Chapter 8

The Symmetric Eigenvalue Problem

§8.1 Properties and Decompositions
§8.2 Power Iterations
§8.3 The Symmetric QR Algorithm
§8.4 Jacobi Methods
§8.5 Tridiagonal Methods
§8.6 Computing the SVD
§8.7 Some Generalized Eigenvalue Problems

The symmetric eigenvalue problem with its rich mathematical structure is one of the most aesthetically pleasing problems in numerical linear algebra. We begin our presentation with a brief discussion of the mathematical properties that underlie this computation. In §8.2 and §8.3 we develop various power iterations eventually focusing on the symmetric QR algorithm.

In §8.4 we discuss Jacobi's method, one of the earliest matrix algorithms to appear in the literature. This technique is of current interest because it is amenable to parallel computation and because under certain circumstances it has superior accuracy.

Various methods for the tridiagonal case are presented in §8.5. These include the method of bisection and a divide and conquer technique.

The computation of the singular value decomposition is detailed in §8.6. The central algorithm is a variant of the symmetric QR iteration that works on bidiagonal matrices.

In the final section we discuss the generalized eigenvalue problem $Ax = \lambda Bx$ for the important case when A is symmetric and B is symmetric positive definite. No suitable analog of the orthogonally-based QZ algorithm (see §7.7) exists for this specially structured, generalized eigenproblem. However, there are several successful methods that can be applied and these are presented along with a discussion of the generalized singular value decomposition.

Before You Begin

Chapter 1, §§2.1-2.5, and §2.7, Chapter 3, §§4.1-4.3, §§5.1-5.5 and §7.1.1 are assumed. Within this chapter there are the following dependencies:

$$\begin{tikzcd}
\S 8.4 & & & & \\
\S 8.1 \arrow[u] \arrow[r] \arrow[d] & \S 8.2 \arrow[r] & \S 8.3 \arrow[r] & \S 8.6 \arrow[r] & \S 8.7 \\
\S 8.5 & & & &
\end{tikzcd}$$

Many of the algorithms and theorems in this chapter have unsymmetric counterparts in Chapter 7. However, except for a few concepts and definitions, our treatment of the symmetric eigenproblem can be studied before reading Chapter 7.

Complementary references include Wilkinson (1965), Stewart (1973), Gourlay and Watson (1973), Hager (1988), Chatelin (1993), Parlett (1980), Stewart and Sun (1990), Watkins (1991), Jennings and McKeowen (1992), and Datta (1995). Some Matlab functions important to this chapter are schur and svd. LAPACK connections include

LAPACK: Symmetric Eigenproblem	
_SYEV	All eigenvalues and vectors
_SYEVD	Same but uses divide and conquer for eigenvectors
_SYEVX	Selected eigenvalues and vectors
_SYTRD	Householder tridiagonalization
_SBTRD	Householder tridiagonalization (A banded)
_SPTRD	Householder tridiagonalization (A in packed storage)
_STEQR	All eigenvalues and vectors of tridiagonal by implicit QR
_STEDC	All eigenvalues and vectors of tridiagonal by divide and conquer
_STERF	All eigenvalues of tridiagonal by root-free QR
_PTEQR	All eigenvalues and eigenvectors of positive definite tridiagonal
_STEBZ	Selected eigenvalues of tridiagonal by bisection
_STEIN	Selected eigenvectors of tridiagonal by inverse iteration

LAPACK: Symmetric-Definite Eigenproblems	
_SYGST	Converts $A - \lambda B$ to $C - \lambda I$ form
_PBSTF	Split Cholesky factorization
_SBGST	Converts banded $A - \lambda B$ to $C - \lambda I$ form via split Cholesky

LAPACK: SVD	
_GESVD	$A = U\Sigma V^T$
_BDSQR	SVD of real bidiagonal matrix
_GEBRD	bidiagonalization of general matrix
_ORGBR	generates the orthogonal transformations
_GBBRD	bidiagonalization of band matrix

LAPACK: The Generalized Singular Value Problem	
_GGSVP	Converts $A^TA - \mu^2B^TB$ to triangular $A_1^TA_1 - \mu^2B_1^TB_1$
_TGSJA	Computes GSVD of a pair of triangular matrices.

8.1 Properties and Decompositions

In this section we set down the mathematics that is required to develop and analyze algorithms for the symmetric eigenvalue problem.

8.1.1 Eigenvalues and Eigenvectors

Symmetry guarantees that all of A's eigenvalues are real and that there is an orthonormal basis of eigenvectors.

Theorem 8.1.1 (Symmetric Schur Decomposition) *If $A \in \mathbb{R}^{n\times n}$ is symmetric, then there exists a real orthogonal Q such that*

$$Q^TAQ \;=\; \Lambda \;=\; \text{diag}(\lambda_1, \ldots, \lambda_n).$$

Moreover, for $k = 1{:}n$, $AQ(:,k) = \lambda_k Q(:,k)$. See Theorem 7.1.3.

Proof. Suppose $\lambda_1 \in \lambda(A)$ and that $x \in \mathbb{C}^n$ is a unit 2-norm eigenvector with $Ax = \lambda_1 x$. Since $\lambda_1 = x^HAx = x^HA^Hx = \overline{x^HAx} = \overline{\lambda_1}$ it follows that $\lambda_1 \in \mathbb{R}$. Thus, we may assume that $x \in \mathbb{R}^n$. Let $P_1 \in \mathbb{R}^{n\times n}$ be a Householder matrix such that $P_1^Tx = e_1 = I_n(:,1)$. It follows from $Ax = \lambda_1 x$ that $(P_1^TAP_1)e_1 = \lambda e_1$. This says that the first column of $P_1^TAP_1$ is a multiple of e_1. But since $P_1^TAP_1$ is symmetric it must have the form

$$P_1^TAP_1 \;=\; \begin{bmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{bmatrix}$$

where $A_1 \in \mathbb{R}^{(n-1)\times(n-1)}$ is symmetric. By induction we may assume that there is an orthogonal $Q_1 \in \mathbb{R}^{(n-1)\times(n-1)}$ such that $Q_1^TA_1Q_1 = \Lambda_1$ is diagonal. The theorem follows by setting

$$Q = P_1\begin{bmatrix} 1 & 0 \\ 0 & Q_1 \end{bmatrix} \qquad \text{and} \qquad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \Lambda_1 \end{bmatrix}$$

and comparing columns in the matrix equation $AQ = Q\Lambda$. □

Example 8.1.1 If

$$A \;=\; \begin{bmatrix} 6.8 & 2.4 \\ 2.4 & 8.2 \end{bmatrix} \qquad \text{and} \qquad Q \;=\; \begin{bmatrix} .6 & -.8 \\ .8 & .6 \end{bmatrix},$$

then Q is orthogonal and $Q^TAQ = \text{diag}(10,5)$.

For a symmetric matrix A we shall use the notation $\lambda_k(A)$ to designate the kth largest eigenvalue. Thus,

$$\lambda_n(A) \le \cdots \le \lambda_2(A) \le \lambda_1(A).$$

It follows from the orthogonal invariance of the 2-norm that A has singular values $\{|\lambda_1(A)|, \ldots, |\lambda_n(A)|\}$ and so

$$\| A \|_2 \;=\; \max\{\, |\lambda_1(A)|\,,\, |\lambda_n(A)|\,\}.$$

The eigenvalues of a symmetric matrix have a "minimax" characterization based on the values that can be assumed by the quadratic form ratio x^TAx/x^Tx.

Theorem 8.1.2 (Courant-Fischer Minimax Theorem) *If $A \in \mathbb{R}^{n\times n}$ is symmetric, then*

$$\lambda_k(A) \;=\; \max_{\dim(S)=k} \;\; \min_{0\ne y\in S} \; \frac{y^TAy}{y^Ty}$$

for $k = 1{:}n$.

Proof. Let $Q^TAQ = \text{diag}(\lambda_i)$ be the Schur decomposition with $\lambda_k = \lambda_k(A)$ and $Q = [\, q_1,\, q_2, \ldots,\, q_n\,]$. Define

$$S_k = \text{span}\{q_1, \ldots, q_k\},$$

the invariant subspace associated with $\lambda_1, \ldots, \lambda_k$. It is easy to show that

$$\max_{\dim(S)=k} \;\; \min_{0\ne y\in S} \; \frac{y^TAy}{y^Ty} \;\ge\; \min_{0\ne y\in S_k} \; \frac{y^TAy}{y^Ty} \;=\; q_k^TAq_k \;=\; \lambda_k(A).$$

To establish the reverse inequality, let S be any k-dimensional subspace and note that it must intersect span$\{q_k, \ldots, q_n\}$, a subspace that has dimension $n-k+1$. If $y_* = \alpha_k q_k + \cdots + \alpha_n q_n$ is in this intersection, then

$$\min_{0\ne y\in S} \; \frac{y^TAy}{y^Ty} \;\le\; \frac{y_*^TAy_*}{y_*^Ty_*} \;\le\; \lambda_k(A).$$

Since this inequality holds for all k-dimensional subspaces,

$$\max_{\dim(S)=k} \;\; \min_{0\ne y\in S} \; \frac{y^TAy}{y^Ty} \;\le\; \lambda_k(A)$$

thereby completing the proof of the theorem. □

If $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite, then $\lambda_n(A) > 0$.

8.1.2 Eigenvalue Sensitivity

An important solution framework for the symmetric eigenproblem involves the production of a sequence of orthogonal transformations $\{Q_k\}$ with the property that the matrices $Q_k^T A Q_k$ are progressively "more diagonal." The question naturally arises, how well do the diagonal elements of a matrix approximate its eigenvalues?

Theorem 8.1.3 (Gershgorin) *Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric and that $Q \in \mathbb{R}^{n\times n}$ is orthogonal. If $Q^T A Q = D + F$ where $D = \mathrm{diag}(d_1, \ldots, d_n)$ and F has zero diagonal entries, then*

$$\lambda(A) \subseteq \bigcup_{i=1}^{n} [d_i - r_i, d_i + r_i]$$

where $r_i = \sum_{j=1}^{n} |f_{ij}|$ *for $i = 1{:}n$. See Theorem 7.2.1.*

Proof. Suppose $\lambda \in \lambda(A)$ and assume without loss of generality that $\lambda \neq d_i$ for $i = 1{:}n$. Since $(D - \lambda I) + F$ is singular, it follows from Lemma 2.3.3 that

$$1 \le \| (D - \lambda I)^{-1} F \|_\infty = \sum_{j=1}^{n} \frac{|f_{kj}|}{|d_k - \lambda|} = \frac{r_k}{|d_k - \lambda|}$$

for some k, $1 \le k \le n$. But this implies that $\lambda \in [d_k - r_k, d_k + r_k]$. □

Example 8.1.2 The matrix

$$A = \begin{bmatrix} 2.0000 & 0.1000 & 0.2000 \\ 0.2000 & 5.0000 & 0.3000 \\ 0.1000 & 0.3000 & -1.0000 \end{bmatrix}$$

has Gerschgorin intervals $[1.7, 2.3]$, $[4.5, 5.5]$, and $[-1.4, -.6]$ and eigenvalues 1.9984, 5.0224, and -1.0208.

The next results show that if A is perturbed by a symmetric matrix E, then its eigenvalues do not move by more than $\| E \|$.

Theorem 8.1.4 (Wielandt-Hoffman) *If A and $A + E$ are n-by-n symmetric matrices, then*

$$\sum_{i=1}^{n} (\lambda_i(A + E) - \lambda_i(A))^2 \le \| E \|_F^2 .$$

Proof. A proof can be found in Wilkinson (1965, pp.104–8) or Stewart and Sun (1991, pp.189–191). See also P8.1.5. □

Example 8.1.3 If

$$A = \begin{bmatrix} 6.8 & 2.4 \\ 2.4 & 8.2 \end{bmatrix} \quad \text{and} \quad E = \begin{bmatrix} .002 & .003 \\ .003 & .001 \end{bmatrix},$$

then $\lambda(A) = \{5, 10\}$ and $\lambda(A+E) = \{4.9988, 10.004\}$ confirming that

$$1.95 \times 10^{-5} = |4.9988 - 5|^2 + |10.004 - 10|^2 \le \| E \|_F^2 = 2.3 \times 10^{-5}.$$

Theorem 8.1.5 *If A and $A+E$ are n-by-n symmetric matrices, then*

$$\lambda_k(A) + \lambda_n(E) \le \lambda_k(A+E) \le \lambda_k(A) + \lambda_1(E) \qquad k = 1{:}n.$$

Proof. This follows from the minimax characterization. See Wilkinson (1965, pp.101–2) or Stewart and Sun (1990, p.203). □

Example 8.1.4 If

$$A = \begin{bmatrix} 6.8 & 2.4 \\ 2.4 & 8.2 \end{bmatrix} \quad \text{and} \quad E = \begin{bmatrix} .002 & .003 \\ .003 & .001 \end{bmatrix},$$

then $\lambda(A) = \{5,\ 10\}$, $\lambda(E) = \{-.0015,\ .0045\}$, and $\lambda(A+E) = \{4.9988,\ 10.0042\}$. confirming that

$$\begin{array}{rcl} 5 - .0015 & \le 4.9988 \le & 5 + .0045 \\ 10 - .0015 & \le 10.0042 \le & 10 + .0045. \end{array}$$

Corollary 8.1.6 *If A and $A+E$ are n-by-n symmetric matrices, then*

$$|\lambda_k(A+E) - \lambda_k(A)| \le \| E \|_2$$

for $k = 1{:}n$.

Proof.

$$|\lambda_k(A+E) - \lambda_k(A)| \le \max\{|\lambda_n(E)|\,,\, |\lambda_1(E)||\} = \| E \|_2. \, \square$$

Several more useful perturbation results follow from the minimax property.

Theorem 8.1.7 (Interlacing Property) *If $A \in \mathbb{R}^{n\times n}$ is symmetric and $A_r = A(1{:}r, 1{:}r)$, then*

$$\lambda_{r+1}(A_{r+1}) \le \lambda_r(A_r) \le \lambda_r(A_{r+1}) \le \cdots \le \lambda_2(A_{r+1}) \le \lambda_1(A_r) \le \lambda_1(A_{r+1})$$

for $r = 1{:}n-1$.

Proof. Wilkinson (1965, pp.103–4). □

Example 8.1.5 If

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 3 & 6 & 10 \\ 1 & 4 & 10 & 20 \end{bmatrix}$$

then $\lambda(A_1) = \{1\}$, $\lambda(A_2) = \{.3820,\ 2.6180\}$, $\lambda(A_3) = \{.1270,\ 1.0000,\ 7.873\}$, and $\lambda(A_4) = \{.0380,\ .4538,\ 2.2034, 26.3047\}$.

Theorem 8.1.8 *Suppose $B = A + \tau cc^T$ where $A \in \mathbb{R}^{n\times n}$ is symmetric, $c \in \mathbb{R}^n$ has unit 2-norm and $\tau \in \mathbb{R}$. If $\tau \geq 0$, then*

$$\lambda_i(B) \in [\lambda_i(A),\ \lambda_{i-1}(A)] \qquad i = 2{:}n$$

while if $\tau \leq 0$ then

$$\lambda_i(B) \in [\lambda_{i+1}(A), \lambda_i(A)], \qquad i = 1{:}n-1\ .$$

In either case, there exist nonnegative $m_1, \ldots, m_n$ such that

$$\lambda_i(B) = \lambda_i(A) + m_i\tau, \qquad i = 1{:}n$$

with $m_1 + \cdots + m_n = 1$.

Proof. Wilkinson (1965, pp.94–97). See also P8.1.8. □

8.1.3 Invariant Subspaces

Many eigenvalue computations proceed by breaking the original problem into a collection of smaller subproblems. The following result is the basis for this solution framework.

Theorem 8.1.9 *Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric and that*

$$Q = \begin{array}{c} [\ Q_1 \quad Q_2\ ] \\ \quad r \quad\ n-r \end{array}$$

is orthogonal. If $\operatorname{ran}(Q_1)$ is an invariant subspace, then

$$Q^TAQ = D = \begin{array}{cc} \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} & \begin{array}{c} r \\ n-r \end{array} \\ \begin{array}{cc} r & n-r \end{array} & \end{array} \qquad (8.1.1)$$

and $\lambda(A) = \lambda(D_1) \cup \lambda(D_2)$. See also Lemma 7.1.2.

Proof. If

$$Q^TAQ = \begin{bmatrix} D_1 & E_{21}^T \\ E_{21} & D_2 \end{bmatrix},$$

then from $AQ = QD$ we have $AQ_1 - Q_1D_1 = Q_2E_{21}$. Since $\text{ran}(Q_1)$ is invariant, the columns of Q_2E_{21} are also in $\text{ran}(Q_1)$ and therefore perpendicular to the columns of Q_2. Thus,

$$0 = Q_2^T(AQ_1 - Q_1D_1) = Q_2^TQ_2E_{21} = E_{21}.$$

and so (8.1.1) holds. It is easy to show

$$\det(A - \lambda I_n) = \det(Q^TAQ - \lambda I_n) = \det(D_1 - \lambda I_r)\det(D_2 - \lambda I_{n-r})$$

confirming that $\lambda(A) = \lambda(D_1) \cup \lambda(D_2)$. □

The sensitivity to perturbation of an invariant subspace depends upon the separation of the associated eigenvalues from the rest of the spectrum. The appropriate measure of separation between the eigenvalues of two symmetric matrices B and C is given by

$$\text{sep}(B,C) = \min_{\substack{\lambda\in\lambda(B)\\ \mu\in\lambda(C)}} |\lambda - \mu|. \tag{8.1.2}$$

With this definition we have

Theorem 8.1.10 *Suppose A and $A+E$ are n-by-n symmetric matrices and that*

$$Q = \begin{array}{c} [\; Q_1 \quad Q_2 \;] \\ \quad r \quad\; n-r \end{array}$$

is an orthogonal matrix such that $\text{ran}(Q_1)$ *is an invariant subspace for A. Partition the matrices Q^TAQ and Q^TEQ as follows:*

$$Q^TAQ = \begin{array}{cc} \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix} & \begin{array}{c} r \\ n-r \end{array} \\ \begin{array}{cc} r & n-r \end{array} & \end{array} \qquad Q^TEQ = \begin{array}{cc} \begin{bmatrix} E_{11} & E_{21}^T \\ E_{21} & E_{22} \end{bmatrix} & \begin{array}{c} r \\ n-r \end{array} \\ \begin{array}{cc} r & n-r \end{array} & \end{array}.$$

If $\text{sep}(D_1, D_2) > 0$ *and*

$$\| E \|_2 \le \frac{\text{sep}(D_1, D_2)}{5},$$

then there exists a matrix $P \in \mathbb{R}^{(n-r)\times r}$ with

$$\| P \|_2 \le \frac{4}{\text{sep}(D_1, D_2)} \| E_{21} \|_2$$

such that the columns of $\hat{Q}_1 = (Q_1 + Q_2P)(I + P^TP)^{-1/2}$ define an orthonormal basis for a subspace that is invariant for $A+E$. See also Theorem 7.2.4.

Proof. This result is a slight adaptation of of Theorem 4.11 in Stewart (1973). The matrix $(I + P^TP)^{-1/2}$ is the inverse of the square root of $I + P^TP$. See §4.2.10. □

Corollary 8.1.11 *If the conditions of the theorem hold, then*

$$\text{dist}(\text{ran}(Q_1), \text{ran}(\hat{Q}_1)) \le \frac{4}{\text{sep}(D_1, D_2)} \| E_{21} \|_2.$$

See also Corollary 7.2.5.

Proof. It can be shown using the SVD that

$$\| P(I + P^TP)^{-1/2} \|_2 \le \| P \|_2. \tag{8.1.3}$$

Since $Q_2^T\hat{Q}_1 = P(I + P^HP)^{-1/2}$ it follows that

$$\begin{aligned} \text{dist}(\text{ran}(Q_1), \text{ran}(\hat{Q}_1)) &= \| Q_2^T\hat{Q}_1 \|_2 = \| P(I + P^HP)^{-1/2} \|_2 \\ &\le \| P \|_2 \le \| E_{21} \|_2/\text{sep}(D_1, D_2). \ \square \end{aligned}$$

Thus, the reciprocal of $\text{sep}(D_1, D_2)$ can be thought of as a condition number that measures the sensitivity of $\text{ran}(Q_1)$ as an invariant subspace.

The effect of perturbations on a single eigenvector is sufficiently important that we specialize the above results to this important case.

Theorem 8.1.12 *Suppose A and $A + E$ are n-by-n symmetric matrices and that*

$$Q = \begin{array}{c} [\ q_1 \quad Q_2\ ] \\ \quad 1 \quad n-1 \end{array}$$

is an orthogonal matrix such that q_1 is an eigenvector for A. Partition the matrices Q^TAQ and Q^TEQ as follows:

$$Q^TAQ = \begin{array}{cc} \left[\begin{array}{cc} \lambda & 0 \\ 0 & D_2 \end{array}\right] & \begin{array}{c} 1 \\ n-1 \end{array} \\ \begin{array}{cc} 1 & n-1 \end{array} & \end{array} \qquad Q^TEQ = \begin{array}{cc} \left[\begin{array}{cc} \epsilon & e^T \\ e & E_{22} \end{array}\right] & \begin{array}{c} 1 \\ n-1 \end{array} \\ \begin{array}{cc} 1 & n-1 \end{array} & \end{array} .$$

If $d = \min_{\mu \in \lambda(D_2)} |\lambda - \mu| > 0$ *and*

$$\| E \|_2 \le \frac{d}{4},$$

then there exists $p \in \mathbb{R}^{n-1}$ satisfying

$$\| p \|_2 \le \frac{4}{d} \| e \|_2$$

such that $\hat{q}_1 = (q_1 + Q_2 p)/\sqrt{1+p^Tp}$ *is a unit 2-norm eigenvector for* $A+E$. *Moreover,*

$$\text{dist}(\text{span}\{q_1\}, \text{span}\{\hat{q}_1\}) \;=\; \sqrt{1-(q_1^T\hat{q}_1)^2} \;\leq\; \frac{4}{d}\,\|\, e \,\|_2.$$

See also Corollary 7.2.6.

Proof. Apply Theorem 8.1.10 and Corollary 8.1.11 with $r = 1$ and observe that if $D_1 = (\lambda)$, then $d = \text{sep}(D_1, D_2)$. □

Example 8.1.6 If $A = \text{diag}(.999,\ 1.001,\ 2.)$, and

$$E \;=\; \begin{bmatrix} 0.00 & 0.01 & 0.01 \\ 0.01 & 0.00 & 0.01 \\ 0.01 & 0.01 & 0.00 \end{bmatrix},$$

then $\hat{Q}^T(A+E)\hat{Q} = \text{diag}(.9899,\ 1.0098,\ 2.0002)$ where

$$\hat{Q} \;=\; \begin{bmatrix} -.7418 & .6706 & .0101 \\ .6708 & .7417 & .0101 \\ .0007 & -.0143 & .9999 \end{bmatrix}$$

is orthogonal. Let $\hat{q}_i = \hat{Q}e_i$, $i = 1,2,3$. Thus, $\hat{q}_i$ is the perturbation of A's eigenvector $q_i = e_i$. A calculation shows that

$$\text{dist}\{\text{span}\{q_1\}, \text{span}\{\hat{q}_1\}\} \;=\; \text{dist}\{\text{span}\{q_2\}, \text{span}\{\hat{q}_2\}\} \;=\; .67$$

Thus, because they are associated with nearby eigenvalues, the eigenvectors q_1 and q_2 cannot be computed accurately. On the other hand, since λ_1 and λ_2 are well separated from λ_3, they define a two-dimensional subspace that is not particularly sensitive as $\text{dist}\{\text{span}\{q_1, q_2\}, \text{span}\{\hat{q}_1, \hat{q}_2\}\} = .01$.

8.1.4 Approximate Invariant Subspaces

If the columns of $Q_1 \in \mathbb{R}^{n\times r}$ are independent and the *residual matrix* $R = AQ_1 - Q_1 S$ is small for some $S \in \mathbb{R}^{r\times r}$, then the columns of Q_1 define an approximate invariant subspace. Let us discover what we can say about the eigensystem of A when in the possession of such a matrix.

Theorem 8.1.13 *Suppose* $A \in \mathbb{R}^{n\times n}$ *and* $S \in \mathbb{R}^{r\times r}$ *are symmetric and that*

$$AQ_1 - Q_1 S \;=\; E_1$$

where $Q_1 \in \mathbb{R}^{n\times r}$ *satisfies* $Q_1^TQ_1 = I_r$. *Then there exist* $\mu_1, \ldots, \mu_r \in \lambda(A)$ *such that*

$$|\mu_k - \lambda_k(S)| \;\leq\; \sqrt{2}\,\|\, E_1 \,\|_2$$

for $k = 1{:}r$.

Proof. Let $Q_2 \in \mathbb{R}^{n\times(n-r)}$ be any matrix such that $Q = [\ Q_1,\ Q_2\ ]$ is orthogonal. It follows that

$$Q^TAQ = \begin{bmatrix} S & 0 \\ 0 & Q_2^TAQ_2 \end{bmatrix} + \begin{bmatrix} Q_1^TE_1 & E_1^TQ_2 \\ Q_2^TE_1 & 0 \end{bmatrix} \equiv B + E$$

and so by using Corollary 8.1.6 we have $|\lambda_k(A) - \lambda_k(B)| \leq \| E \|_2$ for $k = 1{:}n$. Since $\lambda(S) \subseteq \lambda(B)$, there exist $\mu_1, \ldots, \mu_r \in \lambda(A)$ such that

$$|\mu_k - \lambda_k(S)| \leq \| E \|_2$$

for $k = 1{:}r$. The theorem follows by noting that for any $x \in \mathbb{R}^r$ and $y \in \mathbb{R}^{n-r}$ we have

$$\left\| E \begin{bmatrix} x \\ y \end{bmatrix} \right\|_2 \leq \| E_1x \|_2 + \| E_1^TQ_2y \|_2 \leq \| E_1 \|_2 \| x \|_2 + \| E_1 \|_2 \| y \|_2$$

from which we readily conclude that $\| E \|_2 \leq \sqrt{2}\| E_1 \|_2$. □

Example 8.1.7 If

$$A = \begin{bmatrix} 6.8 & 2.4 \\ 2.4 & 8.2 \end{bmatrix}, \qquad Q_1 = \begin{bmatrix} .7994 \\ .6007 \end{bmatrix}, \text{ and } S = (5.1) \in \mathbb{R}$$

then

$$AQ_1 - Q_1S = \begin{bmatrix} -.0828 \\ -.0562 \end{bmatrix} = E_1.$$

The theorem predicts that A has an eigenvalue within $\sqrt{2}\ \| E_1 \|_2 \approx .1415$ of 5.1. This is true since $\lambda(A) = \{5, 10\}$.

