RESIDUAL BOUNDS ON APPROXIMATE EIGENSYSTEMS OF NONNORMAL MATRICES

by

B. N. Parlett and E. Jiang

Memorandum No. UCB/ERL M80/39

5 September 1980

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Residual Bounds on Approximate Eigensystems of Nonnormal Matrices

B. N. Parlett[1] and E. Jiang[2]

1. Math. Dept. and Comp. Sci. Div., U. C. Berkeley, Berkeley, CA 94720

2. Math. Dept., Fudan Univ., Shanghai, China; visiting scholar at U. C. Berkeley.

The first author gratefully acknowledges support by the Office of Naval Research Contract N00014-76-C-0013.

1. INTRODUCTION

This paper is concerned with the following question and its ramifications. The goal is to compute some or all of the eigenvalues of a square matrix B which is not symmetric. There is on hand an approximate column eigenvector x, an approximate row eigenvector y^*, and a number γ. In addition someone has computed the norms of their residual vectors, $\|r\|$ and $\|s^*\|$, where

$$r \equiv Bx - x\gamma, \quad s^* \equiv y^*B - \gamma y^* \ .$$

It turns out that $\|r\|/\|x\| \leq 10^{-5}\ \|B\|$ and $\|s^*\|/\|y^*\| \leq 10^{-6}\ \|B\|$. How good is γ as an approximate eigenvalue of B?

The experts know that nothing very useful can be said. A priori bounds on $|\lambda-\gamma|$ are known (see Section 2), where λ is some eignenvalue of B. Moreover these bounds are best possible, in the sense that equality can be achieved, but either they involve auxiliary terms, such as the matrix which transforms B into its Jordan canonical form, and consequently are virtually impossible to compute, or else they are hopelessly pessimistic in the majority of cases.

This is in sad contrast to the symmetric case where various well known error bounds enjoy three remarkable properties:

(α) they are best possible inferences from the information,

(β) they are computable,

(γ) they are not asymptotic.

By (γ) we mean that the error bounds are not based on perturbation theory where the neglect of certain complicated terms is justified only when the error is "small enough." No, the bounds to which we refer are valid however bad the best approximations may be.

When approximations to several different eigenvectors are known then, again in the symmetric case, the Rayleigh-Ritz procedure tells us how to make the best approximations to eigenvalues and tells us what residuals to compute in order to have error bounds for these approximations.

This theory is too useful to be abandoned and indeed the Ritz procedure is readily generalized as the Galerkin approximation. What is not discussed (to our knowledge) is whether the Galerkin approximations are optimal in any useful sense.

It seems that the only way to extend the optimality properties is by making a radical change in the notion of error. Experience shows that the first time readers are exposed to this viewpoint they often feel that it dodges real difficulties. It does. Rather than scale the vertical face of a mountain some people prefer to take an easier indirect route round the back which quite often leads to the peak.

In the context of the problem addressed at the outset the idea is as follows.

> Ask not for the value of $|\lambda-\gamma|$ but instead ask for $\|B-B'\|$ where B' is the closest matrix to B for which γ is an eigenvalue and x and y^* are its eigenvectors.

This idea has been used with great success by J. H. Wilkinson in a variety of matrix problems and is often called a "backward" error analysis because it casts the errors back as an equivalent change to the original data. If B is not known exactly it sometimes happens that B' is indistinguishable from B and then γ is as good an answer as the data warrants.

Once this change has been made it is not very difficult to see in what way the approximations are optimal but, to the best of our knowledge, the full extension of Rayleigh-Ritz has not been made. A possible reason for

this omission is the fact that only recently have there been serious efforts to compute some eigenvalues of large (n>1000) nonnormal matrices.† It was our desire to find a good stopping criterion for the unsymmetric Lanczos process which prompted this work. All expressions must be computable. Our goal (with apologies to Richard Hamming) is numbers, not insight.

Last but not least the error expressions of § 4 can be combined with computable condition numbers to give (exactly) the first term in the perturbation series for $|\lambda-\gamma|$ in powers of $\|B-B'\|$.

In §2 we sketch the theory we want to generalize and exhibit some traditional error bounds for the nonnormal case. In §3 we collect important preliminary facts. In §4 comes the main result, and in §5 the application to the Lanczos method. We follow Householder conventions in notation: capital letters for matrices, lower case roman letters for column vectors, and lower case greek letters for scalars. However x^* (not x^H) denotes the conjugate transpose of x, and $\lambda_i(B)$ denotes the ith eigenvalue of B for any given ordering. The norms we use are defined in Section 3. One idiosyncracy is to use symmetric letters (A,H,M,...) for symmetric (or Hermitian) matrices.

2. SOME ERROR BOUNDS

Symmetric case

Theorem 1. For any nonzero vector x and any scalar γ there is a an eigenvalue of A such that

$$|\lambda-\gamma| \leq \|Ax-x\gamma\|/\|x\|.$$

†For small matrices there are alternative techniques based on skillfully chosen Gersgorin disks, [Wilkinson, 1965].

Theorem 2. For any nonzero vector x and $\rho \equiv x^*Ax/x^*x$,

$$\|Ax-x\rho\| = \min_{\gamma} \|Ax-x\gamma\| .$$

Theorem 3. (Rayleigh-Ritz). Given any nxk Z satisfying $Z^*Z = I_k$ the best set of k approximate eigenvalues of A that can be derived from A is the spectrum of Z*AZ. The choice H = Z*AZ uniquely minimizes, over all kxk matrices H the residual norm

$$\|AZ-ZH\|_F .$$

If G is the eigenvector matrix of Z*AZ then the columns of ZG are the best set of approximate eigenvectors. $\|C\|_F^2 \equiv$ trace (C*C).

