NORTH-HOLLAND

Fernando's Solution to Wilkinson's Problem: An Application of Double Factorization

Beresford N. Parlett*
Mathematics Department
University of California
Berkeley, California 94720

and

Inderjit S. Dhillon†
Computer Science Division
EECS Department
University of California
Berkeley, California 94720

Submitted by Hans Schneider

ABSTRACT

Suppose that one knows a very accurate approximation σ to an eigenvalue λ of a symmetric tridiagonal matrix T. A good way to approximate the eigenvector $\mathbf{x}$ is to discard an appropriate equation, say the rth, from the system $(T - \sigma I)\mathbf{x} = \mathbf{0}$ and then to solve the resulting underdetermined system in any of several stable ways. However the output $\mathbf{x}$ can be completely inaccurate if r is chosen poorly, and in the absence of a quick and reliable way to choose r, this method has lain neglected for over 35 years. Experts in boundary value problems have known about the special structure of the inverse of a tridiagonal matrix since the 1960s, and their double triangular factorization technique (down and up) gives directly the redundancy of *each* equation and so reveals the set of good choices for r. The relation of double factorization to the eigenvector algorithm of Godunov and his collaborators is described. The results extend to band matrices and to zero entries in eigenvectors, and have uses beyond eigenvector computation.

*Supported by ONR, contract N000014-90-J-1372.
†The author was supported in part by ARPA contract DAAL03-91-C0047 via a subcontract from the University of Tennessee.

655 Avenue of the Americas, New York, NY 10010
0024-3795/97/$17.00
PII S0024-3795(97)00022-0

1. INTRODUCTION

The task that started these investigations is the computation of eigenvectors of a symmetric tridiagonal matrix [(i, j) entry vanishes if $|i - j| > 1$] once the eigenvalues are in hand. This is not a new problem, and there are good programs available in libraries such as LAPACK and NAG. Nevertheless the experts do not consider the situation satisfactory (see [17]); the complexity of the programs seems out of proportion to the difficulty of the task, and the adaptation of the current versions of inverse iteration to parallel mode is frustrating.

Let us briefly sketch the situation. Given an accurate approximation σ to an eigenvalue λ of an $n \times n$ symmetric tridiagonal matrix T, one considers the solution $\mathbf{x}$ to the system of equations

$$(T - \sigma I)\mathbf{x} = \mathbf{b}, \tag{1}$$

where $\mathbf{b}$ is to be chosen wisely. Since $\sigma \neq \lambda$, the best choice for $\mathbf{b}$ is the eigenvector we seek, but that is not a serious option. A feasible alternative is to choose column r of the identity matrix $I = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$. As will become clear in Section 3, choosing $\mathbf{b} = \mathbf{e}_r$ is equivalent to omitting equation r from the homogeneous version of (1). The value $r = n$ was proposed by Wallace Givens in 1957 (see [11]), but no fixed value of r, independent of σ and T, will do.

Here is a quotation from Wilkinson concerning the computation of an eigenvector $\mathbf{u}_k$, in Chapter 5, Section 50, below Equation (50.3) of [27]:

> Hence if the largest component of $\mathbf{u}_k$ is the rth, then it is the rth equation which should be omitted when computing $\mathbf{u}_k$. This result is instructive but not particularly useful, since we will not know *a priori* the position of the largest component of $\mathbf{u}_k$.

Ipsen, in a very readable survey, attributes the idea of omitting one equation of the system to Wilkinson (see Section 7 of [17]), but we suspect that this method was routinely taught in mathematics classes before Wilkinson was born; see [16], [4], and [23]. He was born in 1919, and [16] was published in 1921.

Wilkinson abandoned the hunt for a good value of r and used $\mathbf{b} = PL\mathbf{e}$, where $T - \sigma I = PLU$ denotes triangular factorization with partial pivoting and $\mathbf{e} = \sum_{i=1}^{n} \pm \mathbf{e}_i$; see [26]. However, even this choice fails if some eigenvalues are equal to working accuracy, and he resorted to "tweaking" the computed eigenvalues in such cases.

In private communication to one of us, Wilkinson declared that he would prefer $\mathbf{b} = \mathbf{e}_r$ to $\mathbf{b} = PL\mathbf{e}$ if only he knew a quick, reliable way to choose r

so that the rth entry of the wanted vector is above average, not necessarily the greatest. One goal of this paper is to show that a large entry may be located with little extra expense.

The current LAPACK codes (see [1]) do not use Wilkinson's choice; instead $\mathbf{b}$ is chosen "at random" from an appropriate distribution, but this still makes it difficult to obtain orthogonal eigenvectors for close eigenvalues and difficult to prove "correctness" of the code. The case for this approach is made in [18].

Here ends our review of the situation.

This paper rederives results from the differential equations community concerning double factorization and proceeds to describe and justify K. V. Fernando's solution to Wilkinson's problem of choosing the index r mentioned above. This is an important contribution to the computation of satisfactory eigenvectors, but it is not enough, in our opinion, to ensure mutual orthogonality. See [8] for a different viewpoint.

Section 2 discusses the "obvious" solution to the problem and shows its shortcomings. The new idea, in its simplest form, comes from Theorem 2, which is established in Section 3 along with Theorem 6, which presents accurate ways to compute the determinant. Section 4 reviews some closely related work. First, there is the use of double factorization by experts on boundary value problems for second order equations. Second, Godunov and his coworkers join together a forward and a backward sequence at a well-chosen index to obtain an eigenvector. Section 5 shows how the quantities introduced in Theorem 2 reveal the *envelope* of an eigenvector when the tridiagonal is normal, and goes on to justify the choice of r. Section 6 extends the results to cover breakdown in triangular factorization and the interesting phenomenon of zero entries in eigenvectors. Section 7 extends the results of Section 3 to block tridiagonal matrices.

Theorem 2 has applications to finding the envelope of an invariant subspace and in computing condition numbers; see [22] and [5]. Even in the eigenvector application Theorem 2 has uses beyond finding the best value of r. It allows us to omit the calculation of completely negligible entries, and that enhances efficiency when n becomes large (say $n \gg 100$). An example of this is given in Figure 4.

The reader is expected to know the LDU theorem concerning existence and uniqueness of triangular factorization and the expressions for the pivots, as the diagonal entries of D are often called. In this representation both L and U have 1's on the diagonal. In practice, when division is slow, people sometimes use $(LD)D^{-1}(DU)$ instead of LDU, but the distinction is not important in this paper.

The main notational issue is the representation of submatrices. In MATLAB notation the submatrix of M in rows i through j and columns k through l is given by $M(i:j, k:l)$. This is clear but sometimes too obtrusive. We use

$M^{i:j}$ to denote the principal submatrix $M(i:j, i:j)$. For column vectors we prefer simple lowercase Latin letters $\mathbf{x}, \mathbf{y}, \ldots$ in boldface type, with entries $x(1), x(2), \ldots, x(n)$. For subvectors we use either $\mathbf{x}(i:j)$ or $\mathbf{x}^{i:j}$. Finally, we try to use lowercase Greek letters $\alpha, \beta, \ldots$ for scalars, although matrix entries will be written as $M(i, j)$ or M_{ij}.

One notational innovation is to use $+$ to indicate a process taking rows in increasing order and $-$ to indicate the process going in decreasing order; e.g., LDU is written as $L_+D_+U_+$, while UDL is written as $U_-D_-L_-$.

As usual, $\|\mathbf{v}\| = \|\mathbf{v}\|_2 = \sqrt{\mathbf{v}^*\mathbf{v}}$, while $\|\mathbf{v}\|_\infty = \max_i |v(i)|$. The dimension, or order, of $\mathbf{e}_1$ or of any column of the identity matrix $I = (\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n)$ is given by the context.

A different method to compute the $\mathbf{z}^{(k)}$ of Theorem 2, a technique more prone to overflow, was presented at a Numerical Analysis Day at Stanford University in April 1993, at the Householder XII conference at Lake Arrowhead, California in June 1993, and at the SIAM Applied Linear Algebra meeting at Snowbird, Utah in June 1994, by Fernando and Parlett. The context of all these presentations was the computation of singular vectors of a bidiagonal matrix, and the choice of the "best" k was based on heuristic rules. Theorem 2 was presented at the SIAM conference on Parallel Processing in San Francisco, California in February 1995, by Fernando and Parlett, with help from unpublished notes by I. S. Dhillon. Fernando gave talks on computing an eigenvector, from an alternative, more general viewpoint, at ICIAM 95 in Hamburg, Germany and ILAS 96 in Chemnitz, Germany.

2. A CLASSICAL ANALYSIS

In case a pure mathematician should, by chance, read this material, it seems wise to begin by explaining that the problem discussed here is not as trivial as it may appear at first. It is the computer's limited precision that causes the difficulties.

Anyone who has mastered an introductory course in matrix theory and who has absorbed the significance of the tridiagonal form J (with nonzero values adjacent to the diagonal J is said to be unreduced) might reason as follows.

LEMMA 1. *An eigenvector of an unreduced tridiagonal matrix J cannot have a 0 in the first or last component.*

Proof. Consider the equation for an eigenvector $\mathbf{x}$ ($\neq 0$) associated with an eigenvalue λ,

$$(J - \lambda I)\mathbf{x} = \mathbf{0}. \tag{2}$$

Suppose that $x(1) = 0$. Then the first equation in (2) dictates that $x(2) = (\lambda - J_{11})x(1)/J_{12} = 0$ as well, since $J_{12} \neq 0$. Now the second equation dictates that $x(3)$ is a linear combination of $x(1)$ and $x(2)$ and also vanishes. Proceeding with the remaining equations, in order, it appears that every entry of $\mathbf{x}$ must vanish, in contradiction to the assumption that $\mathbf{x}$ is an eigenvector. So the assumption that $x(1) = 0$ is not tenable. By similar reasoning but taking the equations in reverse order it is untenable that $x(n) = 0$. ■

The preceding argument also shows one way to compute an eigenvector of J. It is valid to set $x(1) = 1$ and to use the first equation of (2) to determine $x(2)$, and the second to determine $x(3)$, using $x(1)$ and $x(2)$. Proceeding as before, the rth equation may be used to determine $x(r + 1)$, and thus $\mathbf{x}$ may be obtained without actually making use of the nth equation, which, says the mathematician, will be satisfied automatically, since the system (2) is singular.

