JOURNAL OF COMPUTATIONAL PHYSICS 28, 82–91 (1978)

On the Householder–Fox Algorithm for Decomposing a Projection

CLEVE B. MOLER*

University of New Mexico, Albuquerque, New Mexico 87131

AND

G. W. STEWART†

University of Maryland, College Park, Maryland 20742

Received October 20, 1976; revised September 20, 1977

The Householder–Fox algorithm uses the Cholesky decomposition to calculate an orthonormal basis for the range of a projection. In this paper it is shown that the algorithm continues to give good results when it is applied to an approximate projection in the presence of rounding error.

1. INTRODUCTION

A real orthogonal projection is a real matrix A satisfying the following two conditions:

$$\begin{aligned} &1.\quad A^T = A \qquad \text{(symmetry)},\\ &2.\quad A^2 = A \qquad \text{(idempotence)}. \end{aligned} \tag{1.1}$$

Applied to a vector x such a matrix produces the orthogonal projection Ax of x onto the column space of A (denoted by $\mathscr{R}(A)$); that is, $x_1 = Ax$ and $x_2 = (I - A)x$ are the unique vectors satisfying

$$\begin{aligned} &1.\quad x = x_1 + x_2\,,\\ &2.\quad x_1 \in \mathscr{R}(A),\\ &3.\quad x_1 \perp x_2\,. \end{aligned} \tag{1.2}$$

Conditions (1.2) are easily seen to follow from (1.1).

* This work was done while the author was a visiting staff member at C Division of Los Alamos Scientific Laboratory.

† This work was supported in part by the Office of Naval Research under Contract No. N00014-76-C-0391.

0021-9991/78/0281–0082$02.00/0

In some applications one is given a projection A and wishes to find an orthonormal basis for the subspace $\mathscr{R}(A)$. For example, if A is known to be of low rank, say $\text{rank}(A) = r$, then A can be represented economically in the form

$$A = QQ^T$$

where the r columns of Q form an orthonormal basis for $\mathscr{R}(A)$. The savings in storage can be substantial if the order n of A is very much greater than r; for A requires $n^2/2$ locations for its storage while Q requires only nr. Projections of low rank arise in the study of the spectra of molecules with high degress of symmetry (cf. the work of Fox and Krohn [3]).

One method for computing Q is to apply various orghogonalizing techniques to the columns of A. For example, one might use Householder transformations with column pivoting to compute a QR factorization of A [5, 7]. However, these techniques do not preserve the symmetry of A. Moreover, there is considerable evidence that when A is sparse, orthogonalization methods can lead to excessive fill-in [2].

A method which is symmetry preserving is to calculate the eigensystem of A [6]. The eigenvalues of A must be either zero or unity, and the eigenvectors corresponding to the eigenvalue unity form a basis for $\mathscr{R}(A)$. However, the method suffers from fill-in problems, and does not directly use the idempotency of A.

Householder and [4] have observed that the Cholesky factorization of a projection gives the required basis for $\mathscr{R}(A)$ directly. The form of the Cholesky decomposition used here is stated in the following theorem, whose proof is usually a constructive technique for calculating it (see section 3).

THEOREM 1.1. *Let A be a positive semidefinite matrix of order n and rank r. Then there is a permutation matrix P and an $n \times r$ lower trapezoidal matrix of rank r such that*

$$P^TAP = LL^T.$$

The process by which the rows and columns of A are rearranged, i.e., the manner in which P is chosen, is called pivoting. For the present we shall assume that the pivoting has been done initially and suppress mention of the matrix P. We shall return to the role of pivoting in the final section of this paper.

The importance of the Cholesky decomposition for our purposes is contained in the following corollary,

COROLLARY 1.2. *Suppose, in addition to the hypotheses of Theorem* 1.1, *that $A^2 = A$. Then*

$$L^TL = I.$$

Proof. From the relation $A = LL^T$, it follows that

$$LL^TLL^T = A^2 = A = LL^T. \tag{1.3}$$

Since the colums of L are inpependent, L has a pseudoinverse $L^\dagger = (L^TL)^{-1}L^T$ satisfying $L^\dagger L = I$. Then from (1.3)

$$L^TL = L^\dagger(LL^TLL^T)\,L^{\dagger T} = L^\dagger(LL^T)\,L^{\dagger T} = I. \quad \blacksquare$$

The import of the corollary is that the columns of L are orthonormal. They of course span $\mathscr{R}(A)$; hence the columns of L form the required basis. However, in practice the algorithm must be used in the presence of errors of various sorts, and it is the purpose of this paper to show that one can still expect to obtain good results.