Theorem 4. (Error Bounds). There are k eigenvalues of A, call them $\lambda_1,\ldots,\lambda_k$, which can be put into one-one correspondence with the k eigenvalues $\theta_1,\ldots,\theta_k$ of any symmetric kxk matrix H so that, given Z,

$$\max|\lambda_i-\theta_i| \leq \|AZ-ZH\|,$$

$$\sum_{i=1}^{k} (\lambda_i-\theta_i)^2 \leq \|AZ-ZH\|_F^2 .$$

Discussion of these classical results, and proofs, can be found in [Parlett, 1980 Chaps. 4 and 11]. More refined results than these are known and some of them may be found in [Kato, 1966] and [Parlett, 1980].

We list an important residual bound which gives considerable insight but is not readily computable.

Theorem 5. If λ is the eigenvalues of A closest to $\rho(x)$ ($\equiv x^*Ax/x^*x$) and δ is the separation of ρ from the next closest eigenvalue then

$$|\lambda-\rho| \leq (\|Ax-x\rho\|/\|x\|)^2/\delta .$$

If λ is simple and z is its eigenvector then

$$|\sin \angle (x,z)| \leq \|Ax-x\rho\|/(\|x\| \delta) .$$

The General Case

We begin with an example which shows that the residual norms of the trio (γ,x,y^*) with respect to B can be tiny despite a large gap separating γ from the spectrum of B.

$$B = (b_{ij})_1^n,\ b_{ij} = \begin{cases} 0, & i \geq j, \\ 1, & i < j\ . \end{cases}$$

$\|B\|_F \approx n/\sqrt{2}$, B's eigenvalues are all 0.

$x^* = (2^{n-2},2^{n-3},\ldots,2,1,1)$, $\|x\|^2 = (4^{n-1}+2)/3$

$y^* = (1,1,2,\ldots,2^{n-3},2^{n-2})$, $\|y\| = \|x\|$.

$Bx-x\cdot 1 = (0,\ldots,0,-1)^*$, $y^*B-1\cdot y^* = (-1,0,\ldots,0)$.

Thus $\|Bx-x\cdot 1\|/\|x\| = \|y^*B-1\cdot y^*\|/\|y^*\| \approx \sqrt{3}/2^{n-1}$,

yet $\min_\lambda |\lambda-\gamma|/\|B\|_F = 1/\|B\|_F \approx \sqrt{2}/n$.

The reader is invited to concoct a similar, but somewhat more complicated example in which the eigenvalues are evenly spaced. Indeed it is not the fact that B above has a Jordan block of order n which causes the phenomenon, rather it is B's excessive departure from normality which is responsible. (A matrix is normal if $BB^* = B^*B$.) Sometimes, but not always, departure from normality is explained by proximity (in matrix space) to a matrix with large Jordan blocks.

With one exception we have found no computable error bounds for the eigenvalues of nonnormal matrices which explicitly involve residual norms. Furthermore, in [Householder, 1972] we read "To the best of my knowledge no one has obtained inclusion theorems that apply to a matrix having nonlinear elementary divisors." An inclusion theorem describes a region of the complex plane guaranteed to contain an eigenvalue (but not all the eigenvalues). So this quotation makes our quest for computable inclusion theorems look hopeless.

Faced with this impasse we give some results on spectral variation and these point up the difficulties. Actually residual norms can be linked to changes in the matrix to good advantage, as is done in Section 4.

Theorem 6 [Ostrowski, 1957]

Let $C = B-E$ with $\mu = \max_{i,j}\{|b_{ij}|,|c_{ij}|\}$, $\delta = \sum_{i=1}^{n}\sum_{j=1}^{n}|e_{ij}|/\mu$. Then

1. To each $\lambda_i(C)$ there is a $\lambda_j(B)$ such that $|\lambda_i(C)-\lambda_j(B)| \leq (n+2)\mu\delta^{1/n}$
2. The eigenvalues of B and C can be paired so that $|\lambda_i(B)-\lambda_i(C)| \leq 2(n+1)^2\mu\delta^{1/n}$, $i = 1,\ldots,n$.

Remark. The nth root of δ must be there but it does dampen one's enthusiasm for a priori error bounds, especially when $n > 100$.

Theorem 7. [Bauer and Fike, 1960]

If C $(=B-E)$ is diagonalizable (= semi-simple), i.e., $C = F\Lambda F^{-1}$, then to each $\lambda_i(B)$ there is a $\lambda_j(C)$ such that

$$|\lambda_i(B)-\lambda_j(C)| \leq \|FEF^{-1}\| \leq \|F\|\,\|F^{-1}\|\,\|E\|,$$

using any matrix norm for which $\|\Lambda\| = \max_i |\lambda_i|$.

Remark. If C is normal then $\|F\| = \|F^{-1}\| = 1$ and the bound is as good as in the symmetric case.

Next we give an extension of the Bauer-Fike Theorem to the general case.

Theorem 8. Let $C = B-E$ and let $J = FBF^{-1}$ be B's Jordan form. To any eigenvalue μ of C there corresponds an eigenvalue λ of B such that

$$\frac{|\lambda-\mu|^m}{(1+|\lambda-\mu|)^{m-1}} \leq \|FEF^{-1}\|$$

where m is the order of the largest Jordan block to which λ belongs. The spectral (or the Frobenius) norm must be used here.

Since this bound is not computable we have relegated the proof to an appendix. It is included because, to the best of our knowledge, it has not appeared before.