It would be equally valid to begin with $\tilde{x}(n) = 1$ and to take the equations in reverse order to compute $\tilde{x}(n - 1), \ldots, \tilde{x}(2), \tilde{x}(1)$ in turn without using the first equation in (2). When normalized in the same way, $\mathbf{x}$ and $\tilde{\mathbf{x}}$ will yield the same eigenvector. Note that the problem has been solved without the bother of computing a triangular factorization.

The method described above is "obvious" and was mentioned by W. Givens in 1957; see [11]. It often gives good results when realized on a computer, but at other times delivers vectors pointing in completely wrong directions.

The preceding analysis is valid in exact arithmetic but is inapplicable to computer work for the following reasons. First, it is rare that an eigenvalue of a tridiagonal (or any other) matrix is representable in limited precision. Consequently the systems such as (2) that are to be solved are not singular, and in (2) the unused equation will not be satisfied automatically even if the solutions of the other equations, in turn, were obtained exactly. The second weakness is that, in a computer, the sequence $x(1), x(2), \ldots$, can overflow. This is a possibility that pure mathematicians do not have to worry about.

It turns out that, for isolated eigenvalues, the natural method gives an excellent approximate eigenvector whenever the first or last entry of the wanted eigenvector is above average in magnitude. Conversely, it gives

disastrous results when those extreme entries are tiny. Wilkinson gives a striking example in Section 52, Chapter 5 of [27].

The purpose of this section was to show that the "obvious" method for computing eigenvectors is not adequate for finite precision arithmetic.

3. DIAGONAL OF THE INVERSE

In basic courses in matrix theory one is taught to solve a system of equations by computing a row echelon form. If the system is singular, at least one row of the echelon form vanishes and the corresponding row of the original system is redundant. The homogeneous system is solved by assigning any values to the "free" variables and backsolving for the rest of them. In general a discarded row is not unique; it need only be a linear combination of the remaining ones.

In practice our system is nearly, but not quite, singular, and a natural modification of the standard procedure is to seek a row that is most nearly redundant and then ignore it while determining a solution $\mathbf{x}$ to the *remaining* homogeneous system. This solution $\mathbf{x}$ will not satisfy the omitted (rth) equation. In other words, faced with the fact that $M\mathbf{x} = \mathbf{0}$ admits only the trivial solution, one finds a suitable r and solves, instead, $M\mathbf{x} = \mathbf{e}_r\delta_r$, where δ_r is the *defect* (or residual norm) of the rth equation. This is what is meant by "omitting the rth equation."

In general it is difficult to find r and to solve the reduced homogeneous system. Fortunately, when M is tridiagonal the omission of row r splits the system into two separate parts. For a modest cost the residual δ_j, *for every choice of* j, can be computed, and that gives an excellent basis for choosing the right r.

The preceding paragraphs are motivation, but the results of this section are for generic tridiagonals that need not be nearly singular. Section 5 returns to the nearly singular case. Theorem 2 is not new (see Section 4), but it deserves more prominence than it has received up to now.

THEOREM 2 (Double factorization). *Let J be a tridiagonal $n \times n$ complex matrix that permits triangular factorization in both increasing and decreasing order of rows:*

$$L_+D_+U_+ = J = U_-D_-L_-. \tag{3}$$

For each k, $1 \leqslant k \leqslant n$, define γ_k and $\mathbf{z}^{(k)}$ by

$$J\mathbf{z}^{(k)} = \mathbf{e}_k\gamma_k, \qquad z^{(k)}(k) = 1. \tag{4}$$

Then

$$\gamma_k = D_+(k) + D_-(k) - J_{kk}. \tag{5}$$

NOTE 1. Note that $\mathbf{z}^{(k)}$ is a multiple of $J^{-1}\mathbf{e}_k$, when J^{-1} exists. In that case, from (4) (see Corollary 3),

$$\gamma_k^{-1} = (J^{-1})_{kk}. \tag{6}$$

Proof. In what follows MATLAB notation will be used for submatrices that are not square, and a more condensed representation, $M^{1:j}$ for $M(1:j,\ 1:j)$, otherwise. In addition, if terms that involve out of range indices are dropped, then the analysis that follows covers the extreme cases $k = 1$ and $k = n$ as well. For brevity write $\mathbf{z}$ for $\mathbf{z}^{(k)}$.

Omit the kth equation from (4), and what remains is two homogeneous systems. Next use the appropriate triangular factorization (3) to write these systems as

$$L_+^{1:k-1}D_+^{1:k-1}U_+(1:k-1, 1:k)\mathbf{z}(1:k) = \mathbf{0}, \tag{7}$$

$$U_-^{k+1:n}D_-^{k+1:n}L_-(k+1:n, k:n)\mathbf{z}(k:n) = \mathbf{0}. \tag{8}$$

By the assumption that the *LDU* and *UDL* factorizations exist the matrices $L_+^{1:k-1}$, $D_+^{1:k-1}$, $U_-^{k+1:n}$, $D_-^{k+1:n}$ must be invertible. Premultiply (7) and (8) by the appropriate inverses to find

$$U_+(1:k-1, 1:k)\mathbf{z}(1:k) = \mathbf{0}, \tag{9}$$

$$L_-(k+1:n, k:n)\mathbf{z}(k:n) = \mathbf{0}. \tag{10}$$

The last equation in (9) shows that

$$1 \cdot z(k-1) + U_{k-1,k}^{+}z(k) = 0. \tag{11}$$

The first equation in (10) shows that

$$L_{k+1,k}^{-}z(k) + 1 \cdot z(k+1) = 0. \tag{12}$$

Recall that $z(k) = 1$, and substitute, from (11) and (12), the values for $z(k-1)$ and $z(k+1)$ into the kth equation of (4) to find, for $k = 2, \ldots, n-1$,

$$\gamma_k = -J_{k,k-1}U^+_{k-1,k} + J_{kk} - J_{k,k+1}L^-_{k+1,k} \tag{13}$$

$$= (J_{kk} - J_{k,k-1}U^+_{k-1,k}) - J_{kk} + (J_{kk} - J_{k,k+1}L^-_{k+1,k})$$

$$= D_+(k) - J_{kk} + D_-(k), \tag{14}$$

as claimed. For $k = 1$ note that $D_+(1) = J_{11}$ and $\gamma_1 = D_-(1)$. For $k = n$ note that $D_-(n) = J_{nn}$ and $\gamma_n = D_+(n)$. Thus (5) holds for $k = 1$ and $k = n$ as well as for $k = 2, \ldots, n-1$. ∎

What follows is a nice summary of the relationships. Perhaps this formulation is new.

COROLLARY 3. *Let J satisfy the hypotheses of Theorem 2. Either J is singular and then $D_+(n) = D_-(1) = \gamma_n = \gamma_1 = 0$ and both $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(n)}$ are in J's null space, or*

$$\operatorname{diag}(J^{-1})^{-1} + \operatorname{diag}(J) = D_+ + D_-. \tag{15}$$

Proof.

$$D_+(n) = \frac{\det J}{\det J^{1:n-1}}, \qquad D_-(1) = \frac{\det J}{\det J^{2:n}}.$$

By (3), $\det J^{1:n-1} \neq 0$, $\det J^{2:n} \neq 0$. Hence when J is singular $D_+(n) = D_-(1) = 0$. Always $D_+(1) = J_{11}$ and $D_-(n) = J_{nn}$, and so, when J is singular,

$$\gamma_n = 0 - J_{nn} + D_-(n) = 0, \qquad \gamma_1 = D_+(1) - J_{11} + 0 = 0,$$

and so $J\mathbf{z}^{(1)} = J\mathbf{z}^{(n)} = \mathbf{0}$.

If J is invertible, then substitute (6) into (5). ∎

In applications it is useful to have several different expressions for γ_k in addition to (14).

COROLLARY 4. *With the notation of Theorem 2, for* $1 < k < n$,

$$\begin{aligned}\gamma_k &= D_+(k) - J_{kk} + D_-(k)\\ &= -L^+_{k,k-1}J_{k-1,k} + J_{kk} - J_{k+1,k}U^-_{k,k+1}\\ &= -J_{k,k-1}U^+_{k-1,k} + J_{kk} - J_{k,k+1}L^-_{k+1,k}\\ &= D_+(k) - U^-_{k,k+1}J_{k+1,k}\\ &= -L^+_{k,k-1}J_{k-1,k} + D_-(k).\end{aligned}$$

For $k = 1$ *and* $k = n$ *omit terms with invalid indices.*

Proof. The first and third expressions are just (5) and (13). The others come from rewriting (13) as

$$\gamma_k = -\frac{J_{k,k-1}J_{k-1,k}}{D_+(k-1)} + J_{kk} - \frac{J_{k,k+1}J_{k+1,k}}{D_-(k+1)} \tag{16}$$

and using $J_{k,k+1} = U^-_{k,k+1}D_-(k+1) = D_+(k)U^+_{k,k+1}$ etc. and $D_-(k) = J_{kk} - J_{k,k+1}J_{k+1,k}/D_-(k+1)$, etc. ■

For eigenvector computation the middle formula in Corollary 4 is of most interest because of the following result, which shows that no divisions are needed to find $\mathbf{z}^{(k)}$ once k is known. In current computers divisions are significantly slower than multiplications.