The eigenvalue bounds in Theorem 8.1.13 depend on $\| AQ_1 - Q_1S \|_2$. Given A and Q_1, the following theorem indicates how to choose S so that this quantity is minimized in the Frobenius norm.

Theorem 8.1.14 *If $A \in \mathbb{R}^{n\times n}$ is symmetric and $Q_1 \in \mathbb{R}^{n\times r}$ has orthonormal columns, then*

$$\min_{S\in\mathbb{R}^{r\times r}} \| AQ_1 - Q_1S \|_F = \| (I - Q_1Q_1^T)AQ_1 \|_F$$

and $S = Q_1^TAQ_1$ is the minimizer.

Proof. Let $Q_2 \in \mathbb{R}^{n\times(n-r)}$ be such that $Q = [\ Q_1,\ Q_2\ ]$ is orthogonal. For any $S \in \mathbb{R}^{r\times r}$ we have

$$\begin{aligned} \| AQ_1 - Q_1S \|_F^2 &= \| Q^TAQ_1 - Q^TQ_1S \|_F^2 \\ &= \| Q_1^TAQ_1 - S \|_F^2 + \| Q_2^TAQ_1 \|_F^2. \end{aligned}$$

Clearly, the minimizing S is given by $S = Q_1^T A Q_1$. □

This result enables us to associate any r-dimensional subspace ran(Q_1), with a set of r "optimal" eigenvalue-eigenvector approximates.

Theorem 8.1.15 *Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric and that $Q_1 \in \mathbb{R}^{n\times r}$ satisfies $Q_1^T Q_1 = I_r$. If*

$$Z^T(Q_1^T A Q_1)Z \;=\; \mathrm{diag}(\theta_1,\ldots,\theta_r) \;=\; D$$

is the Schur decomposition of $Q_1^T A Q_1$ and $Q_1 Z = [\, y_1,\ldots,y_r \,]$, then

$$\| Ay_k - \theta_k y_k \|_2 \;=\; \| (I - Q_1Q_1^T)AQ_1Ze_k \|_2 \;\le\; \| (I - Q_1Q_1^T)AQ_1 \|_2$$

for $k = 1{:}r$.

Proof.

$$Ay_k - \theta_k y_k \;=\; AQ_1Ze_k = Q_1ZDe_k \;=\; (AQ_1 - Q_1(Q_1^TAQ_1))Ze_k.$$

The theorem follows by taking norms. □

In Theorem 8.1.15, the θ_k are called *Ritz values,* the y_k are called *Ritz vectors,* and the (θ_k, y_k) are called *Ritz pairs.*

The usefulness of Theorem 8.1.13 is enhanced if we weaken the assumption that the columns of Q_1 are orthonormal. As can be expected, the bounds deteriorate with the loss of orthogonality.

Theorem 8.1.16 *Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric and that*

$$AX_1 - X_1S \;=\; F_1,$$

where $X_1 \in \mathbb{R}^{n\times r}$ and $S = X_1^TAX_1$. If

$$\| X_1^TX_1 - I_r \|_2 = \tau < 1, \tag{8.1.4}$$

then there exist $\mu_1,\ldots,\mu_r \in \lambda(A)$ such that

$$|\mu_k - \lambda_k(S)| \;\le\; \sqrt{2}\,(\| F_1 \|_2 \;+\; \tau(2+\tau)\| A \|_2)$$

for $k = 1{:}r$.

Proof. Let $X_1 = ZP$ be the polar decomposition of X_1. Recall from §4.2.10 that this means $Z \in \mathbb{R}^{n\times r}$ has orthonormal columns and $P \in \mathbb{R}^{k\times k}$ is a symmetric positive semidefinite matrix that satisfies $P^2 = X_1^TX_1$. Taking norms in the equation

$$\begin{aligned} E_1 \equiv AZ - ZS \;&=\; (AX_1 - X_1S) \;+\; A(Z - X_1) \;-\; (Z - X_1)S \\ &=\; F_1 \;+\; AZ(I - P) \;-\; Z(I - P)X_1^TAX_1 \end{aligned}$$

gives

$$\| E_1 \|_2 \;\le\; \| F_1 \|_2 \;+\; \| A \|_2 \, \| I - P \|_2 \, \left(1 + \| X_1 \|_2^2 \,\right). \tag{8.1.5}$$

Equation (8.1.4) implies that

$$\| X_1 \|_2^2 \;\le 1 + \tau. \tag{8.1.6}$$

Since P is positive semidefinite, $(I + P)$ is nonsingular and so

$$I - P = (I + P)^{-1}(I - P^2) = (I + P)^{-1}(I - X_1^T X_1)$$

which implies $\| I - P \|_2 \le \tau$. By substituting this inequality and (8.1.6) into (8.1.5) we have $\| E_1 \|_2 \le \| F_1 \|_2 + \tau(2 + \tau)\| A \|_2$. The proof is completed by noting that we can use Theorem 8.1.13 with $Q_1 = Z$ to relate the eigenvalues of A and S via the residual E_1. □

8.1.5 The Law of Inertia

The *inertia* of a symmetric matrix A is a triplet of nonnegative integers (m, z, p) where m, z, and p are respectively the number of negative, zero, and positive elements of $\lambda(A)$.

Theorem 8.1.17 (Sylvester Law of Inertia) *If $A \in \mathbb{R}^{n\times n}$ is symmetric and $X \in \mathbb{R}^{n\times n}$ is nonsingular, then A and X^TAX have the same inertia.*

Proof. Suppose for some r that $\lambda_r(A) > 0$ and define the subspace $S_0 \subseteq \mathbb{R}^n$ by

$$S_0 \;=\; \text{span}\{X^{-1}q_1, \ldots, X^{-1}q_r\}, \qquad q_i \ne 0$$

where $Aq_i = \lambda_i(A)q_i$ and $i = 1{:}r$. From the minimax characterization of $\lambda_r(X^TAX)$ we have

$$\lambda_r(X^TAX) \;=\; \max_{\dim(S)=r} \; \min_{y\in S} \; \frac{y^T(X^TAX)y}{y^Ty} \;\ge\; \min_{y\in S_0} \; \frac{y^T(X^TAX)y}{y^Ty}.$$

Since

$$y \in \mathbb{R}^n \;\Rightarrow\; \frac{y^T(X^TX)y}{y^Ty} \;\ge\; \sigma_n(X)^2$$

$$y \in S_0 \;\Rightarrow\; \frac{y^T(X^TAX)y}{y^Ty} \;\ge\; \lambda_r(A)$$

it follows that

$$\lambda_r(X^TAX) \;\ge\; \min_{y\in S_0} \; \left\{ \frac{y^T(X^TAX)y}{y^T(X^TX)y} \, \frac{y^T(X^TX)y}{y^Ty} \right\} \;\ge\; \lambda_r(A)\sigma_n(X)^2.$$

An analogous argument with the roles of A and X^TAX reversed shows that

$$\lambda_r(A) \ \geq \ \lambda_r(X^TAX)\sigma_n(X^{-1})^2 \ = \ \frac{\lambda_r(X^TAX)}{\sigma_1(X)^2}.$$

Thus, $\lambda_r(A)$ and $\lambda_r(X^TAX)$ have the same sign and so we have shown that A and X^TAX have the same number of positive eigenvalues. If we apply this result to $-A$, we conclude that A and X^TAX have the same number of negative eigenvalues. Obviously, the number of zero eigenvalues possessed by each matrix is also the same. □

Example 8.1.8 If $A = \text{diag}(3, 2, -1)$ and

$$X \ = \ \begin{bmatrix} 1 & 4 & 5 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix},$$

then

$$X^TAX \ = \ \begin{bmatrix} 3 & 12 & 15 \\ 12 & 50 & 64 \\ 15 & 64 & 82 \end{bmatrix}$$

and $\lambda(X^TAX) \ = \ \{134.769, \ .3555, \ -.1252\}$.

Problems

P8.1.1 Without using any of the results in this section, show that the eigenvalues of a 2-by-2 symmetric matrix must be real.

P8.1.2 Compute the Schur decomposition of $A = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}$.

P8.1.3 Show that the eigenvalues of a Hermitian matrix ($A^H = A$) are real. For each theorem and corollary in this section, state and prove the corresponding result for Hermitian matrices. Which results have analogs when A is skew-symmetric? (Hint: If $A^T = -A$, then iA is Hermitian.)

P8.1.4 Show that if $X \in \mathbb{R}^{n \times r}$, $r \leq n$, and $\| X^TX - I \| = \tau < 1$, then $\sigma_{min}(X) \geq 1-\tau$.

P8.1.5 Suppose $A, E \in \mathbb{R}^{n \times n}$ are symmetric and consider the Schur decomposition $A + tE = QDQ^T$ where we *assume* that $Q = Q(t)$ and $D = D(t)$ are continuously differentiable functions of $t \in \mathbb{R}$. Show that $\dot{D}(t) \ = \ \text{diag}(Q(t)^TEQ(t))$ where the matrix on the right is the diagonal part of $Q(t)^TEQ(t)$. Establish the Wielandt-Hoffman theorem by integrating both sides of this equation from 0 to 1 and taking Frobenius norms to show that

$$\| D(1) - D(0) \|_F \ \leq \ \int_0^1 \| \text{diag}(Q(t)^TEQ(t) \|_F dt \ \leq \ \| E \|_F .$$

P8.1.6 Prove Theorem 8.1.5.

P8.1.7 Prove Theorem 8.1.7.

P8.1.8 If $C \in \mathbb{R}^{n \times n}$ then the *trace function* $\text{tr}(C) = c_{11} + \cdots + c_{nn}$ equals the sum of C's eigenvalues. Use this to prove Theorem 8.1.8.

P8.1.9 Show that if $B \in \mathbb{R}^{m \times m}$ and $C \in \mathbb{R}^{n \times n}$ are symmetric, then $\text{sep}(B, C) = \min$

$\| BX - XC \|_F$ where the min is taken over all matrices in $\mathbb{R}^{m\times n}$.

P8.1.10 Prove the inequality (8.1.3).

P8.1.11 Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric and $C \in \mathbb{R}^{n\times r}$ has full column rank and assume that $r \ll n$. By using Theorem 8.1.8 relate the eigenvalues of $A + CC^T$ to the eigenvalues of A.

Notes and References for Sec. 8.1

The perturbation theory for the symmetric eigenvalue problem is surveyed in Wilkinson (1965, chapter 2), Parlett (1980, chapters 10 and 11), and Stewart and Sun (1990, chapters 4 and 5). Some representative papers in this well-researched area include

G.W. Stewart (1973). "Error and Perturbation Bounds for Subspaces Associated with Certain Eigenvalue Problems," *SIAM Review 15,* 727–64.

C.C. Paige (1974). "Eigenvalues of Perturbed Hermitian Matrices," *Lin. Alg. and Its Applic* . 8, 1–10.

A. Ruhe (1975). "On the Closeness of Eigenvalues and Singular Values for Almost Normal Matrices," *Lin. Alg. and Its Applic. 11,* 87–94.

W. Kahan (1975). "Spectra of Nearly Hermitian Matrices," *Proc. Amer. Math. Soc. 48,* 11–17.

A. Schonhage (1979). "Arbitrary Perturbations of Hermitian Matrices," *Lin. Alg. and Its Applic. 24,* 143–49.

P. Deift, T. Nanda, and C. Tomei (1983). "Ordinary Differential Equations and the Symmetric Eigenvalue Problem," *SIAM J. Numer. Anal. 20,* 1–22.

D.S. Scott (1985). "On the Accuracy of the Gershgorin Circle Theorem for Bounding the Spread of a Real Symmetric Matrix," *Lin. Alg. and Its Applic. 65,* 147–155

J.-G. Sun (1995). "A Note on Backward Error Perturbations for the Hermitian Eigenvalue Problem," *BIT 35,* 385–393.

R.-C. Li (1996). "Relative Perturbation Theory (I) Eigenvalue and Singular Value Variations," Technical Report UCB//CSD-94-855, Department of EECS, University of California at Berkeley.

R.-C. Li (1996). "Relative Perturbation Theory (II) Eigenspace and Singular Subspace Variations," Technical Report UCB//CSD-94-856, Department of EECS, University of California at Berkeley.

8.2 Power Iterations

Assume that $A \in \mathbb{R}^{n\times n}$ is symmetric and that $U_0 \in \mathbb{R}^{n\times n}$ is orthogonal. Consider the following *QR iteration:*

```
$T_0 = U_0^T A U_0$
for $k = 1, 2, \ldots$
    $T_{k-1} = U_k R_k$    (QR factorization)        (8.2.1)
    $T_k = R_k U_k$
end
```

Since $T_k = R_k U_k = U_k^T (U_k R_k) U_k = U_k^T T_{k-1} U_k$ it follows by induction that

$$T_k = (U_0 U_1 \cdots U_k)^T A (U_0 U_1 \cdots U_k). \tag{8.2.2}$$

Thus, each T_k is orthogonally similar to A. Moreover, the T_k almost always converge to diagonal form and so it can be said that (8.2.1) almost always "converges" to a Schur decomposition of A. In order to establish this remarkable result we first consider the power method and the method of orthogonal iteration.

8.2.1 The Power Method

Given a unit 2-norm $q^{(0)} \in \mathbb{R}^n$, the *power method* produces a sequence of vectors $q^{(k)}$ as follows:

```
for $k = 1, 2, \ldots$
    $z^{(k)} = Aq^{(k-1)}$
    $q^{(k)} = z^{(k)}/\| z^{(k)} \|_2$                (8.2.3)
    $\lambda^{(k)} = [q^{(k)}]^T Aq^{(k)}$
end
```

If $q^{(0)}$ is not "deficient" and A's eigenvalue of maximum modulus is unique, then the $q^{(k)}$ converge to an eigenvector.

Theorem 8.2.1 *Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric and that*

$$Q^T AQ = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$$

where $Q = [\, q_1, \ldots, q_n \,]$ is orthogonal and $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$. Let the vectors q_k be specified by (8.2.3) and define $\theta_k \in [0, \pi/2]$ by

$$\cos(\theta_k) = \left| q_1^T q^{(k)} \right| .$$

If $\cos(\theta_0) \neq 0$, *then*

$$|\sin(\theta_k)| \;\leq\; \tan(\theta_0) \left| \frac{\lambda_2}{\lambda_1} \right|^k \tag{8.2.4}$$

$$|\lambda^{(k)} - \lambda| \;\leq\; |\lambda_1 - \lambda_n| \tan(\theta_0)^2 \left| \frac{\lambda_2}{\lambda_1} \right|^{2k} . \tag{8.2.5}$$

Proof. From the definition of the iteration, it follows that $q^{(k)}$ is a multiple of $A^k q^{(0)}$ and so

$$|\sin(\theta_k)|^2 \;=\; 1 \;-\; \left(q_1^T q^{(k)} \right)^2 \;=\; 1 \;-\; \left(\frac{q_1^T A^k q^{(0)}}{\| A^k q^{(0)} \|_2} \right)^2 .$$

If $q^{(0)}$ has the eigenvector expansion $q^{(0)} = a_1 q_1 + \cdots + a_n q_n$, then

$$|a_1| = |q_1^T q^{(0)}| = \cos(\theta_0) \neq 0,$$

$$a_1^2 + \cdots + a_n^2 = 1,$$

and

$$A^k q^{(0)} = a_1\lambda_1^k q_1 + a_2\lambda_2^k q_2 + \cdots + a_n\lambda_n^k q_n .$$

Thus,

$$\begin{aligned}
|\sin(\theta_k)|^2 &= 1 - \frac{a_1^2\lambda_1^{2k}}{\sum_{i=1}^n a_i^2\lambda_i^{2k}} = \frac{\sum_{i=2}^n a_i^2\lambda_i^{2k}}{\sum_{i=1}^n a_i^2\lambda_i^{2k}} \\
&\le \frac{\sum_{i=2}^n a_i^2\lambda_i^{2k}}{a_1^2\lambda_1^{2k}} = \frac{1}{a_1^2}\sum_{i=2}^n a_i^2\left(\frac{\lambda_i}{\lambda_1}\right)^{2k} \\
&\le \frac{1}{a_1^2}\left(\sum_{i=2}^n a_i^2\right)\left(\frac{\lambda_2}{\lambda_1}\right)^{2k} = \frac{1-a_1^2}{a_1^2}\left(\frac{\lambda_2}{\lambda_1}\right)^{2k} \\
&= \tan(\theta_0)^2\left(\frac{\lambda_2}{\lambda_1}\right)^{2k} .
\end{aligned}$$

This proves (8.2.4). Likewise,

$$\lambda^{(k)} = \left[q^{(k)}\right]^T A q^{(k)} = \frac{\left[q^{(0)}\right]^T A^{2k+1} q^{(0)}}{\left[q^{(0)}\right]^T A^{2k} q^{(0)}} = \frac{\sum_{i=1}^n a_i^2\lambda_i^{2k+1}}{\sum_{i=1}^n a_i^2\lambda_i^{2k}}$$

and so

$$\begin{aligned}
\left|\lambda^{(k)} - \lambda_1\right| &= \left|\frac{\sum_{i=2}^n a_i^2\lambda_i^{2k}(\lambda_i - \lambda_1)}{\sum_{i=1}^n a_i^2\lambda_i^{2k}}\right| \le |\lambda_1 - \lambda_n|\frac{1}{a_1^2}\sum_{i=2}^n a_i^2\left(\frac{\lambda_i}{\lambda_1}\right)^{2k} \\
&\le |\lambda_1 - \lambda_n|\tan(\theta_0)^2\left(\frac{\lambda_2}{\lambda_1}\right)^{2k} . \ \square
\end{aligned}$$

Example 8.2.1 The eigenvalues of

$$A = \begin{bmatrix} -1.6407 & 1.0814 & 1.2014 & 1.1539 \\ 1.0814 & 4.1573 & 7.4035 & -1.0463 \\ 1.2014 & 7.4035 & 2.7890 & -1.5737 \\ 1.1539 & -1.0463 & -1.5737 & 8.6944 \end{bmatrix}$$

are given by $\lambda(A) = \{12, 8, -4, -2\}$. If (8.2.3) is applied to this matrix with $q^{(0)} = [1\,0\,0\,0]^T$, then

k	$\lambda^{(k)}$
1	2.3156
2	8.6802
3	10.3163
4	11.0663
5	11.5259
6	11.7747
7	11.8967
8	11.9534
9	11.9792
10	11.9907

Observe the convergence to $\lambda_1 = 12$ with rate $|\lambda_2/\lambda_1|^{2k} = (8/12)^{2k} = (4/9)^k$.

Computable error bounds for the power method can be obtained by using Theorem 8.1.13. If

$$\| Aq^{(k)} - \lambda^{(k)}q^{(k)} \|_2 = \delta,$$

then there exists $\lambda \in \lambda(A)$ such that $|\lambda^{(k)} - \lambda| \le \sqrt{2}\delta$.

8.2.2 Inverse Iteration

Suppose the power method is applied with A replaced by $(A - \lambda I)^{-1}$. If λ is very close to a distinct eigenvalue of A, then the next iterate vector will be very rich in the corresponding eigendirection:

$$\left.\begin{array}{l} x = \displaystyle\sum_{i=1}^{n} a_i q_i \\[2ex] Aq_i = \lambda_i q_i,\ i = 1{:}n \end{array}\right\} \Rightarrow (A - \lambda I)^{-1}x = \sum_{i=1}^{n} \frac{a_i}{\lambda_i - \lambda} q_i.$$

Thus, if $\lambda \approx \lambda_j$ and a_j is not too small, then this vector has a strong component in the direction of q_j. This process is called *inverse iteration* and it requires the solution of a linear system with matrix of coefficients $A - \lambda I$.

8.2.3 Rayleigh Quotient Iteration

Suppose $A \in \mathbb{R}^{n \times n}$ is symmetric and that x is a given nonzero n-vector. A simple differentiation reveals that

$$\lambda = r(x) \equiv \frac{x^T A x}{x^T x}$$

minimizes $\| (A - \lambda I)x \|_2$. (See also Theorem 8.1.14.) The scalar $r(x)$ is called the *Rayleigh quotient* of x. Clearly, if x is an approximate eigenvector, then r(x) is a reasonable choice for the corresponding eigenvalue.

Combining this idea with inverse iteration gives rise to the *Rayleigh quotient iteration:*

```
$x_0$ given, $\| x_0 \|_2 = 1$
for $k = 0, 1, \ldots$
    $\mu_k = r(x_k)$                                        (8.2.6)
    Solve $(A - \mu_k I)z_{k+1} = x_k$ for $z_{k+1}$
    $x_{k+1} = z_{k+1}/\| z_{k+1} \|_2$
end
```

Example 8.2.2 If (8.2.6) is applied to

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 3 & 6 & 10 & 15 & 21 \\ 1 & 4 & 10 & 20 & 35 & 56 \\ 1 & 5 & 15 & 35 & 70 & 126 \\ 1 & 6 & 21 & 56 & 126 & 252 \end{bmatrix}$$

with $x_0 = [1,\ 1,\ 1,\ 1,\ 1,\ 1]^T/6$, then

k	μ_k
0	153.8333
1	120.0571
2	49.5011
3	13.8687
4	15.4959
5	15.5534

The iteration is converging to the eigenvalue $\lambda = 15.5534732737$.

The Rayleigh quotient iteration almost always converges and when it does, the rate of convergence is cubic. We demonstrate this for the case $n = 2$. Without loss of generality, we may assume that $A = \text{diag}(\lambda_1, \lambda_2)$, with $\lambda_1 > \lambda_2$. Denoting x_k by

$$x_k = \begin{bmatrix} c_k \\ s_k \end{bmatrix} \qquad c_k^2 + s_k^2 = 1$$

it follows that $\mu_k = \lambda_1 c_k^2 + \lambda_1 s_k^2$ in (8.2.6) and

$$z_{k+1} = \frac{1}{\lambda_1 - \lambda_2} \begin{bmatrix} c_k/s_k^2 \\ -s_k/c_k^2 \end{bmatrix}.$$

A calculation shows that

$$c_{k+1} = \frac{c_k^3}{\sqrt{c_k^6 + s_k^6}} \qquad s_{k+1} = \frac{-s_k^3}{\sqrt{c_k^6 + s_k^6}}. \qquad (8.2.7)$$

From these equations it is clear that the x_k converge cubically to either $\text{span}\{e_1\}$ or $\text{span}\{e_2\}$ provided $|c_k| \neq |s_k|$.

Details associated with the practical implementation of the Rayleigh quotient iteration may be found in Parlett (1974).

8.2.4 Orthogonal Iteration

A straightforward generalization of the power method can be used to compute higher-dimensional invariant subspaces. Let r be a chosen integer satisfying $1 \leq r \leq n$. Given an n-by-r matrix Q_0 with orthonormal columns, the method of *orthogonal iteration* generates a sequence of matrices $\{Q_k\} \subseteq \mathbb{R}^{n\times r}$ as follows:

```
for $k = 1, 2, \ldots$
    $Z_k = AQ_{k-1}$                                    (8.2.8)
    $Q_k R_k = Z_k$        (QR factorization)
end
```

Note that if $r = 1$, then this is just the power method. Moreover, the sequence $\{Q_k e_1\}$ is precisely the sequence of vectors produced by the power iteration with starting vector $q^{(0)} = Q_0 e_1$.

In order to analyze the behavior of (8.2.8), assume that

$$Q^T A Q \;=\; D \;=\; \mathrm{diag}(\lambda_i) \qquad |\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n| \tag{8.2.9}$$

is a Schur decomposition of $A \in \mathbb{R}^{n\times n}$. Partition Q and D as follows:

$$Q \;=\; \begin{array}{c} [\; Q_\alpha \quad Q_\beta \;] \\ \;\; r \quad\; n-r \end{array} \qquad\qquad D \;=\; \begin{array}{cc} \left[\begin{array}{cc} D_1 & 0 \\ 0 & D_2 \end{array}\right] & \begin{array}{c} r \\ n-r \end{array} \\ \begin{array}{cc} r & n-r \end{array} & \end{array} \tag{8.2.10}$$

If $|\lambda_r| > |\lambda_{r+1}|$, then

$$D_r(A) \;=\; \mathrm{ran}(Q_\alpha)$$

is the *dominant* invariant subspace of dimension r. It is the unique invariant subspace associated with the eigenvalues $\lambda_1, \ldots, \lambda_r$.

The following theorem shows that with reasonable assumptions, the subspaces $\mathrm{ran}(Q_k)$ generated by (8.2.8) converge to $D_r(A)$ at a rate proportional to $|\lambda_{r+1}/\lambda_r|^k$.

Theorem 8.2.2 *Let the Schur decomposition of $A \in \mathbb{R}^{n\times n}$ be given by (8.2.9) and (8.2.10) with $n \geq 2$. Assume that $|\lambda_r| > |\lambda_{r+1}|$ and that the n-by-r matrices $\{Q_k\}$ are defined by (8.2.8). If $\theta \in [0, \pi/2]$ is specified by*

$$\cos(\theta) \;=\; \min_{\substack{u \in D_r(A) \\ v \in \mathrm{ran}(Q_0)}} \frac{|u^T v|}{\|\, u \,\|_2 \|\, v \,\|_2} \;>\; 0\,,$$

then

$$\mathrm{dist}(D_r(A), \mathrm{ran}(Q_k)) \;\leq\; \tan(\theta)\left|\frac{\lambda_{r+1}}{\lambda_r}\right|^k .$$

See also Theorem 7.3.1.