COROLLARY. Let $\omega(m,\varepsilon)$ be the nonnegative solution to $\xi^m-\varepsilon(\xi^{m-1}+1) = 0$ and let k be the order of the largest Jordan block for B then there is a pairing of the eigenvalues of C and B such that, for $i = 1,\ldots,n$,

$$|\lambda_i(C)-\lambda_i(B)| \leq 2(n+1)\ \omega(k,\|FEF^{-1}\|).$$

Theorem 9. [Henrici, 1962]

Let $0 \neq \Delta(B) = 4\sqrt{\frac{n^3-1}{12}}\ \|BB^*-B^*B\|_F^{1/2}$ and let $\gamma(\eta)$ denote the nonnegative solution of $\xi+\xi^2+\ldots+\xi^n = \eta$. For any complex number μ and any unit vector x there is an eigenvalue λ of B such that

$$|\lambda-\mu| \leq \Delta(B)/\gamma(\Delta(B)/\|Bx-x\mu\|).$$

Remark. This is our only example of a computable residual bound. It also seems to be a counter-example to Householder's remark quoted above. As $\Delta(B) \to 0$ the standard residual bound for normal matrices is recovered. At the other extreme the bound shows that small residuals do not imply accurate eigenvalue approximations for very abnormal matrices. This confirms our numerical example.

Two valuable references for more recent work on perturbations are [Stewart, 1973] and [Chatelin, 1981]. The former concerns bounds on approximate eigenvectors and invariant subspaces, the latter uses residual norms, but their results are more for insight than computation since they involve quantities which are not readily computable.

3. ALTERNATIVE FORMULATIONS

To each eigenvalue λ of a complex (or real) nxn matrix B there corresponds at least one column eigenvector x and at least one row eigenvector y* satisfying

$$Bx = x\lambda, \quad y^*B = \lambda y^* \tag{3.1}$$

We call the set (λ,x,y^*) an <u>eigentriple</u> or <u>eigenelement</u> of B.

Associated with B is its <u>conjugate transpose</u> B*. Abstractly B* can be characterized as the <u>adjoint</u> of B on (or in) Euclidean n-space E^n with inner product $(\cdot,\cdot)$. In other words B* is the unique matrix which satisfies

$$(u,B^*v) = (Bu,v) \quad \text{for } \underline{\text{all}}\ u,\ v \text{ in } E^n.$$

The spectrum of B* is the conjugate of the spectrum of B and we could rewrite (3.1) as

$$Bx = x\lambda, \quad B^*y = y\bar{\lambda}\ . \tag{3.2}$$

We have put down these elementary facts because every author discussing the eigenvalue problem must, for coherence, choose and stick to one of two equivalent formulations, illustrated in (3.1) and (3.2):

(I) one operator and two (dual) spaces; B, E^n, E^n_*,

(II) two (adjoint) operators and one space; B, B*, E^n.

Our choice is (I); E^n is the Euclidean space of column vectors and E^n_* is the (dual) space of row vectors. Although E^n and E^n_* are isomorphic in a trivial way it helps to regard them as distinct in this essay. Formulation I lends itself to concrete matrix notation and we shall use the familiar forms:

$$a,b \in E^n, \quad (a,b) \equiv b^*a \equiv \sum_{i=1}^{n} \bar{\beta}_i \alpha_i, \qquad \|a\| \equiv \sqrt{a^*a} \quad ,$$

$$c^*,d^* \in E_*^n, (c^*,d^*) \equiv d^*c \equiv \sum_{i=1}^{n} \bar{\delta}_i \gamma_i, \qquad \|c^*\| \equiv \sqrt{c^*c} \quad ,$$

The reward for having this inner product structure is that it makes sense to compare vectors in norm or speak of the angle between a and b (or between c* and d*). Consequently adjectives like small and large can be used legitimately and that is essential for work in matrix computations.

There is more than one matrix norm compatible with Euclidean space. We shall confine ourselves to the two extreme unitarily invariant matrix norms. The spectral (or bound) norm is defined by

$$\|B\| \equiv \max_{u \neq 0} \|Bu\|/\|u\| = \sqrt{\lambda_{max}(B^*B)} \quad , \tag{3.3}$$

and the Frobenius (or Hilbert Schmidt) norm is

$$\|B\|_F \equiv \left(\sum_i \sum_j |b_{ij}|^2 \right)^{1/2} = \sqrt{\text{trace } (B^*B)} \quad . \tag{3.4}$$

The latter is easy to compute but yields $\|I\|_F = \sqrt{n}$ for the identity matrix instead of 1. For any B

$$\|B\| \leq \|B\|_F \leq \sqrt{\text{rank}(B)}\, \|B\| \leq \sqrt{n}\, \|B\| \quad . \tag{3.5}$$

In particular, for any square rank one matrix uv*,

$$\|uv^*\|_F = \|uv^*\| = \|u\| \cdot \|v^*\| \quad . \tag{3.6}$$

The matrix norms are needed so that we can speak of the matrix C being close to B in the sense that $\|C-B\|/\|B\|$ (or $\|C-B\|_F/\|B\|_F$) is small compared to 1.

In the symmetric case (B=A) a key role is played by Rayleigh's quotient, which is defined for all nonzero x by

$$\rho(x) = \rho(x;A) \equiv x^*Ax/x^*x \ .$$

The natural extension to the general case is defined for all pairs x,y^* such that $y^*x \neq 0$ by

$$\rho(x,y^*) = \rho(x,y^*;B) \equiv y^*Bx/y^*x \ . \tag{3.7}$$

Many authors call this the generalized Rayleigh quotient but the adjective is superfluous and we will drop it. What matters is that certain crucial properties do carry over from the nice symmetric case, namely

(1) ρ is homogeneous of degree 0, i.e., $\rho(\alpha x,\beta y^*) = \rho(x,y^*)$

(2) ρ is stationary at, and only at, the nondegenerate eigenvalues of B.