COROLLARY 5. *Consider Theorem 2 and a given value of* k, *so that* $\mathbf{z} = \mathbf{z}^{(k)}$, $J\mathbf{z} = \mathbf{e}_k\gamma_k$. *Then* $z(k) = 1$ *and*

$$\begin{aligned}z(j) &= U^+_{j,j+1}z(j+1), \qquad j = k-1,\ldots,1,\\ z(i) &= -L^-_{i,i-1}z(i-1), \qquad i = k+1,\ldots,n.\end{aligned}$$

Proof. These equations are (9) and (10) in expanded form. ■

Corollary 5 shows that there is no need to retain D_+ and D_- in order to compute $\mathbf{z}^{(k)}$. It is interesting that Godunov and his coworkers are actually computing $U_+(i)$ and $1/L_-(i)$, $i = 1,\ldots,n$ when they form left and right

Sturm sequences. However, they do not use Corollary 5, but a much more expensive procedure, to obtain an approximate eigenvector.

Another reward for computing both factorizations is a wide choice of expressions for $\det J$. These offer more accuracy than the usual formula $\prod_{i=1}^{n} D_+(i)$ derived from $J = L_+ D_+ U_+$ when $\det J$ is small. These expressions are well known in the differential equations community; see [20, Theorem 2.3], for example.

THEOREM 6. *Assume the hypothesis of Theorem 2. Then for* $k = 1, \ldots, n$,

$$\det J = D_+(1) \cdots D_+(k-1)\gamma_k D_-(k+1) \cdots D_-(n)$$

$$= D_+(1) \cdots D_+(k-2) \det \begin{bmatrix} D_+(k-1) & J_{k-1,k} \\ J_{k,k-1} & D_-(k) \end{bmatrix}$$

$$\times D_-(k+1) \cdots D_-(n)$$

and

$$\frac{\gamma_k}{\gamma_{k+1}} = \frac{D_+(k)}{D_-(k+1)}.$$

Proof. Apply Cramer's rule for the kth entry of $\mathbf{z}^{(k)}$ where $J\mathbf{z}^{(k)} = \mathbf{e}_k \gamma_k$. The numerator is a determinant whose kth column is $\mathbf{e}_k \gamma_k$. Expand it by column k to find

$$1 = z^{(k)}(k) = \gamma_k \det J^{1:k-1} \det J^{k+1:n} / \det J.$$

Since $J = L_+ D_+ U_+ = U_- D_- L_-$, it follows that

$$\det J^{1:k-1} = D_+(1) \cdots D_+(k-1),$$

$$\det J^{k+1:n} = D_-(k+1) \cdots D_-(n).$$

The second expression comes from the twisted factorization of J:

$$J = \begin{bmatrix} L_+^{1:k-1} & O \\ O & U_-^{k:n} \end{bmatrix} \begin{bmatrix} D_+^{1:k-2} & & \\ & \diamond & \\ & & D_+^{k+1:n} \end{bmatrix} \begin{bmatrix} U_+^{1:k-1} & O \\ O & L_-^{k:n} \end{bmatrix}$$

where

$$\diamond = \begin{bmatrix} D_+(k-1) & J_{k-1,k} \\ J_{k,k-1} & D_-(k) \end{bmatrix}.$$

From the first expression for $\det J$ it follows that

$$\gamma_k D_-(k+1) = \gamma_{k+1} D_+(k),$$

which gives the ratio of consecutive γ's. ∎

COROLLARY 7.

$$\gamma_k = D_+(k) \prod_{i=k+1}^{n} \frac{D_+(i)}{D_-(i)},$$

$$= D_-(k) \prod_{i=1}^{k-1} \frac{D_-(i)}{D_+(i)}.$$

When there is severe cancellation in computing γ_k from any of the formulae in Corollary 4, then it may be possible to take extra care in the evaluation of $\det \diamond$ and win back a few bits of precision. If warranted, the idea may be taken further to use

$$\det \begin{bmatrix} D_+(k-1) & J_{k-1,k} & 0 \\ J_{k,k-1} & J_{kk} & J_{k,k+1} \\ 0 & J_{k+1,k} & D_-(k+1) \end{bmatrix}$$

for the sensitive part of the computation. These details are of great practical importance when J is close to singular, as occurs in iterative methods for finding eigenvalues.

REMARK 8. An attentive reader may be puzzled that Corollary 5 cannot generate an isolated 0 entry in $\mathbf{z}^{(k)}$. If $z(j)$ vanishes [because $J(j, j+1) = 0$], then all entries $z(l)$, $l < j < k$, must vanish too. This is appropriate, since the matrix is reduced when $J_{j,j+1} = 0$. Yet there exist unreduced tridiagonals with isolated $z^{(k)}(j)$ that vanish. The explanation is that the hypothesis (3) does indeed rule out isolated zero entries. Section 6 extends the results of Theorem 2 to cover these important cases. There we see that the hypothesis (3) in Theorem 2 may be weakened, and that is the topic of Section 7.

REMARK 9. For large n there will be many products in the calculation of $z(1)$ or $z(n)$ for a given r from Corollary 5. In general one is concerned about possible overflow, but when r is selected so that $z(r)$ should be a maximal, or nearly maximal, entry of $\mathbf{z}$, then no overflow can occur. Underflow, if it occurs, is harmless here and may be gradual or flushed to zero.

REMARK 10. From (4) we see that the vector $\mathbf{z}^{(r)}$ is annihilated exactly by $J - \mathbf{e}_r \gamma_r \mathbf{e}_r^*$, a rank one perturbation.

4. RELATED WORK

4.1. *Boundary Value Problems for Second Order Differential Equations*

In 1992 in [20] Meurant reviewed a significant portion of the literature on the inverses of band matrices as well as presenting the main ideas in a nice unified framework. The compact representation of the inverse of a tridiagonal was in the literature in the 1960s and began to be linked to double factorization in the 1980s. Theorem 2.3 of [20] gives the quotient product form of $(J^{-1})_{kk}$ as

$$\frac{D_-(k+1)\cdots D_-(n)}{D_+(k)\cdots D_+(n)}.$$

This is Corollary 7 of our Theorem 6. The inexpensive additive form (Theorem 2 and Corollary 4) is included in Theorem 3.1 of [20], and our Theorem 17 extends Theorem 3.1 a little.

The researchers on Meurant's reference list were not interested in computing eigenvectors but in obtaining analytic expressions for elements of the inverse, when possible, and the decay rate in terms of distance from the main diagonal. Even the better procedures in [20] for computing the diagonal of the inverse require n extra products after the forward and backward pivots are in place. So investigators interested in eigenvectors may be forgiven for not seeing that double factorization could be useful to them.

In addition to the papers reviewed in [20] are several on twisted factorization. Here r is selected in advance and elimination starts from the top and the bottom and stops at r. Our Section 5 shows where to make the twist for the eigenvector problem. Henrici (1963) used twisted *LDU* implicitly in deriving optimal Gersgorin bounds, in [14], for unreduced real tridiagonals with some complex eigenvalues. D. Kershaw (1970), in [19], obtained nice bounds on $(J^{-1})_{rr}/J_{rr}$ using twisted *LDU*. Babuska (1972) wanted a specific entry in $J^{-1}\mathbf{b}$ for any $\mathbf{b}$, and his formula (5.29) on p. 62 of [2] is one instance of our Corollary 4. Fischer et al. (1974), in [7], discuss a twisted Toeplitz factorization of Buneman and attribute the adjective "twisted" to Strang [24]. Dongarra et al. (1979), in the LINPACK codes, use twisted factorization meeting in the middle for improved efficiency (see [6]), and the practice has been taken up in parallel computation (see [25]). It was referred to as the BABE algorithm (begin, or burn, at both ends) in [15]. See also [9] for an error analysis.

4.2. The Russian Approach

This section tries to present the essence of Chapters 4 and 5 of the somewhat inscrutable book [13], but in our notation. The part described here appeared in [12] in 1985. In order to convey the main ideas we assume that the tridiagonal matrix $\tilde{J}$ is unreduced, $\tilde{J}_{k,k-1}\tilde{J}_{k-1,k} \neq 0$, $k = 2, \ldots, n$, and that various factorizations exist.

We have, for any σ,

$$J = \tilde{J} - \sigma I = L_+ D_+ U_+ = U_- D_- L_-, \tag{17}$$

from Theorem 2, where $D_+ = \operatorname{diag}(D_+(1), \ldots, D_+(n))$ holds the (forward) pivots. Moreover,

$$\begin{aligned} D_+(1) &= \tilde{J}_{11} - \sigma, \\ D_+(k) &= \tilde{J}_{kk} - \sigma - \frac{\tilde{J}_{k,k-1}\tilde{J}_{k-1,k}}{D_+(k-1)}, \qquad k = 2, \ldots, n. \end{aligned} \tag{18}$$

The authors of [13], whom we abbreviate collectively as GAKK, base their algorithms (for eigenvalues and eigenvectors when $\tilde{J}$ is symmetric) not on the

pivots $D_+(i)$ but on their scaled reciprocals

$$P_k = P_k(\sigma) = \frac{\tilde{J}_{k,k+1}}{D_+(k)}, \qquad k = 1, \ldots, n-1,$$
$$P_n = \frac{1}{D_+(n)}. \tag{19}$$

In our notation, for any σ, we have $P_k = U_+(k)$, $k = 1, \ldots, n-1$, and this connection gives a meaning to $\{P_k\}$ even when σ is far from an eigenvalue. The recurrence for the P_k is

$$P_k = \frac{\tilde{J}_{k,k+1}}{\tilde{J}_{kk} - \sigma - \tilde{J}_{k,k-1}P_{k-1}} \quad \text{for } k = 2, \ldots, n \qquad \left(\tilde{J}_{n,n+1} = 1\right). \tag{20}$$

It holds for $k = 1$ as well if we define $P_0 = 0$. The sequence $\{P_k\}_0^n$ is called by GAKK a *left* Sturm sequence of the first kind. In what follows we abbreviate Sturm sequence by Ss and omit the qualifier "of the first kind," since the second kind is used just for one purpose. The adjective "left" signals that the recurrence (20) is evaluated from the initial value $P_0 = 0$ in increasing order of index k. To emphasize this fact the left Ss is written $\{P_k^{(+)}\}_0^n$. Since (20) is a two-term recurrence, it may be run backward from any final value of P_n. For GAKK a *right* Ss, written $\{P_k^{(-)}\}_0^n$, is one computed, in reverse order, from the special final value $P_n^{(-)} = \infty$. Solving (20) for P_{k-1} yields

$$P_k^{(-)} = P_k^{(-)}(\sigma) = \frac{\tilde{J}_{k+1,k+1} - \sigma - \tilde{J}_{k+1,k+2}/P_{k+1}^{(-)}}{\tilde{J}_{k+1,k}},$$
$$k = n-1, \ldots, 1. \tag{21}$$

If a left Ss is also a right Ss, $P_j^{(+)} = P_j^{(-)}$, it is called a *two-sided* Ss. Since $P_n^{(-)} = P_n^{(+)} = \infty$, Equation (19) shows that $D_+(n) = 0$ and so σ must be an eigenvalue of $\tilde{J}$.