2. Assessment of the Final Results

There are two sources of error in the use of the Householder–Fox algorithm. First the matrix A may not be exactly idempotent (in most applications the symmetry of A is forced by other considerations). We summarize this state of affairs by writing

$$A^2 = A + F, \tag{2.1}$$

where the symmetric matrix F is presumed small.

The second source of error is the rounding error made in the course of the Cholesky reduction of A. The effects of rounding error will be investigated in more detail in Section 3. For the present we will make the reasonable assumption that the computed matrix L satisfies a stability requirement of the form

$$LL^T = A + E,$$

where E is a small matrix of order rounding error (cf. Theorem 3.2 below). Assuming (2.1) and (2.2), we shall in this section give answers to the following two questions:

1. How near are the columns of L to orthonormality?
2. What is $\mathscr{R}(L)$?

We shall answer these questions in terms of norms. Specifically we shall use the Euclidean vector norm defined by

$$\| x \| = (x^Tx)^{1/2}$$

and the spectral matrix norm defined by

$$\| A \| = \sup_{\|x\|=1} \| Ax \|.$$

When A is symmetric, its spectral norm is the maximum of the absolute values of the eigenvalues of A. Also for any matrix X, $\| X \|^2 = \| X^TX \|$.

We begin our development by locating the eignenvalues of the matrix A which for the rest of this paper is assumed to be symmetric. The eigenvalues of a projection

can be only zero and unity, and Theorem 2.1 generalizes this fact by showing that an approximate projection in the sense of (2.1) must have eigenvalues clustering about zero and unity.

THEOREM 2.1. *Let A satisfy* (2.1). *Then the eigenvalues of A lie in one of the two intervals*

$$\left[\frac{1-(1+4\|F\|)^{1/2}}{2}, \frac{1-(1-4\|F\|)^{1/2}}{2}\right] \tag{2.3}$$

and

$$\left[\frac{1+(1-4\|F\|)^{1/2}}{2}, \frac{1+(1+4\|F\|)^{1/2}}{2}\right]. \tag{2.4}$$

In particular

$$\|A\| \leqslant 1 + \|F\|. \tag{2.5}$$

Proof. The eigenvalues of $A^2 - A$ are $\lambda^2 - \lambda$, where λ is an eigenvalue of A. Since $A^2 - A = F$, the eigenvalues of A must satisfy

$$\lambda^2 - \lambda \in [\|F\|, \|F\|],$$

which is equivalent to saying that λ lies in one of the two intervals (2.3) or (2.4). The largest eigenvalue of A cannot be larger than the right-hand end of the interval (2.4), which is bounded by $1 + \|F\|$. This establishes (2.5). ∎

Asymptotically for small F the intervals (2.3) and (2.4) reduce to $[-\|F\|, \|F\|]$ and $[1 - \|F\|, 1 + \|F\|]$.

If A is a projection, then so is $I - A$. If A is an approximate projection in the sense of (2.1), then

$$(I - A)^2 = I - 2A + A^2 = (I - A) + F.$$

Hence $(I - A)$ is an approximate projection, and from Theorem 2.1 we have the following bound:

$$\|I - A\| \leqslant 1 + \|F\|.$$

We are now in a position to answer the first of our questions.

THEOREM 2.2. *Let the matrix A satisfy* (2.1) *and let L satisfy* (2.2). *Suppose that L is of full column rank and satisfies*

$$\epsilon \|L^\dagger\|^2 < \tfrac{1}{2}, \tag{2.6}$$

where

$$\epsilon = \|F\| + \|E\|(2 + 2\|F\| + \|E\|).$$

Then

$$\|L^\dagger\|^2 < (1 - 2\epsilon)^{-1} \tag{2.7}$$

and

$$\|L^T L - I\| < \frac{\epsilon}{1 - 2\epsilon}. \tag{2.8}$$

Proof. From (2.1) and (2.2) it follows that

$$LL^TLL^T = (A + E)^2 = A + E + F - E + EA + AE + E^2.$$

Hence

$$L^TL = I + L^\dagger[F + E(A - I) + AE + E^2]\,L^{T\dagger}. \tag{2.9}$$

It follows from (2.9) and (2.6) that

$$\| L^TL - I \| < \tfrac{1}{2}.$$

In particular no eigenvalue of L^TL can be less than or equal to 1/2, from which it follows that $\| L^\dagger \|^2 < 2$. Again from (2.9)

$$\| L^TL - I \| < 2\epsilon,$$

and from this the bound (2.7) follows. Finally, applying (2.7) to (2.9) gives (2.8) ∎

Condition (2.6) is a requirement that the columns of L be independent. It is not very strong, and if it is satisfied it implies that the columns of L are almost orthonormal in the sense of the inequality (2.8), whose right-hand side is essentially $\| F \| + 2 \| E \|$. Indeed the theorem may be interpreted as saying that the columns of an LL^T decomposition of A cannot be slightly independent without being completely so.