The gradients of ρ are

$$\nabla_x\rho(x,y^*) = y^*[B-\rho(x,y^*)I]/y^*x \ ,$$

$$\nabla_{y^*}\rho(x,y^*) = [B-\rho(x,y^*)I]x/y^*x \ ,$$

and the left hand sides vanish simultaneously only when (λ,x,y^*) is an eigenelement of B and then $\rho(x,y^*) = \lambda$. This condition precludes λ from belonging to Jordan blocks because the eigenvectors x and y^* for such λ satisfy $y^*x = 0$. This anomaly tells us to reformulate the Rayleigh Quotient. If λ belongs to a single Jordan block of order k there are (many) nxk matrices X and Y such that the columns of X and the rows of Y^* span λ's invariant subspaces in E^n and E^n_* and, moreover, $Y^*X = I_k$. It turns out that the kxk matrix

$$\rho(X,Y^*) \equiv Y^*BX \tag{3.8}$$

is similar to λ's Jordan block. In general, for any nxk matrices X, Y satisfying $Y^*X = I_k$ we say that (3.8) defines the Rayleigh quotient of X and Y*.

In the definition (3.7) x and y* are independent variables and this freedom entails that there is no upper bound on $|\rho(x,y^*)|$ in terms of B. However we do have

$$|y^*Bx| \leq \|B\|\|x\|\|y^*\| \ . \tag{3.9}$$

4. OPTIMAL APPROXIMATIONS FROM ROW AND COLUMN SUBSPACES

We are almost ready to exhibit the closest matrix B-E to B for which eigenelement approximations become exact. When the closest matrix is close enough then it can be regarded as a perturbation of B. On the other hand we can just as well regard B as a perturbation of B-E. To be specific let γ be any scalar and let q and p* be any vectors satisfying

$$p^*q = 1 \ . \tag{4.1}$$

We want

$$(B-E)q = q\gamma, \quad p^*(B-E) = \gamma p^* \tag{4.2}$$

so that γ is a simple eigenvalue of B-E. Standard perturbation theory [Wilkinson, 1965] says that there is an eigenvalue λ of B such that

$$|\lambda-\gamma| = \text{cond}(\gamma)\|E\| + O(\|E\|^2) \tag{4.3}$$

as $\|E\| \to 0$. Moreover the condition number of γ is computable,

$$\text{cond}(\gamma) \equiv \frac{\|q\|\cdot\|p^*\|}{p^*q} = \|q\|\cdot\|p^*\| \ . \tag{4.4}$$

Cond(γ) is the secant of the acute angle between p and q. It is also the spectral norm of the spectral projector qp* belonging to γ. Equations (4.3) and (4.4) provide strong incentive for knowing $\|E\|$.

We begin with a special case of the main theorem given below. Define residuals

$$r \equiv B\bar{x} - \bar{x}\gamma \ , \quad s^* \equiv \bar{y}^*B - \gamma\bar{y}^* \tag{4.5}$$

where

$$\bar{y} = y/\|y\| \ , \ \bar{x} = x/\|x\|, \quad \text{and} \quad y^*x = 1 \ .$$

COROLLARY 1. The distance to the closest matrix B-E for which (γ,x,y^*) is an eigentriple is $\|E\|_F$ and

$$\|E\|_F^2 = \|r\|^2 + \|s^*\|^2 - \left[\frac{(\gamma - y^*Bx)}{\|x\| \cdot \|y^*\|}\right]^2$$

Note that r and s* depend on γ.

This result shows that the Rayleigh quotient $\rho(x,y^*)$ is not necessarily the best approximate eigenvalue in the sense of minimizing $\|E\|_F$. Since $\|E\|_F^2$ is a quadratic in γ a little algebra yields the best value $\hat{\gamma}$,

$$\hat{\gamma} = \frac{\rho(x,x^*)+\rho(y,y^*)-\rho(x,y^*)/\|y^*\|^2\|x\|^2}{2-1/\|y^*\|^2\|x\|^2} \tag{4.6}$$

For practical purposes it is worth noting that when both $\|r\|$ and $\|s^*\|$ are small then the best value of γ is close to $\rho(q,p^*)$ and it is not worth going to the trouble of finding the best value.

We now give the analogue of the familiar Rayleigh-Ritz approximations from a subspace.

Let Q and P be any n by m matrices ($m \leq n$) satisfying $P^*Q = I_m$. The columns of Q constitute a basis for the subspace span Q in E^n and the

rows of P* constitute a basis for the subspace span P* in E^n_*. In a sense made precise below the best set of m approximate eigenelements for a matrix B from span Q and span P* are given by the eigenelements of the m by m matrix $\rho(Q,P^*) \equiv P^*BQ$ in the usual way. Specifically to each bi-orthonormal pair of eigenvectors u,v* of $\rho(Q,P^*)$ satisfying

$$P^*BQu = u\theta, \quad v^*P^*BQ = \theta v^*, \quad \theta = v^*P^*BQu,$$

there corresponds the approximate eigenelement of B

$$(\theta, Qu, v^*P^*) \ . \tag{4.7}$$

In the symmetric case (with P = Q) the word optimal is justified by the fact that $\|BQ-Q\Gamma\|_F$ is minimized by, and only by, the choice $\Gamma = P^*BQ$.

In the general case the word optimal is justified by the fact that all m approximate eigenelements from $\rho(Q,P^*)$ are exact eigenelements of that matrix closest to B which satisfies the three natural conditions given in the main theorem.