Proof. By induction it can be shown that

$$A^k Q_0 = Q_k (R_k \cdots R_1)$$

and so with the partitionings (8.2.10) we have

$$\begin{bmatrix} D_1^k & 0 \\ 0 & D_2^k \end{bmatrix} \begin{bmatrix} Q_\alpha^T Q_0 \\ Q_\beta^T Q_0 \end{bmatrix} = \begin{bmatrix} Q_\alpha^T Q_k \\ Q_\beta^T Q_k \end{bmatrix} (R_k \cdots R_1).$$

If

$$Q^T Q_k = [Q_\alpha \, , \, Q_\beta]^T Q_k = \begin{bmatrix} Q_\alpha^T Q_k \\ Q_\beta^T Q_k \end{bmatrix} \equiv \begin{bmatrix} V_k \\ W_k \end{bmatrix},$$

then

$$\begin{aligned} \cos(\theta_{min}) &= \sigma_r(V_0) = \sqrt{1 - \| W_0 \|_2^2} \\ \text{dist}(D_r(A), \text{ran}(Q_k)) &= \| W_k \|_2 \\ D_1^k V_0 &= V_k (R_k \cdots R_1) \\ D_2^k W_0 &= W_k (R_k \cdots R_1) \end{aligned}$$

It follows that V_0 is nonsingular which in turn implies that V_k and $(R_k \cdots R_1)$ are also nonsingular. Thus,

$$\begin{aligned} W_k &= D_2^k W_0 (R_k \cdots R_1)^{-1} = D_2^k W_0 \left(V_k^{-1} D_1^k V_0\right)^{-1} \\ &= D_2^k W_0 V_0^{-1} D_1^{-k} V_k \end{aligned}$$

and so

$$\begin{aligned} \| W_k \|_2 &\le \| D_2^k \|_2 \, \| W_0 \|_2 \, \| V_0^{-1} \|_2 \, \| D_1^{-k} \|_2 \, \| V_k \|_2 \\ &\le |\lambda_{r+1}|^k \sin(\theta) \frac{1}{\cos(\theta)} \frac{1}{|\lambda_r|^k} = \tan(\theta) \left| \frac{\lambda_{r+1}}{\lambda_r} \right|^k . \; \square \end{aligned}$$

Example 8.2.3 If (8.2.8) is applied to the matrix of Example 8.2.1 with $r = 2$ and $Q_0 = I_4(:, 1{:}2)$, then

k	dist($D_2(A)$,ran(Q_k))
1	0.8806
2	0.4091
3	0.1121
4	0.0313
5	0.0106
6	0.0044
7	0.0020
8	0.0010
9	0.0005
10	0.0002

8.2.5 The QR Iteration

Consider what happens when we apply the method of orthogonal iteration (8.2.8) with $r = n$. Let $Q^TAQ = \text{diag}(\lambda_1, \ldots, \lambda_n)$ be the Schur decomposition and assume

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|.$$

If $Q = [\, q_1, \ldots, q_n \,]$ and $Q_k = \left[\, q_1^{(k)}, \ldots, q_n^{(k)} \,\right]$ and

$$\text{dist}(D_i(A), \text{span}\{q_1^{(0)}, \ldots, q_i^{(0)}\}) \; < \; 1 \qquad (8.2.11)$$

for $i = 1{:}n-1$, then it follows from Theorem 8.2.2 that

$$\text{dist}(\text{span}\{q_1^{(k)}, \ldots, q_i^{(k)}\}, \text{span}\{q_1, \ldots, q_i\}) \; = \; 0\left(\left|\frac{\lambda_{i+1}}{\lambda_i}\right|^k\right).$$

for $i = 1{:}n-1$. This implies that the matrices T_k defined by

$$T_k \; = \; Q_k^TAQ_k$$

are converging to diagonal form. Thus, it can be said that the method of orthogonal iteration computes a Schur decomposition if $r = n$ and the original iterate $Q_0 \in \mathbb{R}^{n \times n}$ is not deficient in the sense of (8.2.11).

The QR iteration arises by considering how to compute the matrix T_k directly from its predecessor T_{k-1}. On the one hand, we have from (8.2.1) and the definition of T_{k-1} that

$$T_{k-1} = Q_{k-1}^TAQ_{k-1} = Q_{k-1}^T(AQ_{k-1}) = (Q_{k-1}^TQ_k)R_k.$$

On the other hand,

$$T_k = Q_k^TAQ_k = (Q_k^TAQ_{k-1})(Q_{k-1}^TQ_k) = R_k(Q_{k-1}^TQ_k).$$

Thus, T_k is determined by computing the QR factorization of T_{k-1} and then multiplying the factors together in reverse order. This is precisely what is done in (8.2.1).

Example 8.2.4 If the QR iteration (8.2.1) is applied to the matrix in Example 8.2.1, then after 10 iterations

$$T_{10} = \begin{bmatrix} 11.9907 & -0.1926 & -0.0004 & 0.0000 \\ -0.1926 & 8.0093 & -0.0029 & 0.0001 \\ -0.0004 & -0.0029 & -4.0000 & 0.0007 \\ 0.0000 & 0.0001 & 0.0007 & -2.0000 \end{bmatrix}.$$

The off-diagonal entries of the T_k matrices go to zero as follows:

k	$\|T_k(2,1)\|$	$\|T_k(3,1)\|$	$\|T_k(4,1)\|$	$\|T_k(3,2)\|$	$\|T_k(4,2)\|$	$\|T_k(4,3)\|$
1	3.9254	1.8122	3.3892	4.2492	2.8367	1.1679
2	2.6491	1.2841	2.1908	1.1587	3.1473	0.2294
3	2.0147	0.6154	0.5082	0.0997	0.9859	0.0748
4	1.6930	0.2408	0.0970	0.0723	0.2596	0.0440
5	1.2928	0.0866	0.0173	0.0665	0.0667	0.0233
6	0.9222	0.0299	0.0030	0.0405	0.0169	0.0118
7	0.6346	0.0101	0.0005	0.0219	0.0043	0.0059
8	0.4292	0.0034	0.0001	0.0113	0.0011	0.0030
9	0.2880	0.0011	0.0000	0.0057	0.0003	0.0015
10	0.1926	0.0004	0.0000	0.0029	0.0001	0.0007

Note that a single QR iteration involves $O(n^3)$ flops. Moreover, since convergence is only linear (when it exists), it is clear that the method is a prohibitively expensive way to compute Schur decompositions. Fortunately, these practical difficulties can be overcome as we show in the next section.

Problems

P8.2.1 Suppose $A_0 \in \mathbb{R}^{n\times n}$ is symmetric and positive definite and consider the following iteration:

```
for $k = 1, 2, \ldots$
    $A_{k-1} = G_k G_k^T$        (Cholesky)
    $A_k = G_k^T G_k$
end
```

(a) Show that this iteration is defined. (b) Show that if $A_0 = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ with $a \geq c$ has eigenvalues $\lambda_1 \geq \lambda_2 > 0$, then the A_k converge to $\text{diag}(\lambda_1, \lambda_2)$.

P8.2.2 Prove (8.2.7).

P8.2.3 Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric and define the function $f:\mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ by

$$f\begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{bmatrix} Ax - \lambda x \\ (x^T x - 1)/2 \end{bmatrix}$$

where $x \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$. Suppose x_+ and λ_+ are produced by applying Newton's method to f at the "current point" defined by x_c and λ_c. Give expressions for x_+ and λ_+ assuming that $\| x_c \|_2 = 1$ and $\lambda_c = x_c^T A x_c$.

Notes and References for Sec. 8.2

The following references are concerned with the method of orthogonal iteration (a.k.a. the method of simultaneous iteration):

G.W. Stewart (1969). "Accelerating The Orthogonal Iteration for the Eigenvalues of a Hermitian Matrix," *Numer. Math. 13,* 362–76.

M. Clint and A. Jennings (1970). "The Evaluation of Eigenvalues and Eigenvectors of Real Symmetric Matrices by Simultaneous Iteration," *Comp. J. 13,* 76–80.

H. Rutishauser (1970). "Simultaneous Iteration Method for Symmetric Matrices," *Numer. Math. 16,* 205–23. See also Wilkinson and Reinsch (1971,pp.284–302).

References for the Rayleigh quotient method include

J. Vandergraft (1971). "Generalized Rayleigh Methods with Applications to Finding Eigenvalues of Large Matrices," *Lin. Alg. and Its Applic.* *4*, 353–68.

B.N. Parlett (1974). "The Rayleigh Quotient Iteration and Some Generalizations for Nonnormal Matrices," *Math. Comp.* *28*, 679-93.

R.A. Tapia and D.L. Whitley (1988). "The Projected Newton Method Has Order $1+\sqrt{2}$ for the Symmetric Eigenvalue Problem," *SIAM J. Num. Anal.* *25*, 1376–1382.

S. Batterson and J. Smillie (1989). "The Dynamics of Rayleigh Quotient Iteration," *SIAM J. Num. Anal.* *26*, 624–636.

C. Beattie and D.W. Fox (1989). "Localization Criteria and Containment for Rayleigh Quotient Iteration," *SIAM J. Matrix Anal. Appl.* *10*, 80–93.

P.T.P. Tang (1994). "Dynamic Condition Estimation and Rayleigh-Ritz Approximation," *SIAM J. Matrix Anal. Appl.* *15*, 331–346.

8.3 The Symmetric QR Algorithm

The symmetric QR iteration (8.2.1) can be made very efficient in two ways. First, we show how to compute an orthogonal U_0 such that $U_0^T AU = T$ is tridiagonal. With this reduction, the iterates produced by (8.2.1) are all tridiagonal and this reduces the work per step to $O(n^2)$. Second, the idea of shifts are introduced and with this change the convergence to diagonal form proceeds at a cubic rate. This is far better than having the off-diagonal entries going to to zero like $|\lambda_{i+1}/\lambda_i|^k$ as discussed in §8.2.5.

8.3.1 Reduction to Tridiagonal Form

If A is symmetric, then it is possible to find an orthogonal Q such that

$$Q^T AQ = T \tag{8.3.1}$$

is tridiagonal. We call this the *tridiagonal decomposition* and as a compression of data, it represents a very big step towards diagonalization.

We show how to compute (8.3.1) with Householder matrices. Suppose that Householder matrices $P_1, \ldots, P_{k-1}$ have been determined such that if $A_{k-1} = (P_1 \cdots P_{k-1})^T A(P_1 \cdots P_{k-1})$, then

$$A_{k-1} = \begin{array}{cc} \left[\begin{array}{ccc} B_{11} & B_{12} & 0 \\ B_{21} & B_{22} & B_{23} \\ 0 & B_{32} & B_{33} \end{array}\right] & \begin{array}{c} k-1 \\ 1 \\ n-k \end{array} \\ \begin{array}{ccc} k-1 & 1 & n-k \end{array} & \end{array}$$

is tridiagonal through its first $k-1$ columns. If $\bar{P}_k$ is an order $n-k$ Householder matrix such that $\bar{P}_k B_{32}$ is a multiple of $I_{n-k}(:,1)$ and if $P_k =$

$\text{diag}(I_k, \bar{P}_k)$, then the leading k-by-k principal submatrix of

$$A_k \;=\; P_kA_{k-1}P_k \;=\; \begin{array}{cc} \left[\begin{array}{ccc} B_{11} & B_{12} & 0 \\ B_{21} & B_{22} & B_{23}\bar{P}_k \\ 0 & \bar{P}_kB_{32} & \bar{P}_kB_{33}\bar{P}_k \end{array}\right] & \begin{array}{c} k-1 \\ 1 \\ n-k \end{array} \\ \begin{array}{ccc} k-1 & \quad 1 \quad & n-k \end{array} & \end{array}$$

is tridiagonal. Clearly, if $U_0 = P_1 \cdots P_{n-2}$, then $U_0^TAU_0 = T$ is tridiagonal.

In the calculation of A_k it is important to exploit symmetry during the formation of the matrix $\bar{P}_kB_{33}\bar{P}_k$. To be specific, suppose that $\bar{P}_k$ has the form

$$\bar{P}_k \;=\; I - \beta vv^T \qquad \beta = 2/v^Tv, \quad 0 \neq v \in \mathbb{R}^{n-k}.$$

Note that if $p = \beta B_{33}v$ and $w = p - (\beta p^Tv/2)v$, then

$$\bar{P}_kB_{33}\bar{P}_k \;=\; B_{33} - vw^T - wv^T.$$

Since only the upper triangular portion of this matrix needs to be calculated, we see that the transition from A_{k-1} to A_k can be accomplished in only $4(n-k)^2$ flops.

Algorithm 8.3.1 (Householder Tridiagonalization) Given a symmetric $A \in \mathbb{R}^{n\times n}$, the following algorithm overwrites A with $T = Q^TAQ$, where T is tridiagonal and $Q = H_1 \cdots H_{n-2}$ is the product of Householder transformations.

```
for $k = 1{:}n-2$
    $[v, \beta] = \textbf{house}(A(k+1{:}n, k))$
    $p = \beta A(k+1{:}n, k+1{:}n)v$
    $w = p - (\beta p^Tv/2)v$
    $A(k+1, k) = \| A(k+1{:}n, k) \|_2;\; A(k, k+1) = A(k+1.k)$
    $A(k+1{:}n, k+1{:}n) = A(k+1{:}n, k+1{:}n) - vw^T - wv^T$
end
```

This algorithm requires $4n^3/3$ flops when symmetry is exploited in calculating the rank-2 update. The matrix Q can be stored in factored form in the subdiagonal portion of A. If Q is explicitly required, then it can be formed with an additional $4n^3/3$ flops.

Example 8.3.1

$$\left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & .6 & .8 \\ 0 & .8 & -.6 \end{array}\right]^T \left[\begin{array}{ccc} 1 & 3 & 4 \\ 3 & 2 & 8 \\ 4 & 8 & 3 \end{array}\right] \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & .6 & .8 \\ 0 & .8 & -.6 \end{array}\right] = \left[\begin{array}{ccc} 1 & 5 & 0 \\ 5 & 10.32 & 1.76 \\ 0 & 1.76 & -5.32 \end{array}\right].$$

Note that if T has a zero subdiagonal, then the eigenproblem splits into a pair of smaller eigenproblems. In particular, if $t_{k+1,k} = 0$, then $\lambda(T) =$

$\lambda(T(1{:}k,1{:}k)) \cup \lambda(T(k+1{:}n,k+1{:}n))$. If T has no zero subdiagonal entries, then it is said to be *unreduced.*

Let $\hat{T}$ denote the computed version of T obtained by Algorithm 8.3.1. It can be shown that $\hat{T} = \tilde{Q}^T(A+E)\tilde{Q}$ where $\tilde{Q}$ is exactly orthogonal and E is a symmetric matrix satisfying $\| E \|_F \le c\mathbf{u}\| A \|_F$ where c is a small constant. See Wilkinson (1965, p. 297).

8.3.2 Properties of the Tridiagonal Decomposition

We prove two theorems about the tridiagonal decomposition both of which have key roles to play in the sequel. The first connects (8.3.1) to the QR factorization of a certain *Krylov matrix.* These matrices have the form

$$K(A,v,k) = \left[v,\ Av, \cdots,\ A^{k-1}v\right] \qquad A \in \mathbb{R}^{n\times n},\ v \in \mathbb{R}^n.$$

Theorem 8.3.1 *If $Q^TAQ = T$ is the tridiagonal decomposition of the symmetric matrix $A \in \mathbb{R}^{n\times n}$, then $Q^TK(A,Q(:,1),n) = R$ is upper triangular. If R is nonsingular, then T is unreduced. If R is singular and k is the smallest index so $r_{kk} = 0$, then k is also the smallest index so $t_{k,k-1}$ is zero. See also Theorem 7.4.3.*

Proof. It is clear that if $q_1 = Q(:,1)$, then

$$\begin{aligned} Q^TK(A,Q(:,1),n) &= \left[\,Q^Tq_1,\ (Q^TAQ)(Q^Tq_1),\ldots,(Q^TAQ)^{n-1}(Q^Tq_1)\,\right] \\ &= \left[\,e_1,\ Te_1,\ldots,T^{n-1}e_1\,\right] \;=\; R \end{aligned}$$

is upper triangular with the property that $r_{11} = 1$ and $r_{ii} = t_{21}t_{32}\cdots t_{i,i-1}$ for $i = 2{:}n$. Clearly, if R is nonsingular, then T is unreduced. If R is singular and r_{kk} is its first zero diagonal entry, then $k \ge 2$ and $t_{k,k-1}$ is the first zero subdiagonal entry. □

The next result shows that Q is essentially unique once $Q(:,1)$ is specified.

Theorem 8.3.2 (Implicit Q Theorem) *Suppose $Q = [\,q_1,\ldots,q_n\,]$ and $V = [\,v_1,\ldots,v_n\,]$ are orthogonal matrices with the property that both $Q^TAQ = T$ and $V^TAV = S$ are tridiagonal where $A \in \mathbb{R}^{n\times n}$ is symmetric. Let k denote the smallest positive integer for which $t_{k+1,k} = 0$, with the convention that $k = n$ if T is unreduced. If $v_1 = q_1$, then $v_i = \pm q_i$ and $|t_{i,i-1}| = |s_{i,i-1}|$ for $i = 2{:}k$. Moreover, if $k < n$, then $s_{k+1,k} = 0$. See also Theorem 7.4.2.*

Proof. Define the orthogonal matrix $W = Q^TV$ and observe that $W(:,1) = I_n(:,1) = e_1$ and $W^TTW = S$. By Theorem 8.3.1, $W^TK(T,e_1,k)$ is upper triangular with full column rank. But $K(T,e_1,k)$ is upper triangular and so by the essential uniqueness of the thin QR factorization,

$$W(:,1{:}k) = I_n(:,1{:}k)\mathrm{diag}(\pm 1,\ldots,\pm 1).$$

This says that $Q(:,i) = \pm V(:,i)$ for $i = 1{:}k$. The comments about the subdiagonal entries follows from this since $t_{i+1,i} = Q(:,i+1)^T AQ(:,i)$ and $s_{i+1,i} = V(:,i+1)^T AV(:,i)$ for $i = 1{:}n-1$. □

8.3.3 The QR Iteration and Tridiagonal Matrices

We quickly state four facts that pertain to the QR iteration and tridiagonal matrices. Complete verifications are straight forward.

1. *Preservation of Form.* If $T = QR$ is the QR factorization of a symmetric tridiagonal matrix $T \in \mathbb{R}^{n\times n}$, then Q has lower bandwidth 1 and R has upper bandwidth 2 and it follows that

$$T_+ = RQ = Q^T(QR)Q = Q^TTQ$$

   is also symmetric and tridiagonal.

2. *Shifts.* If $s \in \mathbb{R}$ and $T - sI = QR$ is the QR factorization, then

$$T_+ = RQ + sI = Q^TTQ$$

   is also tridiagonal. This is called a *shifted* QR step.

3. *Perfect Shifts.* If T is unreduced, then the first $n-1$ columns of $T-sI$ are independent regardless of s. Thus, if $s \in \lambda(T)$ and

$$QR = T - sI$$

   is a QR factorization, then $r_{nn} = 0$ and the last column of $T_+ = RQ + sI$ equals $sI_n(:,n) = se_n$.

4. *Cost.* If $T \in \mathbb{R}^{n\times n}$ is tridiagonal, then its QR factorization can be computed by applying a sequence of $n-1$ Givens rotations:

```
for $k = 1{:}n-1$
    $[c,s] = \mathbf{givens}(t_{kk}, t_{k+1,k})$
    $m = \min\{k+2, n\}$
    $T(k{:}k+1, k{:}m) = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}^T T(k{:}k+1, k{:}m)$
end
```

   This requires $O(n)$ flops. If the rotations are accumulated, then $O(n^2)$ flops are needed.

8.3.4 Explicit Single Shift QR Iteration

If s is a good approximate eigenvalue, then we suspect that the $(n, n-1)$ will be small after a QR step with shift s. This is the philosophy behind the following iteration:

```
$T = U_0^T A U_0$        (tridiagonal)
for $k = 0, 1, \ldots$
    Determine real shift $\mu$.                                (8.3.2)
    $T - \mu I = UR$        (QR factorization)
    $T = RU + \mu I$
end
```

If

$$T = \begin{bmatrix} a_1 & b_1 & & \cdots & 0 \\ b_1 & a_2 & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & b_{n-1} \\ 0 & \cdots & & b_{n-1} & a_n \end{bmatrix}.$$

then one reasonable choice for the shift is $\mu = a_n$. However, a more effective choice is to shift by the eigenvalue of

$$T(n-1{:}n, n-1{:}n) = \begin{bmatrix} a_{n-1} & b_{n-1} \\ b_{n-1} & a_n \end{bmatrix}$$

that is closer to a_n. This is known as the *Wilkinson shift* and it is given by

$$\mu = a_n + d - \text{sign}(d)\sqrt{d^2 + b_{n-1}^2} \qquad (8.3.3)$$

where $d = (a_{n-1} - a_n)/2$. Wilkinson (1968b) has shown that (8.3.2) is cubically convergent with either shift strategy, but gives heuristic reasons why (8.3.3) is preferred.

8.3.5 Implicit Shift Version

It is possible to execute the transition from T to $T_+ = RU + \mu I = U^T T U$ without explicitly forming the matrix $T - \mu I$. This has advantages when the shift is much larger than some of the a_i. Let $c = \cos(\theta)$ and $s = \sin(\theta)$ be computed such that

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix}^T \begin{bmatrix} a_1 - \mu \\ b_1 \end{bmatrix} = \begin{bmatrix} \times \\ 0 \end{bmatrix}.$$

If we set $G_1 = G(1,2,\theta)$ then $G_1e_1 = Ue_1$ and

$$T \;\leftarrow\; G_1^TTG_1 \;=\; \begin{bmatrix} \times & \times & + & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 & 0 \\ + & \times & \times & \times & 0 & 0 \\ 0 & 0 & \times & \times & \times & 0 \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \end{bmatrix}.$$

We are thus in a position to apply the Implicit Q theorem provided we can compute rotations $G_2, \ldots, G_{n-1}$ with the property that if $Z = G_1G_2\cdots G_{n-1}$ then $Ze_1 = G_1e_1 = Ue_1$ *and* Z^TTZ is tridiagonal.

Note that the first column of Z and U are identical provided we take each G_i to be of the form $G_i = G(i, i+1, \theta_i)$, $i = 2{:}n-1$. But G_i of this form can be used to chase the unwanted nonzero element "+" out of the matrix $G_1^TTG_1$ as follows:

$$\xrightarrow{G_2} \begin{bmatrix} \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & + & 0 & 0 \\ 0 & \times & \times & \times & 0 & 0 \\ 0 & + & \times & \times & \times & 0 \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \end{bmatrix} \quad \xrightarrow{G_3} \begin{bmatrix} \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 & 0 \\ 0 & \times & \times & \times & + & 0 \\ 0 & 0 & \times & \times & \times & 0 \\ 0 & 0 & + & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \end{bmatrix}$$

$$\xrightarrow{G_4} \begin{bmatrix} \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 & 0 \\ 0 & \times & \times & \times & 0 & 0 \\ 0 & 0 & \times & \times & \times & + \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & + & \times & \times \end{bmatrix} \quad \xrightarrow{G_5} \begin{bmatrix} \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 & 0 \\ 0 & \times & \times & \times & 0 & 0 \\ 0 & 0 & \times & \times & \times & 0 \\ 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times \end{bmatrix}$$

Thus, it follows from the Implicit Q theorem that the tridiagonal matrix Z^TTZ produced by this zero-chasing technique is essentially the same as the tridiagonal matrix T obtained by the explicit method. (We may assume that all tridiagonal matrices in question are unreduced for otherwise the problem decouples.)

Note that at any stage of the zero-chasing, there is only one nonzero entry outside the tridiagonal band. How this nonzero entry moves down the matrix during the update $T \leftarrow G_k^TTG_k$ is illustrated in the following:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c & s & 0 \\ 0 & -s & c & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}^T \begin{bmatrix} a_k & b_k & z_k & 0 \\ b_k & a_p & b_p & 0 \\ z_k & b_p & a_q & b_q \\ 0 & 0 & b_q & a_r \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c & s & 0 \\ 0 & -s & c & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} a_k & b_k & 0 & 0 \\ b_k & a_p & b_p & z_p \\ 0 & b_p & a_q & b_q \\ 0 & z_p & b_q & a_r \end{bmatrix}.$$

Here $(p, q, r) = (k+1, k+2, k+3)$. This update can be performed in about

26 flops once c and s have been determined from the equation $b_k s + z_k c = 0$. Overall we obtain

Algorithm 8.3.2 (Implicit Symmetric QR Step with Wilkinson Shift) Given an unreduced symmetric tridiagonal matrix $T \in \mathbb{R}^{n \times n}$, the following algorithm overwrites T with $Z^T T Z$, where $Z = G_1 \cdots G_{n-1}$ is a product of Givens rotations with the property that $Z^T(T - \mu I)$ is upper triangular and μ is that eigenvalue of T's trailing 2-by-2 principal submatrix closer to t_{nn}.

```
$d = (t_{n-1,n-1} - t_{nn})/2$
$\mu = t_{nn} - t_{n,n-1}^2 / \left(d + \text{sign}(d)\sqrt{d^2 + t_{n,n-1}^2}\right)$
$x = t_{11} - \mu$
$z = t_{21}$
for $k = 1{:}n-1$
    $[\,c, s\,] = \mathbf{givens}(x, z)$
    $T = G_k^T T G_k$,    where $G_k = G(k, k+1, \theta)$
    if $k < n - 1$
        $x = t_{k+1,k}$
        $z = t_{k+2,k}$
    end
end
```

This algorithm requires about $30n$ flops and n square roots. If a given orthogonal matrix Q is overwritten with $QG_1 \cdots G_{n-1}$, then an additional $6n^2$ flops are needed. Of course, in any practical implementation the tridiagonal matrix T would be stored in a pair of n-vectors and not in an n-by-n array.

Example 8.3.2 If Algorithm 8.3.2 is applied to

$$T = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 3 & .01 \\ 0 & 0 & .01 & 4 \end{bmatrix},$$

then the new tridiagonal matrix T is given by

$$T = \begin{bmatrix} .5000 & .5916 & 0 & 0 \\ .5916 & 1.785 & .1808 & 0 \\ 0 & .1808 & 3.7140 & .0000044 \\ 0 & 0 & .0000044 & 4.002497 \end{bmatrix}.$$

Algorithm 8.3.2 is the basis of the symmetric QR algorithm—the standard means for computing the Schur decomposition of a dense symmetric matrix.