Our second question amounts to asking if, having obtained L, we have obtained something useful. This of course will depend on what we originally desired to compute; however, in most applications we are seeking a basis for the column space of an exact projection which we believe to be near A. Now any matrix satisfying (2.1) divides n-space naturally into two complementary subspaces. They are the subspace $\mathscr{A}_1$ spanned by the eigenvectors associated with the eigenvalues clustered about unity and the subspace $\mathscr{A}_0$ spanned by the eigenvectors associated with the eigenvalues clustered about zero. These subspaces are orthogonal complements, and because the eigenvalues associated with the two subspaces are well separated, they are insensitive to small perturbations of A (see [1] for further details). It follows that $\mathscr{A}_1$ must be a good approximation to the column space of any projection near A.

We should like to show that $\mathscr{R}(L)$ is a good approximation to $\mathscr{A}_1$. We shall do this indirectly by showing that the columns of L are almost orthogonal to $\mathscr{A}_0$. Since the columns of L are almost orthonormal $\mathscr{R}(L)$ must be almost orthogonal to $\mathscr{A}_0$ and cannot help being a good approximation to $\mathscr{A}_1$.

THEOREM 2.3. *Under the hypotheses of Theorem* 2.2, *if for any vector* x *with* $\| x \| = 1$

$$\| Ax \| = \delta,$$

then

$$\| L^Tx \| \leqslant \frac{\delta + \| E \|}{(1 - 2\epsilon)^{1/2}}.$$

Proof. From (2.2) we have

$$LL^Tx = (A + E)\,x = Ax + Ex.$$

Hence

$$L^Tx = L^+(Ax + Ex)$$

and

$$\| L^Tx \| \leqslant \| L^+ \| \,(\| Ax \| + \| E \| \,\| x \|) \leqslant \frac{\delta + \| E \|}{(1 - 2\epsilon)^{1/2}}\,. \quad \blacksquare$$

It should be pointed out that, having obtained L, one can approximate the projection onto $\mathscr{A}_0$ by $I - LL^T$. If the dimension of $\mathscr{A}_0$ is very much less than that of $\mathscr{A}_1$, it will pay to decompose $I - A$ to obtain an L spanning $\mathscr{A}_0$ that has fewer columns. The projection for $\mathscr{A}_1$ can then be represented as $I - LL^T$ (however, some care must be taken to insure the orthogonality of the computed projections $(LL^T)\,x$ and $(I - LL^T)\,x$).

3. The Effects of Rounding Error and the Role of Pivoting

The size of the matrix E that describes the effects of rounding error on the computation has played an important role in Section 2. In this section we shall give reasons for expecting E to be quite small. The analysis also makes clear the role of pivoting in computing the decomposition.

We begin with a detailed description of the Cholesky algorithm in its "exterior product" form. The algorithm proceeds in stages. At the kth stage A has been decomposed in the form

$$A = L_kL_k^T + B_k\,,$$

where L_k has k columns and B_k has the form

$$B_k = \begin{pmatrix} 0 & 0 \\ 0 & B_{22}^{(k)} \end{pmatrix} \tag{3.1}$$

with $B_{22}^{(k)}$ of order $n - k$. Denoting by $b_k^{(k)}$ the kth column of B_k and by $\beta_{kk}^{(k)}$ the (k, k) element of B_k, we form

$$B_{k+1} = B_k - \frac{b_k^{(k)}b_k^{(k)T}}{\beta_{kk}^{(k)}}$$

and

$$L_{k+1} = \Big(L_k\,,\ \frac{b_k}{(\beta_{kk}^{(k)})^{1/2}}\Big),$$

It is easily verified that L_{k+1} is lower trapezoidal, that B_{k+1} is zero except for its trailing principal minor of order $n - k - 1$, and that $A = L_{k+1}^TL_{k+1} + B_{k+1}$. Thus the decomposition is advanced one stage. The algorithm terminates when some B_k is negligible.