MAIN THEOREM. B is an n by n complex matrix, Q and P are n by m ($m \leq n$) satisfying $P^*Q = I_m$. To each m by m Γ there is a unique closest matrix B-E, using $\|\cdot\|_F$, such that

(i) $(B-E)Q = Q\Gamma$, [span Q invariant, Γ gives spectrum],

(ii) $P^*(B-E) = \Gamma P^*$, [span P* invariant, Γ gives spectrum].

Only the choice $\Gamma = J \equiv \rho(Q,P^*)$ ensures that

(iii) $P^*(B-E)Q = J$, [preservation of $\rho(Q,P^*)$].

Define residual matrices

$$R = R(\Gamma) \equiv (BQ-Q\Gamma)(Q^*Q)^{-1/2} \ , \tag{4.8}$$
$$S^* = S^*(\Gamma) \equiv (P^*P)^{-1/2}(P^*B-\Gamma P^*) \ .$$

This minimal change E to B is given in (4.15) and

$$\|E\|_F^2 = \|R\|_F^2 + \|S^*\|_F^2 - 4\|Z\|_F^2 \tag{4.9}$$

where Z is defined in (4.12). Note also that when $\Gamma = J$ then $Z = 0$ and

$$\|E\| = \max\{\|R\|, \|S^*\|\} . \tag{4.10}$$

It is convenient to use orthonormal bases for span Q and span P* so we define

$$\bar{Q} = Q(Q^*Q)^{-1/2}, \quad \bar{P} = P(P^*P)^{-1/2} , \tag{4.11}$$

noting that the square root of a positive definite matrix M is uniquely defined as the symmetric positive definite matrix H satisfying $H^2 = M$.

Proof. Define the auxiliary matrix Z by

$$\bar{P}^*R = (P^*P)^{-1/2}(J-\Gamma)(Q^*Q)^{-1/2} \equiv 2Z , \tag{4.12}$$

or

$$S^*\bar{Q} = (P^*P)^{-1/2}(J-\Gamma)(Q^*Q)^{-1/2} \equiv 2Z . \tag{4.13}$$

Conditions (i) and (ii) can be rewritten as

$$E\bar{Q} = R, \quad \bar{P}^*E = S^* . \tag{4.14}$$

Now use (4.12), (4.13) to verify that one solution of (4.14) is

$$E = (R-\bar{P}Z)\bar{Q}^* + \bar{P}(S^*-Z\bar{Q}^*) . \tag{4.15}$$

By linearity of the equations (4.14) all other solutions are E+G with G satisfying

$$G\bar{Q} = 0 , \quad \bar{P}^*G = 0 , \tag{4.16}$$

We abbreviate trace by tr and observe that for any nxn G

$$\mathrm{tr}[(E+G)^*(E+G)] = \mathrm{tr}(E^*E) + 2\mathrm{tr}(E^*G) + \mathrm{tr}(G^*G) \quad . \tag{4.17}$$

Equations (4.15) and (4.16) imply that

$$E^*G = \bar{Q}(R-\bar{P}X)^*G + (S-\bar{Q}Z^*)\bar{P}^* \quad ,$$

$$= \bar{Q}R^*G - 0 + 0 \quad ,$$

Now use $\mathrm{tr}[KL] = \mathrm{tr}[LK]$ and (4.16) to find

$$\mathrm{tr}[E^*G] = \mathrm{tr}[\bar{Q}R^*G] = \mathrm{tr}[R^*G\bar{Q}] = \mathrm{tr}[0] = 0 \quad . \tag{4.18}$$

Thus from (4.17)

$$\|E+G\|_F^2 = \|E\|_F^2 + \|G\|_F^2 \geq \|E\|_F^2 \tag{4.19}$$

and so B-E is the closest matrix satisfying (i) and (ii) and using $\|\cdot\|_F$.

It remains to compute $\|E\|_F$ using (4.12), (4.13), (4.15);

$$E^*E = \bar{Q}(R^*R-Z^*\bar{P}^*R-R^*\bar{P}Z+Z^*Z)\bar{Q}^*$$

$$+ SS^* - \bar{Q}Z^*P^* - PZ\bar{Q}^* + \bar{Q}Z^*Z\bar{Q}^* \quad ,$$

$$= \bar{Q}(R^*R-4Z^*Z)\bar{Q}^* + SS^* \quad .$$

Using $\mathrm{tr}(KL) = \mathrm{tr}(LK)$ yields

$$\mathrm{tr}(E^*E) = \mathrm{tr}(R^*R+S^*S-4Z^*Z) \quad . \tag{4.20}$$

Consider now condition (iii). It forces

$$P^*EQ = 0 \quad . \tag{4.21}$$

Use (4.15) and (4.12), (4.13) to see that

$$\bar{P}^*E\bar{Q} = 2Z = 0 \quad .$$

In other words, $\Gamma = J$. In this case $E = R\bar{Q}^* + \bar{P}S^*$,

and

$$E^*E = \bar{Q}R^*\bar{Q}^* + SS^*$$

$$= \bar{Q}R^*R\bar{Q}^* + \bar{S}(S^*S)\bar{S}^* ,$$

$$= (\bar{Q},\bar{S})\ \mathrm{diag}(R^*R,S^*S)(\bar{Q},\bar{S})^* \qquad (4.22)$$

By (4.13) the matrix $(\bar{Q},\bar{S})$ is orthonormal and the 2m positive eigenvalues of E*E are those of R*R and S*S. This yields (4.10). ¤

The factor 4Z*Z in (4.20) shows that, in general, the choice $\Gamma = P^*BQ$ does not lead to the minimum $\|E\|_F$ such that span Q and span P* are invariant. Nor should it when P* and Q are chosen perversely. The simplest way to see what happens is to consider the inadmissable case when span Q and span P* are each invariant but associated with disjoint parts of B's spectrum. Since P*Q = 0 the Rayleigh quotient is not defined and yet a good choice for Γ is the set of eigenvalues associated with either Q*BQ or P*BP. Then the approximate eigenvalues are exact and either R or S* is the zero matrix.