In terms of tridiagonal factorization one may verify that, for any value of σ,

$$P_k^{(-)} = \frac{D_-(k+1)}{\tilde{J}_{k+1,k}} = \frac{1}{L_-(k)}, \qquad k = n-1, \ldots, 1. \tag{22}$$

This identification is not obvious and is not mentioned in GAKK, but it gives a meaning to $\{P_k^{(-)}\}$ for any value of σ.

It is well known that the Ss consisting of the leading principal minors of $\tilde{J} - \sigma I$ is closely related to entries in λ's eigenvectors—see [27, p. 316, (48.4)]—and similar reasoning, in the present context, shows that for a two-sided Ss

$$P_k = z_k / z_{k+1}, \qquad k = 1, \ldots, n-1, \tag{23}$$

where $\mathbf{z}$ is an eigenvector of $\tilde{J}$ for the eigenvalue σ.

All the relations mentioned so far hold in exact arithmetic. The admirable purpose of GAKK was to produce algorithms with guaranteed accuracy even in finite precision arithmetic, and they give examples to show that, in practice, a two-sided Ss is extremely unlikely to occur even when σ is the closest machine number to an eigenvalue. They use bisection, with care, to produce two close representable values x, y satisfying

$$x \leqslant \lambda \leqslant y \tag{24}$$

for each eigenvalue λ of a symmetric $\tilde{J}$. Their idea, to find the eigenvector, is to combine a left Ss and a right Ss to produce an acceptable approximation to a two-sided Ss. After mapping a Ss of the first kind into one of the second kind and employing careful analysis, they establish the following important result.

Define

$$\varphi_k^{(+)} = \arctan P_k^{(+)} + \nu_k \pi,$$

$$\nu_k = \text{no. of non-positive terms in } P_1^{(+)}, \ldots, P_k^{(+)},$$

$$\varphi_k^{(-)} = \operatorname{arccot} P_k^{(-)} + \mu_k \pi$$

$$\mu = \text{no. of positive terms in } P_{k+1}^{(-)}, \ldots, P_{n-1}^{(-)}.$$

With appropriate directed rounding the *computed* Ss $\{\bar{\varphi}_k^{(+)}(y)\}$ and the *computed* Ss $\{\hat{\varphi}_k^{(-)}(x)\}$ must "cross" for some index $j < n$. In other words,

$$\begin{aligned} \hat{\varphi}_i^{(+)}(y) &\leqslant \hat{\varphi}_i^{(-)}(x), \qquad i < j, \\ \hat{\varphi}_i^{(-)}(x) &\leqslant \hat{\varphi}_i^{(+)}(y), \qquad i \geqslant j. \end{aligned} \tag{25}$$

No overflows occur if the data are normalized in a prescribed manner. GAKK approximate a two-sided Ss by "sewing" together the two one-sided Ss at the index j,

$$0 = P_0^{(+)}(y), P_1^{(+)}(y), \ldots, P_{j-1}^{(+)}(y), P_j^{(-)}(x), \ldots, P_n^{(-)}(x) = \infty. \quad (26)$$

Their main result (Theorem 6.1) is that (26) is a two-sided Ss for some ideal small perturbation of $\tilde{J}$, not generally symmetric, with eigenvalue $(x + y)/2$.

As indicated in (23), the sequence (26) may be used to approximate an eigenvector, and there are several ways to do it. Unfortunately, despite the title of Section 5.8 of [13], no specific algorithm is given or even referenced.

Earlier chapters of [13] suggest that the authors' preferred method for computing an eigenvector $\mathbf{z}$ from a two-sided Ss $\{P_k\}$ is by mapping the Ss into a chain of plane rotations $\{c_k, s_k\}_{k=1}^n$ and then using the representation

$$\begin{aligned} z_1 &= s_2 s_3 \cdots s_n, \\ z_2 &= c_2 s_3 \cdots s_n, \\ z_3 &= c_3 s_4 \cdots s_n, \\ &\vdots \\ z_{n-1} &= c_{n-1} s_n, \\ z_n &= c_n. \end{aligned}$$

This procedure is safe from overflow and is accurate, but it is a rich man's solution to the problem. The recurrence for the cosines and sines is

$$\begin{aligned} c_1 &= 1, \\ t_i &= -(\operatorname{sign} J_{i,i-1}) P_{i-1}/c_{i-1}, \\ c_i &= \left(1 + t_i^2\right)^{-1/2}, \quad s_i = c_i t_i, \qquad i = 2, \ldots, n. \end{aligned}$$

The cost of this procedure could be regarded as the penalty for not knowing the index of an above average entry in the eigenvector.

Now Section 5 of this paper shows that the jth entry (the crossing point) is among the largest in the eigenvector $\mathbf{z}$, and so the remaining entries may be evaluated, by multiplication or division only, setting $z(j) = 1$ and using (23), with no danger of overflow.

We mention, in passing, that in [13] the singular value problem for a bidiagonal is solved by expanding it to the associated symmetric tridiagonal with a zero diagonal. We do not think that this is the most efficient approach.

It is apparent that GAKK anticipated us with the essential idea of joining a forward sequence and a backward sequence at the correct place to overcome the well-known weakness of using either one alone. Not relating their procedures to factorization, they did not see the connection with $\operatorname{diag}((\tilde{J} - \sigma I)^{-1})$, i.e. Theorem 2, and the use of Corollary 5. This comment is not meant to belittle their significant achievement.

In other words, for GAKK the importance of the crossing point is to obtain a very good bound on the perturbation of $\tilde{J}$ that makes (26) a two-sided Ss, whereas for us the crossing point gives the index of a nearly maximal entry of the desired eigenvector.

We wish to say that by running a single recurrence in two directions at different (but close) values $\sigma = x$ and $\sigma = y$ (as GAKK do) it is not easy to see the connection with our double factorization, which uses two recurrences evaluated on the same matrix $\tilde{J} - \sigma I$.

The Russian algorithm is the first work we have seen to "prove correctness" for such an algorithm. It guarantees delivering an eigenvector of a close matrix. However, this alone does not guarantee that computed eigenvectors for close eigenvalues are orthogonal to working accuracy. See Theorem 12 in the next section for the reason.

4.3. *Our Involvement*

In 1994, unaware of [12], [13], or [20], K. V. Fernando proposed computing pivots $\{D_+(k)\}$ using (18) in two ways: as usual, going forward from the initial value $D_+(1) = J_{11} - \sigma$, and secondly, going in reverse from the final value $D_+(n) = 0$. Let us call the second sequence $\mathring{D}_+(k)$, $k = n, n - 1, \ldots, 1$. He proposed "splicing" the two sequences where $|D_+(k) - \mathring{D}_+(k)|$ is minimized. B. N. Parlett pointed out that this difference is the residual norm of a candidate eigenvector, namely

$$\left| D_+(k) - \mathring{D}_+(k) \right| = \left\| (\tilde{J} - \sigma I)\mathbf{z}^{(k)} \right\|,$$

where $(\tilde{J} - \sigma I)\mathbf{z}^{(k)} = \mathbf{e}_k \gamma_k$, $z^{(k)}(k) = 1$. So a minimizing k gives a maximum diagonal entry of $(\tilde{J} - \sigma I)^{-1}$. In other words if j is the minimizing index, then $\mathbf{e}_j$ is one of the best columns of I to start inverse iteration. The reverse pivots $D_-(k)$ carry the same information as $\mathring{D}_+(k)$ since $\mathring{D}_+(k) =$

$J_{k,k+1}J_{k+1,k}/D_-(k+1)$. Compare this with the fourth expression in Corollary 4. Finally I. S. Dhillon formulated Theorem 2 which is of interest independent of the eigenvector problem and has various applications, including those in [20] of which we were unaware at that time.

5. THE EIGENVECTOR CONNECTION

We return to the task of computing an eigenvector $\mathbf{v}$ for an isolated eigenvalue λ of a tridiagonal $\tilde{J}$. Theorem 2 shows how to compute all the γ_k where

$$(\tilde{J} - \sigma I)\mathbf{z}^{(k)} = \mathbf{e}_k\gamma_k, \quad z^{(k)}(k) = 1, \qquad k = 1, \ldots, n.$$

Let σ approximate λ. For any norm $\|\cdot\|$ and any k,

$$\|(\tilde{J} - \sigma I)\mathbf{z}^{(k)}\| = |\gamma_k|,$$

and the smaller is $|\gamma_k|/\|\mathbf{z}^{(k)}\|$, the better is the bound (28) below on $\sin\angle(\mathbf{v}, \mathbf{z}^{(k)})$ when $\tilde{J}$ is normal. Even for nonnormal $\tilde{J}$ it is desirable to have a small quotient, but we will say no more on that case here.

We do not claim that Theorem 2 solves all the problems involved in computing eigenvector approximations. For normal tridiagonals it does solve Wilkinson's problem: how to locate a maximal element of the eigenvector $\mathbf{v}$.

K. V. Fernando's solution (to be slightly modified below) is to choose an index k that minimizes $|\gamma_i|$. We justify it by the link between $\mathbf{v}$ and $\{\gamma_i\}_1^n$ in Lemma 11, which shows that if $|\gamma_r|$ is minimal then $|v(r)|$ is maximal, i.e., $|v(r)| = \|\mathbf{v}\|_\infty$, *provided that* σ is close enough to λ.

