^{-T} \begin{bmatrix} \eta_j \\ \mu_j \end{bmatrix},$$

$$g_j = H_j \begin{bmatrix} \alpha_{j+1} \\ \beta_{j+1} \end{bmatrix},$$

$$a_j = H_j \begin{bmatrix} \beta_{j+1} \\ \delta_{j+1} \end{bmatrix},$$

$$b_j = H_j \begin{bmatrix} \gamma_{j+1} \\ \eta_{j+1} \end{bmatrix}.$$

Lastly, set $g_{n-2} = a_{n-3}$, $g_{n-1} = b_{n-3}$.

Backward Recursion.

Set $g_{n-1} = \begin{bmatrix} \gamma_{n-2} \\ \eta_{n-2} \end{bmatrix}$, $d_n = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $c_n = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $g_{n-2} = \begin{bmatrix} \beta_{n-2} \\ \delta_{n-2} \end{bmatrix}$.

For $j=n-1,n-2,\cdots,3$:

Set $$H_j = \begin{bmatrix} d_{j+1}^T \\ c_{j+1}^T \end{bmatrix} ,$$

$$f_j = H_j^T \begin{bmatrix} \eta_{j-2} \\ \mu_{j-2} \end{bmatrix} ,$$

$$c_j = H_j^T \begin{bmatrix} \delta_{j-2} \\ \eta_{j-2} \end{bmatrix} ,$$

$$d_j = H_j^T \begin{bmatrix} \beta_{j-2} \\ \gamma_{j-2} \end{bmatrix} ,$$

$$g_{j-2} = H_j^{-1} \begin{bmatrix} \alpha_{j-1} \\ \beta_{j-1} \end{bmatrix} .$$

Lastly, set $f_2 = c_3$, $f_1 = d_3$.

These recursions will break down whenever one of the H-matrices is singular. Consider the example of forward recursion:

$$H_j = H_{j-1} \begin{bmatrix} \beta_j & \gamma_j \\ \delta_j & \eta_j \end{bmatrix} . \tag{2.3.4}$$

If j is the first index for which H_j is singular, then $P_{j,j+1,j+2}$ must be such that

$$\begin{bmatrix} \beta_j & \gamma_j \\ \delta_j & \eta_j \end{bmatrix} \tag{2.3.5}$$

is singular. This will clearly be the case if $P_{j,j+1,j+2} = I$, that is $\beta_j = \gamma_j = n_j = 0$. This is the only case, for if $P_{j,j+1,j+2}$ has the form (2.3.1), then from (2.2.9) we may write (2.3.5) as

$$\frac{1}{(\pi_1^2+\pi_2^2+\pi_3^2)} \begin{bmatrix} -2\pi_1\pi_2 & -2\pi_1\pi_3 \\ \pi_1^2+\pi_3^2-\pi_2^2 & -2\pi_2\pi_3 \end{bmatrix} = \begin{bmatrix} -\frac{1}{\lambda}\pi_1\pi_2 & -\frac{1}{\lambda}\pi_1\pi_3 \\ 1-\frac{1}{\lambda}\pi_2^2 & -\frac{1}{\lambda}\pi_2\pi_3 \end{bmatrix} \tag{2.3.6}$$

for suitable π_1,π_2,π_3 and $\lambda = (\pi_1^2+\pi_2^2+\pi_3^2)/2$. If this is singular, then

$$(\frac{1}{\lambda^2})\pi_1\pi_2^2\pi_3+\frac{1}{\lambda}\pi_1\pi_3 - (\frac{1}{\lambda^2})\pi_1\pi_2^2\pi_3 = \frac{1}{\lambda}\pi_1\pi_3 = 0 , \tag{2.3.7}$$

which implies that $\pi_1 = \pi_3 = 0$. But from (2.2.7) we see that except when $P_{j,j+1,j+2} = I$, $|\pi_3|$ must equal $|\kappa| + \sqrt{\kappa^2+\pi_1^2+\pi_2^2}$, which has to be nonzero.

Thus, as would seem reasonable, the forward recursion will be interrupted at those values of j for which anomalously defined reflectors $P_{j,j+1,j+2}$ appear in the product. A similar result holds for backward recursion.

3. Generalizing method C1 of [5]. Method C1 of [5] was originally proposed by Bennet [1] and is not recommended for numerical reasons. Our generalization will suffer from the same problems when the original matrix is nearly singular. We include it here for completeness and because it shows how our algorithms may be easily adapted to vector and parallel machines. Our explanation mimics that of [5] and does not involve any of the concepts of § 2.

Let A be an $n \times n$ symmetric positive definite matrix, and assume the matrices M and D have been computed where M is unit lower triangular and D is diagonal such that

$$A = MDM^T. \tag{3.1}$$

Assume Z is an $n \times 2$ matrix, B is a 2×2 symmetric matrix, and we wish to find a unit lower triangular matrix $\overline{M}$ and a diagonal matrix $\overline{D}$ such that

$$\overline{M}\overline{D}\overline{M}^T = \overline{A} = A + ZBZ^T. \tag{3.2}$$

Let V be an $n \times 2$ matrix such that

$$MV = Z. \tag{3.3}$$

From (3.1) and (3.2), we then have

$$\overline{A} = M(D + VBV^T)M^T. \tag{3.4}$$

If one can find an $\hat{M}$ and a $\hat{D}$ having the same structure as M and D such that

$$\hat{M}\hat{D}\hat{M}^T = D + VBV^T, \tag{3.5}$$

then from (3.4) we may write

$$\overline{M} = M\hat{M} \quad \text{and} \quad \overline{D} = \hat{D}. \tag{3.6}$$

To determine $\hat{M}$ and $\hat{D}$, we equate elements of the jth column of (3.5) to obtain

$$\sum_{i=1}^{j-1} \hat{d}_i \hat{m}_{ji}^{\,2} + \hat{d}_j = d_j + \mathbf{v}_j^T B \mathbf{v}_j \tag{3.7}$$

and for $r = j+1, \cdots, n$

$$\sum_{i=1}^{j-1} \hat{d}_i \hat{m}_{ji} \hat{m}_{ri} + \hat{d}_j \hat{m}_{rj} = \mathbf{v}_r^T B \mathbf{v}_j, \tag{3.8}$$

where $\mathbf{v}_j^T$ denotes the jth row of the matrix V.

Using (3.7) and (3.8) we will show that there exists a $2 \times n$ matrix C, whose jth column will be denoted by $\mathbf{c}_j$ such that

$$\hat{m}_{rj} = \mathbf{v}_r^T \mathbf{c}_j \tag{3.9}$$

for $j = 1, 2, \cdots, n$. Assuming (3.9) for $r < j$, from (3.7) we have

$$\sum_{i=1}^{j-1} \hat{d}_i (\mathbf{v}_j^T \mathbf{c}_i)^2 + \hat{d}_j = d_j + \mathbf{v}_j^T B \mathbf{v}_j, \tag{3.10}$$

which gives us immediately a formula for $\hat{d}_j$. The same assumption and (3.8) suggest

$$\sum_{i=1}^{j-1} \hat{d}_i \mathbf{v}_j^T \mathbf{c}_i \mathbf{v}_r^T \mathbf{c}_i + \hat{d}_j \hat{m}_{rj} = \mathbf{v}_r^T B \mathbf{v}_j,$$

which implies

$$\hat{m}_{rj} = \mathbf{v}_r^T (B\mathbf{v}_j - \sum_{i=1}^{j-1} \hat{d}_i (\mathbf{v}_j^T \mathbf{c}_i) \mathbf{c}_i) / \hat{d}_j. \tag{3.11}$$

If (3.9) is true, (3.11) means that

$$\mathbf{c}_j = (B\mathbf{v}_j - \sum_{i=1}^{j-1} \hat{d}_i \mathbf{v}_j^T \mathbf{c}_i \mathbf{c}_i) / \hat{d}_j. \tag{3.12}$$

Since

$$\mathbf{v}_j^T \mathbf{c}_i \mathbf{c}_i = \mathbf{c}_i \mathbf{c}_i^T \mathbf{v}_j ,$$

equation (3.12) implies that

$$\mathbf{c}_j = (B - \sum_{i=1}^{j-1} \hat{d}_i \mathbf{c}_i \mathbf{c}_i^T) \mathbf{v}_j / \hat{d}_j . \tag{3.13}$$

This completely determines $\hat{M}$.