When $P^*Q = I_m$ then the matrix QP* is the spectral projector of B-E associated with $\rho(Q,P^*)$. Specifically PQ* is one term in a partition of unity

$$I_n = \sum_i Q_i P_i^*$$

with the special property

$$B-E = \sum_i Q_i \rho_i P_i^*, \qquad \rho_i = P_i^*(B-E)Q_i \ .$$

This decomposition is useful if the $\|Q_iP_i^*\|_F$ are not too large; moreover

$$\|QP^*\|_F \le \|Q\|_F\|P^*\|_F \ .$$

Whenever $\|Q\|_F\|P\|_F$ is very large it should be taken as a hint that more columns should be added to P and Q.

5. Application to the Lanczos Algorithm

Simple subspaces from which approximations are sought to eigenelements of an nxn matrix B are Krylov subspaces. Given a pair of starting vectors q_1 and p_1^* satisfying $p_1^*q_1 = 1$ the Lanczos algorithm, by step j, produces two bi-orthonormal nxj matrices Q_j and P_j which satisfy in exact arithmetic,

$$P_j^*Q_j = I_j \ , \tag{5.1}$$

$$BQ_j - Q_jJ_j = (\underline{0},\underline{0},..,\underline{0},q_{j+1}\beta_{j+1}) \ , \tag{5.2}$$

$$P_j^*B - J_jP_j^* = \begin{pmatrix} \underline{0}^* \\ \vdots \\ \gamma_{j+1}p_{j+1}^* \end{pmatrix} \tag{5.3}$$

and J_j is a j by j tridiagonal matrix

$$J_j = \begin{bmatrix} \alpha_1 & \gamma_2 & & \bigcirc \\ \beta_2 & \alpha_2 & \gamma_3 & \\ & \beta_3 & \alpha_3 & \cdot \\ \bigcirc & & \cdot & \cdot \end{bmatrix} \tag{5.4}$$

Theoretically Q_j and P_j^* can be thought of as the result of bi-orthonormaling, via Gram-Schmidt, the two Krylov sequences

$$\{q_1, Bq_1, B^2q_1,\ldots,B^{j-1}q_1\} \ ,$$

$$\{p_1^*, p_1^*B, p_1^*B^2,.., p_1^*B^{j-1}\} \ .$$

The algorithm comes to a halt as soon as $\beta_i\gamma_i = 0$, for some i, but we will assume that such good luck has not occurred by step j.

Section 4 showed that there is good reason to seek approximate eigentriples from the projection of B along span Q_j and span P_j^*, namely

$$J_j = P_j^* B Q_j \tag{5.5}$$

Let us examine a typical eigenvalue θ; say

$$J_j z = z\theta, \quad w^* J_j = \theta w^*, \quad w^* z = 1 \ . \tag{5.6}$$

We suppress the dependence of θ, z, and w^* on j. Now multiply (5.2) and (5.3) by z and w^*, and introduce two important quantities ζ_j and ω_j, the last elements of z and w^*, to find that

$$B(Q_j z) - (Q_j z)\theta = q_{j+1}\beta_{j+1}\zeta_j \ , \tag{5.7}$$

$$(w^* P_j^*)B - \theta(w^* P_j) = \omega_j \gamma_{j+1} p_{j+1}^* \ , \tag{5.8}$$

The approximate eigenvectors are $x = Q_j z, y^* = w^* P_j^*$ and, by applying the main theorem with m = 1 to (5.7) and (5.8) we obtain a computable expression for the error.

COROLLARY. The closest matrix to B with (θ, x, y^*) as an eigenelement is B-E and

$$\|E\| = \max\left\{\frac{|\beta_{j+1}\zeta_j|\,\|q_{j+1}\|}{\|x\|} \ , \ \frac{|\gamma_{j+1}\omega_j|\,\|p_{j+1}^*\|}{\|y^*\|}\right\} \tag{5.9}$$

From (4.15) E is the rank two matrix

$$E = \left(\frac{\beta_{j+1}\zeta_j}{\|x\|^2}\right) q_{j+1} x^* + \left(\frac{\gamma_{j+1}\omega_j}{\|y\|^2}\right) y p_{j+1}^* \ .$$

The object of interest, $|\lambda-\theta|$, is unknown but when $\|E\|$ is small enough then (4.3) applies,

$$|\lambda-\theta| = \mathrm{cond}(\theta)\|E\| + O(\|E\|^2) \tag{5.10}$$

and, by (4.4),

$$\mathrm{cond}(\theta) = \|x\|\ \|y^*\| \ . \tag{5.11}$$

Consequently the Lanczos algorithm should be continued until both $\|E\|$ is small and $\mathrm{cond}(\theta)\cdot\|E\|$ is below the given threshold for accuracy.

For $j << n$ (say $j = 10\sqrt{n}$) the computation of eigenvalues of the tridiagonal J_j is modest compared with the cost of a Lanczos step. In principle x and y^* are computable but this would require the use of the matrices Q_j and P_j; that cost is 2jn operations and probably exceeds the cost of a Lanczos step. However it is advisable to keep in the fast store the quantities $\|Q_j\|_F^2 = \sum_{i=1}^{j} \|q_i\|^2$ and $\|P_j^*\|_F^2 = \sum_{i=1}^{j} \|p_i^*\|^2$, which are easy to update. A computable bound on $\mathrm{cond}(\theta)$ is then

$$\mathrm{cond}(\theta) = \|x\|\|y^*\| \leq \|Q_j\|_F\ \|P_j^*\|_F\ \|z\|\ \|w^*\| \tag{5.12}$$

A way to use the error estimates in practice is described at the end of the section. The error (5.9) continues to hold <u>in practice</u>, apart from roundoff terms, because no use was made of (5.1) which fails completely in finite precision.