Suppose that $\tilde{J}$ is a normal matrix as well as tridiagonal;

$$\tilde{J} = V\Lambda V^*, \qquad V^{-1} = V^*,$$

and

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \qquad V = (\mathbf{v}_1, \ldots, \mathbf{v}_n).$$

LEMMA 11. *Let $\tilde{J} - \sigma I$ be normal and satisfy the hypotheses of Theorem 2 for all σ in some open interval containing λ_j, a simple eigenvalue of $\tilde{J}$. (When $\tilde{J}$ is unreduced, then each λ_j is simple.) Then $\gamma_k = \gamma_k(\sigma)$ for each k. As $\sigma \to \lambda_j$,*

$$\frac{\gamma_k^{-1}}{\sum_{m=1}^n \gamma_m^{-1}} \to |v_j(k)|^2, \qquad k = 1, 2, \ldots, n. \tag{27}$$

Proof. For $\sigma \neq \lambda_j$,

$$(\tilde{J} - \sigma I)^{-1} = V(\Lambda - \sigma I)^{-1} V^* = \sum \mathbf{v}_i \mathbf{v}_i^* (\lambda_i - \sigma)^{-1}.$$

From (6) in Theorem 2, $\gamma_k^{-1} = [(\tilde{J} - \sigma I)^{-1}]_{kk}$. As $\sigma \to \lambda_j$,

$$(\lambda_j - \sigma)(\tilde{J} - \sigma I)^{-1} \to \mathbf{v}_j \mathbf{v}_j^*,$$

$$(\lambda_j - \sigma)\gamma_k^{-1} \to |v_j(k)|^2, \qquad k = 1, 2, \ldots, n$$

$$(\lambda_j - \sigma) \sum_{m=1}^{n} \gamma_m^{-1} \to \|\mathbf{v}_j\|^2 = 1. \qquad \blacksquare$$

The only blemish in this result is that in finite precision arithmetic, in difficult cases, the closest representable value to λ_j may not be close enough to reveal the limit. The vector of normalized $\{|\gamma_m|^{-1/2}\}_{m=1}^n$ shows the envelope of $\mathbf{v}_j$ to visual accuracy. This is illustrated in Figure 1. An unexpected practical benefit of Lemma 11 for large matrices is that it indicates when the support of $\mathbf{v}_j$ is a small subset of $\{1, 2, \ldots, n\}$. Figure 2 indicates that entries $42, 43, \ldots, 105$ of a certain eigenvector may be set to zero with no loss in accuracy. In other words, large γ-values carry useful information.

Finite precision arithmetic and nonunique minimizers oblige us to modify the simple response to Wilkinson's question. If the computed value of γ_k is 0, it is not always the case that $|v(k)| = \|\mathbf{v}\|_\infty$. Here is an example: $\gamma_k = 0$, but a change of one unit in the last place of J_{kk} yielded $\gamma_k \approx 10^{-14}$. Another index j had $\gamma_j \approx 10^{-16}$ and $|v(j)| = \|\mathbf{v}\|_\infty$, whereas $|v(k)| < \|\mathbf{v}\|_\infty$. There are several ways to replace exact zeros before seeking a minimum. Theorem 6 gives one way:

$$\gamma_k = D_+(k)\gamma_{k+1}/D_-(k+1),$$

and Remark 10 suggests

$$\text{if} \quad \gamma_k = 0 \quad \text{then} \quad \gamma_k \leftarrow \text{macheps} \cdot J_{kk}.$$

There is no need to locate the maximal entry in $\mathbf{v}$. Any index k satisfying

$$|v(k)| > \tfrac{1}{2}\|\mathbf{v}\|_\infty$$

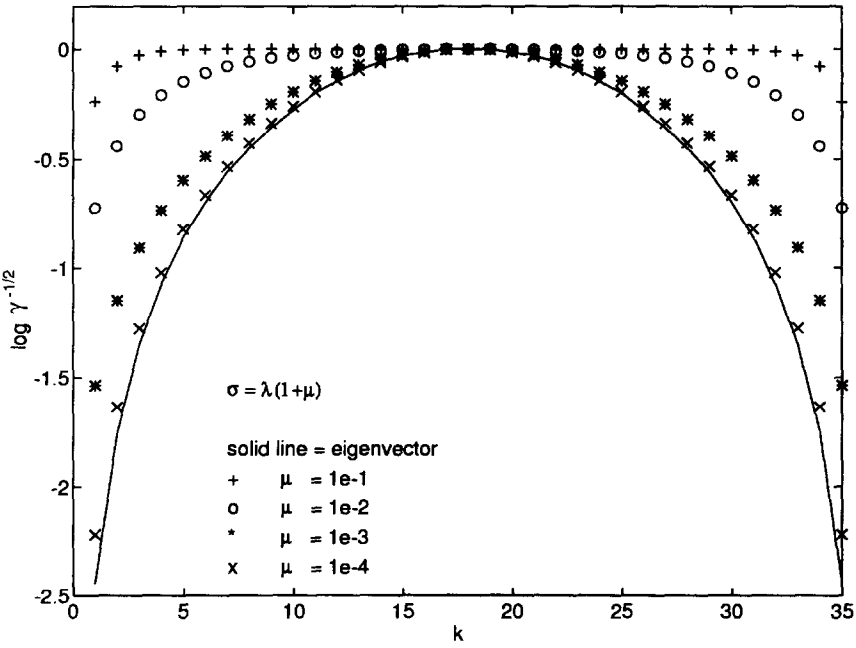

FIG. 1. Convergence of $(1/\gamma)^{1/2}$, $n = 35$.

would be satisfactory. The advantage of having a set of acceptable indices is subtle: it is important to select distinct indices for adjacent and close eigenvalues. It is quite easy for $\|\mathbf{v}\|_\infty$ to be attained at the same index (but not uniquely) for two adjacent eigenvalues, and then it is vital to have a choice.

Finally, if $\min_i |\gamma_i|$ is not small enough, then this process should be halted: perhaps σ is not accurate enough to warrant inverse iteration.

Although this paper does not propose an eigenvector algorithm, the remainder of this section is devoted to an analysis of how good $\mathbf{z}^{(k)}$ is as an approximate eigenvector. A usable definition of "isolated eigenvalue" emerges, and for such eigenvalues the error bounds of Lemma 13 are close to equalities.

It is current practice to compute approximate eigenvalues (of tridiagonals) very accurately by the best current techniques and then turn to inverse iteration for the eigenvectors; but see [12] and [13] for an exception. This approach reveals the not too clustered eigenvalues and guarantees for them that $|\gamma_k|/\|\mathbf{z}^{(k)}\|$ is tiny for at least one k when $\tilde{J}$ is normal. To demonstrate this recall the sin Θ theorem of Davis and Kahan (see [21, Chapter 11]), valid for all Hermitian matrices.

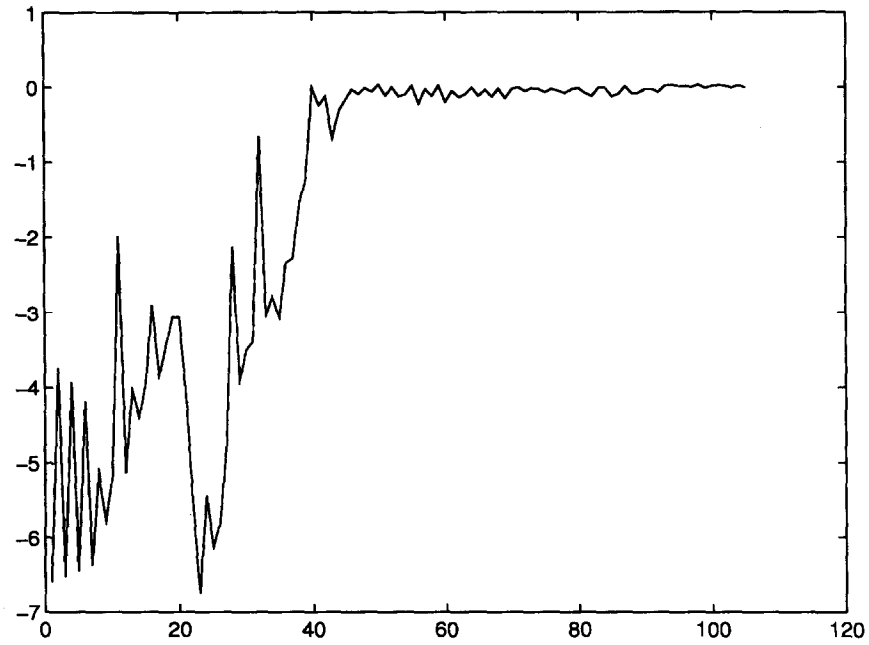

FIG. 2. $\log \gamma$; negligible eigenvector entries from entry 42 to 105.

THEOREM 12 (Davis and Kahan). *Let $A = A^*$ have an isolated eigenvalue λ with normalized eigenvector $\mathbf{v}$. Consider $\mathbf{y}$, $\mathbf{y}^*\mathbf{y} = 1$, and real σ closer to λ than to any other eigenvalue. Then*

$$|\sin \angle(\mathbf{v}, \mathbf{y})| \leqslant \frac{\|A\mathbf{y} - \mathbf{y}\sigma\|_2}{\operatorname{gap}(\sigma)}$$

where $\operatorname{gap}(\sigma) = \min\{|\mu - \sigma| : \mu \neq \lambda,\ \mu \in \operatorname{spectrum}(A)\}$.