To determine $\hat{M}$ and $\hat{D}$ iteratively, we introduce

$$A_j = B - \sum_{i=1}^{j-1} \hat{d}_i \mathbf{c}_i \mathbf{c}_i^T . \tag{3.14}$$

We note that the A_j are 2×2 symmetric matrices. A_j, $\hat{d}_j$, and $\mathbf{c}_j$ can be determined using the following algorithm:

```
Set $A_1 \equiv B$.
For $j = 1, 2, \cdots, n$
    Set $\hat{d}_j = d_j + \mathbf{v}_j^T A_j \mathbf{v}_j$.
    Set $\mathbf{c}_j = (A_j \mathbf{v}_j) / \hat{d}_j$.
    Set $A_{j+1} = A_j - \hat{d}_j \mathbf{c}_j \mathbf{c}_j^T$.
```

Thus $\hat{M}$ looks like

$$\begin{bmatrix} 1 & & & & & \\ \mathbf{v}_2^T \mathbf{c}_1 & 1 & & & & \\ \mathbf{v}_3^T \mathbf{c}_1 & \mathbf{v}_3^T \mathbf{c}_2 & 1 & & & \\ . & . & \mathbf{v}_4^T \mathbf{c}_3 & 1 & & \\ . & . & . & . & 1 & \\ \mathbf{v}_n^T \mathbf{c}_1 & \mathbf{v}_n^T \mathbf{c}_2 & \mathbf{v}_n^T \mathbf{c}_3 & . & . & 1 \end{bmatrix} , \tag{3.15}$$

$$\hat{\mathbf{m}}_1 = V \mathbf{c}_1 - \mathbf{e}_1 \mathbf{v}_1^T \mathbf{c}_1 + \mathbf{e}_1 ,$$

and

$$M \hat{\mathbf{m}}_1 = (M \mathbf{V} - \mathbf{m}_1 \mathbf{v}_1^T) \mathbf{c}_1 + \mathbf{m}_1 .$$

In general,

$$M \hat{\mathbf{m}}_i = (M \mathbf{V} - \sum_{k=1}^{i} \mathbf{m}_k \mathbf{v}_k^T) \mathbf{c}_i + \mathbf{m}_i .$$

Notice that $MV = Z$ and define $W^{(i)} = Z - \sum_{k=1}^{i-1} \mathbf{m}_k \mathbf{v}_k^T$. Then $W^{(i)}$ is an n×2 matrix whose first i rows are zero, $W^{(i)} = W^{i-1} - \mathbf{m}_{i-1} \mathbf{v}_{i-1}^T$ and

$$M \hat{\mathbf{m}}_i = W^{(i+1)} \mathbf{c}_i + \mathbf{m}_i . \tag{3.16}$$

Recall that since M is unit lower triangular, $\mathbf{v}_1^T$ is just $\mathbf{z}_1^T = \mathbf{w}_1^{(1)T}$. Moreover, $\mathbf{v}_2^T$ is the second row of $Z - \mathbf{m}_1 \mathbf{v}_1^T$, which is the second row of $\mathbf{w}^{(2)}$. In general,

$$\mathbf{v}_i^T = \mathbf{w}_i^{(i)T} . \tag{3.17}$$

We are thus lead to the following algorithm:

Algorithm GC1

```
Set $A_1 \equiv B$, $W^{(1)} \equiv Z$.
For $j=1, \cdots, n$
    Let $\mathbf{v}_j^T = \mathbf{w}_j^{(j)T}$, i.e., the $j$th row of $V$.
    Set $\bar{d}_j = d_j + \mathbf{v}_j^T A_j \mathbf{v}_j$.
    Set $\mathbf{c}_j = (A_j \mathbf{v}_j)/\bar{d}_j$.
    Set $A_{j+1} = A_j - \bar{d}_j \mathbf{c}_j \mathbf{c}_j^T$.
    For $r = j+1, \cdots, n$
(3.18)      Set $\mathbf{w}_r^{(j+1)T} = \mathbf{w}_r^{(j)T} - m_{rj}\mathbf{v}_j^T$.
(3.19)      Set $\bar{m}_{rj} = m_{rj} + \mathbf{w}_r^{(j+1)T}\mathbf{c}_j$.
```

In (3.18) both columns of the matrix $W^{(j+1)}$ can be determined simultaneously. Equation (3.19) may be thought of as a rank 2 correction to the columns of M. Both (3.18) and (3.19) are just the formulae given in Algorithm B of § 1. Asymptotically, two applications of C1 in [5] and one of GC1 both require $2n^2 + O(n)$ multiplications, which means that on a scalar machine there should be no advantage in using GC1. In fact, for a 100×100 problem on the Sequent with no multiprocessing, the ratio of two applications of C1 to one of GC1 is about 1.06. However, when multiprocessing is enabled, the ratio climbs to 1.44. This is somewhat less than that reported for two applications of Algorithm C of § 1 over one of Algorithm D for the Sequent with multiprocessing, because more $O(n)$ work is involved. For the other computers considered in § 8, vector memory references are the dominant cost and the speedup is around 1.2.

4. Generalizing C2. Of the stable methods in [5], method C2 requires the least number of arithmetic operations. It is based on the fact that the orthogonal reduction of a rank 1 correction of the identity to a lower triangular matrix produces a matrix of special form that can be characterized by three vectors. Using Lemma 1 of § 2, we show that a rank 2 correction of a symmetric positive definite matrix produces a problem involving a rank 2 correction of the identity matrix. Reducing this rank 2 modified matrix to a lower triangular matrix also produces a special matrix, but in our situation we need one vector $\mathbf{g}$, and two n×2 matrices, V and J. The special lower triangular matrix $L(V,J,\mathbf{g})$ has the form

$$L(V,J,\mathbf{g}) = \begin{bmatrix} g_1 & & & & \\ \mathbf{v}_2^T\mathbf{j}_1 & g_2 & & & \\ \mathbf{v}_3^T\mathbf{j}_1 & \mathbf{v}_3^T\mathbf{j}_2 & g_3 \cdot & & \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \mathbf{v}_n^T\mathbf{j}_1 & \mathbf{v}_n^T\mathbf{j}_2 & \mathbf{v}_n^T\mathbf{j}_3 & \cdot\ \cdot & g_n \end{bmatrix}. \tag{4.1}$$

As in § 3, let A be an n×n symmetric positive definite matrix, and assume the matrices M and D have been computed, where M is unit lower triangular and D is diagonal such that

$$A = MDM^T. \tag{4.2}$$

Assume Z is an n×2 matrix, B is a 2×2 symmetric matrix, and we wish to find a unit lower triangular matrix $\bar{M}$ and a diagonal matrix $\bar{D}$ such that

$$\overline{M}\overline{D}\overline{M}^T = \overline{A} = A + ZBZ^T. \tag{4.3}$$

Let V be an $n\times 2$ matrix such that

$$MD^{1/2}V = Z. \tag{4.4}$$

Then from (4.3) we have

$$\overline{A} = MD^{1/2}(I + VBV^T)D^{1/2}M^T. \tag{4.5}$$

Now Lemma 1 of § 2 implies that there exists a 2×2 matrix C such that (4.5) can be written as

$$\overline{A} = MD^{1/2}(I + VCV^T)(I + VC^TV^T)D^{1/2}M^T. \tag{4.6}$$

We will show that there is a sequence of Householder transformations P such that $(I + VCV^T)P$ is a special lower triangular matrix $\hat{M} = L(V,J,\mathbf{g})$, as in (4.1). Now

$$D^{1/2}\hat{M} = L(D^{1/2}V,J,D^{1/2}\mathbf{g}) = L(H,R,\mathbf{e})GD^{1/2},$$

where

$$\begin{aligned} &G = diag(g_i) \\ &\mathbf{h}_s^T = d_s^{1/2}\mathbf{v}_s^T \qquad s = 1,2,\cdots,n \\ &\mathbf{r}_s^T = \mathbf{j}_s^T/(d_s^{1/2}g_s) \qquad s = 1,2,\cdots,n. \end{aligned} \tag{4.7}$$

Thus

$$\overline{M} = ML(H,R,\mathbf{e})$$

and

$$\overline{D} = GDG. \tag{4.8}$$

Because $L(H,R,\mathbf{e})$ has the same structure as (3.15), the same algorithm used to form $\overline{M}$ in § 3 can be used here.

The important point about this algorithm is that $\overline{D}$, computed using (4.8), will always be nonnegative.

The remainder of this section will be spent filling in the details and verifying that there exists a sequence of standard Householder transformations $P = P_1P_2 \cdots P_n$ such that

$$(I + VCV^T)P = L(V,J,\mathbf{g})$$

for some J and $\mathbf{g}$. Let us begin by partitioning V as

$$V^T = (\mathbf{y} : W^{(1)T}). \tag{4.9}$$

Note that $\mathbf{y}$ is a two-element vector. The (1,1) element of $I + VCV^T$ is

$$\theta = 1 + \mathbf{y}^TC\mathbf{y}.$$

Now P_1, which may be written as $I - \mathbf{u}\mathbf{u}^T/\tau$, must reduce the vector $(\theta : \mathbf{x}^T)$ to a multiple of $\mathbf{e}_1^T$. Here

$$\mathbf{x} = W^{(1)}C\mathbf{y} \equiv W^{(1)}\mathbf{f}. \tag{4.10}$$

If P_1 is a Householder transformation, then

$$P_1 = (I - \frac{1}{\tau}\mathbf{u}\mathbf{u}^T),$$

$$\mathbf{u}^T=(u_1:\mathbf{x}^T)$$

and

$$u_1=\theta+g_1,$$
$$g_1^2=\theta^2+\mathbf{x}^T\mathbf{x},$$
$$\tau=g_1u_1.$$

We claim that there exists a vector $\mathbf{j}^{(1)}$ of length 2 and a 2×2 matrix C_2 such that

$$(I+VCV^T)(I-\frac{1}{\tau}\mathbf{u}\mathbf{u}^T)=\begin{bmatrix} -g_1 & \\ W^{(1)}\mathbf{j}^{(1)} & I+W^{(1)}C_2W^{(1)T}\end{bmatrix}. \tag{4.11}$$

The proof of the claim is as follows: The last $n-1$ elements of the first column of the left-hand side of (4.11) are given by

$$(W^{(1)}C\mathbf{y}\ :\ I+W^{(1)}CW^{(1)T})\begin{bmatrix}1-\frac{1}{\tau}u_1^2\\ -\frac{1}{\tau}u_1\mathbf{x}\end{bmatrix}=W^{(1)}C\mathbf{y}(1-\frac{1}{\tau}u_1^2-\frac{1}{\tau}u_1)-\frac{u_1}{\tau}W^{(1)}CW^{(1)T}\mathbf{x}.$$