The algorithm cannot be carried out in the form described above if $\beta_{kk}^{(k)}$ is not positive. However, in this event it may happen that there is an integer $l_k \geqslant k$ such that $\beta_{l_k l_k}^{(k)}$ is positive. Let P_k denote the permutation matrix obtained by interchanging rows k and l_k of the identity matrix and consider the decomposition

$$P_k A P_k^T = (P_k L_k)(P_k L_k)^T + P_k B P_k^T.$$

The matrix $P_k L_k$ is still lower trapezoidal, and the matrix $P_k B_k P_k^T$ still has the form (3.1). However, the (k, k)th element of $P_k B_k P_k^T$ is $\beta_{l_k l_k}^{(k)}$, and the decomposition of $P_k A P_k^T$ can proceed as usual. This process of interchanging an acceptable element into the (k, k)th position of B_k is the pivoting process mentioned in Section 1.

It is still conceivable that no diagonal element of B_k is positive. We shall show that this is not likely to happen unless B_k is itself negligible. We begin by proving a theorem about the diagonal elements of nearly idempotent matrices.

THEOREM 3.1. *Let the symmetric matrix A of order n satisfy $A^2 = A + F$, where*

$$\gamma \equiv \frac{1 - (1 - 4 \| F \|)^{1/2}}{2} < \frac{1}{2n}.$$

Then either $\| A \| \leqslant \gamma$ or there is a diagonal element α_{ii} of A that satisfies

$$\alpha_{ii} \geqslant \frac{1}{n} - \gamma > \frac{1}{2n}. \tag{3.2}$$

Proof. As was observed in Theorem 2.1, the eigenvalues of A lie in the nonoverlapping intervals $[-\gamma, \gamma]$ and $[1 - \gamma, 1 + \gamma]$. By perturbing the eigenvalues in the first interval to zero and those in the second interval to unity we obtain a matrix $A + G$ whose eigenvalues are either zero or unity; i.e., $A + G$ is a projection. Moreover, $\| G \| \leqslant \gamma$. Now if $\| A \| > \gamma$, then one of the eigenvalues of A must lie in the interval $[1 - \gamma, 1 + \gamma]$, and $A + G$ must have unity for an eigenvalue. The trace of $A + G$ is the sum of the eigenvalues of $A + G$; hence

$$\sum_{i=1}^{n} (\alpha_{ii} + \gamma_{ii}) \geqslant 1.$$

Thus there is a diagonal element $\alpha_{ii} + \gamma_{ii}$ of $A + G$ satisfying

$$\alpha_{ii} + \gamma_{ii} > 1/n. \tag{3.3}$$

Since $\gamma_{ii} \leqslant \| G \|$, (3.3) implies (3.2) ∎

It must be noted that the term $1/n$ in (3.2) is an extreme lower bound and can be replaced by p/n, where p is the number of eigenvalues of A in the interval $[1 - \gamma, 1 + \gamma]$.

Theorem 3.1 shows that there is always a reasonable pivot element to start the reduction. To show that it can be completed, we shall show that the matrices B_k are also nearly idempotent, after which Theorem 3.1 applies to give us the required pivot element. In addition to the usual assumption that $A^2 = A + F$, we shall take account of rounding error by supposing that the computed L_k and B_k satisfy

$$L_k L_k^T + B_k = A + E_k .$$

For notational convenience we shall drop the subscripts k during the analysis.

Let B be partitioned as in (3.1), and let A, E, and L be partitioned conformally:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = (A_1 , A_2),$$

$$E = \begin{pmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{pmatrix} = (E_1 , E_2),$$

$$L = \begin{pmatrix} L_1 \\ L_2 \end{pmatrix}.$$

Assume the L_1 is nonsingular and set

$$\lambda = \| L_1^{-1} \|.$$

Now

$$LL_1^T = A_1 + E_1 .$$

Hence

$$\begin{aligned} L_1 L^T L L_1^T &= A_1^T A_1 + A_1^T E_1 + E_1^T A_1 + E_1^T E_1 \\ &= (A_{11} + E_{11}) + F_{11} - E_{11} + A_1^T E_1 + E_1^T A_1 + E_1^T E_1 \end{aligned}$$

or

$$L^T L = I + L_1^{-1}(F_{11} - E_{11} + A_1^T E_1 + E_1^T A_1 + E_1^T E_1) L_1^T \equiv I + G, \quad (3.4)$$

where

$$\| G \| \leqslant \lambda^2[\| F \| + \| E \|(3 + 2 \| F \| + \| E \|)].$$

It also follows from (3.4) that

$$\| L^T L \| \leqslant 1 + \| G \|$$

and

$$\| L \| \leqslant 1 + \tfrac{1}{2} \| G \|.$$

We next obtain a bound for $L - AL$. We have

$$\begin{aligned} AL L_1^T &= A(A_1 + E_1) = A_1 + F + AE_1 \\ &= (A_1 + E_1) + F_1 + (A - I) E_1 . \end{aligned}$$