This result also applies to skew-Hermitian matrices since $S^* = -S$ implies that $(iS) = (iS)^*$, $i^2 = -1$. So, for those $\tilde{J}$ that are Hermitian or skew-Hermitian, Theorem 12 with $\mathbf{y} = \mathbf{z}^{(k)}/\|\mathbf{z}^{(k)}\|$ yields

$$\begin{aligned} |\sin \angle(\mathbf{v}, \mathbf{z}^{(k)})| &\leqslant \frac{\|(\tilde{J} - \sigma I)\mathbf{z}^{(k)}\|}{\|\mathbf{z}^{(k)}\|_2 \operatorname{gap}(\sigma)} \\ &= \frac{1}{\operatorname{gap}(\sigma)} \cdot \frac{|\gamma_k|}{\|\mathbf{z}^{(k)}\|_2} \end{aligned} \tag{28}$$

$$\leqslant \frac{|\sigma - \lambda|}{\text{gap}(\sigma)} \cdot \frac{1}{|v(k)|}. \tag{29}$$

The last inequality follows from Lemma 13, given below. From (29) for all k such that $|v(k)| = \|\mathbf{v}\|_\infty$

$$|\sin \angle(\mathbf{v}, \mathbf{z}^{(k)})| \leqslant \frac{|\sigma - \lambda|}{\text{gap}(\sigma) \cdot \|\mathbf{v}\|_\infty}. \tag{30}$$

These bounds will be used later.

The next result replaces the limits of Lemma 11 with error bounds. Note that the eigenvalues may be complex.

LEMMA 13. *Let $\bar{J} - \sigma I$ be normal and satisfy the hypotheses of Theorem 2. Define γ_k and $\mathbf{z}^{(k)}$ by $(\bar{J} - \sigma I)\mathbf{z}^{(k)} = \mathbf{e}_k \gamma_k$, $z^{(k)}(k) = 1$. If λ_j is a simple eigenvalue of $\bar{J}$ and*

$$|\sigma - \lambda_j| < |\sigma - \lambda_i|, \qquad i \neq j,$$

then

$$\frac{|\gamma_k|}{\|\mathbf{z}^{(k)}\|} = \frac{|\sigma - \lambda_j|}{|v_j(k)|}\left[1 + \left(|v_j(k)|^{-2} - 1\right)\mathscr{A}_2\right]^{-1/2} \qquad \text{for all admissible } k$$

$$\leqslant \frac{|\sigma - \lambda_j|}{\|\mathbf{v}_j\|_\infty} \leqslant \sqrt{n}\,|\sigma - \lambda_j| \qquad \text{for at least one } k,$$

$$\gamma_k = \frac{\lambda_j - \sigma}{|v_j(k)|^2}\left[1 + \left(|v_j(k)|^{-2} - 1\right)\mathscr{A}_1\right]^{-1}.$$

Here $\mathscr{A}_2$ is a weighted arithmetic mean of $\{|(\sigma - \lambda_j)/(\sigma - \lambda_i)|^2,\ i \neq j\}$, $0 < \mathscr{A}_2 < [|\lambda_j - \sigma|/\text{gap}(\sigma)]^2$, and $\mathscr{A}_1$ is the same weighted arithmetic mean of $\{\lambda_j - \sigma/\lambda_i - \sigma,\ i \neq j\}$, $0 \leqslant |\mathscr{A}_1| < \lambda_j - \sigma|/\text{gap}(\sigma)$, where $\text{gap}(\sigma) = \min_{i \neq j} |\lambda_i - \sigma|$.

Proof. For simplicity write $\mathbf{z} = \mathbf{z}^{(k)}$, $V = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ and use $\mathbf{z}$'s eigenvector expansion to find

$$\mathbf{z} = (\tilde{J} - \sigma I)^{-1}\mathbf{e}_k\gamma_k,$$

$$\|\mathbf{z}\|^2 = |\gamma_k|^2\mathbf{e}_k^* V(\bar{\Lambda} - \bar{\sigma} I)^{-1}(\Lambda - \sigma I)^{-1}V^*\mathbf{e}_k$$

$$= |\gamma_k|^2 \sum_{i=1}^{n} \frac{|v_i(k)|^2}{|\sigma - \lambda_i|^2},$$

$$\gamma_k^{-1} = \mathbf{e}_k^*(\tilde{J} - \sigma I)^{-1}\mathbf{e}_k,$$

$$= \sum_{i=1}^{n} \frac{|v_i(k)|^2}{\lambda_i - \sigma}.$$

Extract the jth term to find

$$\left(\frac{\|\mathbf{z}\|}{|\gamma_k|}\right)^2 = \left(\frac{|v_j(k)|}{|\sigma - \lambda_j|}\right)^2 \left(1 + \sum_{i \neq j} \left|\frac{v_i(k)}{v_j(k)}\right|^2 \left|\frac{\sigma - \lambda_j}{\sigma - \lambda_i}\right|^2\right), \tag{31}$$

$$\gamma_k^{-1} = \frac{|v_j(k)|^2}{\lambda_j - \sigma}\left(1 + \sum_{i \neq j} \left|\frac{v_i(k)}{v_j(k)}\right|^2 \frac{\lambda_j - \sigma}{\lambda_i - \sigma}\right). \tag{32}$$

Since

$$\sum_{i \neq j} \left|\frac{v_i(k)}{v_j(k)}\right|^2 = \frac{1 - |v_j(k)|^2}{|v_j(k)|^2},$$

the $\Sigma_{i \neq j}$ terms in (31) and (32) may be written as

$$\left(|v_j(k)|^{-2} - 1\right)\mathscr{A}_2, \qquad 0 < \mathscr{A}_2 < \left(\frac{|\lambda_j - \sigma|}{\text{gap}(\sigma)}\right)^2,$$

$$\left(|v_j(k)|^{-2} - 1\right)\mathscr{A}_1, \qquad |\mathscr{A}_1| \leqslant \frac{|\lambda_j - \sigma|}{\text{gap}(\sigma)},$$

where

$$\mathscr{A}_2 = \sum_{i \neq j} w_i \left| \frac{\sigma - \lambda_j}{\sigma - \lambda_i} \right|^2, \qquad 1 = \sum_{i \neq j} w_i, \quad w_i \geqslant 0,$$

$$\mathscr{A}_1 = \sum_{i \neq j} w_i \frac{\lambda_j - \sigma}{\lambda_i - \sigma}, \qquad \text{same } w_i, \quad \text{gap}(\sigma) = \min_{i \neq j} |\lambda_i - \sigma|,$$

and yield the equalities in Lemma 13. For all γ_k with $\gamma_k \neq 0$, (31) yields

$$\frac{|\gamma_k|}{\|\mathbf{z}^{(k)}\|} \leqslant \frac{|\sigma - \lambda_j|}{|v_j(k)|}.$$

Thus for all k with $|v_j(k)|$ above average ($\geqslant n^{-1/2}$),

$$\frac{|\gamma_k|}{\|\mathbf{z}^{(k)}\|} \leqslant \sqrt{n}\,|\sigma - \lambda_j|$$

but the best choices of k have $|v_j(k)| = \|\mathbf{v}_j\|_\infty$. ∎

Neither $\mathbf{z}^{(k)}$ nor $\mathbf{v}$ is known at the stage when the $\{\gamma_i\}$ have just been computed. Although $|\lambda - \sigma|$ is not known precisely, the bisection technique yields tight bounds, and we expect to have $|\lambda - \sigma| = O(\text{macheps} \cdot |\lambda|)$. Moreover, Lemma 13 shows that if

$$\frac{|\lambda_j - \sigma|}{\text{gap}(\sigma)} \leqslant \frac{1}{2(n-1)}, \tag{33}$$

then γ_k is within 50% of the value $(\lambda_j - \sigma)/|v_j(k)|^2$ when $|v_j(k)|$ is above average. Suppose that m is the minimizing argument for $\{|\gamma_k|\}$ but $|v_j(m)| < |v_j(r)| = \|\mathbf{v}_j\|_\infty$. Then

$$\left(1 - \tfrac{1}{2}\right)\frac{|\lambda_j - \sigma|}{|v_j(m)|^2} \leqslant |\gamma_m| \leqslant |\gamma_r| \leqslant \left(1 + \tfrac{1}{2}\right)\frac{|\lambda_j - \sigma|}{|v_j(r)|^2},$$

so that $|v_j(m)| \geqslant (1/\sqrt{3})\|\mathbf{v}\|_\infty$, and this is acceptable. Thus (33) is a reasonable requirement for calling an eigenvalue isolated.

When (33) holds, then (29) yields

$$|\sin \angle(\mathbf{v}, \mathbf{z}^{(j)})| \leqslant \frac{|\lambda - \sigma|}{\text{gap}(\sigma)} \cdot \frac{\sqrt{3}}{\|\mathbf{v}\|_\infty} < \sqrt{3n}\,\frac{|\lambda - \sigma|}{\text{gap}(\sigma)}$$

and we are going to use $\mathbf{z}^{(j)}$ only when $|\lambda - \sigma|/\text{gap}(\sigma)$ is $O(\varepsilon)$. We have noted, and so has Jesse Barlow [3], that a simple recurrence will yield all values of $\|\mathbf{z}^{(k)}\|$ for $O(n)$ operations. Consequently it would be feasible to minimize $|\gamma_k|/\|\mathbf{z}^{(k)}\|$ to obtain better bounds on $\sin \angle(\mathbf{v}, \mathbf{z}^{(k)})$: (28) $\leqslant$ (29). At present we feel that the extra expense is not warranted.

6. ZERO PIVOTS

Triangular factorization is said to fail, or not exist, if a zero pivot $D_+(j)$ or $D_-(j)$ is encountered prematurely. The last pivot is allowed to vanish, because it does not occur as a denominator in the computation.

One of the attractions of an unreduced tridiagonal matrix is that the damage done by a zero pivot is localized. Indeed, if ∞ is added to the number system, then triangular factorization cannot break down and the algorithm always maps J into unique triplets L, D, U. There is no need to spoil the inner loop with tests. It is no longer true that $LDU = J$, but equality does hold for all entries except for those at or adjacent to any infinite pivot.

It is possible to work with signed ∞ (affine geometry) or unsigned ∞ (the complex plane); it will be easiest for our purposes to use the unsigned ∞. Thus $+1/0 = -1/0 = \infty$.