Algorithm 8.3.3 (Symmetric QR Algorithm) Given $A \in \mathbb{R}^{n\times n}$ (symmetric) and a tolerance *tol* greater than the unit roundoff, this algorithm computes an approximate symmetric Schur decomposition $Q^TAQ = D$. A is overwritten with the tridiagonal decomposition.

```
Use Algorithm 8.3.1, compute the tridiagonalization
    $T = (P_1 \cdots P_{n-2})^T A (P_1 \cdots P_{n-2})$.
    Set $D = T$ and if $Q$ is desired, form $Q = P_1 \cdots P_{n-2}$. See §5.1.6.
until $q = n$
    For $i = 1{:}n-1$, set $d_{i+1,i}$ and $d_{i,i+1}$ to zero if
        $|d_{i+1,i}| = |d_{i,i+1}| \le tol(|d_{ii}| + |d_{i+1,i+1}|)$
    Find the largest $q$ and the smallest $p$ such that if

        $$D = \begin{array}{cc} \left[\begin{array}{ccc} D_{11} & 0 & 0 \\ 0 & D_{22} & 0 \\ 0 & 0 & D_{33} \end{array}\right] & \begin{array}{c} p \\ n-p-q \\ q \end{array} \\ \begin{array}{ccc} p & n-p-q & q \end{array} & \end{array}$$

    then $D_{33}$ is diagonal and $D_{22}$ is unreduced.
    if $q < n$
        Apply Algorithm 8.3.2 to $D_{22}$:
            $D = \text{diag}(I_p, \bar{Z}, I_q)^T D \ \text{diag}(I_p, \bar{Z}, I_q)$
        If $Q$ is desired, then $Q = Q \ \text{diag}(I_p, \bar{Z}, I_q)$.
    end
end
```

This algorithm requires about $4n^3/3$ flops if Q is not accumulated and about $9n^3$ flops if Q is accumulated.

Example 8.3.3 Suppose Algorithm 8.3.3 is applied to the tridiagonal matrix

$$A = \left[\begin{array}{cccc} 1 & 2 & 0 & 0 \\ 2 & 3 & 4 & 0 \\ 0 & 4 & 5 & 6 \\ 0 & 0 & 6 & 7 \end{array}\right]$$

The subdiagonal entries change as follows during the execution of Algorithm 8.3.3:

Iteration	a_{21}	a_{32}	a_{43}
1	1.6817	3.2344	.8649
2	1.6142	2.5755	.0006
3	1.6245	1.6965	10^{-13}
4	1.6245	1.6965	converg.
5	1.5117	.0150	
6	1.1195	10^{-9}	
7	.7071	converg.	
8	converg.		

Upon completion we find $\lambda(A) = \{-2.4848, .7046, 4.9366, 12.831\}$.

The computed eigenvalues $\hat{\lambda}_i$ obtained via Algorithm 8.3.3 are the exact eigenvalues of a matrix that is near to A, i.e., $Q_0^T(A+E)Q_0 = \text{diag}(\hat{\lambda}_i)$ where $Q_0^TQ_0 = I$ and $\| E \|_2 \approx \mathbf{u}\| A \|_2$. Using Corollary 8.1.6 we know that the absolute error in each $\hat{\lambda}_i$ is small in the sense that $|\hat{\lambda}_i - \lambda_i| \approx \mathbf{u}\| A \|_2$. If $\hat{Q} = [\, \hat{q}_1, \ldots, \hat{q}_n \,]$ is the computed matrix of orthonormal eigenvectors, then the accuracy of $\hat{q}_i$ depends on the separation of λ_i from the remainder of the spectrum. See Theorem 8.1.12.

If all of the eigenvalues and a few of the eigenvectors are desired, then it is cheaper not to accumulate Q in Algorithm 8.3.3. Instead, the desired eigenvectors can be found via inverse iteration with T. See §8.2.2. Usually just one step is sufficient to get a good eigenvector, even with a random initial vector.

If just a few eigenvalues and eigenvectors are required, then the special techniques in §8.5 are appropriate.

It is interesting to note the connection between Rayleigh quotient iteration and the symmetric QR algorithm. Suppose we apply the latter to the tridiagonal matrix $T \in \mathbb{R}^{n\times n}$ with shift $\sigma = e_n^TTe_n = t_{nn}$ where $e_n = I_n(:,n)$. If $T - \sigma I = QR$, then we obtain $T = RQ + \sigma I$. From the equation $(T - \sigma I)Q = R^T$ it follows that

$$(T - \sigma I)q_n = r_{nn}e_n,$$

where q_n is the last column of the orthogonal matrix Q. Thus, if we apply (8.2.6) with $x_0 = e_n$, then $x_1 = q_n$.

8.3.6 Orthogonal Iteration with Ritz Acceleration

Recall from §8.2.4 that an orthogonal iteration step involves a matrix-matrix product and a QR factorization:

$$\begin{aligned} Z_k &= A\tilde{Q}_{k-1} \\ \tilde{Q}_kR_k &= Z_k \quad \text{(QR factorization)} \end{aligned}$$

Theorem 8.1.14 says that we can minimize $\| A\tilde{Q}_k - \tilde{Q}_kS \|_F$ by setting $S = S_k \equiv \tilde{Q}_k^TA\tilde{Q}_k$. If $U_k^TS_kU_k = D_k$ is the Schur decomposition of $S_k \in \mathbb{R}^{r\times r}$ and $Q_k = \tilde{Q}_kU_k$, then

$$\| AQ_k - Q_kD_k \|_F = \| A\tilde{Q}_k - \tilde{Q}_kS_k \|_F$$

showing that the columns of Q_k are the best possible basis to take after k steps from the standpoint of minimizing the residual. This defines the *Ritz acceleration* idea:

```
$Q_0 \in \mathbb{R}^{n\times p}$ given with $Q_0^TQ_0 = I_p$
for $k = 1, 2, \ldots$
    $Z_k = AQ_{k-1}$
    $\tilde{Q}_kR_k = Z_k$    (QR factorization)
    $S_k = \tilde{Q}_k^TA\tilde{Q}_k$                                    (8.3.6)
    $U_k^TS_kU_k = D_k$    (Schur decomposition)
    $Q_k = \tilde{Q}_kU_k$
end
```

It can be shown that if

$$D_k \;=\; \mathrm{diag}(\theta_1^{(k)}, \ldots, \theta_r^{(k)}) \qquad |\theta_1^{(k)}| \geq \cdots \geq |\theta_r^{(k)}|$$

then

$$|\theta_i^{(k)} - \lambda_i(A)| \;=\; O\left(\left|\frac{\lambda_{r+1}}{\lambda_i}\right|^k\right) \qquad i = 1{:}r$$

Recall that Theorem 8.2.2 says the eigenvalues of $\tilde{Q}_k^TA\tilde{Q}_k$ converge with rate $|\lambda_{r+1}/\lambda_r|^k$. Thus, the Ritz values converge at a more favorable rate. For details, see Stewart (1969).

Example 8.3.4 If we apply (8.3.6) with

$$A \;=\; \begin{bmatrix} 100 & 1 & 1 & 1 \\ 1 & 99 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad Q_0 \;=\; \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

then

k	$\mathrm{dist}\{D_2(A), Q_k\}$
0	$.2 \times 10^{-1}$
1	$.5 \times 10^{-3}$
2	$.1 \times 10^{-4}$
3	$.3 \times 10^{-6}$
4	$.8 \times 10^{-8}$

Clearly, convergence is taking place at the rate $(2/99)^k$.

Problems

P8.3.1 Suppose λ is an eigenvalue of a symmetric tridiagonal matrix T. Show that if λ has algebraic multiplicity k, then at least $k-1$ of T's subdiagonal elements are zero.

P8.3.2 Suppose A is symmetric and has bandwidth p. Show that if we perform the shifted QR step $A - \mu I = QR$, $A = RQ + \mu I$, then A has bandwidth p.

P8.3.3 Suppose $B \in \mathbb{R}^{n\times n}$ is upper bidiagonal with diagonal entries $d(1{:}n)$ and superdiagonal entries $f(1{:}n-1)$. State and prove a singular value version of Theorem 8.3.1.

P8.3.4 Let $A = \begin{bmatrix} w & x \\ x & z \end{bmatrix}$ be real and suppose we perform the following shifted QR step: $A - zI = UR$, $\bar{A} = RU + zI$. Show that if $\bar{A} = \begin{bmatrix} \bar{w} & \bar{x} \\ \bar{x} & \bar{z} \end{bmatrix}$ then

$$\begin{aligned}\bar{w} &= w + x^2(w-z)/[(w-z)^2 + x^2]\\ \bar{z} &= z - x^2(w-z)/[(w-z)^2 + x^2]\\ \bar{x} &= -x^3/[(w-z)^2 + x^2].\end{aligned}$$

P8.3.5 Suppose $A \in \mathbb{C}^{n\times n}$ is Hermitian. Show how to construct unitary Q such that $Q^HAQ = T$ is real, symmetric, and tridiagonal.

P8.3.6 Show that if $A = B + iC$ is Hermitian, then $M = \begin{bmatrix} B & -C \\ C & B \end{bmatrix}$ is symmetric. Relate the eigenvalues and eigenvectors of A and M.

P8.3.7 Rewrite Algorithm 8.2.2 for the case when A is stored in two n-vectors. Justify the given flop count.

P8.3.8 Suppose $A = S + \sigma uu^T$ where $S \in \mathbb{R}^{n\times n}$ is skew-symmetric ($A^T = -A$, $u \in \mathbb{R}^n$ has unit 2-norm, and $\sigma \in \mathbb{R}$. Show how to compute an orthogonal Q such that Q^TAQ is tridiagonal and $Q^Tu = I_n(:,1) = e_1$.

Notes and References for Sec. 8.3

The tridiagonalization of a symmetric matrix is discussed in

R.S. Martin and J.H. Wilkinson (1968). "Householder's Tridiagonalization of a Symmetric Matrix," *Numer. Math. 11,* 181-95. See also Wilkinson and Reinsch (1971, pp.212–26).

H.R. Schwartz (1968). "Tridiagonalization of a Symmetric Band Matrix," *Numer. Math. 12,* 231–41. See also Wilkinson and Reinsch (1971, pp.273–83).

N.E. Gibbs and W.G. Poole, Jr. (1974). "Tridiagonalization by Permutations," *Comm. ACM 17,* 20–24.

The first two references contain Algol programs. Algol procedures for the explicit and implicit tridiagonal QR algorithm are given in

H. Bowdler, R.S. Martin, C. Reinsch, and J.H. Wilkinson (1968). "The QR and QL Algorithms for Symmetric Matrices," *Numer. Math. 11,* 293–306. See also Wilkinson and Reinsch (1971, pp.227–40).

A. Dubrulle, R.S. Martin, and J.H. Wilkinson (1968). "The Implicit QL Algorithm," *Numer. Math. 12,* 377-83. see also Wilkinson and Reinsch (1971, pp.241–48).

The "QL" algorithm is identical to the QR algorithm except that at each step the matrix $T - \lambda I$ is factored into a product of an orthogonal matrix and a lower triangular matrix. Other papers concerned with these methods include

G.W. Stewart (1970). "Incorporating Original Shifts into the QR Algorithm for Symmetric Tridiagonal Matrices," *Comm. ACM 13,* 365–67.

A. Dubrulle (1970). "A Short Note on the Implicit QL Algorithm for Symmetric Tridiagonal Matrices," *Numer. Math. 15,* 450.

Extensions to Hermitian and skew-symmetric matrices are described in

D. Mueller (1966). "Householder's Method for Complex Matrices and Hermitian Matrices," *Numer. Math. 8,* 72–92.

R.C. Ward and L.J. Gray (1978). "Eigensystem Computation for Skew-Symmetric and A Class of Symmetric Matrices," *ACM Trans. Math. Soft. 4,* 278–85.

The convergence properties of Algorithm 8.2.3 are detailed in Lawson and Hanson (1974, Appendix B), as well as in

J.H. Wilkinson (1968b). "Global Convergence of Tridiagonal QR Algorithm With Origin Shifts," *Lin. Alg. and Its Applic. 1,* 409–20.

T.J. Dekker and J.F. Traub (1971). "The Shifted QR Algorithm for Hermitian Matrices," *Lin. Alg. and Its Applic. 4,* 137–54.

W. Hoffman and B.N. Parlett (1978). "A New Proof of Global Convergence for the Tridiagonal QL Algorithm," *SIAM J. Num. Anal. 15,* 929–37.

S. Batterson (1994). "Convergence of the Francis Shifted QR Algorithm on Normal Matrices," *Lin. Alg. and Its Applic. 207,* 181–195.

For an analysis of the method when it is applied to normal matrices see

C.P. Huang (1981). "On the Convergence of the QR Algorithm with Origin Shifts for Normal Matrices," *IMA J. Num. Anal. 1,* 127–33.

Interesting papers concerned with shifting in the tridiagonal QR algorithm include

F.L. Bauer and C. Reinsch (1968). "Rational QR Transformations with Newton Shift for Symmetric Tridiagonal Matrices," *Numer. Math. 11,* 264–72. See also Wilkinson and Reinsch (1971, pp.257–65).

G.W. Stewart (1970). "Incorporating Origin Shifts into the QR Algorithm for Symmetric Tridiagonal Matrices," *Comm. Assoc. Comp. Mach. 13,* 365–67.

Some parallel computation possibilities for the algorithms in this section are discussed in

S. Lo, B. Philippe, and A. Sameh (1987). "A Multiprocessor Algorithm for the Symmetric Tridiagonal Eigenvalue Problem," *SIAM J. Sci. and Stat. Comp. 8,* s155-s165.

H.Y. Chang and M. Salama (1988). "A Parallel Householder Tridiagonalization Strategy Using Scattered Square Decomposition," *Parallel Computing 6,* 297-312.

Another way to compute a specified subset of eigenvalues is via the rational QR algorithm. In this method, the shift is determined using Newton's method. This makes it possible to "steer" the iteration towards desired eigenvalues. See

C. Reinsch and F.L. Bauer (1968). "Rational QR Transformation with Newton's Shift for Symmetric Tridiagonal Matrices," *Numer. Math. 11,* 264–72. See also Wilkinson and Reinsch (1971, pp.257–65).

Papers concerned with the symmetric QR algorithm for banded matrices include

R.S. Martin and J.H. Wilkinson (1967). "Solution of Symmetric and Unsymmetric Band Equations and the Calculation of Eigenvectors of Band Matrices," *Numer. Math. 9,* 279–301. See also See also Wilkinson and Reinsch (1971, pp.70–92).

R.S. Martin, C. Reinsch, and J.H. Wilkinson (1970). "The QR Algorithm for Band Symmetric Matrices," *Numer. Math. 16,* 85–92. See also Wilkinson and Reinsch (1971, pp.266–72).

8.4 Jacobi Methods

Jacobi methods for the symmetric eigenvalue problem attract current attention because they are inherently parallel. They work by performing a sequence of orthogonal similarity updates $A \leftarrow Q^TAQ$ with the property that each new A, although full, is "more diagonal" than its predecessor. Eventually, the off-diagonal entries are small enough to be declared zero.

After surveying the basic ideas behind the Jacobi approach we develop a parallel Jacobi procedure.

8.4.1 The Jacobi Idea

The idea behind Jacobi's method is to systematically reduce the quantity

$$\text{off}(A) \;=\; \sqrt{\sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}^2}\;,$$

i.e., the"norm" of the off-diagonal elements. The tools for doing this are rotations of the form

$$J(p,q,\theta) \;=\; \begin{array}{cc} \left[\begin{array}{ccccccc} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{array}\right] & \begin{array}{c} \\ \\ p \\ \\ q \\ \\ \\ \end{array} \\ \begin{array}{ccccccc} & & p & & q & & \end{array} & \end{array}$$

which we call *Jacobi rotations*. Jacobi rotations are no different from Givens rotations, c.f. §5.1.8. We submit to the name change in this section to honor the inventor.

The basic step in a Jacobi eigenvalue procedure involves (1) choosing an index pair (p,q) that satisfies $1 \le p < q \le n$, (2) computing a cosine-sine pair (c,s) such that

$$\left[\begin{array}{cc} b_{pp} & b_{pq} \\ b_{qp} & b_{qq} \end{array}\right] = \left[\begin{array}{cc} c & s \\ -s & c \end{array}\right]^T \left[\begin{array}{cc} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{array}\right] \left[\begin{array}{cc} c & s \\ -s & c \end{array}\right] \qquad (8.4.1)$$

is diagonal, and (3) overwriting A with $B = J^TAJ$ where $J = J(p,q,\theta)$. Observe that the matrix B agrees with A except in rows and columns p

and q. Moreover, since the Frobenius norm is preserved by orthogonal transformations we find that

$$a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2 \;=\; b_{pp}^2 + b_{qq}^2 + 2b_{pq}^2 \;=\; b_{pp}^2 + b_{qq}^2$$

and so

$$\begin{aligned} \text{off}(B)^2 &= \| B \|_F^2 - \sum_{i=1}^{n} b_{ii}^2 \qquad (8.4.2)\\ &= \| A \|_F^2 - \sum_{i=1}^{n} a_{ii}^2 \;+\; (a_{pp}^2 + a_{qq}^2 - b_{pp}^2 - b_{qq}^2)\\ &= \text{off}(A)^2 - 2a_{pq}^2 \,. \end{aligned}$$

It is in this sense that A moves closer to diagonal form with each Jacobi step.

Before we discuss how the index pair (p,q) can be chosen, let us look at the actual computations associated with the (p,q) subproblem.

8.4.2 The 2-by-2 Symmetric Schur Decomposition

To say that we diagonalize in (8.4.1) is to say that

$$0 \;=\; b_{pq} \;=\; a_{pq}(c^2 - s^2) + (a_{pp} - a_{qq})cs. \qquad (8.4.3)$$

If $a_{pq} = 0$, then we just set $(c,s) = (1,0)$. Otherwise define

$$\tau \;=\; \frac{a_{qq} - a_{pp}}{2a_{pq}} \quad \text{and} \quad t \;=\; s/c$$

and conclude from (8.4.3) that $t = \tan(\theta)$ solves the quadratic

$$t^2 + 2\tau t - 1 = 0\,.$$

It turns out to be important to select the smaller of the two roots,

$$t \;=\; -\tau \pm \sqrt{1+\tau^2}$$

whereupon c and s can be resolved from the formulae

$$c = 1/\sqrt{1+t^2} \qquad s = tc\,.$$

Choosing t to be the smaller of the two roots ensures that $|\theta| \le \pi/4$ and has the effect of minimizing the difference between B and A because

$$\| B - A \|_F^2 \;=\; 4(1-c)\sum_{\substack{i=1\\ i\neq p,q}}^{n} (a_{ip}^2 + a_{iq}^2) \;+\; 2a_{pq}^2/c^2$$

We summarize the 2-by-2 computations as follows:

Algorithm 8.4.1 Given an n-by-n symmetric A and integers p and q that satisfy $1 \le p < q \le n$, this algorithm computes a cosine-sine pair $(c\,,\ s)$ such that if $B = J(p,q,\theta)^T A J(p,q,\theta)$ then $b_{pq} = b_{qp} = 0$.

```
function: $[c\,,\ s] = \mathbf{sym.schur2}(A,p,q)$
    if $A(p,q) \neq 0$
        $\tau = (A(q,q) - A(p,p))/(2A(p,q))$
        if $\tau \ge 0$
            $t = 1/(\tau + \sqrt{1+\tau^2})$;
        else
            $t = -1/(-\tau + \sqrt{1+\tau^2})$;
        end
        $c = 1/\sqrt{1+t^2}$
        $s = tc$
    else
        $c = 1$
        $s = 0$
    end
```

8.4.3 The Classical Jacobi Algorithm

As we mentioned above, only rows and columns p and q are altered when the (p,q) subproblem is solved. Once **sym.schur2** determines the 2-by-2 rotation, then the update $A \leftarrow J(p,q,\theta)^T A J(p,q,\theta)$ can be implemented in $6n$ flops if symmetry is exploited.

How do we choose the indices p and q? From the standpoint of maximizing the reduction of off(A) in (8.4.2), it makes sense to choose (p,q) so that a_{pq}^2 is maximal. This is the basis of the *classical* Jacobi algorithm.

Algorithm 8.4.2 (Classical Jacobi) Given a symmetric $A \in \mathbb{R}^{n\times n}$ and a tolerance $tol > 0$, this algorithm overwrites A with V^TAV where V is orthogonal and off$(V^TAV) \le tol\| A \|_F$.

```
$V = I_n$; $eps = tol\| A \|_F$
while off$(A) > eps$
    Choose $(p,q)$ so $|a_{pq}| = \max_{i\neq j} |a_{ij}|$.
    $(c\,,\ s) = \mathbf{sym.schur2}(A,p,q)$
    $A = J(p,q,\theta)^T A J(p,q,\theta)$
    $V = VJ(p,q,\theta)$
end
```

Since $|a_{pq}|$ is the largest off-diagonal entry, off$(A)^2 \le N(a_{pq}^2 + a_{qp}^2)$ where

$N = n(n-1)/2$. From (8.4.2) it follows that

$$\text{off}(B)^2 \le \left(1 - \frac{1}{N}\right) \text{off}(A)^2 .$$

By induction, if $A^{(k)}$ denotes the matrix A after k Jacobi updates, then

$$\text{off}(A^{(k)})^2 \le \left(1 - \frac{1}{N}\right)^k \text{off}(A^{(0)})^2 .$$

This implies that the classical Jacobi procedure converges at a linear rate.

However, the asymptotic convergence rate of the method is considerably better than linear. Schonhage (1964) and van Kempen (1966) show that for k large enough, there is a constant c such that

$$\text{off}(A^{(k+N)}) \le c \cdot \text{off}(A^{(k)})^2$$

i.e., quadratic convergence. An earlier paper by Henrici (1958) established the same result for the special case when A has distinct eigenvalues. In the convergence theory for the Jacobi iteration, it is critical that $|\theta| \le \pi/4$. Among other things this precludes the possibility of "interchanging" nearly converged diagonal entries. This follows from the formulae $b_{pp} = a_{pp} - ta_{pq}$ and $b_{qq} = a_{qq} + ta_{pq}$, which can be derived from equations (8.4.1) and the definition $t = \sin(\theta)/\cos(\theta)$.

It is customary to refer to N Jacobi updates as a *sweep*. Thus, after a sufficient number of iterations, quadratic convergence is observed when examining off(A) after every sweep.

Example 8.4.1 Applying the classical Jacobi iteration to

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 3 & 6 & 10 \\ 1 & 4 & 10 & 20 \end{bmatrix}$$

we find

sweep	$O(\text{off}(A))$
0	10^{2}
1	10^{1}
2	10^{-2}
3	10^{-11}
4	10^{-17}

There is no rigorous theory that enables one to predict the number of sweeps that are required to achieve a specified reduction in off(A). However, Brent and Luk (1985) have argued heuristically that the number of sweeps is proportional to $\log(n)$ and this seems to be the case in practice.

8.4.4 The Cyclic-by-Row Algorithm

The trouble with the classical Jacobi method is that the updates involve $O(n)$ flops while the search for the optimal (p, q) is $O(n^2)$. One way to address this imbalance is to fix the sequence of subproblems to be solved in advance. A reasonable possibility is to step through all the subproblems in row-by-row fashion. For example, if $n = 4$ we cycle as follows:

$$(p, q) = (1,2), (1,3), (1,4), (2,3), (2,4), (3,4), (1,2), \ldots$$

This ordering scheme is referred to as *cyclic-by-row* and it results in the following procedure:

Algorithm 8.4.3 (Cyclic Jacobi) Given a symmetric $A \in \mathbb{R}^{n \times n}$ and a tolerance $tol > 0$, this algorithm overwrites A with $V^T AV$ where V is orthogonal and $\text{off}(V^T AV) \le tol\| A \|_F$.

```
$V = I_n$
$eps = tol\| A \|_F$
while $\text{off}(A) > eps$
    for $p = 1{:}n-1$
        for $q = p+1{:}n$
            $(c\,,\ s) = \textbf{sym.schur2}(A, p, q)$
            $A = J(p, q, \theta)^T A J(p, q, \theta)$
            $V = VJ(p, q, \theta)$
        end
    end
end
```

Cyclic Jacobi converges also quadratically. (See Wilkinson (1962) and van Kempen (1966).) However, since it does not require off-diagonal search, it is considerably faster than Jacobi's original algorithm.

Example 8.4.2 If the cyclic Jacobi method is applied to the matrix in Example 8.4.1 we find

Sweep	$O(\text{off}(A))$
0	10^2
1	10^1
2	10^{-1}
3	10^{-6}
4	10^{-16}

8.4.5 Error Analysis

Using Wilkinson's error analysis it is possible to show that if r sweeps are needed in Algorithm 8.4.3 then the computed d_i satisfy

$$\sum_{i=1}^{n}(d_i - \lambda_i)^2 \;\le\; (\delta + k_r)\| A \|_F \mathbf{u}$$

for some ordering of A's eigenvalues λ_i. The parameter k_r depends mildly on r.

Although the cyclic Jacobi method converges quadratically, it is not generally competitive with the symmetric QR algorithm. For example, if we just count flops, then 2 sweeps of Jacobi is roughly equivalent to a complete QR reduction to diagonal form with accumulation of transformations. However, for small n this liability is not very dramatic. Moreover, if an approximate eigenvector matrix V is known, then V^TAV is almost diagonal, a situation that Jacobi can exploit but not QR.

Another interesting feature of the Jacobi method is that it can a compute the eigenvalues with small *relative* error if A is positive definite. To appreciate this point, note that the Wilkinson analysis cited above coupled the §8.1 perturbation theory ensures that the computed eigenvalues $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_n$ satisfy

$$\frac{|\hat{\lambda}_i - \lambda_i(A)|}{\lambda_i(A)} \;\approx\; \mathbf{u}\frac{\| A \|_2}{\lambda_i(A)} \;\le\; \mathbf{u}\kappa_2(A).$$

However, a refined, componentwise error analysis by Demmel and Veseliĉ (1992) shows that in the positive definite case,

$$\frac{|\hat{\lambda}_i - \lambda_i(A)|}{\lambda_i(A)} \;\approx\; \mathbf{u}\kappa_2(D^{-1}AD^{-1}). \tag{8.4.4}$$

where $D = \text{diag}(\sqrt{a_{11}},\ldots,\sqrt{a_{nn}})$ and this is generally a much smaller approximating bound. The key to establishing this result is some new perturbation theory and a demonstration that if A_+ is a computed Jacobi update obtained from the current matrix A_c, then the eigenvalues of A_+ are relatively close to the eigenvalues of A_c in the sense of (8.4.4). To make the whole thing work in practice, the termination criteria is not based upon the comparison of off(A) with $\mathbf{u}\| A \|_F$ but rather on the size of each $|a_{ij}|$ compared to $\mathbf{u}\sqrt{a_{ii}a_{jj}}$. This work is typical of a new genre of research concerned with high-accuracy algorithms based upon careful, componentwise error analysis. See Mathias (1995).