Thus $\mathbf{j}^{(1)}$ in (4.11) is given by

$$\mathbf{j}^{(1)}=(1-\frac{1}{\tau}u_1(1+u_1))C\mathbf{y}-\frac{u_1}{\tau}CW^{(1)T}W^{(1)}C\mathbf{y},$$

which simplifies to

$$\mathbf{j}^{(1)}=-\frac{1}{g_1}(2I+CV^TVC)\mathbf{y}. \tag{4.12}$$

To determine C_2, we note that the bottom $(n-1)\times(n-1)$ submatrix of the left-hand side of (4.10) is given by

$$(W^{(1)}C\mathbf{y}\ :\ I+W^{(1)}CW^{(1)T})\begin{bmatrix}-\frac{1}{\tau}u_1\mathbf{x}^T\\ I-1/\tau\mathbf{x}\mathbf{x}^T\end{bmatrix}=-\frac{1}{\tau}u_1W^{(1)}C\mathbf{y}\mathbf{x}^T+I+W^{(1)}CW^{(1)T}$$
$$-\frac{1}{\tau}\mathbf{x}\mathbf{x}^T-\frac{1}{\tau}W^{(1)}CW^{(1)T}\mathbf{x}\mathbf{x}^T. \tag{4.13}$$

Since $\mathbf{x}=W^{(1)}\mathbf{f}=W^{(1)}C\mathbf{y}$, the right-hand side of (4.13) must be

$$I+W^{(1)}(-\frac{1}{\tau}u_1\mathbf{f}\mathbf{f}^T+C-\frac{1}{\tau}\mathbf{f}\mathbf{f}^T-\frac{1}{\tau}CW^{(1)T}W^{(1)}\mathbf{f}\mathbf{f}^T)W^{(1)T}. \tag{4.14}$$

If one assigns the matrix within the parentheses of (4.14) to C_2, our claim is true. Since the $(n-1)x(n-1)$ submatrix has the same structure as the original matrix, one can proceed in the same fashion. Let us partition the matrix V as

$$V=\begin{bmatrix}\mathbf{v}_1^T\\ \mathbf{v}_2^T\\ \cdot\\ \cdot\\ \mathbf{v}_n^T\end{bmatrix}. \tag{4.15}$$

Let $W^{(i)}$ represent the last $n-i$ rows of V. Then there exists a matrix C_{i+1} such that

$$(I+VCV^{T)}P_1P_2\cdots P_i=\begin{bmatrix} g_1 & & & \\ & g_2 & & \\ & & \cdot & \\ W^{(1)}\mathbf{j}^{(1)} & W^{(2)}\mathbf{j}^{(2)} & \cdot & g_i \\ & & & W^{(i)}C_{i+1}W^{(i)T} \end{bmatrix}.$$

In Algorithm GC2, given below, we attempt to use the notation given above, with a few additions. Rather than computing $\mathbf{f}$ of (4.10), we compute $\hat{\mathbf{f}}=D^{1/2}\mathbf{f}$. Since $\mathbf{x}$ never appears by itself, but always as part of the expression $\mathbf{x}^T\mathbf{x}$, we compute $\bar{x}=\mathbf{x}^T\mathbf{x}$. Similarly, the W's never appear by themselves, and we let $\overline{W}_s=W_s^TW_s$. Moreover, although we have developed everything from (4.9) through (4.15) in terms of V and J, we really need formulae in terms of H and R of (4.7).

Algorithm GC2

```
Solve $MH=Z$. The matrix $H$ is $n$×2 and is the one given in (4.7).
Compute $C$ using Lemma 1 of § 2 and set $C_1=C$.
Set $W=Z$ and $\overline{W}_0=H^TD^{-1}H$.
For $s=1,\cdots,n$
    Set $\hat{\mathbf{f}}=C_s\begin{bmatrix} h_{s1} \\ h_{s2} \end{bmatrix}$.
    Set $\theta=1+(h_{s1}\hat{f}_1+h_{s2}\hat{f}_2)/d_s$.
    Set $\overline{W}_s=\overline{W}_{s-1}-\frac{1}{d_s}\begin{bmatrix} h_{s1} \\ h_{s2} \end{bmatrix}\begin{bmatrix} h_{s1} \\ h_{s2} \end{bmatrix}^T$.
    Set $\mathbf{y}=\overline{W}_s\hat{\mathbf{f}}$.

    Set $\mathbf{k}=\underline{C}\mathbf{y}$.
    Set $\bar{x}=\hat{\mathbf{f}}^T\mathbf{y}/d_s$.
    Set $g=(\theta^2+\bar{x})^{1/2}$.
    Set $\mathbf{r}_s=\frac{1}{g^2d_s}((1+\theta)\hat{\mathbf{f}}+\mathbf{k})$.
    Set $\mathbf{q}=\frac{-1.0}{gd_s(\theta+g)}((1+\theta+g)\hat{\mathbf{f}}+\mathbf{k})$.
    Set $\underline{C}_{s+1}=C_s+\mathbf{q}\hat{\mathbf{f}}^T$.
    Set $\bar{d}_s=g^2d_s$.
    For $t=s+1,\cdots,n$
        Set $\underline{\mathbf{w}}_t^{(s+1)T}=\mathbf{w}_t-l_{ts}\mathbf{h}_j^T$.
        Set $\bar{l}_{ts}=\bar{l}_{ts}+\mathbf{w}_t^{(s+1)T}\mathbf{r}_s$.
```

Algorithm GC2 requires $3n^2+O(n)$ multiplications and n square roots, which is a saving of n square roots over two applications of C2 of [5]. On the Sequent with no multiprocessing, the speedup over two applications of C2 on a 100×100 problem was 1.05. As in the case of GC1 and C1, there was a much greater speedup with multiprocessing. For the same problem set with multiprocessing enabled, the speedup was 1.39. The speedup in general was less than that found in C1-GC1 for most of the vectorizing machines because of the amount of $O(n)$ work in GC2. In general, that work is four times the amount required in C2 because of the number of matrix by vector multiplications with 2×2 matrices.

For a rank k correction, as opposed to a rank 2 correction, this $O(n)$ work would be increased by k^2. Thus the true operation count can be stated as $(3kn^2)/2+O(k^2 n)$ multiplications. Obviously, one would not want k to be too large.

5. Generalizing C3. In this section we generalize algorithm C3 of [5] to update the Cholesky factor of

$$\bar{A} = A + ZBZ^T \tag{5.1}$$

when the Cholesky factorization of $A = LL^T$ has already been determined, where L is lower triangular. We will let $\bar{L}\bar{L}^T$ denote the Cholesky factorization we wish to compute. We will also generalize a slightly modified version of C3 for negative updates, which is given in [8] and analyzed in [2], because it leads to a more economical version when B is negative definite and indefinite. We distinguish three distinct cases: (a) B is positive definite, (b) B is negative definite, and (c) B is indefinite. As before, we assume that $\bar{A}$ is positive definite.

Case (a): B is positive definite. If B of (5.1) is positive definite, then there exists an upper triangular matrix R such that

$$B = R^T R. \tag{5.2}$$

Let Y be the n×2 matrix defined by

$$Y = ZR^T \tag{5.3}$$

and let C be the matrix

$$C = (L:Y). \tag{5.4}$$

Thus C will have the form

$$\begin{bmatrix} x & & & & & x & x \\ x & x & & & & x & x \\ x & x & x & & & x & x \\ x & x & x & x & & x & x \\ x & x & x & x & x & x & x \end{bmatrix}.$$

Assume P is a sequence of standard Householder transformations $P = P_1 P_2 \cdots P_n$ that operates on C from the right and reduces it to the lower triangular matrix $\bar{L}$. If $C^{(0)} = C$ and $C^{(i)} = C^{(i-1)} P_i$. The transformation P_i is a transformation in planes i, $n+1$, and $n+2$ designed to annihilate $c_{i,n+1}^{(i-1)}$ and $c_{i,n+2}^{(i-1)}$. Then

$$\begin{aligned} \bar{L}\bar{L}^T &= CPP^T C^T = YY^T + LL^T \\ &= ZR^T RZ^T + LL^T \\ &= ZBZ^T + LL^T. \end{aligned}$$

Therefore $\bar{L}$ is the Cholesky factor of A. Thus we have the following:

Algorithm GC3 for positive definite B

```
Determine the Cholesky factorization $B=R^TR$, as in (5.2).
Set $Y=ZR^T$.
Form $C^{(0)}=(L:Y)$.
For $i=1, \cdots ,n$
    Find $P_i$ to annihilate $c_{i,n+1}^{(i-1)}$ and $c_{i,n+2}^{(i-1)}$.
    Set $C^{(i)}=C^{(i-1)}P_i$.
```

Asymptotically this algorithm requires $5n^2/2+O(n)$ multiplications and additions and n square roots. If two rank 1 corrections had been used with P composed of a sequence of Householder transformations requiring three multiplications per vector, then asymptotically $3n^2+O(n)$ multiplications + $2n$ square roots would have been required. Thus even on scalar machines there would have been an improvement. Indeed on the Sequent without multiprocessing for a 100×100 problem, we see a speedup of 1.30. When multiprocessing was enabled, the speedup jumped to a surprisingly large 1.67. Large speedups were also seen on other machines, as the table in § 8 indicates.

This algorithm can be easily generalized to the case where B is rank k and positive definite. The matrix C would have k nonzero superdiagonals. The innermost loop would consist of a rank k vector statement followed by $k+1$ rank 1 statements.