If we allowed off diagonal entries to vanish, in which case J is said to be *reduced*, then we might encounter

$$L(k+1,k) = \frac{J(k+1,k)}{D(k)} = \frac{0}{0},$$

and that would be a genuine breakdown. So we insist on unreduced J.

Let us examine what happens when $D(k-1) = 0$. In turn

$$L(k,k-1) = \frac{J(k,k-1)}{D(k-1)} = \infty,$$

$$U(k-1,k) = \frac{J(k-1,k)}{D(k-1)} = \infty,$$

$$D(k) = J(k,k) - L(k,k-1)J(k-1,k) = \infty,$$

$$L(k+1,k) = \frac{J(k+1,k)}{D(k)} = 0,$$

$$\bullet \quad U(k,k+1) = \frac{J(k,k+1)}{D(k)} = 0,$$

$$D(k+1) = J(k+1,k+1) - L(k+1,k)J(k,k+1)$$
$$= J(k+1,k+1).$$

Unless $J(k+1,k+1) = 0$ the factorization proceeds normally until the next zero pivot is encountered. We have placed $\bullet$ against entries that are not computed when a simple $L\tilde{U}$ factorization is used. Here $\tilde{U} = DU$ in the finite case.

When the product LDU is formed in the case given above, then various strange expressions such as $0 \cdot \infty$ and $\infty + \infty$ arise, and we designate them by NaN (not a number). We discover that $LDU = J$ except in row and column k. Note that $D(k) = \infty$.

It is important for our later results to show that when J is singular then

$$D_+(k) = \infty \quad \text{if, and only if,} \quad D_-(k) = \infty,$$

where the notation follows Section 3.

It turns out that infinite pivots correspond to zero entries in eigenvectors and so have a legitimate role in the theory.

THEOREM 14. *Let J be $n \times n$, tridiagonal, unreduced, and singular. For each k, $1 < k < n$, $J^{1:k-1}$ is singular if, and only if, $J^{k+1:n}$ is singular. They are singular if, and only if, $z(k) = 0$ whenever $J\mathbf{z} = \mathbf{0}$.*

Proof. Write

$$\mathbf{z} = \begin{pmatrix} \mathbf{z}_+ \\ z(k) \\ \mathbf{z}_- \end{pmatrix},$$

and partition $J\mathbf{z} = \mathbf{0}$ conformably. Thus

$$J^{1:k-1}\mathbf{z}_+ + J_{k-1,k}z(k)\mathbf{e}_{k-1} = \mathbf{0}, \tag{34}$$

$$\mathbf{e}_1 J_{k+1,k} z(k) + J^{k+1:n}\mathbf{z}_- = \mathbf{0}, \tag{35}$$

and $\mathbf{z}_+(1) \neq 0$, $\mathbf{z}_+(n) \neq 0$ by Lemma 1 in Section 2.

If $z(k) = 0$, then (34) shows that $\mathbf{z}_+$ $(\neq 0)$ is in $J^{1:k-1}$'s null space, and (35) shows that $\mathbf{z}_-$ $(\neq 0)$ is in $J^{k+1:n}$'s null space. So both matrices are singular.

Now consider the converse, $z(k) \neq 0$. Since J is unreduced, rank $J = n - 1$ and its null space is one dimensional. So the system

$$J\mathbf{z} = \mathbf{0}, \qquad z(k) = 1$$

has a unique solution. Thus both (34) and (35) are inhomogeneous equations with unique solutions. Thus $J^{1:k-1}$ and $J^{k+1:n}$ are invertible. ■

COROLLARY 15. *Let J be $n \times n$, tridiagonal, unreduced, and singular. Let the triangular factorization algorithm applied to J in both increasing and decreasing order of rows yield unique matrices L_+, D_+, U_+, U_-, D_-, and L_-. Then, for $j = 1, 2, \ldots, n$,*

$$D_+(j) = \infty \quad \textit{iff} \quad D_-(j) = \infty.$$

Proof.

$$\begin{aligned} D_+(j) = \infty \quad &\Leftrightarrow \quad D_+(j-1) = 0 \\ &\Leftrightarrow \quad J^{1:j-1} \quad \text{singular} \\ &\Leftrightarrow \quad J^{j+1:n} \quad \text{singular (by Theorem 14)} \\ &\Leftrightarrow \quad D_-(j+1) = 0 \\ &\Leftrightarrow \quad D_-(j) = \infty. \end{aligned}$$

■

In Theorem 2 of Section 3 the value of γ_k was fixed by the condition $z(k) = 1$ imposed on the solution of $J\mathbf{z} = \mathbf{e}_k\gamma_k$. When J is singular there is a nonzero solution to $J\mathbf{z} = \mathbf{0}$ and the attempted normalization $z(k) = 1$ is valid, even if not wise, in all cases *except when* $z(k) = 0$.

An appropriate signal that an infeasible normalization has been imposed is that $\gamma_k = \text{NaN}$, and that is precisely what the formulae in Corollary 4 deliver whenever J is singular and $D_+(k) = D_-(k) = \infty$. In these cases, in exact arithmetic, $D_+(n) = 0$ and $\gamma_n = 0$ as well as $D_-(1) = 0$ and $\gamma_1 = 0$. Thus, in the search for a minimum value of $|\gamma_j|$, indices j that have $\gamma_j = \text{NaN}$ will never be selected.

The good news is that by computing all the $\{\gamma_j\}$ it is known *in advance* whether or not $\mathbf{z}^{(k)}$ has a zero entry. In the generic case, with no zeros, the algorithm given in Corollary 3 in Section 3 may be used free of any tests for invalid operations. In the exceptional case the following procedure may be used.

ALGORITHM 1 (Vectors with zeros).

$$z(j) = \begin{cases} -U^+_{j,j+1} z(j+1), & z(j+1) \neq 0 \\ -\dfrac{J_{j+1,j+2} z(j+2)}{J_{j+1,j}} & \text{otherwise} \end{cases}, \qquad j = k-1, \ldots, 1$$

$$z(i) = \begin{cases} -L^-_{i,i-1} z(i-1), & z(i-1) \neq 0 \\ -\dfrac{J_{i-1,i-2} z(i-2)}{J_{i-1,i}} & \text{otherwise} \end{cases}, \qquad i = k+1, \ldots, n.$$

This algorithm will not touch any infinite values in U_+ or L_-.

When J is not singular, Theorem 2 continues to hold, if ∞ is allowed, in the following sense.

COROLLARY 16 (of Theorem 6). *If J is unreduced and tridiagonal, and ∞ is represented, then*

$$D_+(k-1) = 0 \quad \textit{implies} \quad (J^{-1})_{kk} = 0 \quad (\gamma_k = \infty),$$

$$D_-(k) = 0 \quad \textit{implies} \quad (J^{-1})_{k-1,k-1} = 0 \quad (\gamma_{k-1} = \infty).$$

Proof. Use the twisted factorization in the proof of Theorem 6 that introduces the 2×2 matrix

$$\Delta = \begin{bmatrix} D_+(k-1) & J_{k-1,k} \\ J_{k,k-1} & D_-(k) \end{bmatrix}.$$

Invert J, and observe that there is a simple expression for the (k,k) and $(k-1, k-1)$ entries:

$$(J^{-1})_{kk} = (\Delta^{-1})_{2,2} = \frac{D_+(k-1)}{\det \Delta}.$$

If J is unreduced and $D_+(k-1) = 0$, then $\det \Delta = -J_{k,k}J_{k-1,k} \neq 0$. This establishes the first assertion. Similarly

$$(J^{-1})_{k-1,k-1} = \frac{D_-(k)}{\det \Delta}. \qquad \blacksquare$$

Figures 3 and 4 show a striking instance of Theorem 14 for the matrix W_{21}^+ and the pair of eigenvalues close to 6. Each horizontal line of the figure corresponds to one value of k; eigenvalues of $W^{1:k-1}$ are marked by +, and eigenvalues of $W^{k+1:n}$ are marked by ○. Theorem 14 implies that if an eigenvector has a zero entry in position k, then a ○ and a + must coincide on the eigenvalue in line k. Indeed, in Figure 4 (an enlarged picture of Figure 3 near 6), when $k = 11$ this is precisely what happens. For neighboring values of k the Ritz values are not particularly close to eigenvalues, and after $k = 11$ the ○ is replaced by a + in the interval $(\lambda_{12}, \lambda_{13})$. If $\mathbf{v}$ is a normalized eigenvector with eigenvalue λ, then $v(k)^2$ is proportional to the product of the distances of λ from the + and ○ points on line k.

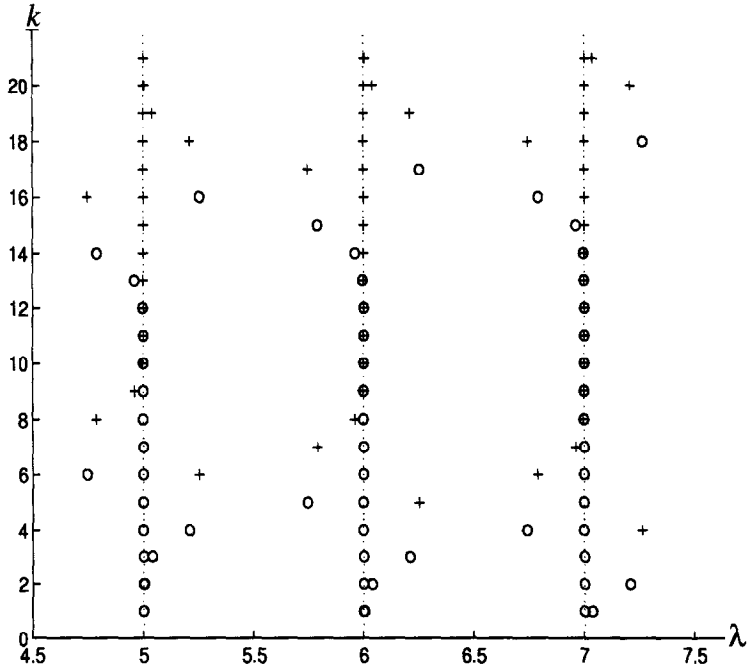

FIG. 3. Ritz values of W_{21}^+ for λ near 6.