Hence

$$\begin{aligned} AL &= L + [F_1 + (A - I) E_1] L_1^{-T} \\ &\equiv L + H, \end{aligned}$$

where

$$\| H \| \leqslant \lambda[\| F \| + \| E \|(1 + \| F \|)].$$

Finally since $B = A + E - LL^T$,

$$\begin{aligned} B^2 &= A^2 - ALL^T - LL^TA + LL^TLL^T + AE + EA + E^2 - ELL^T - LL^TE \\ &= (A + E) - (L + H) L^T - L(L^T + H^T) + L(I + G) L^T \\ &\quad + F + (A - I) E + EA + E^2 - ELL^T - LL^TE \\ &= B - HL^T - LH^T + LGL^T + F + (A - I) E + EA + E^2 \\ &\quad - ELL^T - LL^TE \\ &\equiv B + K, \end{aligned}$$

where

$$\begin{aligned} \| K \| \leqslant{}& \| H \| (2 + \| G \|) + \| G \| (1 + \| G \|) \\ &+ \| F \| + \| E \| (4 + 2 \| F \| + \| E \| + 2 \| G \|). \end{aligned}$$

If we ignore terms of the second order in the bound for $\| K \|$ we obtain the asymptotic bound

$$\| K \| \lesssim \lambda^2(\| F \| + 3 \| E \|) + 2\lambda(\| F \| + \| E \|) + (\| F \| + 4 \| E \|),$$

in which the first term will generally dominate. Since $L_1L_1^T = A_{11} + E_{11}$, the number λ^2 is an estimate of $\| A_{11}^{-1} \|$. This explains the role of pivoting in the algorithm. Not only is pivoting necessary to insure that one stage of the algorithm can be carried out, but it is also necessary to keep small diagonal elements from appearing in L_1. For if this unhappy circumstance occurs, then λ must be large and we cannot guarantee the successful conclusion of the algorithm. Note, however, that if E and F are small we can hope to tolerate rather small diagonal elements, which gives us considerable freedom in the choice of pivot elements.

We have not yet given a quantitative assessment of the effects of rounding error on our algorithm. We cite a well-known theorem [7, 8].

THEOREM 3.2. *Let the algorithm described above be carried out in t-digit binary floating-point arithmetic. Let*

$$\beta_k = \max\{\beta_{ij}^{(l)} : i, j = 1,..., n; l = 1,..., k - 1\}.$$

Then

$$\| E_k \| \leqslant f(n) \beta_k 2^{-t}.$$

The function $f(n)$ depends on the details of the arithmetic used; but it is certainly less than $O(n^2)$ with a modest order constant. The critical factor is the number β_k, which measures the growth of the elements of the matrices B_k. Since $\beta_k \leqslant 1 + \| K_k \|$, the above analysis applies to show that, provided we have maintained a reasonable degree of nonsignularity in the matrices $L_1^{(k)}$, rounding error should have a negligible effect on the algorithm.

To summarize, this is a remarkably stable algorithm. Although we cannot guarantee that the $L_1^{(k)}$ will have small inverses, we think that it is extremely unlikely that anything untoward will happen if a reasonable pivoting strategy (e.g., choosing the largest diagonal element) is adopted. The cautious user can monitor the β_k as the B_k are computed, after which Theorems 3.2 and 2.2 will enable him to assess his results. A particularly attractive feature of the algorithm is the latitude in pivoting strategies that the bound on $\| K \|$ suggests are available to the user. For example, the user might compromise the size of his pivots to preserve sparsity in very large problems and hope to get away with it. Experiments by Fox and Krohn [3] in which the pivot order is fixed initially tend to confirm this view.

REFERENCES

1. C. DAVIS AND W. KAHAN, *SIAM J. Numer. Anal.* 7 (1970), 1–46.
2. I. S. DUFF, *Computing* **13** (1974), 239–248.
3. K. FOX AND B. KROHN, "Computation of Cubic Harmonics," *J. Comput. Phys.* 25 (1977), 386–408.
4. A. S. HOUSEHOLDER AND K. FOX, *J. Comput. Phys.* **8** (1971), 292–294.
5. C. L. LAWSON AND R. J. HANSON, "Solving Least Squares Problems," Prentice-Hall, Englewood Cliffs, N.J. 1974.
6. 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, New York, 1976.
7. G. W. STEWART, "Introduction to Matrix Computations," Academic, New York, 1973.
8. J. H. WILKINSON, "The Algebraic Eigenvalue Problem," Oxford Univ. Press (Clarendon), London, 1965.