8.4.6 Parallel Jacobi

Perhaps the most interesting distinction between the QR and Jacobi approaches to the symmetric eigenvalue problem is the rich inherent parallelism of the latter algorithm. To illustrate this, suppose $n = 4$ and group the six subproblems into three *rotation sets* as follows:

$$\begin{aligned} rot.set(1) &= \{(1,2),(3,4)\} \\ rot.set(2) &= \{(1,3),(2,4)\} \\ rot.set(3) &= \{(1,4),(2,3)\} \end{aligned}$$

Note that all the rotations within each of the three rotation sets are "nonconflicting." That is, subproblems (1,2) and (3,4) can be carried out in parallel. Likewise the (1,3) and (2,4) subproblems can be executed in parallel as can subproblems (1,4) and (2,3). In general, we say that

$$(i_1,j_1),(i_2,j_2),\ldots,(i_N,j_N) \qquad N = (n-1)n/2$$

is a *parallel ordering* of the set $\{(i,j) \mid 1 \le i < j \le n\}$ if for $s = 1{:}n-1$ the rotation set $rot.set(s) = \{\ (i_r,j_r) : r = 1 + n(s-1)/2{:}ns/2\ \}$ consists of nonconflicting rotations. This requires n to be even, which we assume throughout this section. (The odd n case can be handled by bordering A with a row and column of zeros and being careful when solving the subproblems that involve these augmented zeros.)

A good way to generate a parallel ordering is to visualize a chess tournament with n players in which everybody must play everybody else exactly once. In the $n = 8$ case this entails 7 "rounds." During round one we have the following four games:

1	3	5	7
2	4	6	8

$$rot.set(1) = \{\ (1,2),(3,4),(5,6),(7,8)\ \}$$

i.e., 1 plays 2, 3 plays 4, etc. To set up rounds 2 through 7, player 1 stays put and players 2 through 8 embark on a merry-go-round:

1	2	3	5
4	6	8	7

$$rot.set(2) = \{(1,4),(2,6),(3,8),(5,7)\}$$

1	4	2	3
6	8	7	5

$$rot.set(3) = \{(1,6),(4,8),(2,7),(3,5)\}$$

1	6	4	2
8	7	5	3

$$rot.set(4) = \{(1,8),(6,7),(4,5),(2,3)\}$$

1	8	6	4
7	5	3	2

$$rot.set(5) = \{(1,7),(5,8),(3,6),(2,4)\}$$

1	7	8	6
5	3	2	4

$$rot.set(6) = \{(1,5),(3,7),(2,8),(4,6)\}$$

1	5	7	8
3	2	4	6

$$rot.set(7) = \{(1,3),(2,5),(4,7),(6,8)\}$$

We can encode these operations in a pair of integer vectors $top(1{:}n/2)$ and $bot(1{:}n/2)$. During a given round $top(k)$ plays $bot(k)$, $k = 1{:}n/2$. The pairings for the next round is obtained by updating *top* and *bot* as follows:

```
function: [$new.top, new.bot$] = music($top, bot, n$)
    $m = n/2$
    for $k = 1{:}m$
        if $k = 1$
            $new.top(1) = 1$
        else if $k = 2$
            $new.top(k) = bot(1)$
        elseif $k > 2$
            $new.top(k) = top(k-1)$
        end
        if $k = m$
            $new.bot(k) = top(k)$
        else
            $new.bot(k) = bot(k+1)$
        end
    end
```

Using **music** we obtain the following parallel order Jacobi procedure.

Algorithm 8.4.4 (Parallel Order Jacobi) Given a symmetric $A \in \mathbb{R}^{n\times n}$ and a tolerance $tol > 0$, this algorithm overwrites A with V^TAV where V is orthogonal and $\text{off}(V^TAV) \leq tol\| A \|_F$. It is assumed that n is even.

```
$V = I_n$
$eps = tol\| A \|_F$
$top = 1{:}2{:}n; bot = 2{:}2{:}n$
while $\text{off}(A) > eps$
    for $set = 1{:}n-1$
        for $k = 1{:}n/2$
            $p = \min(top(k), bot(k))$
            $q = \max(top(k), bot(k))$
            $(c\,,\, s)$ = sym.schur2($A, p, q$)
            $A = J(p,q,\theta)^T A J(p,q,\theta)$
            $V = VJ(p,q,\theta)$
        end
        [$top, bot$] = music($top, bot, n$)
    end
end
```

Notice that the k-loop steps through $n/2$ independent, nonconflicting subproblems.

8.4.7 A Ring Procedure

We now discuss how Algorithm 8.4.4 could be implemented on a ring of p processors. We assume that $p = n/2$ for clarity. At any instant, Proc(μ) houses two columns of A and the corresponding V columns. For example, if $n = 8$ then here is how the column distribution of A proceeds from step to step:

	Proc(1)	Proc(2)	Proc(3)	Proc(4)
Step 1:	[1 2]	[3 4]	[5 6]	[7 8]
Step 2:	[1 4]	[2 6]	[3 8]	[5 7]
Step 3:	[1 6]	[4 8]	[2 7]	[3 5]
		etc.		

The ordered pairs denote the indices of the housed columns. The first index names the *left* column and the second index names the *right* column. Thus, the *left* and *right* columns in Proc(3) during step 3 are 2 and 7 respectively.

Note that in between steps, the columns are shuffled according to the permutation implicit in **music** and that nearest neighbor communication prevails. At each step, each processor oversees a single subproblem. This involves (a) computing an orthogonal $V_{small} \in \mathbb{R}^{2\times 2}$ that solves a local 2-by-2 Schur problem, (b) using the 2-by-2 V_{small} to update the two housed columns of A and V, (c) sending the 2-by-2 V_{small} to all the other processors, and (d) receiving the V_{small} matrices from the other processors and updating the local portions of A and V accordingly. Since A is stored by column, communication is necessary to carry out the V_{small} updates because they effect rows of A. For example, in the second step of the $n = 8$ problem, Proc(2) must receive the 2-by-2 rotations associated with subproblems (1,4), (3,8), and (5,7). These come from Proc(1), Proc(3), and Proc(4) respectively. In general, the sharing of the rotation matrices can be conveniently implemented by circulating the 2-by-2 V_{small} matrices in "merry go round" fashion around the ring. Each processor copies a passing 2-by-2 V_{small} into its local memory and then appropriately updates the locally housed portions of A and V.

The termination criteria in Algorithm 8.4.4 poses something of a problem in a distributed memory environment in that the value of off($\cdot$) and $\| A \|_F$ require access to all of A. However, these global quantities can be computed during the V matrix merry-go-round phase. Before the circulation of the V's begins, each processor can compute its contribution to $\| A \|_F$ and off($\cdot$) . These quantities can then be summed by each processor if they are placed on the merry-go-round and read at each stop. By the end of one revolution each processor has its own copy of $\| A \|_F$ and off($\cdot$).

8.4.8 Block Jacobi Procedures

It is usually the case when solving the symmetric eigenvalue problem on a p-processor machine that $n \gg p$. In this case a block version of the Jacobi algorithm may be appropriate. Block versions of the above procedures are straightforward. Suppose that $n = rN$ and that we partition the n-by-n matrix A as follows:

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & & \vdots \\ A_{N1} & \cdots & A_{NN} \end{bmatrix}.$$

Here, each A_{ij} is r-by-r. In block Jacobi the (p,q) subproblem involves computing the $2r$-by-$2r$ Schur decomposition

$$\begin{bmatrix} V_{pp} & V_{pq} \\ V_{qp} & V_{qq} \end{bmatrix}^T \begin{bmatrix} A_{pp} & A_{pq} \\ A_{qp} & A_{qq} \end{bmatrix} \begin{bmatrix} V_{pp} & V_{pq} \\ V_{qp} & V_{qq} \end{bmatrix} = \begin{bmatrix} D_{pp} & O \\ O & D_{qq} \end{bmatrix}$$

and then applying to A the block Jacobi rotation made up of the V_{ij} . If we call this block rotation V then it is easy to show that

$$\text{off}(V^TAV)^2 = \text{off}(A)^2 - \left(2\| A_{pq} \|_F^2 + \text{off}(A_{pp})^2 + \text{off}(A_{qq})^2\right).$$

Block Jacobi procedures have many interesting computational aspects. For example, there are many ways to solve the subproblems and the choice appears to be critical. See Bischof (1987).

Problems

P8.4.1 Let the scalar γ be given along with the matrix

$$A = \begin{bmatrix} w & x \\ x & z \end{bmatrix}.$$

It is desired to compute an orthogonal matrix

$$J = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}$$

such that the (1, 1) entry of J^TAJ equals γ. Show that this requirement leads to the equation

$$(w-\gamma)\tau^2 - 2x\tau + (z-\gamma) = 0,$$

where $\tau = c/s$. Verify that this quadratic has real roots if γ satisfies $\lambda_2 \le \gamma \le \lambda_1$, where λ_1 and λ_2 are the eigenvalues of A.

P8.4.2 Let $A \in \mathbb{R}^{n\times n}$ be symmetric. Give an algorithm that computes the factorization

$$Q^TAQ = \gamma I + F$$

where Q is a product of Jacobi rotations, $\gamma = \text{trace}(A)/n$, and F has zero diagonal entries. Discuss the uniqueness of Q.

P8.4.3 Formulate Jacobi procedures for (a) skew symmetric matrices and (b) complex

Hermitian matrices.

P8.4.4 Partition the n-by-n real symmetric matrix A as follows:

$$A = \begin{array}{cc} \left[\begin{array}{cc} a & v^T \\ v & A_1 \end{array}\right] & \begin{array}{c} 1 \\ n-1 \end{array} \\ \begin{array}{cc} 1 & n-1 \end{array} & \end{array}$$

Let Q be a Householder matrix such that if $B = Q^TAQ$, then $B(3{:}n, 1) = 0$. Let $J = J(1, 2, \theta)$ be determined such that if $C = J^TBJ$, then $c_{12} = 0$ and $c_{11} \geq c_{22}$. Show $c_{11} \geq a + \| v \|_2$. La Budde (1964) formulated an algorithm for the symmetric eigenvalue probem based upon repetition of this Householder-Jacobi computation.

P8.4.5 Organize function **music** so that it involves minimum workspace.

P8.4.6 When implementing cyclic Jacobi, it is sensible to skip the annihilation of a_{pq} if its modulus is less than some small, sweep-dependent parameter, because the net reduction in off(A) is not worth the cost. This leads to what is called the *threshold Jacobi method.* Details concerning this variant of Jacobi's algorithm may be found in Wilkinson (1965, p.277). Show that appropriate thresholding can guarantee convergence.

Notes and References for Sec. 8.4

Jacobi's original paper is one of the earliest references found in the numerical analysis literature

C.G.J. Jacobi (1846). "Uber ein Leichtes Verfahren Die in der Theorie der Sacularstroungen Vorkommendern Gleichungen Numerisch Aufzulosen," *Crelle's J. 30,* 51–94.

Prior to the QR algorithm, the Jacobi technique was the standard method for solving dense symmetric eigenvalue problems. Early attempts to improve upon it include

M. Lotkin (1956). "Characteristic Values of Arbitrary Matrices," *Quart. Appl. Math. 14,* 267–75.

D.A. Pope and C. Tompkins (1957). "Maximizing Functions of Rotations: Experiments Concerning Speed of Diagonalization of Symmetric Matrices Using Jacobi's Method," *J. ACM 4,* 459–66.

C.D. La Budde (1964). "Two Classes of Algorithms for Finding the Eigenvalues and Eigenvectors of Real Symmetric Matrices," *J. ACM 11,* 53–58.

The computational aspects of Jacobi method are described in Wilkinson (1965,p.265). See also

H. Rutishauser (1966). "The Jacobi Method for Real Symmetric Matrices," *Numer. Math. 9,* 1–10. See also Wilkinson and Reinsch (1971, pp. 202–11).

N. Mackey (1995). "Hamilton and Jacobi Meet Again: Quaternions and the Eigenvalue Problem," *SIAM J. Matrix Anal. Applic. 16,* 421–435.

The method is also useful when a nearly diagonal matrix must be diagonalized. See

J.H. Wilkinson (1968). "Almost Diagonal Matrices with Multiple or Close Eigenvalues," *Lin. Alg. and Its Applic. 1,* 1–12.

Establishing the quadratic convergence of the classical and cyclic Jacobi iterations has attracted much attention:

P. Henrici (1958). "On the Speed of Convergence of Cyclic and Quasicyclic Jacobi Methods for Computing the Eigenvalues of Hermitian Matrices," *SIAM J. Appl. Math. 6,* 144–62.

E.R. Hansen (1962). "On Quasicyclic Jacobi Methods," *ACM J. 9,* 118–35.

J.H. Wilkinson (1962). "Note on the Quadratic Convergence of the Cyclic Jacobi Process," *Numer. Math. 6*, 296–300.

E.R. Hansen (1963). "On Cyclic Jacobi Methods," *SIAM J. Appl. Math. 11*, 448–59.

A. Schonhage (1964). "On the Quadratic Convergence of the Jacobi Process," *Numer. Math. 6*, 410–12.

H.P.M. van Kempen (1966). "On Quadratic Convergence of the Special Cyclic Jacobi Method," *Numer. Math. 9*, 19–22.

P. Henrici and K. Zimmermann (1968). "An Estimate for the Norms of Certain Cyclic Jacobi Operators," *Lin. Alg. and Its Applic. 1*, 489–501.

K.W. Brodlie and M.J.D. Powell (1975). "On the Convergence of Cyclic Jacobi Methods," *J. Inst. Math. Applic. 15*, 279–87.

Detailed error analyses that establish imprtant componentwise error bounds include

J. Barlow and J. Demmel (1990). "Computing Accurate Eigensystems of Scaled Diagonally Dominant Matrices," *SIAM J. Numer. Anal. 27*, 762–791.

J.W. Demmel and K. Veselič (1992). "Jacobi's Method is More Accurate than QR," *SIAM J. Matrix Anal. Appl. 13*, 1204–1245.

Z. Drmač (1994). *The Generalized Singular Value Problem*, Ph.D. Thesis, FernUniversitat, Hagen, Germany.

W.F. Mascarenhas (1994). "A Note on Jacobi Being More Accurate than QR," *SIAM J. Matrix Anal. Appl. 15*, 215–218.

R. Mathias (1995). "Accurate Eigensystem Computations by Jacobi Methods," *SIAM J. Matrix Anal. Appl. 16*, 977–1003.

Attempts have been made to extend the Jacobi iteration to other classes of matrices and to push through corresponding convergence results. The case of normal matrices is discussed in

H.H. Goldstine and L.P. Horowitz (1959). "A Procedure for the Diagonalization of Normal Matrices," *J. Assoc. Comp. Mach. 6*, 176–95.

G. Loizou (1972). "On the Quadratic Convergence of the Jacobi Method for Normal Matrices," *Comp. J. 15*, 274–76.

A. Ruhe (1972). "On the Quadratic Convergence of the Jacobi Method for Normal Matrices," *BIT 7*, 305–13.

See also

M.H.C. Paardekooper (1971). "An Eigenvalue Algorithm for Skew Symmetric Matrices," *Numer. Math. 17*, 189–202.

D. Hacon (1993). "Jacobi's Method for Skew-Symmetric Matrices," *SIAM J. Matrix Anal. Appl. 14*, 619–628.

Essentially, the analysis and algorithmic developments presented in the text carry over to the normal case with minor modification. For non-normal matrices, the situation is considerably more difficult. Consult

J. Greenstadt (1955). "A Method for Finding Roots of Arbitrary Matrices," *Math. Tables and Other Aids to Comp. 9*, 47–52.

C.E. Froberg (1965). "On Triangularization of Complex Matrices by Two Dimensional Unitary Tranformations," *BIT 5*, 230–34.

J. Boothroyd and P.J. Eberlein (1968). "Solution to the Eigenproblem by a Norm-Reducing Jacobi-Type Method (Handbook)," *Numer. Math. 11*, 1–12. See also Wilkinson and Reinsch (1971, pp.327–38).

A. Ruhe (1968). On the Quadratic Convergence of a Generalization of the Jacobi Method to Arbitrary Matrices," *BIT 8*, 210–31.

A. Ruhe (1969). "The Norm of a Matrix After a Similarity Transformation," *BIT 9*, 53–58.

P.J. Eberlein (1970). "Solution to the Complex Eigenproblem by a Norm-Reducing Jacobi-type Method," *Numer. Math. 14,* 232–45. See also Wilkinson and Reinsch (1971, pp.404–17).

C.P. Huang (1975). "A Jacobi-Type Method for Triangularizing an Arbitrary Matrix," *SIAM J. Num. Anal. 12,* 566–70.

V. Hari (1982). "On the Global Convergence of the Eberlein Method for Real Matrices," *Numer. Math. 39,* 361–370.

G.W. Stewart (1985). "A Jacobi-Like Algorithm for Computing the Schur Decomposition of a Nonhermitian Matrix," *SIAM J. Sci. and Stat. Comp. 6,* 853–862.

W-W. Lin and C.W. Chen (1991). "An Acceleration Method for Computing the Generalized Eigenvalue Problem on a Parallel Computer," *Lin.Alg. and Its Applic. 146,* 49–65.

Jacobi methods for complex symmetric matrices have also been developed. See

J.J. Seaton (1969). "Diagonalization of Complex Symmetric Matrices Using a Modified Jacobi Method," *Comp. J. 12,* 156–57.

P.J. Eberlein (1971). "On the Diagonalization of Complex Symmetric Matrices," *J. Inst. Math. Applic. 7,* 377–83.

P. Anderson and G. Loizou (1973). "On the Quadratic Convergence of an Algorithm Which Diagonalizes a Complex Symmetric Matrix," *J. Inst. Math. Applic. 12,* 261–71.

P. Anderson and G. Loizou (1976). "A Jacobi-Type Method for Complex Symmetric Matrices (Handbook)," *Numer. Math. 25,* 347–63.

Although the symmetric QR algorithm is generally much faster than the Jacobi method, there are special settings where the latter technique is of interest. As we illustrated, on a parallel-computer it is possible to perform several rotations concurrently, thereby accelerating the reduction of the off-diagonal elements. See

A. Sameh (1971). "On Jacobi and Jacobi-like Algorithms for a Parallel Computer," *Math. Comp. 25,* 579–90.

J.J. Modi and J.D. Pryce (1985). "Efficient Implementation of Jacobi's Diagonalization Method on the DAP," *Numer. Math. 46,* 443–454.

D.S. Scott, M.T. Heath, and R.C. Ward (1986). "Parallel Block Jacobi Eigenvalue Algorithms Using Systolic Arrays," *Lin. Alg. and Its Applic. 77,* 345–356.

P.J. Eberlein (1987). "On Using the Jacobi Method on a Hypercube," in *Hypercube Multiprocessors*, ed. M.T. Heath, SIAM Publications, Philadelphia.

G. Shroff and R. Schreiber (1989). "On the Convergence of the Cyclic Jacobi Method for Parallel Block Orderings," *SIAM J. Matrix Anal. Appl. 10*, 326–346.

M.H.C. Paardekooper (1991). "A Quadratically Convergent Parallel Jacobi Process for Diagonally Dominant Matrices with Nondistinct Eigenvalues," *Lin.Alg. and Its Applic. 145*, 71–88.

8.5 Tridiagonal Methods

In this section we develop special methods for the symmetric tridiagonal eigenproblem. The tridiagonal form

$$T = \begin{bmatrix} a_1 & b_1 & & \cdots & 0 \\ b_1 & a_2 & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & b_{n-1} \\ 0 & \cdots & & b_{n-1} & a_n \end{bmatrix} \qquad (8.5.1)$$

can be obtained by Householder reduction (cf. §8.3.1). However, symmetric tridiagonal eigenproblems arise naturally in many settings.

We first discuss bisection methods that are of interest when selected portions of the eigensystem are required. This is followed by the presentation of a divide and conquer algorithm that can be used to acquire the full symmetric Schur decomposition in a way that is amenable to parallel processing.

8.5.1 Eigenvalues by Bisection

Let T_r denote the leading r-by-r principal submatrix of the matrix T in (8.5.1). Define the polynomials $p_r(x) = \det(T_r - xI)$, $r = 1{:}n$. A simple determinantal expansion shows that

$$p_r(x) \;=\; (a_r - x)p_{r-1}(x) - b_{r-1}^2 p_{r-2}(x) \qquad (8.5.2)$$

for $r = 2{:}n$ if we set $p_0(x) = 1$. Because $p_n(x)$ can be evaluated in $O(n)$ flops, it is feasible to find its roots using the method of bisection. For example, if $p_n(y)p_n(z) < 0$ and $y < z$, then the iteration

```
while $|y - z| > \epsilon(|y| + |z|)$
    $x = (y + z)/2$
    if $p_n(x)p_n(y) < 0$
        $z = x$
    else
        $y = x$
    end
end
```

is guaranteed to terminate with $(y + z)/2$ an approximate zero of $p_n(x)$, i.e., an approximate eigenvalue of T. The iteration converges linearly in that the error is approximately halved at each step.

8.5.2 Sturm Sequence Methods

Sometimes it is necessary to compute the kth largest eigenvalue of T for some prescribed value of k. This can be done efficiently by using the bisection idea and the following classical result:

Theorem 8.5.1 (Sturm Sequence Property) *If the tridiagonal matrix in (8.5.1) has no zero subdiagonal entries, then the eigenvalues of T_{r-1} strictly separate the eigenvalues of T_r:*

$$\lambda_r(T_r) < \lambda_{r-1}(T_{r-1}) < \lambda_{r-1}(T_r) < \cdots < \lambda_2(T_r) < \lambda_1(T_{r-1}) < \lambda_1(T_r).$$

Moreover, if $a(\lambda)$ denotes the number of sign changes in the sequence

$$\{\, p_0(\lambda),\, p_1(\lambda), \ldots,\, p_n(\lambda)\, \}$$

then $a(\lambda)$ equals the number of T's eigenvalues that are less than λ. Here, the polynomials $p_r(x)$ are defined by (8.5.2) and we have the convention that $p_r(\lambda)$ has the opposite sign of $p_{r-1}(\lambda)$ if $p_r(\lambda) = 0$.

Proof. It follows from Theorem 8.1.7 that the eigenvalues of T_{r-1} weakly separate those of T_r. To prove that the separation must be strict, suppose that $p_r(\mu) = p_{r-1}(\mu) = 0$ for some r and μ. It then follows from (8.5.2) and the assumption that T is unreduced that $p_0(\mu) = p_1(\mu) = \cdots = p_r(\mu) = 0$, a contradiction. Thus, we must have strict separation.

The assertion about $a(\lambda)$ is established in Wilkinson (1965, 300-301). We mention that if $p_r(\lambda) = 0$, then its sign is assumed to be opposite the sign of $p_{r-1}(\lambda)$. □

Example 8.5.1 If

$$T = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 3 & -1 \\ 0 & 0 & -1 & 4 \end{bmatrix}$$

then $\lambda(T) \approx \{.254,\ 1.82,\ 3.18,\ 4.74\}$. The sequence

$$\{\, p_0(2),\, p_1(2),\, p_2(2),\, p_3(2),\, p_4(2)\, \} \;=\; \{\, 1,\ -1,\ -1,\ 0,\ 1\, \}$$

confirms that there are two eigenvalues less than $\lambda = 2$.

Suppose we wish to compute $\lambda_k(T)$. From the Gershgorin theorem (Theorem 8.1.3) it follows that $\lambda_k(T) \in [y, z]$ where

$$y = \min_{1 \le i \le n} a_i - |b_i| - |b_{i-1}| \qquad z = \max_{1 \le i \le n} a_i + |b_i| + |b_{i-1}|$$

if we define $b_0 = b_n = 0$. With these starting values, it is clear from the Sturm sequence property that the iteration

```
while $|z - y| > \mathbf{u}(|y| + |z|)$
        $x = (y + z)/2$
        if $a(x) \geq n - k$                                    (8.5.3)
                $z = x$
        else
                $y = x$
        end
end
```

produces a sequence of subintervals that are repeatedly halved in length but which always contain $\lambda_k(T)$.

Example 8.5.2 If (8.5.3) is applied to the matrix of Example 8.5.1 with $k = 3$, then the values shown in the following table are generated:

y	z	x	$a(x)$
0.0000	5.0000	2.5000	2
0.0000	2.5000	1.2500	1
1.2500	2.5000	1.3750	1
1.3750	2.5000	1.9375	2
1.3750	1.9375	1.6563	1
1.6563	1.9375	1.7969	1

We conclude from the output that $\lambda_3(T) \in [\,1.7969,\ 1.9375\,]$. Note: $\lambda_3(T) \approx 1.82$.

During the execution of (8.5.3), information about the location of other eigenvalues is obtained. By systematically keeping track of this information it is possible to devise an efficient scheme for computing "contiguous" subsets of $\lambda(T)$, e.g., $\lambda_k(T), \lambda_{k+1}(T), \ldots, \lambda_{k+j}(T)$. See Barth, Martin, and Wilkinson (1967).

If selected eigenvalues of a general symmetric matrix A are desired, then it is necessary first to compute the tridiagonalization $T = U_0^T T U_0$ before the above bisection schemes can be applied. This can be done using Algorithm 8.3.1 or by the Lanczos algorithm discussed in the next chapter. In either case, the corresponding eigenvectors can be readily found via inverse iteration since tridiagonal systems can be solved in $O(n)$ flops. See §4.3.6 and §8.2.2.

In those applications where the original matrix A already has tridiagonal form, bisection computes eigenvalues with small relative error, regardless of their magnitude. This is in contrast to the tridiagonal QR iteration, where the computed eigenvalues $\tilde{\lambda}_i$ can be guaranteed only to have small absolute error: $|\tilde{\lambda}_i - \lambda_i(T)| \approx \mathbf{u}\| T \|_2$

Finally, it is possible to compute specific eigenvalues of a symmetric matrix by using the LDL^T factorization (see §4.2) and exploiting the Sylvester inertia theorem (Theorem 8.1.17). If

$$A - \mu I \;=\; LDL^T \qquad A = A^T \in \mathbb{R}^{n\times n}$$

is the LDL^T factorization of $A - \mu I$ with $D = \text{diag}(d_1, \ldots, d_n)$, then the number of negative d_i equals the number of $\lambda_i(A)$ that are less than μ. See Parlett (1980, p.46) for details.