Case (b): B is negative definite. When B is negative definite we present two algorithms. The first generalizes C3 of [5] for negative updates. The second is faster. It is a generalization of row-removal method 3 of [8], which we will call RRM3. Algorithm RRM3 is also called version 3 in [2] and is analyzed there. It may be thought of as a modification of C3, and we derive our generalization of RRM3 from our generalization of C3.

Our generalization of C3 for negative updates is more complicated than our generalization of C3 for positive updates. In the first place, we will now define the $n\times 2$ matrix Y as the solution to

$$LY=Z, \tag{5.5}$$

where again L is the Cholesky factor of A. We then need to prove a lemma about the positive definiteness of a certain matrix, the Cholesky factorization of which is central to the algorithm for determining the Cholesky factorization of $\bar{A}$.

LEMMA. *If B is a symmetric, negative definite matrix, Y is defined in* (5.5), *and the matrices A and $\bar{A}$ defined in* (5.1) *are symmetric positive definite, then the matrix $N=-B^{-1}-Y^TY$ is positive definite.*

Proof. If B is negative definite, then $-B$ is symmetric positive definite and there exists a matrix R such that

$$-B=R^TR. \tag{5.6}$$

From (5.5) $Y=L^{-1}Z$, where L is the Cholesky factor of A. Let

$$J=L^{-1}ZR^T. \tag{5.7}$$

Note that

$$\begin{aligned}\bar{A}&=A+ZBZ^T\\&=LL^T+ZBZ^T\\&=L(I-L^{-1}ZR^TRZ^TL^{-T})L^T\\&=L(I-JJ^T)L^T\end{aligned}$$

from (5.7). Because $I-JJ^T$ is congruent to $\bar{A}$, $I-JJ^T$ has positive eigenvalues because $\bar{A}$ does. This means that all the eigenvalues of JJ^T are less than 1. Since the nonzero eigenvalues of JJ^T are also the nonzero eigenvalues of J^TJ, the eigenvalues of J^TJ are all less than 1, which means that $I-J^TJ$ is positive definite. Now

$$\begin{aligned} N &= -B^{-1}-Y^TY=(R^TR)^{-1}-(L^{-1}Z)^TL^{-1}Z \\ &= R^{-1}(I-J^TJ)R^{-T} \end{aligned}$$

from (5.7). Because $I-J^TJ$ is congruent to N and because it is positive definite, the matrix N must be positive definite and we have proved our lemma. □

Since N is positive definite, one can find an upper triangular matrix Q such that

$$Q^TQ=-B^{-1}-Y^TY. \tag{5.8}$$

Let P be a sequence of modified Householder transformations

$$P=P_{1,n+1,n+2}P_{2,n+1,n+2}\cdots P_{n,n+1,n+2}, \tag{5.9}$$

which when operated on the left of

$$\begin{bmatrix} Y \\ Q \end{bmatrix} \tag{5.10}$$

annihilates the first n rows of that matrix, i.e.,

$$P\begin{bmatrix} Y \\ Q \end{bmatrix}=\begin{bmatrix} O \\ S \end{bmatrix},$$

where S is a 2×2 matrix. Because P is an orthogonal matrix,

$$S^TS=Y^TY+Q^TQ=Y^TY+(-B^{-1}-Y^TY)=-B^{-1}.$$

Thus

$$-B=(S^TS)^{-1}. \tag{5.11}$$

The matrix P may also be constructed as a sequence of standard Householder transformations of the form

$$P=P_1^{n+2}P_1^{n+1}P_2^{n+2}P_2^{n+1}\cdots P_n^{n+2}P_n^{n+1}. \tag{5.12}$$

Because P is composed of matrices that touch only specific planes, it is easy to show that there exists an $n\times n$ lower triangular matrix $\bar{L}$ and a 2×2 matrix F such that

$$P\begin{bmatrix} L^T \\ 0 \end{bmatrix}=\begin{bmatrix} \bar{L}^T \\ F^T \end{bmatrix}. \tag{5.13}$$

Now

$$\begin{bmatrix} Y^T & Q^T \\ L & 0 \end{bmatrix}P^TP\begin{bmatrix} Y & L^T \\ Q & 0 \end{bmatrix}=\begin{bmatrix} 0 & S^T \\ \bar{L} & F \end{bmatrix}\begin{bmatrix} 0 & \bar{L}^T \\ S & F^T \end{bmatrix}. \tag{5.14}$$

Equating both sides of (5.14) we get

$$Z=LY=FS,$$

which implies that

$$F=ZS^{-1}. \tag{5.15}$$

Equation (5.14) also gives us that

$$LL^T = \overline{L}\overline{L}^T + FF^T,$$

which means that

$$\overline{L}\overline{L}^T = LL^T - FF^T = A - ZS^{-1}S^{-T}Z^T = A + ZBZ^T.$$

Thus $\overline{L}$ is the Cholesky factor of $\overline{A}$.

We are then lead to the following algorithm

Algorithm GC3N1 for negative definite B

```
Solve $LY = Z$ for $Y$.
Find $Q$ such that $Q^TQ = -B^{-1} - Y^TY$.
Set $L^{(n)} = L$, $F^{(n)} = 0$.
Set $U^{(n)} = \begin{bmatrix} Y \\ Q \end{bmatrix}$.
For $k = n, n-1, \cdots, 1$
    Find the modified Householder transformation $P_{k,n+1,n+2}$
        to annihilate the $k$th row of $U^{(k)}$ using
        row $n+1$ and $n+2$ of $U^{(k)}$.
    Set $U^{(k-1)} = P_{k,n+1,n+2}U^{(k)}$.
    Set
(5.16)    $\begin{bmatrix} L^{(k-1)T} \\ F^{(k-1)T} \end{bmatrix} = P_{k,n+1,n+2} \begin{bmatrix} L^{(k)T} \\ F^{(k)T} \end{bmatrix}$.
```

The speedup of GC3 over C3 depends on how one constructs P. Using the modified Householder transformation discussed in § 2, on the Sequent without multiprocessing on a 100×100 example the speedup was 1.31, while with multiprocessing the speedup was 1.52.

Lawson and Hanson[8] and Bojanczyk, Brent, Van Dooren, and de Hoog [2] present another version of C3 when the update is negative. We would like to mimic the modification in [2] of C3 for the rank 2 update case because it (1) leads to an algorithm that does not need the computation of Y in (5.5) or Q in (5.8) and (2) uses standard Householder transformations rather than modified transformations.

From (5.13) we see that

$$P^1_{n+1,n+2}P^2_{n+1,n+2} \cdots P^n_{n+1,n+2} \begin{bmatrix} L^T \\ 0 \end{bmatrix} = \begin{bmatrix} \overline{L}^T \\ F^T \end{bmatrix},$$

where from (5.11) and (5.15) $F = ZS^{-1}$ and $-B = S^{-1}S^{-T}$. In this version we will use the fact that we can compute F and L and try to determine some way of computing the P's as a sequence of 3 plane orthogonal transformations to get us $\overline{L}$. In Algorithm GC3N1 we note that $\overline{L} = L^{(0)}$ and $F = F^{(0)}$. Moreover the first k rows of $F^{(k)}$ are zero. Since we know F, we could almost perform the algorithm with the indices running forwards rather than backwards and choose $P^k_{n+1,n+2}$ as an orthogonal transformation in planes k, $n+1$, and $n+2$ to annihilate the kth row of $F^{(k-1)}$ to produce $F^{(k)}$, as in

$$\begin{bmatrix} L^{(k)T} \\ F^{(k)T} \end{bmatrix} = P^{k^T}_{n+1,n+2} \begin{bmatrix} L^{(k-1)T} \\ F^{(k-1)T} \end{bmatrix}. \tag{5.17}$$

Unfortunately we do not have the kth column of $L^{(k-1)}$, which is actually the kth column of $\bar{L}$. All we have is the kth column of L, which is the kth column of $L^{(k)}$. In other words, we wish to apply a 3-plane transformation in which we know the input in two planes and the output in the third plane and wish to determine the input in one plane and the output in the other two. Since all the transformations are linear, we may proceed as follows:

Let $P_{n+1,n+2}^{k}$ be a standard Householder transformation of the form

$$I-\beta\mathbf{u}\mathbf{u}^{T},$$

where $\beta=\mathbf{u}\mathbf{u}^{T}/2$ and as one would expect

$$u_2=f_{1k}^{(k-1)}, \tag{5.18}$$

$$u_3=f_{2k}^{(k-1)} \tag{5.19}$$

but

$$u_1=(l_{kk}{}^2-u_2{}^2-u_3{}^2)^{1/2}+\|l_{kk}\|. \tag{5.20}$$

Then from (5.16)

$$\mathbf{l}_k{}^T=\bar{\mathbf{l}}_k{}^T-\beta u_1(u_1\bar{\mathbf{l}}_k{}^T+u_2\mathbf{f}_1^{(k-1)T}+u_3\mathbf{f}_2^{(k-1)T}),$$

which means that

$$\bar{\mathbf{l}}_k{}^T=1/(1-\beta u_1^2)(\mathbf{l}_k{}^T+\beta u_1(u_2\mathbf{f}_1^{(k-1)T}+u_3\mathbf{f}_2^{(k-1)T})). \tag{5.21}$$

From (5.17) we can then compute $F^{(k)}$ as

$$F^{(k)T}=F^{(k-1)T}-\beta\begin{bmatrix}u_2\\u_3\end{bmatrix}(u_2\mathbf{f}_1^{(k-1)T}+u_3\mathbf{f}_2^{(k-1)T}+u_1\bar{\mathbf{l}}_k{}^T). \tag{5.22}$$

We are thus lead to the following algorithm:

Algorithm GC3N2 for negative definite B

```
Determine the Cholesky factorization of $-B=R^TR$, where $R$ is upper triangular.
Set $F^{(0)}=ZR^T$.
For $k=1,2,\cdots,n$
    Determine $\mathbf{u}$ as in (5.18), (5.19), and (5.20).
    Compute $\bar{\mathbf{l}}_k^T$ as in (5.21).
    Compute $F^{(k)}$ from (5.22).
```

Of course, one can modify (5.18) and (5.19) slightly, so that P has the form of (2.2.2), to eliminate one multiplication in the inner loop of GC3N2. Algorithms GC3N1 and GC3N2 with the P of (2.2.2) both require $(7n^2)/2+O(n)$ multiplications, but algorithm GC3N2 requires fewer vector memory references. In § 8 we compare the negative updating algorithms of C3 and of RRM3 and their generalizations. We looked at rather small problems on each machine together with one or more problems that either might be considered a good large size for that particular machine or is at a point where the ratios of two instantiations of a rank 1 algorithm to a rank 2 algorithm seemed to stabilize. The data corroborates the point in [2] that RRM3 is less costly than C3 for negative updates, although in theory both algorithms require the same number of operations asymptotically. Table 8.3 suggests that GC3N2 is the preferred algorithm. Because the code to generate the standard Householder transformations is simpler than that to compute the modified ones in GC3N1, Algorithm GC3N2 should be cheaper than GC3N1. Comparing the times on the Sequent without multiprocessing and with multiprocessing, we see that the rank 2 algorithms seem to be able to take much better advantage of multiprocessing than the rank 1 algorithms. In general, the ratios for vector and multiprocessing machines seem to favor the rank 2 algorithms more heavily than the theoretical multiplication count would suggest, which is one of our points.

Case(c): B is indefinite. When B is indefinite and rank 2, we have found no generalization based entirely on method C3 of [5]. However one can combine the ideas for a positive update of C3 with those for a negative update in [2] to obtain a rank 2 method.

Since B in (5.1) is symmetric, one can find an eigendecomposition of B as

$$B = QDQ^T,$$

where Q is an orthogonal matrix and D is diagonal. If B is indefinite, one may assume $d_{11} > 0$ and $d_{22} < 0$. If

$$Y = ZQ \begin{bmatrix} d_{11}^{1/2} & \\ & (-d_{22})^{1/2} \end{bmatrix}$$

and

$$G = A + \mathbf{y}_1 \mathbf{y}_1^T, \tag{5.23}$$

then from (5.1)

$$\bar{A} = G - \mathbf{y}_2 \mathbf{y}_2^T. \tag{5.24}$$

One could use C3 of [5] on (5.23) and proceed as follows:

```
Form the matrix
    $C^{(1)} = \begin{bmatrix} L^T \\ \mathbf{y}_1^T \end{bmatrix}$
For $i = 1, \cdots, n$
    Find a transformation $P_i$ in planes $i$ and $n+1$
        to annihilate $c^{(i)}_{n+1,i}$
    Set $C^{(i+1)} = P_i C^{(i)}$
```

One could then apply the algorithm in [2] on (5.24) and continue as follows

```
Set $\mathbf{y}_2^{(0)} = \mathbf{y}_2$
For $i = 1, \cdots, n$
    Find a transformation $Q_i$
    to annihilate $\tilde{c}^{(i)}_{n+2,i}$ and such that
    $\begin{bmatrix} c_{ii}^{(n)} \\ 0 \end{bmatrix} = Q_i \begin{bmatrix} \bar{l}_{ii} \\ y_{2i}^{(i-1)} \end{bmatrix}$
    Determine $\mathbf{y}_2^{(i)}$ and $\bar{\mathbf{l}}_i$ such that
    $\begin{bmatrix} \mathbf{c}_i^{(n)T} \\ \mathbf{y}_2^{(i)T} \end{bmatrix} = Q_i \begin{bmatrix} \bar{\mathbf{l}}_i^T \\ \mathbf{y}_2^{(i-1)T} \end{bmatrix}.$
```

Because Q_i and P_i are the only transformations affecting row i, one could apply Q_i as soon as P_i is complete. In fact, one could really combine them. Considering both the P's and Q's as Givens transformations of the form

$$\begin{bmatrix} c & -s \\ s & c \end{bmatrix},$$

one is lead to the following algorithm:

Algorithm GC3I for indefinite B

```
Determine the eigendecomposition of $B$ as in (5.23).
Form $Y$ as in (5.24).
For $i=1, \cdots ,n$
        Set $r_p=(l_{ii}^2+y_{i1}^{(i-1)\,2})^{1/2}$.
        Set $r_q=(r_p^2-y_{i2}^{(i-1)\,2})^{1/2}$.
        Set $c_p=l_{ii}/r_p$ and $s_p=-y_{i1}^{(i-1)}/r_p$.
        Set $c_q=r_q/r_p$ and $s_q=-y_{i2}^{(i-1)}/r_p$.
        Set $e=-s_p/c_q, f=c_p/c_q$ and $g=s_q/c_q$.
        Set $\mathbf{y}_1^{(i)\,T}=c_p\mathbf{y}_1^{(i-1)\,T}+s_p\mathbf{l}_i^T$.
        Set $\overline{\mathbf{l}}_i^T=e\mathbf{y}_1^{(i-1)\,T}+f\mathbf{l}_i^T+g\mathbf{y}_2^{(i-1)\,T}$.
        Set $\mathbf{y}_2^{(\mathbf{i})\,T}=c_q\mathbf{y}_2^{(\mathbf{i-1})\,T}+s_q\overline{\mathbf{l}}_i^T$.
```

Asymptotically, this algorithm requires $7n^2/2+O(n)$ operations. Since the operation count for GC3I and C3 for a positive update followed by C3 for a negative update, one is not surprised by the table in § 8 that shows that the speedup for indefinite B is less than that for positive definite B. Again, multiprocessing on the Sequent produces a tremendous difference in the speedup.

6. Generalizing C4. Algorithm C4 of [5] is the most expensive, but probably the least complicated theoretically. It is based on the fact that one can find a sequence of planar orthogonal transformations that reduces a vector to a multiple of $\mathbf{e}_1$ and, when it is applied on the right to the lower triangular Cholesky factor of a matrix A, leaves all the elements above the first superdiagonal zero. To find the Cholesky factor of the updated matrix, one then changes the first column of this lower Hessenberg matrix and finds a sequence of transformations that will reduce it back to lower triangular form.

In our generalization we work not with one n-vector but with an $n\times2$ matrix. Using the modified Householder transformations given in § 2.2, this matrix can be reduced to a 2×2 matrix. When these transformations are applied on the right to the lower triangular Cholesky factor of the original matrix, the result is a matrix with zeros above the "second" superdiagonal. To find the Cholesky of the updated matrix, one changes the first two columns of the transformed matrix and reduces it back to lower triangular form. The specifics of the algorithm are as follows: Let A be an $n\times n$ symmetric positive definite matrix and assume the lower triangular matrix L has been computed such that

$$A=LL^T. \tag{6.1}$$

Assume Z is an $n\times2$ matrix, B is a 2×2 symmetric matrix and we wish to find a lower triangular matrix $\overline{L}$ such that

$$\bar{L}\bar{L}^T = \bar{A} = A + ZBZ^T. \tag{6.2}$$

Let V be an $n \times 2$ matrix such that

$$LV = Z. \tag{6.3}$$

From (6.1) and (6.2), we then have

$$\bar{A} = LL^T + ZBZ^T = L(I + VBV^T)L^T. \tag{6.4}$$

Now construct a product Q of reflectors according to (2.3.2) such that

$$QV = E = \begin{bmatrix} K \\ 0 \end{bmatrix}, \tag{6.5}$$

where K is a 2×2 matrix. Then from (6.4) we get

$$\bar{A} = LQ^T Q(I + VBV^T)Q^T QL^T. \tag{6.6}$$

Now

$$Q(I + VBV^T)Q^T = I + EBE^T. \tag{6.7}$$

Since $\bar{A}$ is positive definite, by Sylvester's Law of Inertia, $I + VBV^T$ must be positive definite as well as the 2×2 matrix $I + KBK^T$. Hence there exists a lower triangular matrix $\hat{L}$ such that

$$\hat{L}\hat{L}^T = I + KBK^T. \tag{6.8}$$

Let J be the $n \times n$ matrix

$$J = \begin{bmatrix} \hat{L} & 0 \\ 0 & I \end{bmatrix}.$$

Then from (6.6), (6.7), and (6.8)

$$\bar{A} = LQ^T JJ^T QL^T. \tag{6.9}$$

Let

$$H = LQ^T.$$

Because Q is composed of modified Householders as in (2.3.2), H will be zero above its second superdiagonal. Because multiplying H by J on the right only affects the first two columns of H, the product

$$\bar{H} = HJ \tag{6.10}$$

will have zeros above the second superdiagonal, i.e., $\bar{H}$ will have the form

$$\begin{bmatrix} x & x & x & & & & \\ x & x & x & x & & & \\ x & x & x & x & x & & \\ x & x & x & x & x & x & \\ x & x & x & x & x & x & x \end{bmatrix}.$$

Assume P is a sequence of standard Householder transformations that operates on H from the right and reduces it to the lower triangular matrix $\bar{L}$. Then from (6.9)

$$\bar{A} = \bar{H}\bar{H}^T = \bar{H}PP^T\bar{H}^T = \bar{L}\bar{L}^T. \tag{6.11}$$

Thus we have the following algorithm:

Algorithm GC4

(1) Solve $LV=Z$ for V.