<u>Persistance of Ritz Values</u>

The eigenvalues of the tridiagonal J_k are, in general, distinct from those of J_{k-1} (In the symmetric case the two sets interlace each other.) Nevertheless for large enough k some of these Ritz values change by negligible amounts when k increases. Which ones? Certainly among them are the Ritz values which have already stabilized at eigenvalues of B. This fact is a straightforward extension of the symmetric version by C. C. Paige. The key observation is that the corresponding eigenvectors do not change much as k increases.

We shall need some extra notation. For any vector v let $\tilde{v}$ denote the vector $\binom{v}{0}$ where 0 elements are appended to v to make $\tilde{v}$ have the appropriate dimension for the context in which it is used.

COROLLARY OF MAIN THEOREM.

Let (θ,z,w^*) be an eigentriple of the tridiagonal matrix J_j of (5.4) with $w^*z = 1$. Then, for all $k > j$, $(\theta,\tilde{z},\tilde{w}^*)$ is an eigentriple of $J_k - G_k$ and

$$G_k = \left(\frac{\beta_{j+1}\zeta_j}{\|z\|^2}\right)\tilde{e}_{j+1}\tilde{z}^* + \left(\frac{\gamma_{j+1}\omega_j}{\|w^*\|^2}\right)\tilde{w}e^*_{j+1}$$

Moreover

$$\|G_k\| = \max\left\{\frac{|\beta_{j+1}\zeta_j|}{\|z\|}, \frac{|\gamma_{j+1}\omega_j|}{\|w^*\|}\right\} \quad ; \quad \|G_k\|_F^2 = \frac{|\beta_{j+1}\zeta_j|^2}{\|z\|^2} + \frac{|\gamma_{j+1}\omega_j|^2}{\|w^*\|^2}$$

and both values are independent of k.

Proof. The residual vectors for $\tilde{z}$ with respect to J_k is readily verified to be

$$\|\tilde{z}\|\tilde{r} \equiv (J_k\tilde{z} - \tilde{z}\theta) = \left(\begin{array}{c|c} J_j & 0 \\ \hline 0 & \hat{J} \end{array}\right)\binom{z}{0} - \binom{z}{0}\theta = \binom{J_jz - z\theta}{0} = \binom{0}{0}$$

where

$$\text{¤} = \beta_{j+1}(e_j^*z) = \beta_{j+1}\zeta_j \quad ,$$

and $\|\tilde{z}\| = \|z\|$. Similarly for $\tilde{s}^*$. The corollary results from using these expressions for the residuals in the main theorem in Section 4. ¤

The application is immediate and remarkable. When $|\zeta_j|$ and $|\omega_j|$ have dwindled sufficiently so that $\text{cond}(\theta)\ \|G_j\| \leq \text{tol}$ then θ is a stabilization (or condensation) point for the Lanczos process. There is

a matrix within tol of J_k which has θ in its spectrum for all $k \geq j$.

Use of Error Estimates

At the jth step of the Lanczos process there is freedom in the choice of β_{j+1} and γ_{j+1} although the quantities $\beta_{j+1}\gamma_{j+1}$, $|\beta_{j+1}|\|q_{j+1}\|$, $|\gamma_{j+1}|\|p^*_{j+1}\|$ are all determined. In the corollary given above we may pretend that β and γ are scaled so that

$$\|G_k\| = \max\{\frac{|\beta_{j+1}\zeta_j|}{\|z\|}, \frac{|\gamma_{j+1}\omega_j|}{\|w^*\|}\} ,$$

$$= \{\frac{|\beta_{j+1}\gamma_{j+1}\zeta_j\omega_j|}{\|z\|\ \|w^*\|}\}^{1/2} . \qquad (5.13)$$

With respect to J_j the condition number of θ is

$$\text{cond}(\theta;J_j) = \|z\|\ \|w^*\| . \qquad (5.14)$$

Moreover, for all $k > j$,

$$\text{cond}(\theta;J_k-G_k) = \|\tilde{z}\|\ \|\tilde{w}^*\| = \text{cond}(\theta;J_j) .$$

Here is an economical way to test for termination when seeking an eigenvalue to within tol$\cdot\|B\|$.

1. Advance the Lanczos process, computing appropriate eigenvalues of J at each step, until the one (or ones) of interest stabilizes to the required accuracy. Call it θ and let the step number be j.
2. If $\text{cond}(\theta;J_j)$ is too big then abandon the assumption that θ approximates a simple eigenvalue of B. Find a cluster of eigenvalues of J_j and associated eigenvector matrices $Z = (z_1,z_2,..)$ and $W = (W_1,W_2,...)$ so that $W^*Z = I$ and $\|Z\|_F\ \|W^*\|_F$ is acceptable. Then use the more general estimates of the main theorem with Q_jZ in place of Q and P_jW in place of P, W^*J_jZ in place of J.

3. If $\mathrm{cond}(\theta;J_j)$ is acceptable (?) then test the statement

$$|\beta_{j+1}\gamma_{j+1}\zeta_j\omega_j|\ \|z\|\|w^*\| \leq (\mathrm{tol}\ \|J_j\|_F^2) \tag{5.20}$$

The left side is $\mathrm{cond}(\theta,J_k-G_k)\ \|G_k\|$, for $k > j$.