8.5.3 Eigensystems of Diagonal Plus Rank-1 Matrices

Our next method for the symmetric tridiagonal eigenproblem requires that we be able to compute efficiently the eigenvalues and eigenvectors of a matrix of the form $D + \rho zz^T$ where $D \in \mathbb{R}^{n\times n}$is diagonal, $z \in \mathbb{R}^z$, and $\rho \in \mathbb{R}$. This problem is important in its own right and the key computations rest upon the following pair of results.

Lemma 8.5.2 *Suppose* $D = \text{diag}(d_1, \ldots, d_n) \in \mathbb{R}^{n\times n}$ *has the property that* $d_1 > \cdots > d_n$. *Assume that* $\rho \neq 0$ *and that* $z \in \mathbb{R}^n$ *has no zero components. If*

$$(D + \rho zz^T)v = \lambda v \qquad v \neq 0$$

then $z^T v \neq 0$ *and* $D - \lambda I$ *is nonsingular.*

Proof. If $\lambda \in \lambda(D)$, then $\lambda = d_i$ for some i and thus

$$0 = e_i^T[(D - \lambda I)v + \rho(z^T v)z] = \rho(z^T v)z_i.$$

Since ρ and z_i are nonzero we must have $0 = z^T v$ and so $Dv = \lambda v$. However, D has distinct eigenvalues and therefore, $v \in \text{span}\{e_i\}$. But then $0 = z^T v = z_i$, a contradiction. Thus, D and $D + \rho zz^T$ do not have any common eigenvalues and $z^T v \neq 0$. □

Theorem 8.5.3 *Suppose* $D = \text{diag}(d_1, \ldots, d_n) \in \mathbb{R}^{n\times n}$ *and that the diagonal entries satisfy* $d_1 > \cdots > d_n$. *Assume that* $\rho \neq 0$ *and that* $z \in \mathbb{R}^n$ *has no zero components. If* $V \in \mathbb{R}^{n\times n}$ *is orthogonal such that*

$$V^T(D + \rho zz^T)V = \text{diag}(\lambda_1, \ldots, \lambda_n)$$

with $\lambda_1 \geq \cdots \geq \lambda_n$ *and* $V = [\, v_1, \ldots, v_n \,]$, *then*

(a) *The* λ_i *are the* n *zeros of* $f(\lambda) = 1 + \rho z^T(D - \lambda I)^{-1}z$.

(b) *If* $\rho > 0$, *then* $\lambda_1 > d_1 > \lambda_2 > \cdots > d_n$.
If $\rho < 0$, *then* $d_1 > \lambda_1 > d_2 > \cdots > d_n > \lambda_n$.

(c) *The eigenvector* v_i *is a multiple of* $(D - \lambda_i I)^{-1}z$.

Proof. If $(D + \rho zz^T)v = \lambda v$, then

$$(D - \lambda I)v + \rho(z^T v)z = 0. \qquad (8.5.4)$$

We know from Lemma 8.5.2 that $D - \lambda I$ is nonsingular. Thus,

$$v \in \text{span}\{(D-\lambda I)^{-1}z\}$$

thereby establishing (c). Moreover, if we apply $z^T(D-\lambda I)^{-1}$ to both sides of equation (8.5.4) we obtain

$$z^T v\left(1 + \rho z^T(D-\lambda I)^{-1}z\right) = 0 .$$

By Lemma 8.5.2, $z^T v \neq 0$ and so this shows that if $\lambda \in \lambda(D + \rho zz^T)$, then $f(\lambda) = 0$. We must show that all the zeros of f are eigenvalues of $D + \rho zz^T$ and that the interlacing relations (b) hold.

To do this we look more carefully at the equations

$$\begin{aligned} f(\lambda) &= 1 + \rho\left(\frac{z_1^2}{d_1 - \lambda} + \cdots + \frac{z_n^2}{d_n - \lambda}\right) \\ f'(\lambda) &= \rho\left(\frac{z_1^2}{(d_1 - \lambda)^2} + \cdots + \frac{z_n^2}{(d_n - \lambda)^2}\right) \end{aligned}$$

Note that f is monotone in between its poles. This allows us to conclude that if $\rho > 0$, then f has precisely n roots, one in each of the intervals

$$(d_n, d_{n-1}), \ldots, (d_2, d_1), (d_1, \infty).$$

If $\rho < 0$ then f has exactly n roots, one in each of the intervals

$$(-\infty, d_n), (d_n, d_{n-1}), \ldots, (d_2, d_1).$$

In either case, it follows that the zeros of f are precisely the eigenvalues of $D + \rho vv^T$.□

The theorem suggests that to compute V we (a) find the roots $\lambda_1, \ldots, \lambda_n$ of f using a Newton-like procedure and then (b) compute the columns of V by normalizing the vectors $(D - \lambda_i I)^{-1}z$ for $i = 1{:}n$. The same plan of attack can be followed even if there are repeated d_i and zero z_i.

Theorem 8.5.4 *If $D = \text{diag}(d_1, \ldots, d_n)$ and $z \in \mathbb{R}^n$, then there exists an orthogonal matrix V_1 such that if $V_1^T D V_1 = \text{diag}(\mu_1, \ldots, \mu_n)$ and $w = V_1^T z$ then*

$$\mu_1 > \mu_2 > \cdots > \mu_r \geq \mu_{r+1} \geq \cdots \geq \mu_n ,$$

$w_i \neq 0$ for $i = 1{:}r$, and $w_i = 0$ for $i = r + 1{:}n$.

Proof. We give a constructive proof based upon two elementary operations. (a) Suppose $d_i = d_j$ for some $i < j$. Let $J(i, j, \theta)$ be a Jacobi rotation in the (i, j) plane with the property that the jth component of $J(i, j, \theta)^T z$ is zero. It is not hard to show that $J(i, j, \theta)^T D J(i, j, \theta) = D$. Thus, we can zero a component of z if there is a repeated d_i. (b) If $z_i = 0$,

$z_j \neq 0$, and $i < j$, then let P be the identity with columns i and j interchanged. It follows that P^TDP is diagonal, $(P^Tz)_i \neq 0$, and $(P^Tz)_j = 0$. Thus, we can permute all the zero z_i to the "bottom." Clearly, repetition of (a) and (b) eventually renders the desired canonical structure. V_1 is the product of the rotations. □

See Barlow (1993) and the references therein for a discussion of the solution procedures that we have outlined above.

8.5.4 A Divide and Conquer Method

We now present a divide-and-conquer method for computing the Schur decomposition

$$Q^TTQ \;=\; \Lambda \;=\; \text{diag}(\lambda_1,\ldots,\lambda_n) \qquad Q^TQ \;=\; I \tag{8.5.5}$$

for tridiagonal T that involves (a) "tearing" T in half, (b) computing the the Schur decompositions of the two parts, and (c) combining the two half-sized Schur decompositions into the required full size Schur decomposition. The overall procedure, developed by Dongarra and Sorensen (1987), is suitable for parallel computation.

We first show how T can be "torn" in half with a rank-one modification. For simplicity, assume $n = 2m$. Define $v \in \mathbb{R}^n$ as follows

$$v \;=\; \begin{bmatrix} e_m^{(m)} \\ \theta e_1^{(m)} \end{bmatrix}. \tag{8.5.6}$$

Note that for all $\rho \in \mathbb{R}$ the matrix $\tilde{T} = T - \rho vv^T$ is identical to T except in its "middle four" entries:

$$\tilde{T}(m{:}m+1, m{:}m+1) \;=\; \begin{bmatrix} a_m - \rho & b_m - \rho\theta \\ b_m - \rho\theta & a_{m+1} - \rho\theta^2 \end{bmatrix}.$$

If we set $\rho\theta = b_m$ then

$$T \;=\; \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix} + \rho vv^T$$

where

$$T_1 \;=\; \begin{bmatrix} a_1 & b_1 & & \cdots & 0 \\ b_1 & a_2 & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & b_{m-1} \\ 0 & \cdots & & b_{m-1} & \tilde{a}_m \end{bmatrix},$$

$$T_2 \;=\; \begin{bmatrix} \tilde{a}_{m+1} & b_{m+1} & & \cdots & 0 \\ b_{m+1} & a_{m+2} & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & b_{n-1} \\ 0 & \cdots & & b_{n-1} & a_n \end{bmatrix},$$

and $\tilde{a}_m = a_m - \rho$ and $\tilde{a}_{m+1} = a_{m+1} - \rho\theta^2$.

Now suppose that we have m-by-m orthogonal matrices Q_1 and Q_2 such that $Q_1^TT_1Q_1 = D_1$ and $Q_2^TT_2Q_2 = D_2$ are each diagonal. If we set

$$U \;=\; \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix},$$

then

$$U^TTU \;=\; U^T\left(\begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix} + \rho vv^T\right)U \;=\; D + \rho zz^T$$

where

$$D \;=\; \begin{bmatrix} D_1 & 0 \\ 0 & D_2 \end{bmatrix}$$

is diagonal and

$$z \;=\; U^Tv \;=\; \begin{bmatrix} Q_1^Te_m \\ \theta Q_2^Te_1 \end{bmatrix}.$$

Comparing these equations we see that the effective synthesis of the two half-sized Schur decompositions requires the quick and stable computation of an orthogonal V such that

$$V^T(D + \rho zz^T)V \;=\; \Lambda \;=\; \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$$

which we discussed in §8.5.3.

8.5.5 A Parallel Implementation

Having stepped through the tearing and synthesis operations, we are ready to illustrate the overall process and how it can be implemented on a multiprocessor. For clarity, assume that $n = 8N$ for some positive integer N and that three levels of tearing are performed. We can depict this with a binary tree as shown in Fig. 8.5.1. The indices are specified in binary. Fig. 8.5.2 depicts a single node and should be interpreted to mean that the eigensystem for the tridiagonal $T(b)$ is obtained from the eigensystems of the tridiagonals $T(b0)$ and $T(b1)$. For example, the eigensystems for the N-by-N matrices $T(110)$ and $T(111)$ are combined to produce the eigensystem for the $2N$-by-$2N$ tridiagonal matrix $T(11)$.

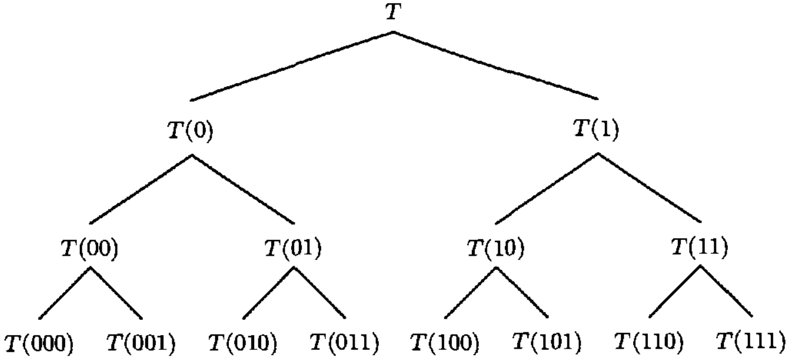

FIGURE 8.5.1 *Computation Tree*

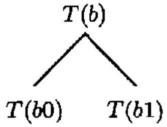

FIGURE 8.5.2 *Synthesis at a Node*

With tree-structured algorithms there is always the danger that parallelism is lost as the tree is "climbed" towards the root, but this is not the case in our problem. To see this suppose we have 8 processors and that the first task of Proc(b) is to compute the Schur decomposition of $T(b)$ where $b = 000, 001, 010, 011, 100, 101, 110, 111$. This portion of the computation is perfectly load balanced and does not involve interprocessor communication. (We are ignoring the Theorem 8.5.4 deflations, which are unlikely to cause significant load imbalance.)

At the next level there are four gluing operations to perform: $T(00)$, $T(01)$, $T(10)$, $T(11)$. However, each of these computations neatly subdivides and we can assign two processors to each task. For example, once the secular equation that underlies the $T(00)$ synthesis is known to both Proc(000) and Proc(001), then they each can go about getting half of the eigenvalues and corresponding eigenvectors. Likewise, 4 processors can each be assigned to the $T(0)$ and $T(1)$ problem. All 8 processors can participate in computing the eigensystem of T. Thus, at every level full parallelism can be maintained because the eigenvalue/eigenvector computations are independent of one another.

Problems

P8.5.1 Suppose λ is an eigenvalue of a symmetric tridiagonal matrix T. Show that if λ has algebraic multiplicity k, then at least $k-1$ of T's subdiagonal elements are zero.

P8.5.2 Give an algorithm for determining ρ and θ in (8.5.6) with the property that $\theta \in \{-1, 1\}$ and $\min\{\, |a_r - \rho|, |a_{r+1} - \rho| \,\}$ is maximized.

P8.5.3 Let $p_r(\lambda) = \det(T(1{:}r, 1{:}r) - \lambda I_r)$ where T is given by (8.5.1). Derive a recursion for evaluating $p_n'(\lambda)$ and use it to develop a Newton iteration that can compute eigenvalues of T.

P8.5.4 What communication is necessary between the processors assigned to a particular T_b? Is it possible to share the work associated with the processing of repeated d_i and zero z_i ?

P8.5.5 If T is positive definite, does it follow that the matrices T_1 and T_2 in §8.5.4 are positive definite?

P8.5.6 Suppose that

$$A = \begin{bmatrix} D & v \\ v^T & d_{nn} \end{bmatrix}$$

where $D = \text{diag}(d_1, \ldots, d_{n-1})$ has distinct diagonal entries and $v \in \mathbb{R}^{n-1}$ has no zero entries. (a) Show that if $\lambda \in \lambda(A)$, then $D - \lambda I_{n-1}$ is nonsingular. (b) Show that if $\lambda \in \lambda(A)$, then λ is a zero of

$$f(\lambda) = \lambda + \sum_{k=1}^{n-1} \frac{v_k^2}{d_k - \lambda} - d_n.$$

P8.5.7 Suppose $A = S + \sigma u u^T$ where $S \in \mathbb{R}^{n \times n}$ is skew-symmetric, $u \in \mathbb{R}^n$, and $\sigma \in \mathbb{R}$. Show how to compute an orthogonal Q such that $Q^T A Q = T + \sigma e_1 e_1^T$ where T is tridiagonal and skew-symmetric and e_1 is the first column of I_n.

P8.5.8 It is known that $\lambda \in \lambda(T)$ where $T \in \mathbb{R}^{n \times n}$ is symmetric and tridiagonal with no zero subdiagonal entries. Show how to compute $x(1{:}n-1)$ from the equation $Tx = \lambda x$ given that $x_n = 1$.

Notes and References for Sec. 8.5

Bisection/ Strum sequence methods are discussed in

W. Barth, R.S. Martin, and J.H. Wilkinson (1967). "Calculation of the Eigenvalues of a Symmetric Tridiagonal Matrix by the Method of Bisection," *Numer. Math. 9,* 386–93. See also Wilkinson and Reinsch (1971, 249–256).

K.K. Gupta (1972). "Solution of Eigenvalue Problems by Sturm Sequence Method," *Int. J. Numer. Meth. Eng. 4,* 379–404.

Various aspects of the divide and conquer algorithm discussed in this section is detailed in

G.H. Golub (1973). "Some Modified Matrix Eigenvalue Problems," *SIAM Review 15,* 318–44.

J.R. Bunch, C.P. Nielsen, and D.C. Sorensen (1978). "Rank-One Modification of the Symmetric Eigenproblem," *Numer. Math. 31,* 31–48.

J.J.M. Cuppen (1981). "A Divide and Conquer Method for the Symmetric Eigenproblem," *Numer. Math. 36,* 177–95.

J.J. Dongarra and D.C. Sorensen (1987). "A Fully Parallel Algorithm for the Symmetric Eigenvalue Problem," *SIAM J. Sci. and Stat. Comp. 8*, S139–S154.

S. Crivelli and E.R. Jessup (1995). "The Cost of Eigenvalue Computation on Distributed Memory MIMD Computers," *Parallel Computing 21*, 401–422.

The very delicate computations required by the method are carefully analyzed in

J.L. Barlow (1993). "Error Analysis of Update Methods for the Symmetric Eigenvalue Problem," *SIAM J. Matrix Anal. Appl. 14*, 598–618.

Various generalizations to banded symmetric eigenproblems have been explored.

P. Arbenz, W. Gander, and G.H. Golub (1988). "Restricted Rank Modification of the Symmetric Eigenvalue Problem: Theoretical Considerations," *Lin. Alg. and Its Applic. 104*, 75–95.

P. Arbenz and G.H. Golub (1988). "On the Spectral Decomposition of Hermitian Matrices Subject to Indefinite Low Rank Perturbations with Applications," *SIAM J. Matrix Anal. Appl. 9*, 40–58.

A related divide and conquer method based on the "arrowhead" matrix (see P8.5.7) is given in

M. Gu and S.C. Eisenstat (1995). "A Divide-and-Conquer Algorithm for the Symmetric Tridiagonal Eigenproblem," *SIAM J. Matrix Anal. Appl. 16*, 172–191.

8.6 Computing the SVD

There are important relationships between the singular value decomposition of a matrix A and the Schur decompositions of the symmetric matrices A^TA, AA^T, and $\begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$. Indeed, if

$$U^TAV \;=\; \operatorname{diag}(\sigma_1,\ldots,\sigma_n)$$

is the SVD of $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$, then

$$V^T(A^TA)V \;=\; \operatorname{diag}(\sigma_1^2,\ldots,\sigma_n^2) \in \mathbb{R}^{n\times n} \tag{8.6.1}$$

and

$$U^T(AA^T)U \;=\; \operatorname{diag}(\sigma_1^2,\ldots,\sigma_n^2,\underbrace{0,\ldots,0}_{m-n}) \in \mathbb{R}^{m\times m} \tag{8.6.2}$$

Moreover, if

$$U = \begin{array}{c} [\; U_1 \quad U_2 \;] \\ \;\;\; n \quad m-n \end{array}$$

and we define the orthogonal matrix $Q \in \mathbb{R}^{(m+n)\times(m+n)}$ by

$$Q = \frac{1}{\sqrt{2}}\begin{bmatrix} V & V & 0 \\ U_1 & -U_1 & \sqrt{2}\,U_2 \end{bmatrix}$$

then

$$Q^T \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix} Q = \text{diag}(\sigma_1, \ldots, \sigma_n, -\sigma_1, \ldots, -\sigma_n, \underbrace{0, \ldots, 0}_{m-n}). \quad (8.6.3)$$

These connections to the symmetric eigenproblem allow us to adapt the mathematical and algorithmic developments of the previous sections to the singular value problem. Good references for this section include Lawson and Hanson (1974) and Stewart and Sun (1990).

8.6.1 Perturbation Theory and Properties

We first establish perturbation results for the SVD based on the theorems of §8.1. Recall that $\sigma_i(A)$ denotes the ith largest singular value of A.

Theorem 8.6.1 *If $A \in \mathbb{R}^{m\times n}$, then for $k = 1{:}\min\{m, n\}$*

$$\sigma_k(A) = \max_{\substack{\dim(S)=k \\ \dim(T)=k}} \min_{\substack{x\in S \\ y\in T}} \frac{y^T Ax}{\| x \|_2 \| y \|_2} = \max_{\dim(S)=k} \min_{x\in S} \frac{\| Ax \|_2}{\| x \|_2}.$$

Note that in this expression $S \subseteq \mathbb{R}^n$ and $T \subseteq \mathbb{R}^m$ are subspaces.

Proof. The right-most characterization follows by applying Theorem 8.1.2 to A^TA. The remainder of the proof we leave as an exercise. □

Corollary 8.6.2 *If A and $A+E$ are in $\mathbb{R}^{m\times n}$ with $m \geq n$, then for $k = 1{:}n$*

$$|\sigma_k(A+E) - \sigma_k(A)| \leq \sigma_1(E) = \| E \|_2.$$

Proof. Apply Corollary 8.1.6 to

$$\begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & (A+E)^T \\ A+E & 0 \end{bmatrix}. \; \square$$

Example 8.6.1 If

$$A = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \quad \text{and} \quad A+E = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6.01 \end{bmatrix}$$

then $\sigma(A) = \{9.5080, .7729\}$ and $\sigma(A+E) = \{9.5145, .7706\}$. It is clear that for $i = 1{:}2$ we have $|\sigma_i(A+E) - \sigma_i(A)| \leq \| E \|_2 = .01$.

Corollary 8.6.3 *Let $A = [\, a_1, \ldots, a_n \,] \in \mathbb{R}^{m\times n}$ be a column partitioning with $m \geq n$. If $A_r = [\, a_1, \ldots, a_r \,]$, then for $r = 1{:}n-1$*

$$\sigma_1(A_{r+1}) \geq \sigma_1(A_r) \geq \sigma_2(A_{r+1}) \geq \cdots \geq \sigma_r(A_{r+1}) \geq \sigma_r(A_r) \geq \sigma_{r+1}(A_{r+1}).$$

Proof. Apply Corollary 8.1.7 to A^TA. □

This last result says that by adding a column to a matrix, the largest singular value increases and the smallest singular value is diminished.

Example 8.3.2

$$A = \begin{bmatrix} 1 & 6 & 11 \\ 2 & 7 & 12 \\ 3 & 8 & 13 \\ 4 & 9 & 14 \\ 5 & 10 & 15 \end{bmatrix} \quad \Rightarrow \quad \left\{ \begin{array}{lcl} \sigma(A_1) & = & \{7.4162\} \\ \sigma(A_2) & = & \{19.5377, 1.8095\} \\ \sigma(A_3) & = & \{35.1272, 2.4654, 0.0000\} \end{array} \right.$$

thereby confirming Corollary 8.6.3.

The next result is a Wielandt-Hoffman theorem for singular values:

Theorem 8.6.4 *If A and $A+E$ are in $\mathbb{R}^{m\times n}$ with $m \geq n$, then*

$$\sum_{k=1}^{n} (\sigma_k(A+E) - \sigma_k(A))^2 \;\leq\; \| E \|_F^2 .$$

Proof. Apply Theorem 8.1.4 to $\begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$ and $\begin{bmatrix} 0 & (A+E)^T \\ A+E & 0 \end{bmatrix}$. □

Example 8.6.3 If

$$A = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \quad \text{and} \quad A+E = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6.01 \end{bmatrix}$$

then

$$\sum_{k=1}^{2} (\sigma_k(A+E) - \sigma_k(A))^2 \;=\; .472 \times 10^{-4} \;\leq\; 10^{-4} \;=\; \| E \|_F^2 .$$

See Example 8.6.1.

For $A \in \mathbb{R}^{m\times n}$ we say that the k-dimensional subspaces $S \subseteq \mathbb{R}^n$ and $T \subseteq \mathbb{R}^m$ form a *singular subspace pair* if $x \in S$ and $y \in T$ imply $Ax \in T$ and $A^Ty \in S$. The following result is concerned with the perturbation of singular subspace pairs.

Theorem 8.6.5 *Let $A, E \in \mathbb{R}^{m\times n}$ with $m \geq n$ be given and suppose that $V \in \mathbb{R}^{n\times n}$ and $U \in \mathbb{R}^{m\times m}$ are orthogonal. Assume that*

$$V = \begin{array}{c} [\; V_1 \quad V_2 \;] \\ \quad r \quad n-r \end{array} \qquad\qquad U = \begin{array}{c} [\; U_1 \quad U_2 \;] \\ \quad r \quad m-r \end{array}$$

and that ran(V_1) *and* ran(U_1) *form a singular subspace pair for A. Let*

$$U^HAV \;=\; \begin{array}{cc} \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} & \begin{array}{c} r \\ m-r \end{array} \\ \begin{array}{cc} r & n-r \end{array} & \end{array}$$

$$U^H E V \;=\; \begin{array}{cc} \left[\begin{array}{cc} E_{11} & E_{12} \\ E_{21} & E_{22} \end{array}\right] & \begin{array}{c} r \\ m-r \end{array} \\ \begin{array}{cc} r & n-r \end{array} & \end{array}$$

and assume that

$$\delta \;=\; \min_{\substack{\sigma\in\sigma(A_{11}) \\ \gamma\in\sigma(A_{22})}} |\sigma \;-\; \gamma| \;>\; 0.$$

If

$$\| E \|_F \;\leq\; \frac{\delta}{4},$$

then there exist matrices $P \in \mathbb{R}^{(n-r)\times r}$ *and* $Q \in \mathbb{R}^{(m-r)\times r}$ *satisfying*

$$\left\| \left[\begin{array}{c} Q \\ P \end{array} \right] \right\|_F \leq 4\,\frac{\| E \|_F}{\delta}$$

such that $\operatorname{ran}(V_1 + V_2 Q)$ *and* $\operatorname{ran}(U_1 + U_2 P)$ *is a singular subspace pair for* $A + E$.

Proof. See Stewart (1973), Theorem 6.4. □

Roughly speaking, the theorem says that $O(\epsilon)$ changes in A can alter a singular subspace by an amount ϵ/δ, where δ measures the separation of the relevant singular values.

Example 8.6.4 The matrix $A = \operatorname{diag}(2.000,\, 1.001,\, .999) \in \mathbb{R}^{4\times 3}$ has singular subspace pairs $(\operatorname{span}\{v_i\}, \operatorname{span}\{u_i\})$ for i = 1, 2, 3 where $v_i = e_i^{(3)}$ and $u_i = e_i^{(4)}$. Suppose

$$A + E \;=\; \left[\begin{array}{ccc} 2.000 & .010 & .010 \\ .010 & 1.001 & .010 \\ .010 & .010 & .999 \\ .010 & .010 & .010 \end{array} \right]$$

The corresponding columns of the matrices

$$\hat{U} \;=\; [\,\hat{u}_1\;\hat{u}_2\;\hat{u}_3\,] \;=\; \left[\begin{array}{rrr} .9999 & -.0144 & .0007 \\ .0101 & .7415 & .6708 \\ .0101 & .6707 & -.7616 \\ .0051 & .0138 & -.0007 \end{array} \right]$$

$$\hat{V} \;=\; [\,\hat{v}_1\;\hat{v}_2\;\hat{v}_3\,] \;=\; \left[\begin{array}{rrr} .9999 & -.0143 & .0007 \\ .0101 & .7416 & .6708 \\ .0101 & .6707 & -.7416 \end{array} \right]$$

define singular subspace pairs for $A+E$. Note that the pair $\{\operatorname{span}\{\hat{v}_i\}, \operatorname{span}\{\hat{u}_i\}\}$, is close to $\{\operatorname{span}\{v_i\}, \operatorname{span}\{u_i\}\}$ for $i = 1$ but not for $i = 2$ or 3. On the other hand, the singular subspace pair $\{\operatorname{span}\{\hat{v}_2, \hat{v}_3\}, \operatorname{span}\{\hat{u}_2, \hat{u}_3\}\}$ is close to $\{\operatorname{span}\ \{v_2, v_3\}, \operatorname{span}\{u_2, u_3\}\}$.