(2) Find Q, a product of generalized Householder transformations that reduce V to a 2×2 matrix, and apply these transformations on the right to L to form H.

(3) Determine $\hat{L}$ in (6.8) and form $\overline{H}$ as in (6.10).

(4) Reduce $\overline{H}$ to lower triangular form using a sequence of standard Householder transformations.

Asymptotically, Algorithm GC4 requires $n^2+O(n)$ multiplications for step (1), $5n^2/2+O(n)$ multiplications for step (2), $O(n)$ multiplications for step (3), and $5n^2/2+O(n)$ multiplications for step (4) for a total of $11n^2+O(n)$ multiplications. In algorithm C4 of [5], if one uses planar Householder transformations requiring three multiplications for each vector application, the total operation count for two instances of Algorithm C4 is $12n^2+O(n)$ multiplications. Thus one would not expect a huge speedup on a scalar machine, but for a 100×100 problem we did obtain a speedup of 1.29 on the Sequent without multiprocessing.

However, on a parallel machine, one should be able to do a bit better using GC4. First of all in step (1), one can solve the lower triangular system with two right-hand sides simultaneously. Then in steps (2) and (4) the basic step has the general form

```
For $j=1, \cdots, n$
    Compute numbers $a,b,c,e,f$.
    For $i=j, \cdots, n$
        Set $d=a\times x_{ij}+b\times x_{i,j+1}+c\times x_{i,j+2}$.
        Set $x_{i,j}=x_{i,j}+d$.
        Set $x_{i,j+1}=x_{i,j+1}+d\times e$.
        Set $x_{i,j+2}=x_{i,j+2}+d\times f$.
```

which gives much room for parallelization and vectorization. On the Sequent with multiprocessing, the speedup for a 100×100 problem was 1.53, a sizable increase from the ratio without multiprocessing. As given in § 8, many machines had ratios for two applications of C4 of [5] versus one application of GC4 in the range of 1.4.

7. Generalizing Algorithm C5.

In this section we develop an algorithm that takes advantage of the special structure of the Q and P matrices of § 6. Let A be an $n\times n$ symmetric positive definite matrix, and assume the lower triangular matrix L has been computed such that

$$A=LL^T. \tag{7.1}$$

Assume Z is an $n\times 2$ matrix, B is a 2×2 symmetric matrix, and we wish to find a lower triangular matrix $\overline{L}$ such that

$$\overline{L}\overline{L}^T=\overline{A}=A+ZBZ^T. \tag{7.2}$$

Let V be an $n\times 2$ matrix such that

$$LV=Z. \tag{7.3}$$

From (7.1) and (7.2), we have

$$\bar{A}=LL^T+ZBZ^T=L(I+VBV^T)L^T. \tag{7.4}$$

From Lemma 1 of § 2, we know there exists a symmetric matrix C such that

$$I+VBV^T=(I+VCV^T)(I+VC^TV^T) \tag{7.5}$$

and hence from (7.4)

$$\bar{A}=L(I+VCV^T)(I+VC^TV^T)L^T. \tag{7.6}$$

As in § 6, let Q be a product of modified Householder transformations, as in (2.3.2), such that

$$QV=E=\begin{bmatrix}K\\0\end{bmatrix}, \tag{7.7}$$

where K is a 2×2 matrix. Let

$$S=Q(I+VC^TV^T)=Q+\begin{bmatrix}K\\0\end{bmatrix}CV^T. \tag{7.8}$$

Then from (7.6)

$$\bar{A}=L(I+VCV^{T)}Q^TQ(I+VC^TV^T)L^T=LS^TSL^T. \tag{7.9}$$

Let us investigate the structure of S. Since we showed in § 2 that Q was zero below the second subdiagonal, S must be zero below the second subdiagonal. Moreover, its third through nth rows are exactly those rows of Q and possess the special structure of that matrix, i.e., the structure of (2.3.3).

From (7.7) we have

$$V=Q^T\begin{bmatrix}K\\0\end{bmatrix}, \tag{7.10}$$

which means that the first two rows of Q are linear combinations of the columns of V, at least when V has rank 2. But from (2.3.3) these rows look like

$$\begin{bmatrix}\mathbf{f}_1^TG^T\\\mathbf{f}_2^TG^T\end{bmatrix},$$

where G is an n×2 matrix and $\mathbf{f}_1$ and $\mathbf{f}_2$ are each two-vectors. Thus G and V span the same subspace, which means there is a 2×2 matrix M such that

$$G^T=MV^T. \tag{7.11}$$

In fact, because of the form of S, it must be another special matrix of the form $H(V,\bar{F},\bar{\gamma})$ with

$$\begin{aligned}\bar{\mathbf{f}}_i^T&=\mathbf{f}_i^TM \quad \text{for } i>2,\\ \bar{\mathbf{f}}_1^T&=\mathbf{f}_1^TM+\mathbf{k}_1^TC^T,\\ \bar{\mathbf{f}}_2^T&=\mathbf{f}_2^TM+\mathbf{k}_2^TC^T.\end{aligned} \tag{7.12}$$

Thus S has the following form:

$$(7.13)\qquad \begin{bmatrix}
\bar{\alpha}_1 & \bar{\mathbf{f}}_1^T\mathbf{v}_2 & \bar{\mathbf{f}}_1^T\mathbf{v}_3 & \cdots & \bar{\mathbf{f}}_1^T\mathbf{v}_{n-2} & \bar{\mathbf{f}}_1^T\mathbf{v}_{n-1} & \bar{\mathbf{f}}_1^T\mathbf{v}_n \\
\bar{\beta}_1 & \bar{\mathbf{f}}_2^T\mathbf{v}_2 & \bar{\mathbf{f}}_2^T\mathbf{v}_3 & \cdots & \bar{\mathbf{f}}_2^T\mathbf{v}_{n-2} & \bar{\mathbf{f}}_2^T\mathbf{v}_{n-1} & \bar{\mathbf{f}}_2^T\mathbf{v}_n \\
\gamma_1 & \bar{\mathbf{f}}_3^T\mathbf{v}_2 & \bar{\mathbf{f}}_3^T\mathbf{v}_3 & \cdots & \bar{\mathbf{f}}_3^T\mathbf{v}_{n-2} & \bar{\mathbf{f}}_3^T\mathbf{v}_{n-1} & \bar{\mathbf{f}}_3^T\mathbf{v}_n \\
 & \gamma_2 & \bar{\mathbf{f}}_4^T\mathbf{v}_3 & \cdots & \bar{\mathbf{f}}_4^T\mathbf{v}_{n-2} & \bar{\mathbf{f}}_4^T\mathbf{v}_{n-1} & \bar{\mathbf{f}}_4^T\mathbf{v}_n \\
 & & \gamma_3 & \cdots & \cdot & \cdot & \cdot \\
 & & & \cdots & \cdot & \cdot & \cdot \\
 & & & \cdot\;\cdot & \cdot & \cdot & \cdot \\
 & & & \gamma_{n-3} & \bar{\mathbf{f}}_{n-1}^T\mathbf{v}_{n-2} & \bar{\mathbf{f}}_{n-1}^T\mathbf{v}_{n-1} & \bar{\mathbf{f}}_{n-1}^T\mathbf{v}_n \\
 & & & & \gamma_{n-2} & \eta_{n-2} & \mu_{n-2}
\end{bmatrix}.$$

Now consider reducing S to triangular form using a sequence P of standard Householder transformations, each of which is designed to annihilate two elements of a three-vector. Thus

$$P=P_{n-1}P_{n-2}\cdots P_1$$

where P_i for $i\le n-2$ has the form

$$P_i=\begin{bmatrix} I_{i-1} & 0 & 0 \\ 0 & I-\mathbf{u}^{(i)}\mathbf{w}^{(i)T} & 0 \\ 0 & 0 & I_{n-i+2}\end{bmatrix}.$$

Now consider applying the middle submatrix of P_i to a vector

$$\left[\bar{\mathbf{f}}_i^T\mathbf{v}_j \quad \bar{\mathbf{f}}_{i+1}^T\mathbf{v}_j \quad \bar{\mathbf{f}}_{i+2}^T\mathbf{v}_j\right]^T.$$

We would get a vector of the form

$$\left[\hat{\mathbf{f}}_i^T\mathbf{v}_j \quad \hat{\mathbf{f}}_{i+1}^T\mathbf{v}_j \quad \hat{\mathbf{f}}_{i+2}^T\mathbf{v}_j\right]^T,$$

where now

$$(7.14)\qquad \begin{aligned}
\hat{\mathbf{f}}_i^T &=(1-u_1^{(i)}w_1^{(i)})\bar{\mathbf{f}}_i^T-u_1^{(i)}w_2^{(i)}\bar{\mathbf{f}}_{i+1}^T-u_1^{(i)}w_3^{(i)}\bar{\mathbf{f}}_{i+2}^T.\\
\hat{\mathbf{f}}_{i+1}^T &=u_2^{(i)}w_1^{(i)}\bar{\mathbf{f}}_i^T+(1-u_2^{(i)}w_2^{(i)})\bar{\mathbf{f}}_{i+1}^T-u_2^{(i)}w_3^{(i)}\bar{\mathbf{f}}_{i+2}^T.\\
\hat{\mathbf{f}}_{i+2}^T &=u_3^{(i)}w_1^{(i)}\bar{\mathbf{f}}_i^T-u_3^{(i)}w_2^{(i)}\bar{\mathbf{f}}_{i+1}^T+(1-u_3^{(i)}w_3^{(i)})\bar{\mathbf{f}}_{i+2}^T.
\end{aligned}$$