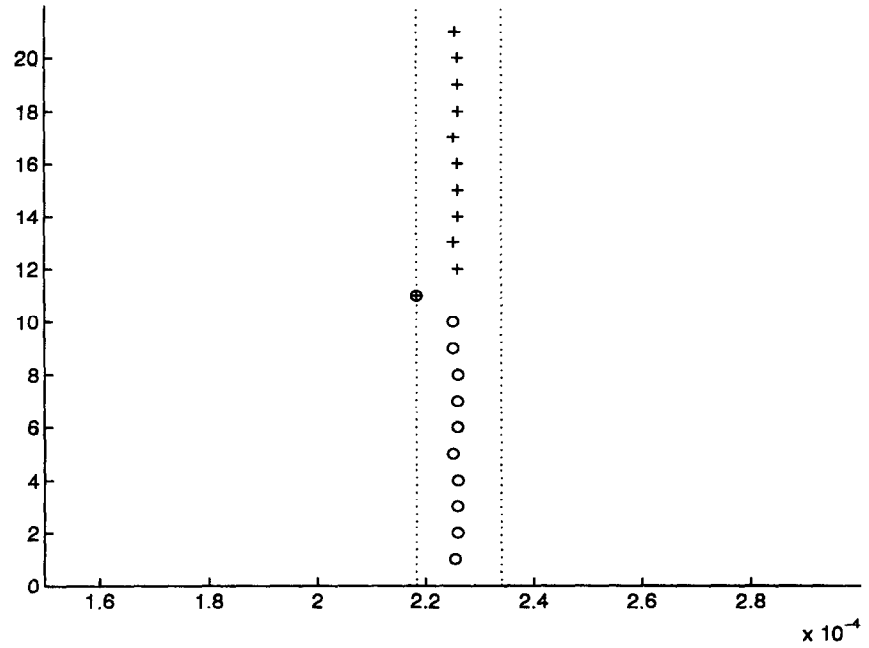

FIG. 4. Blowup of Figure 2 near 6 (read the x-axis as $6 + x$).

7. BLOCK TRIDIAGONALS

If an arithmetic system lacks the symbol ∞, it is possible to extend Theorem 2 by using blocks in the LDU factorization. If J is unreduced, then there always exists a factorization

$$L_+D_+U_+ = J = U_-D_-L_-$$

if the D's are allowed to have 2×2 and 1×1 blocks along diagonal; no larger blocks are needed. However, Theorem 2 extends beyond this case to band matrices and to any block tridiagonal matrix. Thus D_+ and D_- are direct sums of square blocks; L_+ and U_+ conform with D_+, and L_- and U_- conform with D_-.

THEOREM 17. *Let square J permit block triangular factorization in both increasing and decreasing order of indices*

$$L_+D_+U_+ = J = U_-D_-L_-.$$

There is no requirement that the block structures of D_+ and D_- be conformable. However, for any corresponding blocks k and l such that $D_+(k)$ and $D_-(l)$ are conformable and m by m, define the $m \times m$ matrix Γ by equations

$$J\begin{pmatrix} Z^+ \\ I_m \\ Z^- \end{pmatrix} = \begin{pmatrix} O \\ \Gamma \\ O \end{pmatrix}, \qquad Z = \begin{pmatrix} Z^+ \\ I_m \\ Z^- \end{pmatrix}. \tag{36}$$

If $J_{.,.}$ denotes the $m \times m$ block of J conformable with $D_+(k)$ and $D_-(l)$, then

$$\Gamma = D_+(k) - J_{.,.} + D_-(l).$$

The proof is so similar to the proof of Thoerem 2 that we omit it. We have allowed for the fact that D_+ and D_- need not have the same number of blocks.

To use Theorem 17 to approximate an eigenvector suppose that J is nearly singular. Compute all well-defined Γ, and find one with a minimal singular value. Call it $\mathring{\Gamma}$. Let

$$\mathring{\Gamma}\mathbf{v} = \mathbf{u}\sigma_{\min}, \qquad \|\mathbf{u}\| = \|\mathbf{v}\| = 1,$$

define the minimal singular triple $(\sigma_{\min}, \mathbf{u}, \mathbf{v})$ of $\mathring{\Gamma}$. Then, from (36)

$$J(Z\mathbf{v}) = \begin{pmatrix} \mathbf{0} \\ \mathbf{u} \\ \mathbf{0} \end{pmatrix} \sigma_{\min}.$$

If $\sigma_{\min}$ is small enough, then $Z\mathbf{v}$ is a good initial approximation to an eigenvector of J.

It is not hard to verify that, for an unreduced even order J, if $\operatorname{diag}(J) = 0$ then $\operatorname{diag}(J^{-1}) = 0$. In this situation a block factorization with 2×2 blocks is needed to ensure that $J = L_+D_+U_+ = U_-D_-L_-$. It then turns out that

$$D_+ = D_- = \text{block diag}(J) = \text{block diag}(J^{-1})^{-1}.$$

In this situation the set of Γ-matrices in Theorem 17 may give no guidance for computing an eigenvector. That is not quite true, because we may infer

that our eigenvalue approximation, 0, is not closer to one eigenvalue than to any other, and that is useful information. In fact, the unreduced J's with $\mathrm{diag}(J) = 0$ have eigenvalues in $\pm$ pairs, and any tiny $\pm$ pairs may be found efficiently by the method described in [22].

We are glad to acknowledge helpful discussions with the LAPACK group at UC, Berkeley, in particular Ming Gu and Velvel Kahan. We thank Gene Golub for the Babuska reference, and Vince Fernando for the references to Crandall and Strang, Henrici, and Holzer. Finally we thank A. Malyshev for drawing our attention to the work in Russia [12, 13].

REFERENCES

1 E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK Users' Guide, Release 2.0*, SIAM, Philadelphia, 1995.

2 I. Babuska, Numerical stability in problems of linear algebra, *SIAM J. Numer. Anal.* 9:53–77 (1972).

3 J. Barlow, Private communication, 1996.

4 S. H. Crandall and W. G. Strang, An improvement of the Holzer table based on a suggestion of Rayleigh's, *J. Appl. Mech.* 24:228–230 (1957); Discussion, *J. Appl. Mech.* 25:160–161 (1958).

5 I. S. Dhillon, Reliable computation of the condition number of a tridiagonal matrix in $O(n)$ time, *SIAM J. Matrix Anal. Appl.*, to appear.

6 J. J. Dongarra et al., *LINPACK*, SIAM, Philadelphia, 1979.

7 D. Fischer, G. H. Golub, O. Hald, C. Leiva, and O. Widlund, On Fourier Toeplitz methods for separable elliptic problems, *Math. Comp.* 28:349–368 (1974).

8 K. V. Fernando, On computing an eigenvector of a tridiagonal matrix, Part 1: Basic Results, *SIAM J. Matrix Anal. Appl.*, to appear.

9 K. V. Fernando, Accurate BABE Factorization of Tridiagonal Matrices for Eigenproblems, Technical Report TR5/95, NAG Ltd., U.K., 1996.

10 W. Givens, Numerical computation of the characteristic values of a real symmetric matrix, Report ORNL-1574, Oak Ridge National Lab., 1954.

11 W. Givens, The characteristic value–vector problem, *J. Assoc. Comput. Mach.* 4:298–307 (1957).

12 S. K. Godunov, V. I. Kostin, and A. D. Mitchenko, Computation of an eigenvector of symmetric tridiagonal matrices, *Siberian Math. J.* 26:71–85 (1985).

13 S. K. Godunov, A. G. Antonov, O. P. Kiriljuk, and V. I. Kostin, *Guaranteed Accuracy in Numerical Linear Algebra*, Kluwer Academic, 1993. (A revised translation of a Russian text first published in 1988 in Novosibirsk.)

14 P. Henrici, Bounds for eigenvalues of certain tridiagonal matrices, *SIAM J.*, 11:281–290 (1963).

15 N. J. Higham, Efficient algorithms for computing the condition number of a tridiagonal matrix, *SIAM J. Sci. Statist. Comput.* 7:150–165 (1986).
16 H. Holzer, *Die Berechnung der Drehschwingungen*, Springer-Verlag, Berlin, 1921.
17 I. Ipsen, A history of inverse iteration, in *Helmut Wielandt, Mathematische Werke, Vol. II: Matrix Theory and Analysis* (B. Huppert and H. Schneider, Eds.), Walter de Guyter, Berlin, 1995.
18 E. Jessup and I. Ipsen, Improving the accuracy of inverse iteration, *SIAM J. Sci. Statist. Comput.* 13:570–572 (1992).
19 D. Kershaw, Inequalities on elements of the inverse of a certain tridiagonal matrix, *Math. Comp.* 24:155–158 (1970).
20 G. Meurant, A review on the inverse of symmetric tridiagonal and block tridiagonal matrices, *SIAM J. Matrix Anal. Appl.* 13:707–728 (1992).
21 B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, N.J., 1980.
22 B. N. Parlett, Invariant subspaces for tightly clustered eigenvalues of tridiagonals, *BIT* 36:542–562 (1996).
23 R. V. Southwell, *Relaxation Methods in Engineering Science*, Clarendon, Oxford, 1940.
24 W. G. Strang, Implicit difference methods for initial boundary value problems, *J. Math. Anal. Appl.* 16:188–198 (1966).
25 H. A. van der Vorst, Analysis of a parallel solution method for tridiagonal linear systems, *Parallel Computing* 5:303–311 (1987).
26 J. H. Wilkinson, The calculation of the eigenvectors of codiagonal matrices, *Computer J.* 1:90–96 (1958).
27 J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon, Oxford, 1965.

Received 17 September 1995; final manuscript accepted 17 February 1997