8.6.2 The SVD Algorithm

We now show how a variant of the QR algorithm can be used to compute the SVD of an $A \in \mathbb{R}^{m \times n}$ with $m \geq n$. At first glance, this appears straightforward. Equation (8.6.1) suggests that we

- form $C = A^TA$,
- use the symmetric QR algorithm to compute $V_1^TCV_1 = \text{diag}(\sigma_i^2)$,
- apply QR with column pivoting to AV_1 obtaining $U^T(AV_1)\Pi = R$.

Since R has orthogonal columns, it follows that $U^TA(V_1\Pi)$ is diagonal. However, as we saw in Example 5.3.2, the formation of A^TA can lead to a loss of information. The situation is not quite so bad here, since the original A is used to compute U.

A preferable method for computing the SVD is described in Golub and Kahan (1965). Their technique finds U and V simultaneously by *implicitly* applying the symmetric QR algorithm to A^TA. The first step is to reduce A to upper bidiagonal form using Algorithm 5.4.2:

$$U_B^TAV_B = \begin{bmatrix} B \\ 0 \end{bmatrix} \qquad B = \begin{bmatrix} d_1 & f_1 & & \cdots & 0 \\ 0 & d_2 & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & f_{n-1} \\ 0 & \cdots & & 0 & d_n \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

The remaining problem is thus to compute the SVD of B. To this end, consider applying an implicit-shift QR step (Algorithm 8.3.2) to the tridiagonal matrix $T = B^TB$:

- Compute the eigenvalue λ of

$$T(m{:}n, m{:}n) = \begin{bmatrix} d_m^2 + f_{m-1}^2 & d_mf_m \\ d_mf_m & d_n^2 + f_m^2 \end{bmatrix} \qquad m = n - 1$$

  that is closer to $d_n^2 + f_m^2$.

- Compute $c_1 = \cos(\theta_1)$ and $s_1 = \sin(\theta_1)$ such that

$$\begin{bmatrix} c_1 & s_1 \\ -s_1 & c_1 \end{bmatrix}^T \begin{bmatrix} d_1^2 - \lambda \\ d_1f_1 \end{bmatrix} = \begin{bmatrix} \times \\ 0 \end{bmatrix}$$

  and set $G_1 = G(1, 2, \theta_1)$.

- Compute Givens rotations $G_2, \ldots, G_{n-1}$ so that if $Q = G_1 \cdots G_{n-1}$ then Q^TTQ is tridiagonal and $Qe_1 = G_1e_1$.

Note that these calculations require the explicit formation of B^TB, which, as we have seen, is unwise from the numerical standpoint.

Suppose instead that we apply the Givens rotation G_1 above to B directly. Illustrating with the $n = 6$ case this gives

$$B \leftarrow BG_1 = \begin{bmatrix} \times & \times & 0 & 0 & 0 & 0 \\ + & \times & \times & 0 & 0 & 0 \\ 0 & 0 & \times & \times & 0 & 0 \\ 0 & 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times \end{bmatrix}.$$

We then can determine Givens rotations U_1, V_2, $U_2, \ldots, V_{n-1}$, and U_{n-1} to chase the unwanted nonzero element down the bidiagonal:

$$B \leftarrow U_1^TB = \begin{bmatrix} \times & \times & + & 0 & 0 & 0 \\ 0 & \times & \times & 0 & 0 & 0 \\ 0 & 0 & \times & \times & 0 & 0 \\ 0 & 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times \end{bmatrix}$$

$$B \leftarrow BV_2 = \begin{bmatrix} \times & \times & 0 & 0 & 0 & 0 \\ 0 & \times & \times & 0 & 0 & 0 \\ 0 & + & \times & \times & 0 & 0 \\ 0 & 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times \end{bmatrix}$$

$$B \leftarrow U_2^TB = \begin{bmatrix} \times & \times & 0 & 0 & 0 & 0 \\ 0 & \times & \times & + & 0 & 0 \\ 0 & 0 & \times & \times & 0 & 0 \\ 0 & 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times \end{bmatrix}$$

and so on. The process terminates with a new bidiagonal $\bar{B}$ that is related to B as follows:

$$\bar{B} = (U_{n-1}^T \cdots U_1^T)B(G_1V_2 \cdots V_{n-1}) = \bar{U}^TB\bar{V}.$$

Since each V_i has the form $V_i = G(i, i+1, \theta_i)$ where $i = 2{:}n-1$, it follows that $\bar{V}e_1 = Qe_1$. By the implicit Q theorem we can assert that $\bar{V}$ and Q are essentially the same. Thus, we can implicitly effect the transition from T to $\bar{T} = \bar{B}^T\bar{B}$ by working directly on the bidiagonal matrix B.

Of course, for these claims to hold it is necessary that the underlying tridiagonal matrices be unreduced. Since the subdiagonal entries of B^TB are of the form d_{i-1}, f_i, it is clear that we must search the bidiagonal band for zeros. If $f_k = 0$ for some k, then

$$B = \begin{array}{cc} \left[\begin{array}{cc} B_1 & 0 \\ 0 & B_2 \end{array}\right] & \begin{array}{c} k \\ n-k \end{array} \\ \begin{array}{cc} k & n-k \end{array} & \end{array}$$

and the original SVD problem decouples into two smaller problems involving the matrices $B_1$and B_2. If $d_k = 0$ for some $k < n$, then premultiplication by a sequence of Givens transformations can zero f_k. For example, if $n = 6$ and $k = 3$, then by rotating in row planes (3,4), (3,5), and (3,6) we can zero the entire third row:

$$B = \left[\begin{array}{cccccc} \times & \times & 0 & 0 & 0 & 0 \\ 0 & \times & \times & 0 & 0 & 0 \\ 0 & 0 & 0 & \times & 0 & 0 \\ 0 & 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times \end{array}\right] \xrightarrow{(3,4)} \left[\begin{array}{cccccc} \times & \times & 0 & 0 & 0 & 0 \\ 0 & \times & \times & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & + & 0 \\ 0 & 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times \end{array}\right]$$

$$\xrightarrow{(3,5)} \left[\begin{array}{cccccc} \times & \times & 0 & 0 & 0 & 0 \\ 0 & \times & \times & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & + \\ 0 & 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times \end{array}\right] \xrightarrow{(3,6)} \left[\begin{array}{cccccc} \times & \times & 0 & 0 & 0 & 0 \\ 0 & \times & \times & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \times & \times & 0 \\ 0 & 0 & 0 & 0 & \times & \times \\ 0 & 0 & 0 & 0 & 0 & \times \end{array}\right]$$

If $d_n = 0$, then the last column can be zeroed with a series of column rotations in planes $(n-1, n)$, $(n-2, n), \ldots, (1, n)$. Thus, we can decouple if $f_1 \cdots f_{n-1} = 0$ or $d_1 \cdots d_n = 0$.

Algorithm 8.6.1 (Golub-Kahan SVD Step) Given a bidiagonal matrix $B \in \mathbb{R}^{m\times n}$ having no zeros on its diagonal or superdiagonal, the following algorithm overwrites B with the bidiagonal matrix $\bar{B} = \bar{U}^T B\bar{V}$ where $\bar{U}$ and $\bar{V}$ are orthogonal and $\bar{V}$ is essentially the orthogonal matrix that would be obtained by applying Algorithm 8.3.2 to $T = B^TB$.

```
Let $\mu$ be the eigenvalue of the trailing 2-by-2 submatrix of $T = B^TB$
    that is closer to $t_{nn}$.
$y = t_{11} - \mu$
$z = t_{12}$
for $k = 1{:}n-1$
    Determine $c = \cos(\theta)$ and $s = \sin(\theta)$ such that
        $\begin{bmatrix} y & z \end{bmatrix} \begin{bmatrix} c & s \\ -s & c \end{bmatrix} = \begin{bmatrix} * & 0 \end{bmatrix}$
    $B = BG(k, k+1, \theta)$
    $y = b_{kk}$; $z = b_{k+1,k}$
    Determine $c = \cos(\theta)$ and $s = \sin(\theta)$ such that
        $\begin{bmatrix} c & s \\ -s & c \end{bmatrix}^T \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} * \\ 0 \end{bmatrix}$
    $B = G(k, k+1, \theta)^T B$
    if $k < n-1$
        $y = b_{k,k+1}$; $z = b_{k,k+2}$
    end
end
```

An efficient implementation of this algorithm would store B's diagonal and superdiagonal in vectors $a(1{:}n)$ and $f(1{:}n-1)$ respectively and would require $30n$ flops and $2n$ square roots. Accumulating U requires $6mn$ flops. Accumulating V requires $6n^2$ flops.

Typically, after a few of the above SVD iterations, the superdiagonal entry f_{n-1} becomes negligible. Criteria for smallness within B's band are usually of the form

$$\begin{aligned} |f_i| &\leq \epsilon(\,|d_i| + |d_{i+1}|\,) \\ |d_i| &\leq \epsilon \,\| B \| \end{aligned}$$

where ϵ is a small multiple of the unit roundoff and $\|\cdot\|$ is some computationally convenient norm.

Combining Algorithm 5.4.2 (bidiagonalization), Algorithm 8.6.1, and the decoupling calculations mentioned earlier gives

Algorithm 8.6.2 (The SVD Algorithm) Given $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ and ϵ, a small multiple of the unit roundoff, the following algorithm overwrites A with $U^TAV = D + E$, where $U \in \mathbb{R}^{m\times n}$ is orthogonal, $V \in \mathbb{R}^{n\times n}$ is orthogonal, $D \in \mathbb{R}^{m\times n}$ is diagonal, and E satisfies $\| E \|_2 \approx \mathbf{u}\| A \|_2$.

Use Algorithm 5.4.2 to compute the bidiagonalization

$$\begin{bmatrix} B \\ 0 \end{bmatrix} \leftarrow (U_1 \cdots U_n)^T A (V_1 \cdots V_{n-2})$$

```
until $q = n$
    Set $b_{i,i+1}$ to zero if $|b_{i,i+1}| \le \epsilon(|b_{ii}| + |b_{i+1,i+1}|)$
        for any $i = 1{:}n-1$.
    Find the largest $q$ and the smallest $p$ such that if

        $B = \begin{array}{c} \left[\begin{array}{ccc} B_{11} & 0 & 0 \\ 0 & B_{22} & 0 \\ 0 & 0 & B_{33} \end{array}\right] \begin{array}{c} p \\ n-p-q \\ q \end{array} \\ \begin{array}{ccc} p & n-p-q & q \end{array} \end{array}$

        then $B_{33}$ is diagonal and $B_{22}$ has nonzero superdiagonal.
    if $q < n$
        if any diagonal entry in $B_{22}$ is zero, then zero
            the superdiagonal entry in the same row.
        else
            Apply Algorithm 8.6.1 to $B_{22}$,
            $B = \text{diag}(I_p, U, I_{q+m-n})^T B \text{diag}(I_p, V, I_q)$
        end
    end
end
```

The amount of work required by this algorithm and its numerical properties are discussed in §5.4.5 and §5.5.8.

Example 8.6.5 If Algorithm 8.6.2 is applied to

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

then the superdiagonal elements converge to zero as follows:

Iteration	$O(\|a_{21}\|)$	$O(\|a_{32}\|)$	$O(\|a_{43}\|)$
1	10^0	10^0	10^0
2	10^0	10^0	10^0
3	10^0	10^0	10^0
4	10^0	10^{-1}	10^{-2}
5	10^0	10^{-1}	10^{-8}
6	10^0	10^{-1}	10^{-27}
7	10^0	10^{-1}	converg.
8	10^0	10^{-4}	
9	10^{-1}	10^{-14}	
10	10^{-1}	converg.	
11	10^{-4}		
12	10^{-12}		
13	converg.		

Observe the cubic-like convergence.

8.6.3 Jacobi SVD Procedures

It is straightforward to adapt the Jacobi procedures of §8.4 to the SVD problem. Instead of solving a sequence of 2-by-2 symmetric eigenproblems, we solve a sequence of 2-by-2 SVD problems. Thus, for a given index pair (p, q) we compute a pair of rotations such that

$$\begin{bmatrix} c_1 & s_1 \\ -s_1 & c_1 \end{bmatrix}^T \begin{bmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{bmatrix} \begin{bmatrix} c_2 & s_2 \\ -s_2 & c_2 \end{bmatrix} = \begin{bmatrix} d_p & 0 \\ 0 & d_q \end{bmatrix}.$$

See P8.6.8. The resulting algorithm is referred to as *two-sided* because each update involves a pre- and post-multiplication.

A *one-sided* Jacobi algorithm involves a sequence of pairwise column orthogonalizations. For a given index pair (p, q) a Jacobi rotation $J(p, q, \theta)$ is determined so that columns p and q of $AJ(p.q, \theta)$ are orthogonal to each other. See P8.6.8. Note that this corresponds to zeroing the (p, q) and (q, p) entries in A^TA. Once AV has sufficiently orthogonal columns, the rest of the SVD (U and Σ) follows from column scaling: $AV = U\Sigma$.

Problems

P8.6.1 Show that if $B \in \mathbb{R}^{n\times n}$ is an upper bidiagonal matrix having a repeated singular value, then B must have a zero on its diagonal or superdiagonal.

P8.6.2 Give formulae for the eigenvectors of $\begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$ in terms of the singular vectors of $A \in \mathbb{R}^{m\times n}$ where $m \geq n$.

P8.6.3 Give an algorithm for reducing a complex matrix A to *real* bidiagonal form using complex Householder transformations.

P8.6.4 Relate the singular values and vectors of $A = B + iC$ ($B, C \in \mathbb{R}^{m\times n}$) to those of $\begin{bmatrix} B & -C \\ C & B \end{bmatrix}$.

P8.6.5 Complete the proof of Theorem 8.6.1.

P8.6.6 Assume that $n = 2m$ and that $S \in \mathbb{R}^{n\times n}$ is skew-symmetric and tridiagonal. Show that there exists a permutation $P \in \mathbb{R}^{n\times n}$ such that P^TSP has the following form:

$$P^TSP = \begin{array}{cc} \begin{bmatrix} 0 & -B^T \\ B & 0 \end{bmatrix} & \begin{array}{c} m \\ m \end{array} \\ \begin{array}{cc} m & m \end{array} & \end{array}.$$

Describe B. Show how to compute the eigenvalues and eigenvectors of S via the SVD of B. Repeat for the case $n = 2m + 1$.

P8.6.7 (a) Let

$$C = \begin{bmatrix} w & x \\ y & z \end{bmatrix}$$

be real. Give a stable algorithm for computing c and s with $c^2 + s^2 = 1$ such that

$$B = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} C$$

is symmetric. (b) Combine (a) with the Jacobi trigonometric calculations in the text to obtain a stable algorithm for computing the SVD of C. (c) Part (b) can be used to develop a Jacobi-like algorithm for computing the SVD of $A \in \mathbb{R}^{n\times n}$. For a given (p,q) with $p < q$, Jacobi transformations $J(p,q,\theta_1)$ and $J(p,q,\theta_2)$ are determined such that if

$$B = J(p,q,\theta_1)^T A J(p,q,\theta_2),$$

then $b_{pq} = b_{qp} = 0$. Show

$$\text{off}(B)^2 = \text{off}(A)^2 - b_{pq}^2 - b_{qp}^2.$$

How might p and q be determined? How could the algorithm be adapted to handle the case when $A \in \mathbb{R}^{m\times n}$ with $m > n$?

P8.6.8 Let x and y be in $\mathbb{R}^m$ and define the orthogonal matrix Q by

$$Q = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}.$$

Give a stable algorithm for computing c and s such that the columns of $[x,y]Q$ are orthogonal to each other.

P8.6.9 Suppose $B \in \mathbb{R}^{n\times n}$ is upper bidiagonal with $b_{nn} = 0$. Show how to construct orthogonal U and V (product of Givens rotations) so that $U^T BV$ is upper bidiagonal with a zero nth column.

P8.6.10 Suppose $B \in \mathbb{R}^{n\times n}$ is upper bidiagonal with diagonal entries $d(1{:}n)$ and superdiagonal entries $f(1{:}n-1)$. State and prove a singular value version of Theorem 8.5.1.

Notes and References for Sec. 8.6

The mathematical properties of the SVD are discussed in Stewart and Sun (1990) as well as

A.R. Amir-Moez (1965). *Extremal Properties of Linear Transformations and Geometry of Unitary Spaces*, Texas Tech University Mathematics Series, no. 243, Lubbock, Texas.

G.W. Stewart (1973). "Error and Perturbation Bounds for Subspaces Associated with Certain Eigenvalue Problems," *SIAM Review 15,* 727–64.

P.A. Wedin (1972). "Perturbation Bounds in Connection with the Singular Value Decomposition," *BIT 12,* 99–111.

G.W. Stewart (1979). "A Note on the Perturbation of Singular Values," *Lin. Alg. and Its Applic. 28,* 213–16.

G.W. Stewart (1984). "A Second Order Perturbation Expansion for Small Singular Values," *Lin. Alg. and Its Applic. 56,* 231–236.

R.J. Vaccaro (1994). "A Second-Order Perturbation Expansion for the SVD," *SIAM J. Matrix Anal. Applic. 15*, 661–671.

The idea of adapting the symmetric QR algorithm to compute the SVD first appeared in

G.H. Golub and W. Kahan (1965). "Calculating the Singular Values and Pseudo-Inverse of a Matrix," *SIAM J. Num. Anal. Ser. B 2,* 205–24.

and then came some early implementations:

P.A. Businger and G.H. Golub (1969). "Algorithm 358: Singular Value Decomposition of a Complex Matrix," *Comm. Assoc. Comp. Mach. 12,* 564–65.

G.H. Golub and C. Reinsch (1970). "Singular Value Decomposition and Least Squares Solutions," *Numer. Math. 14,* 403–20. See also Wilkinson and Reinsch (1971, 134–51).

Interesting algorithmic developments associated with the SVD appear in

J.J.M. Cuppen (1983). "The Singular Value Decomposition in Product Form," *SIAM J. Sci. and Stat. Comp. 4,* 216–222.

J.J. Dongarra (1983). "Improving the Accuracy of Computed Singular Values," *SIAM J. Sci. and Stat. Comp. 4,* 712–719.

S. Van Huffel, J. Vandewalle, and A. Haegemans (1987). "An Efficient and Reliable Algorithm for Computing the Singular Subspace of a Matrix Associated with its Smallest Singular Values," *J. Comp. and Appl. Math. 19,* 313–330.

P. Deift, J. Demmel, L.-C. Li, and C. Tomei (1991). "The Bidiagonal Singular Value Decomposition and Hamiltonian Mechanics," *SIAM J. Num. Anal. 28,* 1463–1516.

R. Mathias and G.W. Stewart (1993). "A Block QR Algorithm and the Singular Value Decomposition," *Lin. Alg. and Its Applic. 182,* 91–100.

Å. Björck, E. Grimme, and P. Van Dooren (1994). "An Implicit Shift Bidiagonalization Algorithm for Ill-Posed Problems," *BIT 34,* 510–534.

The Polar decomposition of a matrix can be computed immediately from its SVD. However, special algorithms have been developed just for this purpose.

N.J. Higham (1986). "Computing the Polar Decomposition—with Applications," *SIAM J. Sci. and Stat. Comp. 7,* 1160–1174.

N.J. Higham and P. Papadimitriou (1994). "A Parallel Algorithm for Computing the Polar Decomposition," *Parallel Comp. 20,* 1161–1173.

Jacobi methods for the SVD fall into two categories. The two-sided Jacobi algorithms repeatedly perform the update $A \leftarrow U^TAV$ producing a sequence of iterates that are increasingly diagonal.

E.G. Kogbetliantz (1955). "Solution of Linear Equations by Diagonalization of Coefficient Matrix," *Quart. Appl. Math. 13,* 123–132.

G.E. Forsythe and P. Henrici (1960). "The Cyclic Jacobi Method for Computing the Principal Values of a Complex Matrix," *Trans. Amer. Math. Soc. 94,* 1–23.

C.C. Paige and P. Van Dooren (1986). "On the Quadratic Convergence of Kogbetliantz's Algorithm for Computing the Singular Value Decomposition," *Lin. Alg. and Its Applic. 77,* 301–313.

J.P. Charlier and P. Van Dooren (1987). "On Kogbetliantz's SVD Algorithm in the Presence of Clusters," *Lin. Alg. and Its Applic. 95,* 135–160.

Z. Bai (1988). "Note on the Quadratic Convergence of Kogbetliantz's Algorithm for Computing the Singular Value Decomposition," *Lin. Alg. and Its Applic. 104,* 131–140.

J.P. Charlier, M. Vanbegin, P. Van Dooren (1988). "On Efficient Implementation of Kogbetliantz's Algorithm for Computing the Singular Value Decomposition," *Numer. Math. 52,* 279–300.

K.V. Fernando (1989). "Linear Convergence of the Row Cyclic Jacobi and Kogbetliantz methods," *Numer. Math. 56,* 73–92.

The one-sided Jacobi SVD procedures repeatedly perform the update $A \leftarrow AV$ producing a sequence of iterates with columns that are increasingly orthogonal.

J.C. Nash (1975). "A One-Sided Tranformation Method for the Singular Value Decomposition and Algebraic Eigenproblem," *Comp. J. 18,* 74–76.

P.C. Hansen (1988). "Reducing the Number of Sweeps in Hestenes Method," in *Singular Value Decomposition and Signal Processing,* ed. E.F. Deprettere, North Holland.

K. Veselič and V. Hari (1989). "A Note on a One-Sided Jacobi Algorithm," *Numer. Math. 56,* 627–633.

Numerous parallel implementations have been developed.

F.T. Luk (1980). "Computing the Singular Value Decomposition on the ILLIAC IV," *ACM Trans. Math. Soft. 6,* 524–39.

R.P. Brent and F.T. Luk (1985). "The Solution of Singular Value and Symmetric Eigenvalue Problems on Multiprocessor Arrays," *SIAM J. Sci. and Stat. Comp. 6*, 69–84.

R.P. Brent, F.T. Luk, and C. Van Loan (1985). "Computation of the Singular Value Decomposition Using Mesh Connected Processors," *J. VLSI Computer Systems 1*, 242–270.

F.T. Luk (1986). "A Triangular Processor Array for Computing Singular Values," *Lin. Alg. and Its Applic. 77*, 259–274.

M. Berry and A. Sameh (1986). "Multiprocessor Jacobi Algorithms for Dense Symmetric Eigenvalue and Singular Value Decompositions," in *Proc. International Conference on Parallel Processing*, 433–440.

R. Schreiber (1986). "Solving Eigenvalue and Singular Value Problems on an Undersized Systolic Array," *SIAM J. Sci. and Stat. Comp. 7*, 441–451.

C.H. Bischof and C. Van Loan (1986). "Computing the SVD on a Ring of Array Processors," in *Large Scale Eigenvalue Problems*, eds. J. Cullum and R. Willoughby, North Holland, 51-66.

C.H. Bischof (1987). "The Two-Sided Block Jacobi Method on Hypercube Architectures," in *Hypercube Multiprocessors*, ed. M.T. Heath, SIAM Press, Philadelphia.

C.H. Bischof (1989). "Computing the Singular Value Decomposition on a Distributed System of Vector Processors," *Parallel Computing 11*, 171–186.

S. Van Huffel and H. Park (1994). "Parallel Tri- and Bidiagonalization of Bordered Bidiagonal Matrices," *Parallel Computing 20*, 1107–1128.

B. Lang (1996). "Parallel Reduction of Banded Matrices to Bidiagonal Form," *Parallel Computing 22*, 1–18.

The divide and conquer algorithms devised for for the symmetric eigenproblem have SVD analogs:

E.R. Jessup and D.C. Sorensen (1994). "A Parallel Algorithm for Computing the Singular Value Decomposition of a Matrix," *SIAM J. Matrix Anal. Appl. 15*, 530–548.

M. Gu and S.C. Eisenstat (1995). "A Divide-and-Conquer Algorithm for the Bidiagonal SVD," *SIAM J. Matrix Anal. Appl. 16*, 79–92.

Careful analyses of the SVD calculation include

J.W. Demmel and W. Kahan (1990). "Accurate Singular Values of Bidiagonal Matrices," *SIAM J. Sci. and Stat. Comp. 11*, 873–912.

K.V. Fernando and B.N. Parlett (1994). "Accurate Singular Values and Differential qd Algorithms," *Numer. Math. 67*, 191–230.

S. Chandrasekaren and I.C.F. Ipsen (1994). "Backward Errors for Eigenvalue and Singular Value Decompositions," *Numer. Math. 68*, 215–223.

High accuracy SVD calculation and connections among the Cholesky, Schur, and singular value computations are discussed in

J.W. Demmel and K. Veselič (1992). "Jacobi's Method is More Accurate than QR," *SIAM J. Matrix Anal. Appl. 13*, 1204–1245.

R. Mathias (1995). "Accurate Eigensystem Computations by Jacobi Methods," *SIAM J. Matrix Anal. Appl. 16*, 977–1003.

8.7 Some Generalized Eigenvalue Problems

Given a symmetric matrix $A \in \mathbb{R}^{n\times n}$ and a symmetric positive definite $B \in \mathbb{R}^{n\times n}$, we consider the problem of finding a nonzero vector x and a scalar λ so $Ax = \lambda Bx$. This is the *symmetric-definite generalized eigenproblem.* The scalar λ can be thought of as a *generalized eigenvalue.* As λ varies, $A - \lambda B$ defines a *pencil* and our job is to determine

$$\lambda(A, B) = \{\, \lambda \mid \det(A - \lambda B) = 0 \,\}.$$

A symmetric-definite generalized eigenproblem can be transformed to an equivalent problem with a congruence transformation:

$$A - \lambda B \text{ is singular} \quad \Leftrightarrow \quad (X^TAX) - \lambda(X^TBX) \text{ is singular}$$

Thus, if X is nonsingular, then $\lambda(A,B) = \lambda(X^TAX, X^TBX)$.