The matrix P_{n-1} is simply

$$\begin{bmatrix} I_{n-2} & 0 & 0 \\ 0 & c & s \\ 0 & s & -c\end{bmatrix}.$$

Thus if $PS=R$, an upper triangular matrix, R would look like

$$\begin{bmatrix} \hat{\gamma}_1 & \hat{\mathbf{f}}_1^T \mathbf{v}_2 & \hat{\mathbf{f}}_1^T \mathbf{v}_3 & \cdot\ \cdot & \cdot & \hat{\mathbf{f}}_1^T \mathbf{v}_{n-1} & \hat{\mathbf{f}}_1^T \mathbf{v}_n \\ & \hat{\gamma}_2 & \hat{\mathbf{f}}_2^T \mathbf{v}_3 & \cdot\ \cdot & \cdot & \hat{\mathbf{f}}_2^T \mathbf{v}_{n-1} & \hat{\mathbf{f}}_2^T \mathbf{v}_n \\ & & \hat{\gamma}_3 & \cdot\ \cdot & \cdot & \cdot & \cdot \\ & & & \cdot\ \cdot & \cdot & \cdot & \cdot \\ & & & \ \ \cdot & \cdot & \hat{\mathbf{f}}_{n-3} \mathbf{v}_{n-1} & \hat{\mathbf{f}}_{n-3}^T \mathbf{v}_n \\ & & & & \hat{\gamma}_{n-2} & \psi_1 & \psi_2 \\ & & & & & \hat{\gamma}_{n-1} & \psi_3 \\ & & & & & & \hat{\gamma}_n \end{bmatrix}, \tag{7.15}$$

which might be generated as follows:

(7.16)

```
Set $\delta_1 = \bar{\mathbf{f}}_1, \theta_1 = \bar{\mathbf{f}}_2$, and $\hat{\gamma}_1 = (\gamma_1^2 + \alpha_1^2 + \beta_1^2)^{1/2}$ .
For $i = 1, \cdots, n-3$
        Set $\phi = w_1^{(i)} \delta_i + w_2^{(i)} \theta_i + w_3^{(i)} \bar{\mathbf{f}}_{i+2}$.
        Set $\hat{\mathbf{f}}_i = \delta_i - u_1^{(i)} \phi$.
        Set $\delta_{i+1} = \theta_i - u_2^{(i)} \phi$.
        Set $\theta_{i+1} = \bar{\mathbf{f}}_{i+2} - u_3^{(i)} \phi$.
        Set $\hat{\gamma}_{i+1} = (\gamma_{i+1}^2 + \delta_{i+1}^T \mathbf{v}_{i+1}^2 + \theta_{i+1}^T \mathbf{v}_{i+1}^2)^{1/2}$.
Set $\phi = w_1^{(n-2)} \delta_{n-2} + w_2^{(n-2)} \theta_{n-2}$.
Set $\sigma_1 = \phi^T \mathbf{v}_{n-1} + w_3^{(n-2)} \eta_{n-2}$.
Set $\sigma_2 = \phi^T \mathbf{v}_n + w_3^{(n-2)} \mu_{n-2}$.
Set $\psi_1 = \delta_{n-2}^T \mathbf{v}_{n-1} - u_1^{(n-2)} \sigma_1$.
Set $\psi_2 = \delta_{n-2}^T \mathbf{v}_n - u_1^{(n-2)} \sigma_2$.
Set $\hat{\gamma}_{n-1} = ((\theta_{n-1}^T \mathbf{v}_{n-1} - u_2^{(n-2)} \sigma_1)^2 + (\eta_{n-2} - u_3^{(n-2)} \sigma_1)^2)^{1/2}$.
Set $\psi_3 = c(\theta_{n-1}^T \mathbf{v}_n - u_2^{(n-2)} \sigma_2) + s(\mu_{n-2} - u_3^{(n-2)} \sigma_2)$.
Set $\hat{\gamma}_n = s(\theta_{n-1}^T \mathbf{v}_n - u_2^{(n-2)} \sigma_2) - c(\mu_{n-2} - u_3^{(n-2)} \sigma_2)$.
```

From (7.9) we have

$$\bar{A} = LS^T SL^T = LR^T RL^T,$$

which means that we need to determine the matrix $\bar{L} = LR^T$. Because of the special form of R, we may use an algorithm similar to that used in GC1, with the modification that we are now working with general lower triangular matrices rather than unit lower triangular matrices. Thus our generalization of C5 would be as follows:

Algorithm GC5

```
(1) Set $W^{(1)}=Z$ and solve $LV=Z$ for $V$.
(2) Solve for $C$ of (7.5).
(3) Find the transformations that reduce $V$ to upper triangular form; this
determines $F$ and $G$ and $\gamma$ of (2.3.3).
(4) Find $M$ such that $VM^T=G$ and perform the transformations of
(7.12).
(5) For $i=1, \cdots ,n$
        (a) Find $\hat{\mathbf{f}}_i$ and $\hat{\gamma}_i$ using the algorithm of (7.16).
        (b) For $j=i+1, \cdots ,n$
                Set $\mathbf{w}_j^{(i+1)T}=\mathbf{w}_j^{(i)T}-l_{ji}\mathbf{v}_i^T$.
        (c) For $j=i+1, \cdots ,n$
                Set $\bar{l}_{ji}=l_{ji}\hat{\gamma}_i+\mathbf{w}_j^{(i+1)T}\hat{\mathbf{f}}_i$.
```

Asymptotically, steps (1), (5b), and (5c) require $7n^2/2+O(n)$ operations. All the other steps require $O(n)$ operations or fewer. This is $n^2/2$ operations fewer than two applications of C5 and many fewer than GC4. On the Sequent without multiprocessing, for a 100×100 problem the ratio of two applications of C5 of [5] to one of GC5 was 1.25. The improvement was considerably less with multiprocessing than in other methods because of the large number of scalar $O(n)$ operations in steps (3) and (5a). With multiprocessing the speedup was 1.34.

8. Summary. In this section we present our computational evidence. We implemented in FORTRAN the algorithms given in [5] for rank 1 corrections and the ones given in § § 3 – 7 and ran them on a variety of machines.

The VAX 750 had floating point acceleration. The Convex was run with the – O2 option specified, which turns on scalar optimization and vectorization. Although we were using a Cray XMP, we were limited to one processor. On the Alliant, both vectorization and concurrency were specified. The Sequent gives the user complete control of multiprocessing and only performs multiprocessing on command from the user. Our configuration had 30 processors that may be run in parallel.

Tables 8.1, 8.2, 8.3, and 8.4 give our results on several machines for each size averaged over 100 problems. We give raw times as well as speedup ratios so the reader can compare algorithms as well as speedups. For each machine we give results for $n=50$ and $n=100$ and for one or more larger values of n, which either might be considered "large" for that particular machine or were at a point that the speedup ratios changed very little. In Table 8.3 we give results for both versions of C3 and their generalizations for negative definite updates. The unparenthesized numbers are for C3 and GC3N1 and the parenthesized numbers are for the algorithm in [8] and [2] and its generalization GC3N2. Note that the rank 1 algorithms in [5] and those presented in the earlier sections of this paper all vectorize and that all the algorithms are asymptotically $O(n^2)$. The only nonvector operations are of $O(n)$. Sometimes on vector machines, these scalar operations dominate the cost of the algorithm.

Our data indicates that on each machine the appropriate version of GC3 is the method of choice for rank 2 updates. In all instances GC3N2 is preferable to GC3N1.

TABLE 8.1
Experience on several machines on C1-GC1, C2-GC2.

	n	two rank-1	rank 2	ratio
C1-GC1 multiplications		$2n^2+O(n)$	$2n^2+O(n)$	1
Vax 750	50	.149	.116	1.28
	100	.573	.452	1.27
	200	2.25	1.77	1.28
Convex	50	.00315	.00246	1.28
	100	.00776	.00615	1.26
	400	.0780	.0608	1.28
Cray XMP	50	.000265	.000195	1.36
	100	.000673	.000524	1.28
	400	.00603	.00517	1.17
	800	.0211	.0185	1.14
Alliant	50	.00403	.00288	1.40
	800	.348	.243	1.43
	1600	1.30	.905	1.43
	100	.00931	.00665	1.40
Sequent nomp	50	.125	.121	1.03
	100	.483	.456	1.06
	200	1.84	1.72	1.07
Sequent mp	50	.106	.0712	1.49
	100	.238	.164	1.44
	200	.604	.450	1.34
C2-GC2 multiplications		$3n^2+O(n)$	$3n^2+O(n)$	1
Vax 750	50	.259	.209	1.24
	100	.928	.716	1.30
	200	3.59	2.78	1.29
Convex	50	.00564	.00518	1.09
	100	.0128	.0119	1.08
	400	.102	.0955	1.07
Cray	50	.000735	.000516	1.42
	100	.00167	.00123	1.36
	400	.0117	.00957	1.22
	800	.0367	.0316	1.16
Alliant	50	.00789	.00682	1.16
	100	.0174	.0152	1.15
	800	.549	.440	1.25
	1600	1.97	1.62	1.22
Sequent nomp	50	.223	.220	1.02
	100	.798	.763	1.05
	200	3.03	2.85	1.07
Sequent mp	50	.220	.154	1.42
	100	.481	.347	1.39
	200	1.16	.861	1.35

TABLE 8.2
Experience on several machines on C3-GC3.