4. If (5.20) fails then continue the Lanczos process, otherwise test $\mathrm{cond}(\theta;B-E)\cdot\|E\|$ using (5.9), (5.11), (5.12), namely test

$$\max\{|\beta_{j+1}\zeta_j|\|q_{j+1}\|\ \|P_j^*\|_F\ \|w^*\|,\ |\gamma_{j+1}\omega_j|\|p_{j+1}^*\|\ \|Q_j\|_F\ \ \|z\|\}$$

$$\leq \mathrm{tol}\ \|J_j\|_F \tag{5.21}$$

5. If (5.21) fails then continue the Lanczos process, otherwise compute $x = Q_j z$, $y^* = w^*P_j^*$, and deliver $\max\{|\beta_{j+1}\zeta_j|\ \|q_{j+1}\|\ \|y^*\|,\ |\gamma_{j+1}\omega_j|\ \|p_{j+1}^*\|\ \|x\|\}$ as the error estimate for θ.

Comments

(i) We do not know what numerical values should be used to discriminate between simple eigenvalues and perturbations of multiple ones. Please consult [Golub and Wilkinson, 1976], [A. Ruhe, 1970], [Varah, 1970].

(ii) The costly formation of x and y* is not made until θ is acceptable.

(iii) These tests may be used in finite precision. The failure of $P_j^*Q_j = I_j$ means that J_j is not the optimal projection of B. Nevertheless it is still a good approximation.

APPENDIX

Proof of Theorem 8. $B = FJF^{-1}$, $C = B-E$.

We begin by immitating the proof of the Bauer-Fike theorem. By definition of μ, $B-E-\mu I$ is singular. If μ is an eigenvalue of B the bound holds trivially. So we consider the contrary case when the matrix

$$D \equiv J-\mu I$$

is invertible. We know that

$$F(D-F^{-1}EF)F^{-1}$$

is singular. It follows that $\|D^{-1}F^{-1}EF\| \geq 1$ for any norm in which $\|I\| = 1$. Thus

$$1/\|D^{-1}\| \leq \|F^{-1}EF\| \leq \|F^{-1}\|\|F\|\,\|E\| \ .$$

It remains to estimate $\|D^{-1}\|$. We focus on the spectral norm and recall that D is a direct sum of Jordan blocks and so $1/\|D^{-1}\|$ is the smallest singular value among all the blocks. For a typical block the smallest singular value σ satisfies $\sigma^2 = \lambda_1(T) = \lambda_{min}(T)$ for a tridiagonal matrix

$$T = \begin{bmatrix} 1+\delta^2 & \delta & & & \\ \delta & 1+\delta^2 & \delta & & \\ & \delta & \cdot & \cdot & \\ & & \cdot & 1+\delta^2 & \delta \\ & & & \delta & \delta^2 \end{bmatrix}$$

Here $\delta = |\lambda-\mu|$, for some eigenvalue λ of B and T is diagonally similar to LL* where L is a typical block of D. T is positive

definite, but only just so. An asymptotically correct lower bound on its smallest eigenvalue can be obtained quite simply from the observation that $\det T = |\det LL^*| = |\det L|^2 = \delta^{2m}$. By Gersgorin's Theorem all T's eigenvalues satisfy $\lambda_i < 1 + \delta^2 + 2\delta = (1+\delta)^2$. Thus

$$\begin{aligned}\lambda_1 &= \lambda_1 \ldots \lambda_m / (\lambda_2 \ldots \lambda_m) \ , \\ &= \det T / (\lambda_2 \ldots \lambda_m) \ , \\ &> \delta^{2m} / (1+\delta)^{2m-2} \ .\end{aligned}$$

It follows that $1/\|D^{-1}\|$ is greater than the smallest of the expressions $\delta^m/(1+\delta)^{m-1}$ over all the blocks. So for some λ and its associated largest block size m,

$$\frac{|\lambda-\mu|^m}{(1+|\lambda-\mu|)^{m-1}} \leq \frac{1}{\|D^{-1}\|} \leq \|FEF^{-1}\| \ . \qquad \square$$

REFERENCES

F. L. Bauer and C. T. Fike, "Norms and exclusion theorems," Numer. Math. 2, 137-141 (1960).

F. Chatelin, "Spectral approximations for linear operators," (in preparation).

P. Henrici, "Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices," Numer. Math. 4, 24-40 (1962).

A. Householder, "Lectures on Numerical Algebra," Math. Assoc. of Amer. (1972).

T. Kato, "Perturbation Theory for linear operators," Springer-Verlag, (1966).

B. Parlett, "The symmetric eigenvalue problem," (Prentice Hall Inc., New Jersey) (1980).

A. Ostrowski, "On the continuity of characteristic roots as functions of the elements of a matrix," J. ber. Deutsch. Math.-Verein. Abt. 1, 40-42 (1957).

G. W. Stewart III, "Error and perturbation bounds for subspaces associated with certain eigenvalue probelms," SIAM Review 15, 727-764 (1973).

J. H. Wilkinson, "The algebraic eigenvalue problem," Oxford Univ. Press, London and New York, (1965).

J. S. Vandergraft, "Generalized Rayleigh methods with applications for finding eigenvalues of large matrices," J. Lin. Alg. Appls. 4, 353-368 (1971).

G. Golub and J. H. Wilkinson, "Ill-conditioned eigensystems and the computation of the Jordan canonical form," SIAM Review, 18, 578-619 (1976).

A. Ruhe, "An algorithm for the numerical determination of the structure of a general matrix, BIT, 10, 196-216 (1970).

J. M. Varah, "Rigorous machine bounds for the eigensystem of a general complex matrix," Math. Comp., 22, 793-801 (1970).