In this section we present various structure-preserving procedures that solve such eigenproblems through the careful selection of X. The related generalized singular value decomposition problem is also discussed.

8.7.1 Mathematical Background

We seek is a stable, efficient algorithm that computes X such that X^TAX and X^TBX are both in "canonical form." The obvious form to aim for is diagonal form.

Theorem 8.7.1 *Suppose A and B are n-by-n symmetric matrices, and define $C(\mu)$ by*

$$C(\mu) \;=\; \mu A + (1-\mu)B \qquad \mu \in \mathbb{R}\,. \tag{8.7.1}$$

If there exists a $\mu \in [0,1]$ such that $C(\mu)$ is non-negative definite and

$$\text{null}(C(\mu)) \;=\; \text{null}(A) \cap \text{null}(B)$$

then there exists a nonsingular X such that both X^TAX and X^TBX are diagonal.

Proof. Let $\mu \in [0,1]$ be chosen so that $C(\mu)$ is non-negative definite with the property that $\text{null}(C(\mu)) = \text{null}(A) \cap \text{null}(B)$. Let

$$Q_1^TC(\mu)Q_1 \;=\; \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \qquad D = \text{diag}(d_1,\ldots,d_k),\; d_i > 0$$

be the Schur decomposition of $C(\mu)$ and define $X_1 = Q_1\text{diag}(D^{-1/2}, I_{n-k})$. If $A_1 = X_1^TAX_1$, $B_1 = X_1^TBX_1$, and $C_1 = X_1^TC(\mu)X_1$, then

$$C_1 \;=\; \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} \;=\; \mu A_1 + (1-\mu)B_1.$$

Since $\text{span}\{e_{k+1},\ldots,e_n\} = \text{null}(C_1) = \text{null}(A_1)\cap\text{null}(B_1)$ it follows that A_1 and B_1 have the following block structure:

$$A_1 = \begin{array}{cc} \left[\begin{array}{cc} A_{11} & 0 \\ 0 & 0 \end{array}\right] & \begin{array}{c} k \\ n-k \end{array} \\ \begin{array}{cc} k & n-k \end{array} & \end{array} \qquad B_1 = \begin{array}{cc} \left[\begin{array}{cc} B_{11} & 0 \\ 0 & 0 \end{array}\right] & \begin{array}{c} k \\ n-k \end{array} \\ \begin{array}{cc} k & n-k \end{array} & \end{array} .$$

Moreover $I_k = \mu A_{11} + (1-\mu)B_{11}$.

Suppose $\mu \neq 0$. It then follows that if $Z^TB_{11}Z = \text{diag}(b_1,\ldots,b_k)$ is the Schur decomposition of B_{11} and we set $X = X_1\text{diag}(Z, I_{n-k})$ then

$$X^TBX = \text{diag}(b_1,\ldots,b_k,0,\ldots,0) \equiv D_B$$

and

$$\begin{aligned} X^TAX &= \frac{1}{\mu}X^T\left(C(\mu)-(1-\mu)B\right)X \\ &= \frac{1}{\mu}\left(\left[\begin{array}{cc} I_k & 0 \\ 0 & 0 \end{array}\right] - (1-\mu)D_B\right) \equiv D_A . \end{aligned}$$

On the other hand, if $\mu = 0$, then let $Z^TA_{11}Z = \text{diag}(a_1,\ldots,a_k)$ be the Schur decomposition of A_{11} and set $X = X_1\text{diag}(Z, I_{n-k})$. It is easy to verify that in this case as well, both X^TAX and X^TBX are diagonal. □

Frequently, the conditions in Theorem 8.7.1 are satisfied because either A or B is positive definite.

Corollary 8.7.2 *If $A - \lambda B \in \mathbb{R}^{n\times n}$ is symmetric-definite, then there exists a nonsingular $X = [\, x_1,\ldots,x_n \,]$ such that*

$$X^TAX = \text{diag}(a_1,\ldots,a_n) \quad \textit{and} \quad X^TBX = \text{diag}(b_1,\ldots,b_n) .$$

Moreover, $Ax_i = \lambda_i Bx_i$ for $i = 1{:}n$ where $\lambda_i = a_i/b_i$.

Proof. By setting $\mu = 0$ in Theorem 8.7.1 we see that symmetric-definite pencils can be simultaneously diagonalized. The rest of the corollary is easily verified. □

Example 8.7.1 If

$$A = \left[\begin{array}{cc} 229 & 163 \\ 163 & 116 \end{array}\right] \quad \text{and} \quad B = \left[\begin{array}{cc} 81 & 59 \\ 59 & 43 \end{array}\right]$$

then $A - \lambda B$ is symmetric-definite and $\lambda(A,B) = \{5, -1/2\}$. If

$$X = \left[\begin{array}{cc} 3 & -5 \\ -4 & 7 \end{array}\right]$$

then $X^TAX = \text{diag}(5,-1)$ and $X^TBX = \text{diag}(1,2)$.

Stewart (1979) has worked out a perturbation theory for symmetric pencils $A - \lambda B$ that satisfy

$$c(A,B) \;=\; \min_{\|x\|_2=1} \; (x^TAx)^2 + (x^TBx)^2 \;>\; 0 \qquad (8.7.2)$$

The scalar $c(A,B)$ is called the *Crawford number* of the pencil $A - \lambda B$.

Theorem 8.7.3 *Suppose $A - \lambda B$ is an n-by-n symmetric-definite pencil with eigenvalues*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n.$$

Suppose E_A and E_B are symmetric n-by-n matrices that satisfy

$$\epsilon^2 \;=\; \| \, E_A \, \|_2^2 \;+ \| \, E_B \, \|_2^2 \;\; < \;\; c(A,B)\,.$$

Then $(A + E_A) - \lambda(B + E_B)$ is symmetric-definite with eigenvalues

$$\mu_1 \geq \cdots \geq \mu_n$$

that satisfy

$$|\arctan(\lambda_i) - \arctan(\mu_i)| \;\leq\; \arctan(\epsilon/c(A,B))$$

for $i = 1{:}n$.

Proof. See Stewart (1979). □

8.7.2 Methods for the Symmetric-Definite Problem

Turning to algorithmic matters, we first present a method for solving the symmetric-definite problem that utilizes both the Cholesky factorization and the symmetric QR algorithm.

Algorithm 8.7.1 Given $A = A^T \in \mathbb{R}^{n \times n}$ and $B = B^T \in \mathbb{R}^{n \times n}$ with B positive definite, the following algorithm computes a nonsingular X such that $X^TBX = I_n$ and $X^TAX = \text{diag}(a_1, \ldots, a_n)$.

```
Compute the Cholesky factorization $B = GG^T$
        using Algorithm 4.2.2.
Compute $C = G^{-1}AG^{-T}$.
Use the symmetric QR algorithm to compute the Schur
        decomposition $Q^TCQ = \text{diag}(a_1, \ldots, a_n)$.
Set $X = G^{-T}Q$.
```

This algorithm requires about $14n^3$ flops. In a practical implementation, A can be overwritten by the matrix C. See Martin and Wilkinson (1968c) for details. Note that

$$\lambda(A,B) = \lambda(A,GG^T) = \lambda(G^{-1}AG^{-T},I) = \lambda(C) = \{a_1,\ldots,a_n\}.$$

If $\hat{a}_i$ is a computed eigenvalue obtained by Algorithm 8.7.1, then it can be shown that $\hat{a}_i \in \lambda(G^{-1}AG^{-T} + E_i)$, where $\| E_i \|_2 \approx \mathbf{u}\| A \|_2 \| B^{-1} \|_2$. Thus, if B is ill-conditioned, then $\hat{a}_i$ may be severely contaminated with roundoff error even if a_i is a well-conditioned generalized eigenvalue. The problem, of course, is that in this case, the matrix $C = G^{-1}AG^{-T}$ can have some very large entries if B, and hence G, is ill-conditioned. This difficulty can sometimes be overcome by replacing the matrix G in Algorithm 8.7.1 with $VD^{-1/2}$ where $V^TBV = D$ is the Schur decomposition of B. If the diagonal entries of D are ordered from smallest to largest, then the large entries in C are concentrated in the upper left-hand corner. The small eigenvalues of C can then be computed without excessive roundoff error contamination (or so the heuristic goes). For further discussion, consult Wilkinson (1965, pp.337–38).

Example 8.7.2 If

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 3 & 5 & 6 \end{bmatrix} \quad \text{and} \quad G = \begin{bmatrix} .001 & 0 & 0 \\ 1 & .001 & 0 \\ 2 & 1 & .001 \end{bmatrix}$$

and $B = GG^T$, then the two smallest eigenvalues of $A - \lambda B$ are

$$a_1 = -0.619402940600584 \qquad a_2 = 1.627440079051887.$$

If 17-digit floating point arithmetic is used, then these eigenvalues are computed to full machine precision when the symmetric QR algorithm is applied to $fl(D^{-1/2}V^TAVD^{-1/2})$, where $B = VDV^T$ is the Schur decomposition of B. On the other hand, if Algorithm 8.7.1 is applied, then

$$\hat{a}_1 = -0.619373517376444 \qquad \hat{a}_2 = 1.627516601905228.$$

The reason for obtaining only four correct significant digits is that $\kappa_2(B) \approx 10^{18}$.

The condition of the matrix X in Algorithm 8.7.1 can sometimes be improved by replacing B with a suitable convex combination of A and B. The connection between the eigenvalues of the modified pencil and those of the original are detailed in the proof of Theorem 8.7.1.

Other difficulties concerning Algorithm 8.7.1 revolve around the fact that $G^{-1}AG^{-T}$ is generally full even when A and B are sparse. This is a serious problem, since many of the symmetric-definite problems arising in practice are large and sparse.

Crawford (1973) has shown how to implement Algorithm 8.7.1 effectively when A and B are banded. Aside from this case, however, the simultaneous diagonalization approach is impractical for the large, sparse symmetric-definite problem.

An alternative idea is to extend the Rayleigh quotient iteration (8.4.4) as follows:

```
$x_0$ given with $\| x_0 \|_2 = 1$
for $k = 0, 1, \ldots$
    $\mu_k = x_k^T A x_k / x_k^T B x_k$
    Solve $(A - \mu_k B) z_{k+1} = B x_k$ for $z_{k+1}$.                (8.7.3)
    $x_{k+1} = z_{k+1} / \| z_{k+1} \|_2$
end
```

The mathematical basis for this iteration is that

$$\lambda = \frac{x^T A x}{x^T B x} \tag{8.7.4}$$

minimizes

$$f(\lambda) = \| Ax - \lambda Bx \|_B \tag{8.7.5}$$

where $\| \cdot \|_B$ is defined by $\|z\|_B^2 = z^T B^{-1} z$. The mathematical properties of (8.7.3) are similar to those of (8.4.4). Its applicability depends on whether or not systems of the form $(A - \mu B)z = x$ can be readily solved. A similar comment pertains to the following generalized orthogonal iteration:

```
$Q_0 \in \mathbb{R}^{n \times p}$ given with $Q_0^T Q_0 = I_p$
for $k = 1, 2, \ldots$
    Solve $B Z_k = A Q_{k-1}$ for $Z_k$.                (8.7.6)
    $Z_k = Q_k R_k$   (QR factorization)
end
```

This is mathematically equivalent to (7.3.4) with A replaced by $B^{-1}A$. Its practicality depends on how easy it is to solve linear systems of the form $Bz = y$.

Sometimes A and B are so large that neither (8.7.3) nor (8.7.6) can be invoked. In this situation, one can resort to any of a number of gradient and coordinate relaxation algorithms. See Stewart (1976) for an extensive guide to the literature.

8.7.3 The Generalized Singular Value Problem

We conclude with some remarks about symmetric pencils that have the form $A^T A - \lambda B^T B$ where $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times n}$. This pencil underlies the *generalized singular value decomposition* (GSVD), a decomposition that is useful in several constrained least squares problems. (Cf. §12.1.) Note that by Theorem 8.7.1 there exists a nonsingular $X \in \mathbb{R}^{n \times n}$ such that $X^T(A^T A)X$ and $X^T(B^T B)X$ are both diagonal. The value of the GSVD is that these diagonalizations can be achieved without forming A^TA and B^TB.

Theorem 8.7.4 (Generalized Singular Value Decomposition) *If we have $A \in \mathbb{R}^{m\times n}$ with $m \geq n$ and $B \in \mathbb{R}^{p\times n}$, then there exist orthogonal $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{p\times p}$ and an invertible $X \in \mathbb{R}^{n\times n}$ such that*

$$U^TAX = C = \operatorname{diag}(c_1,\ldots,c_n) \qquad c_i \geq 0$$

and

$$V^TBX = S = \operatorname{diag}(s_1,\ldots,s_q) \qquad s_i \geq 0$$

where $q = \min(p,n)$.

Proof. The proof of this decomposition appears in Van Loan (1976). We present a more constructive proof along the lines of Paige and Saunders (1981). For clarity we assume that $\operatorname{null}(A) \cap \operatorname{null}(B) = \{0\}$ and $p \geq n$. We leave it to the reader to extend the proof so that it covers theses cases.

Let

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} R \qquad (8.7.6)$$

be a QR factorization with $Q_1 \in \mathbb{R}^{m\times n}$, $Q_2 \in \mathbb{R}^{p\times n}$, and $R \in \mathbb{R}^{n\times n}$. Paige and Saunders show that the SVD's of Q_1 and Q_2 are related in the sense that

$$Q_1 = UCW^T \qquad Q_2 = VSW^T \qquad (8.7.7)$$

Here, U,V, and W are orthogonal, $C = \operatorname{diag}(c_i)$ with $0 \leq c_1 \leq \cdots \leq c_n$, $S = \operatorname{diag}(s_i)$ with $s_1 \geq \cdots \geq s_n$, and $C^TC + S^TS = I_n$. The decomposition (8.7.7) is a variant of the CS decomposition in §2.6 and from it we conclude that $A = Q_1R = UC(W^TR)$ and $B = Q_2R = VS(W^TR)$. The theorem follows by setting $X = (W^TR)^{-1}$, $D_A = C$, and $D_B = S$. The invertibility of R follows from our assumption that $\operatorname{null}(A) \cap \operatorname{null}(B) = \{0\}$. □

The elements of the set $\sigma(A,B) \equiv \{c_1/s_1,\ldots,c_n/s_q\}$ are referred to as the *generalized singular values* of A and B. Note that $\sigma \in \sigma(A,B)$ implies that $\sigma^2 \in \lambda(A^TA, B^TB)$. The theorem is a generalization of the SVD in that if $B = I_n$, then $\sigma(A,B) = \sigma(A)$.

Our proof of the GSVD is of practical importance since Stewart (1983) and Van Loan (1985) have shown how to stably compute the CS decomposition. The only tricky part is the inversion of W^TR to get X. Note that the columns of $X = [\,x_1,\ldots,x_n\,]$ satisfy

$$s_i^2A^TAx_i = c_i^2B^TBx_i \qquad i = 1{:}n$$

and so if $s_i \neq 0$ then $A^TAx_i = \sigma_i^2B^TBx_i$ where $\sigma_i = c_i/s_i$. Thus, the x_i are aptly termed the *generalized singular vectors* of the pair (A,B).

In several applications an orthonormal basis for some designated generalized singular vector subspace space span$\{x_{i_1},\ldots,x_{i_k}\}$ is required. We show how this can be accomplished without any matrix inversions or cross products:

- Compute the QR factorization

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} R.$$

- Compute the CS decomposition

$$Q_1 = UCW^T \qquad Q_2 = VSW^T$$

  and order the diagonals of C and S so that

$$\{c_1/s_1,\ldots,c_k/s_k\} = \{c_{i_1}/s_{i_1},\ldots,c_{i_k}/s_{i_k}\}.$$

- Compute orthogonal Z and upper triangular T so $TZ = W^TR$. (See P8.7.5.) Note that if $X^{-1} = W^TR = TZ$, then $X = Z^TT^{-1}$ and so the first k rows of Z are an orthonormal basis for span$\{x_1,\ldots,x_k\}$.

Problems

P8.7.1 Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric and $G \in \mathbb{R}^{n\times n}$ is lower triangular and nonsingular. Give an efficient algorithm for computing $C = G^{-1}AG^{-T}$.

P8.7.2 Suppose $A \in \mathbb{R}^{n\times n}$ is symmetric and $B \in \mathbb{R}^{n\times n}$ is symmetric positive definite. Give an algorithm for computing the eigenvalues of AB that uses the Cholesky factorization and the symmetric QR algorithm.

P8.7.3 Show that if C is real and diagonalizable, then there exist symmetric matrices A and B, B nonsingular, such that $C = AB^{-1}$. This shows that symmetric pencils $A - \lambda B$ are essentially general.

P8.7.4 Show how to convert an $Ax = \lambda Bx$ problem into a generalized singular value problem if A and B are both symmetric and non-negative definite.

P8.7.5 Given $Y \in \mathbb{R}^{n\times n}$ show how to compute Householder matrices $H_2,\ldots,H_n$ so that $YH_n\cdots H_2 = T$ is upper triangular. Hint: H_k zeros out the kth row.

P8.7.6 Suppose

$$\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}\begin{bmatrix} y \\ z \end{bmatrix} = \lambda\begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix}\begin{bmatrix} y \\ z \end{bmatrix}$$

where $A \in \mathbb{R}^{m\times n}$, $B_1 \in \mathbb{R}^{m\times m}$, and $B_2 \in \mathbb{R}^{n\times n}$. Assume that B_1 and B_2 are positive definite with Cholesky triangles G_1 and G_2 respectively. Relate the generalized eigenvalues of this problem to the singular values of $G_1^{-1}AG_2^{-T}$

P8.7.7 Suppose A and B are both symmetric positive definite. Show how to compute $\lambda(A,B)$ and the corresponding eigenvectors using the Cholesky factorization and CS decomposition.

Notes and References for Sec. 8.7

An excellent survey of computational methods for symmetric-definite pencils is given in

G.W. Stewart (1976). "A Bibliographical Tour of the Large Sparse Generalized Eigenvalue Problem," in *Sparse Matrix Computations* , ed., J.R. Bunch and D.J. Rose, Academic Press, New York.

Some papers of particular interest include

R.S. Martin and J.H. Wilkinson (1968c). "Reduction of a Symmetric Eigenproblem $Ax = \lambda Bx$ and Related Problems to Standard Form," *Numer. Math. 11,* 99–110.

G. Peters and J.H. Wilkinson (1969). "Eigenvalues of $Ax = \lambda Bx$ with Band Symmetric A and B," *Comp. J. 12,* 398-404.

G. Fix and R. Heiberger (1972). "An Algorithm for the Ill-Conditioned Generalized Eigenvalue Problem," *SIAM J. Num. Anal. 9,* 78–88.

C.R. Crawford (1973). "Reduction of a Band Symmetric Generalized Eigenvalue Problem," *Comm. ACM 16,* 41–44.

A. Ruhe (1974). "SOR Methods for the Eigenvalue Problem with Large Sparse Matrices," *Math. Comp.* 28, 695–710.

C.R. Crawford (1976). "A Stable Generalized Eigenvalue Problem," *SIAM J. Num. Anal. 13,* 854–60.

A. Bunse-Gerstner (1984). "An Algorithm for the Symmetric Generalized Eigenvalue Problem," *Lin. Alg. and Its Applic. 58,* 43–68.

C.R. Crawford (1986). "Algorithm 646 PDFIND: A Routine to Find a Positive Definite Linear Combination of Two Real Symmetric Matrices," *ACM Trans. Math. Soft. 12,* 278–282.

C.R. Crawford and Y.S. Moon (1983). "Finding a Positive Definite Linear Combination of Two Hermitian Matrices," *Lin. Alg. and Its Applic. 51,* 37–48.

W. Shougen and Z. Shuqin (1991). "An Algorithm for $Ax = \lambda Bx$ with Symmetric and Positive Definite A and B," *SIAM J. Matrix Anal. Appl. 12,* 654–660.

K. Li and T-Y. Li (1993). "A Homotopy Algorithm for a Symmetric Generalized Eigenproblem," *Numerical Algorithms 4,* 167–195.

K. Li, T-Y. Li, and Z. Zeng (1994). "An Algorithm for the Generalized Symmetric Tridiagonal Eigenvalue Problem," *Numerical Algorithms 8,* 269–291.

H. Zhang and W.F. Moss (1994). "Using Parallel Banded Linear System Solvers in Generalized Eigenvalue Problems," *Parallel Computing 20,* 1089–1106

The simultaneous reduction of two symmetric matrices to diagonal form is discussed in

A. Berman and A. Ben-Israel (1971). "A Note on Pencils of Hermitian or Symmetric Matrices," *SIAM J. Applic. Math. 21,* 51–54.

F. Uhlig (1973). "Simultaneous Block Diagonalization of Two Real Symmetric Matrices," *Lin. Alg. and Its Applic. 7,* 281–89.

F. Uhlig (1976). "A Canonical Form for a Pair of Real Symmetric Matrices That Generate a Nonsingular Pencil," *Lin. Alg. and Its Applic. 14,* 189–210.

K.N. Majinder (1979). "Linear Combinations of Hermitian and Real Symmetric Matrices," *Lin. Alg. and Its Applic. 25,* 95–105.

The perturbation theory that we presented for the symmetric-definite problem was taken from

G.W. Stewart (1979). "Perturbation Bounds for the Definite Generalized Eigenvalue Problem," *Lin. Alg. and Its Applic. 23,* 69–86.

See also

L. Elsner and J. Guang Sun (1982). "Perturbation Theorems for the Generalized Eigenvalue Problem,; *Lin. Alg. and its Applic. 48,* 341-357.

J. Guang Sun (1982). "A Note on Stewart's Theorem for Definite Matrix Pairs," *Lin. Alg. and Its Applic. 48,* 331–339.

J. Guang Sun (1983). "Perturbation Analysis for the Generalized Singular Value Problem," *SIAM J. Numer. Anal. 20,* 611–625.

C.C. Paige (1984). "A Note on a Result of Sun J.-Guang: Sensitivity of the CS and GSV Decompositions," *SIAM J. Numer. Anal. 21,* 186–191.

The generalized SVD and some of its applications are discussed in

C.F. Van Loan (1976). "Generalizing the Singular Value Decomposition," *SIAM J. Num. Anal. 13,* 76–83.

C.C. Paige and M. Saunders (1981). "Towards A Generalized Singular Value Decomposition," *SIAM J. Num. Anal. 18,* 398–405.

B. Kågström (1985). "The Generalized Singular Value Decomposition and the General $A - \lambda B$ Problem," *BIT 24,* 568-583.

Stable methods for computing the CS and generalized singular value decompositions are described in

G.W. Stewart (1983). "A Method for Computing the Generalized Singular Value Decomposition," in *Matrix Pencils* , ed. B. Kågström and A. Ruhe, Springer-Verlag, New York, pp. 207–20.

C.F. Van Loan (1985). "Computing the CS and Generalized Singular Value Decomposition," *Numer. Math. 46,* 479–492.

M.T. Heath, A.J. Laub, C.C. Paige, and R.C. Ward (1986). "Computing the SVD of a Product of Two Matrices," *SIAM J. Sci. and Stat. Comp. 7,* 1147-1159.

C.C. Paige (1986). "Computing the Generalized Singular Value Decomposition," *SIAM J. Sci. and Stat. Comp. 7,* 1126–1146.

L.M. Ewerbring and F.T. Luk (1989). "Canonical Correlations and Generalized SVD: Applications and New Algorithms," *J. Comput. Appl. Math. 27,* 37–52.

J. Erxiong (1990). "An Algorithm for Finding Generalized Eigenpairs of a Symmetric Definite Matrix Pencil," *Lin.Alg. and Its Applic. 132,* 65–91.

P.C. Hansen (1990). "Relations Between SVD and GSVD of Discrete Regularization Problems in Standard and General Form," *Lin.Alg. and Its Applic. 141,* 165–176.

H. Zha (1991). "The Restricted Singular Value Decomposition of Matrix Triplets," *SIAM J. Matrix Anal. Appl. 12,* 172–194.

B. De Moor and G.H. Golub (1991). "The Restricted Singular Value Decomposition: Properties and Applications," *SIAM J. Matrix Anal. Appl. 12,* 401–425.

V. Hari (1991). "On Pairs of Almost Diagonal Matrices," *Lin. Alg. and Its Applic. 148,* 193–223.

B. De Moor and P. Van Dooren (1992). "Generalizing the Singular Value and QR Decompositions," *SIAM J. Matrix Anal. Appl. 13,* 993–1014.

H. Zha (1992). "A Numerical Algorithm for Computing the Restricted Singular Value Decomposition of Matrix Triplets," *Lin.Alg. and Its Applic. 168,* 1–25.

R-C. Li (1993). "Bounds on Perturbations of Generalized Singular Values and of Associated Subspaces," *SIAM J. Matrix Anal. Appl. 14,* 195–234.

K. Veselič (1993). "A Jacobi Eigenreduction Algorithm for Definite Matrix Pairs," *Numer. Math. 64,* 241–268.

Z. Bai and H. Zha (1993). "A New Preprocessing Algorithm for the Computation of the Generalized Singular Value Decomposition," *SIAM J. Sci. Comp. 14,* 1007–1012.

L. Kaufman (1993). "An Algorithm for the Banded Symmetric Generalized Matrix Eigenvalue Problem," *SIAM J. Matrix Anal. Appl. 14,* 372–389.

G.E. Adams, A.W. Bojanczyk, and F.T. Luk (1994). "Computing the PSVD of Two 2×2 Triangular Matrices," *SIAM J. Matrix Anal. Appl. 15,* 366–382.

Z. Drmač (1994). *The Generalized Singular Value Problem,* Ph.D. Thesis, FernUniversitat, Hagen, Germany.

R-C. Li (1994). "On Eigenvalue Variations of Rayleigh Quotient Matrix Pencils of a Definite Pencil," *Lin. Alg. and Its Applic. 208/209,* 471–483.