	n	two rank-1	rank 2	ratio
C3-GC3	positive definite			
multiplications		$3n^2+O(n)$	$5n^2/2+O(n)$	1.2
Vax 750	50	.213	.137	1.56
	100	.751	.472	1.59
	200	2.77	1.79	1.55
Convex	50	.0047	.0032	1.45
	100	.0106	.00738	1.43
	400	.0795	.0581	1.37
Cray XMP	50	.000564	.000337	1.67
	100	.00127	.000779	1.63
	400	.0105	.0728	1.44
	800	.0335	.0245	1.37
Alliant	50	.00661	.00434	1.52
	100	.0154	.0104	1.47
	800	.501	.356	1.41
	1600	1.84	1.32	1.39
Sequent nomp	50	.215	.163	1.32
	100	.789	.598	1.30
	200	2.94	2.27	1.30
Sequent mp	50	.139	.082	1.70
	100	.326	.196	1.66
	200	.834	.520	1.60
C3-GC3	indefinite			
multiplications		$7n^2/2+O(n)$	$7n^2/2+O(n)$	1
Vax 750	50	.206	.158	1.31
	100	.722	.554	1.30
	200	2.71	2.04	1.33
Convex	50	.00415	.00379	1.09
	100	.00932	.00853	1.09
	40	.00717	.0652	1.10
Cray XMP	50	.000543	.000406	1.34
	100	.00130	.00102	1.28
	400	.0102	.00899	1.14
	800	.0339	.0311	1.09
Alliant	50	.00568	.00408	1.39
	100	.0134	.00985	1.32
	800	.459	.344	1.33
	1600	1.70	1.28	1.33
Sequent nomp	50	.200	.175	1.14
	100	.738	.649	1.14
	200	2.83	2.49	1.13
Sequent mp	50	.152	.100	1.52
	100	.352	.241	1.46
	200	.858	.610	1.41

TABLE 8.3
Experience on several machines on C3-GC3N1, [2]-GC3N2 *for negative definite B matrices.*

	n	2C3([2])	GC3N1(GC3N2)	ratios
multiplications		$4n^2+O(n)$	$7n^2/2+O(n)$	1.2
Vax	50	.292(.197)	.217(.147)	1.35(1.34)
	100	1.06(.703)	.770(.527)	1.37(1.34)
	200	4.00(2.64)	2.90(2.02)	1.38(1.31)
Convex	50	.00591(.00372)	.00430(.00303)	1.38(1.22)
	100	.0133(.00824)	.00993(.00688)	1.34(1.20)
	400	.107(.0644)	.0833(.0638)	1.28(1.20)
Cray	50	.000831(.000488)	.000586(.000324)	1.42(1.50)
	100	.00195(.00118)	.00148(.000835)	1.32(1.42)
	400	.0139(.0101)	.0119(.00797)	1.16(1.27)
	800	.0447(.0338)	.0409(.0267)	1.09(1.27)
Alliant	50	.00912(.00472)	.00650(.004000)	1.40(1.19)
	100	.0209(.0111)	.0153(.00952)	1.36(1.16)
	800	.644(.428)	.479(.333)	1.34(1.29)
	1600	2.43(1.58)	1.75(1.24)	1.40(1.28)
Sequent nomp	50	.363(.186)	.283(.166)	1.29(1.12)
	100	1.36(.687)	1.03(.623)	1.31(1.10)
	200	5.22(2.62)	3.92(2.40)	1.33(1.09)
Sequent mp	50	.243(.140)	.158(.0867)	1.54(1.62)
	100	.543(.322)	.357(.205)	1.52(1.57)
	200	1.32(.805)	.886(.534)	1.48(1.51)

One of the major thrusts of this project was the hope that for a rank 2 (or more) update one could take advantage of a multiprocessing environment. Our data indicates that this is possible to an extent. We obtained rather respectable speedups in many cases. Our success was limited by the fact that many of the generalizations had large $0(k^2n)$ scalar operations, for a rank k update. Thus for a rank 2 update the number of scalar operations was at least four times the number for the rank 1 case. This was particularly true of GC2 and GC5.

This project has several aspects. First there was the mathematical aspect. In § 2 we proved several theorems that were the foundation of the generalizations of the two fastest algorithms of the stable algorithms given in [5]. Throughout the paper we needed to prove propositions about the positive definiteness of a variety of matrices. To make three of the generalizations cost effective, we developed the concept of a modified Householder that would eliminate elements in two vectors at once using a rank 1 correction of the identity. There was the algorithmic aspect of generating the algorithms. Lastly there was the computational aspect of trying to take advantage of new computing environments, which ideally should lend themselves to such a project.

TABLE 8.4
Experience on several machines for C4-GC4, C5-GC5.

	n	two rank-1	rank 2	ratio
C4-GC4 multiplications		$12\,n^2+O(n)$	$11n^2+O(n)$	1.09
Vax 750	50	.657	.500	1.31
	100	2.43	1.84	1.31
	200	9.13	7.00	1.30
Convex	50	.0112	.00761	1.47
	100	.0263	.0181	1.46
	400	.216	.152	1.42
Cray XMP	50	.00132	.000828	1.60
	100	.00316	.00209	1.51
	400	.0243	.0194	1.25
	800	.0779	.0661	1.18
Alliant	50	.0156	.0112	1.39
	100	.0344	.0256	1.35
	800	.997	.751	1.33
	1600	3.71	2.79	1.33
Sequent nomp	50	.766	.600	1.28
	100	2.91	2.26	1.29
	200	11.3	8.78	1.28
Sequent mp	50	.379	.239	1.58
	100	.891	.583	1.53
	200	2.29	1.57	1.46
C5-GC5 multiplications		$4n^2+O(n)$	$\frac{7}{2}n^2+O(n)$	1.14
Vax 750	50	.330	.250	1.35
	100	1.10	.818	1.35
	200	4.00	2.95	1.36
Convex	50	.00800	.00693	1.15
	100	.0175	.0155	1.13
	400	.126	.111	1.13
Cray XMP	50	.00103	.000776	1.33
	100	.00232	.00179	1.30
	400	.0149	.0123	1.21
	800	.0456	.380	1.20
Alliant	50	.0105	.00876	1.20
	100	.0219	.0187	1.17
	800	.600	.4453	1.35
	1600	2.16	1.56	1.38
Sequent nomp	50	.274	.239	1.15
	100	.978	.781	1.25
	200	3.69	2.80	1.32
Sequent mp	50	.248	.184	1.34
	100	.555	.413	1.34
	200	1.34	.986	1.36

REFERENCES

[1] J. M. Bennet, *Triangular factors of modified matrices*, Numer. Math., 7 (1965). pp. 217-221.

[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., 9 (1987), pp. 210-222.

[3] A.W. Bojanczyk, R. P. Brent, and F.R. de Hoog, QR *factorization of Toeplitz matrices*, Numer. Math., 49 (1986), pp. 81-94.

[4] O.E. Bonlund and Th. L. Johnson, *A* QR-*factorization of partitioned matrices*, Comput. Methods Appl. Mech. Engrg., 3 (1974), pp. 153-172.

[5] P.E. Gill, G.H. Golub, W. Murray, and M. A. Saunders, *Methods for modifying matrix factorizations*, Math. Comp., 28 (1975), pp. 505-535.

[6] D. Goldfarb, *Factorized variable metric methods for unconstrained optimization*, Math. Comp., 30 (1976), pp. 796-811.

[7] L. Kaufman, *The generalized Householder transformation and sparse matrices*, Linear Algebra Appl. 90 (1987), pp. 221-234.

[8] C.L. Lawson and R.J. Hanson, *Solving Least Squares Problems*, Prentice Hall, Englewood-Cliffs, NJ, 1974.

[9] C.B. Moler and G. W. Stewart, *An algorithm for the generalized matrix eigenvalue problem*, SIAM J. Numer. Anal., 10 (1973), pp. 241-256.

[10] R. Schreiber and B. Parlett, *Block reflectors: Theory and computation*, SIAM J. Numer. Anal., 25 (1988), pp. 189-205.

SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 4, p. 593, October 1989

013

ERRATUM: Block Kronecker Products and Block Norm Matrices in Large-Scale Systems Analysis*

DAVID C. HYLAND† AND EMMANUEL G. COLLINS, JR.†

In the above paper[1] the last part to the proof of property (A.1) should read as follows:

Substituting (2.11) and (2.12) into (2.10) shows that vecb (ADB) may be expressed as an r-partitioned vector where the qth-partition has dimension $n \times n_q$ and is given by

$$(2.13) \qquad [\operatorname{vecb}(ADB)]_q = \sum_i \begin{bmatrix} \sum_j (B_{iq}^T \otimes A_{1j}) \operatorname{vec}(D_{ji}) \\ \sum_j (B_{iq}^T \otimes A_{2j}) \operatorname{vec}(D_{ji}) \\ \vdots \\ \sum_j (B_{iq}^T \otimes A_{rj}) \operatorname{vec}(D_{ji}) \end{bmatrix}.$$

When we use the definition of $B_{iq}^T \circledast A$ (see (2.5)), it follows that (2.13) is equivalent to

$$(2.14) \qquad [\operatorname{vecb}(ADB)]_q = \sum_i (B_{iq}^T \circledast A) \begin{bmatrix} \operatorname{vec}(D_{1i}) \\ \operatorname{vec}(D_{2i}) \\ \vdots \\ \operatorname{vec}(D_{ri}) \end{bmatrix}.$$

Property (A.1) follows from (2.14). □

* Received by the editors April 24, 1989; accepted for publication May 18, 1989.

† Harris Corporation, Government Aerospace Systems Division, Melbourne, Florida 32902 (bernstein %x102a @ trantor. harris-atd. com).

[1] SIAM J. Matrix Anal. Appl., 10 (1989), pp. 18–29.