5

Eigenvalues and eigenvectors of a symmetric matrix

5.1 Introduction

Eigenvalue problems for symmetric matrices arise in all areas of applied science. The terminology *eigenvalue* comes from the German word *Eigenwert* which means proper or characteristic value. The concept of eigenvalue first appeared in an article on systems of linear differential equations by the French mathematician d'Alembert[1] in the course of studying the motion of a string with masses attached to it at various points.

Let us recall from Chapter 2 the definition of eigenvalue and eigenvector.

Definition 5.1 *Suppose that $A \in \mathbb{R}^{n\times n}$. A complex number λ for which the set of linear equations*

$$A\boldsymbol{x} = \lambda\boldsymbol{x} \tag{5.1}$$

has a nontrivial solution $\boldsymbol{x} \in \mathbb{C}^n_* = \mathbb{C}^n \setminus \{\mathbf{0}\}$ is called an **eigenvalue** *of A; the associated solution $\boldsymbol{x} \in \mathbb{C}^n_*$ is called an* **eigenvector** *of A (corresponding to λ).*

[1] Jean le Rond d'Alembert (17 November 1717, Paris, France – 29 October 1783, Paris, France) was abandoned as a newly born child on the steps of the church of St Jean le Rond in Paris and spent his early life in a home for homeless children. d'Alembert was the central mathematical figure among the French Encyclopedists in the period 1751–1772; the Encyclopedia, edited by Jean Diderot, comprised 28 volumes. D'Alembert made a number of significant contributions to the dynamics of rigid bodies, hydrodynamics, aerodynamics, the three-body problem, and the theory of vibrating strings.

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

In order to motivate the discussion that will follow, we begin with two familiar elementary examples.

In considering the rotation of a rigid body $\Omega \subset \mathbb{R}^3$, the *inertia matrix* is the 3×3 symmetric matrix

$$J = \begin{pmatrix} I_{xx} & -I_{xy} & -I_{xz} \\ -I_{yx} & I_{yy} & -I_{yz} \\ -I_{zx} & -I_{zy} & I_{zz} \end{pmatrix}$$

whose diagonal elements are the moments of inertia about the axes,

$$I_{xx} = \int_\Omega (y^2 + z^2)\,\mathrm{d}\Omega, \quad I_{yy} = \int_\Omega (z^2 + x^2)\,\mathrm{d}\Omega, \quad I_{zz} = \int_\Omega (x^2 + y^2)\,\mathrm{d}\Omega,$$

and whose off-diagonal elements are defined by the corresponding products of inertia

$$I_{xy} = I_{yx} = \int_\Omega xy\,\mathrm{d}\Omega\,,$$
$$I_{yz} = I_{zy} = \int_\Omega yz\,\mathrm{d}\Omega\,,$$
$$I_{zx} = I_{xz} = \int_\Omega zx\,\mathrm{d}\Omega\,.$$

Then, the eigenvectors of the inertia matrix are the directions of the *principal axes of inertia* of the body, about which free steady rotation is possible, and the eigenvalues are the *principal moments of inertia* about these axes.

A second example, which involves matrices of any order, arises in the solution of systems of linear ordinary differential equations of the form

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = A\boldsymbol{x}\,,$$

where $\boldsymbol{x}$ is a vector of n elements, each of which is a function of the independent variable t, and A is an $n \times n$ matrix whose elements are constants. If A were a diagonal matrix, with diagonal elements $a_{ii} = \lambda_i$, $i = 1, 2, \ldots, n$, the solution of this system would be straightforward, as each of the equations could be solved separately, giving

$$x_i(t) = x_i(0)\exp(\lambda_i t)\,, \qquad i = 1, 2, \ldots, n\,.$$

When A is not a diagonal matrix, suppose that we can find a nonsingular matrix M such that

$$M^{-1}AM = D\,,$$

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

where D is a diagonal matrix. Then, on letting

$$\boldsymbol{y} = M^{-1}\boldsymbol{x}\,,$$

we easily see that

$$\frac{\mathrm{d}\boldsymbol{y}}{\mathrm{d}t} = M^{-1}AM\boldsymbol{y} = D\boldsymbol{y}\,.$$

The solution of this system of differential equations is straightforward, as we have just seen, and we then find that

$$x_i = (My)_i = \sum_{j=1}^{n} M_{ij}y_j(0)\exp(\lambda_j t)\,,$$

where $\lambda_j = d_{jj}$ is one of the diagonal elements of D. The numbers λ_j, $j = 1, 2, \ldots, n$, are the eigenvalues of the matrix $A \in \mathbb{R}^{n\times n}$, and the columns of M are the eigenvectors of A, so the solution of this system of differential equations requires the calculation of the eigenvalues and eigenvectors of the matrix A.

In systems of differential equations of this kind the matrix A is not necessarily symmetric. In that case, the problem is more difficult; if the eigenvalues of A are not distinct there may not exist a complete set of linearly independent eigenvectors, and then the matrix M will not exist.[1]

In this chapter, we shall develop numerical algorithms for the solution of the algebraic eigenvalue problem (5.1), assuming throughout that $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix. As has been noted above, the analogous problem for a nonsymmetric matrix is more involved, and will not be considered here.[2]

Throughout this chapter, the set of all real-valued symmetric matrices of order n will be denoted by $\mathbb{R}^{n\times n}_{\mathrm{sym}}$; thus, given a matrix $A = (a_{ij})$,

$$A \in \mathbb{R}^{n\times n}_{\mathrm{sym}} \qquad \Leftrightarrow \qquad A \in \mathbb{R}^{n\times n} \quad \& \quad a_{ij} = a_{ji}\,, \quad i, j = 1, 2, \ldots, n\,.$$

We begin with a reminder of some fundamental properties.

[1] Consider, for example,

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}.$$

This matrix has one eigenvalue of multiplicity 2, $\lambda_{1/2} = 1$, and only one (linearly independent) eigenvector, $(1, 0)^{\mathrm{T}}$.

[2] The reader is referred to the last four chapters of J.H. Wilkinson's monograph, *The Algebraic Eigenvalue Problem,* The Clarendon Press, Oxford University Press, New York, 1988.

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

Theorem 5.1 *Suppose that $A \in \mathbb{R}^{n\times n}_{\rm sym}$; then, the following statements are valid.*

(i) *There exist n linearly independent eigenvectors $\boldsymbol{x}^{(i)} \in \mathbb{R}^n$ and corresponding eigenvalues $\lambda_i \in \mathbb{R}$ such that $A\boldsymbol{x}^{(i)} = \lambda_i \boldsymbol{x}^{(i)}$ for all $i = 1, 2, \ldots, n$.*

(ii) *The function*

$$\lambda \mapsto \det(A - \lambda I) \tag{5.2}$$

is a polynomial of degree n with leading term $(-1)^n\lambda^n$, called the **characteristic polynomial of** *A. The eigenvalues of A are the zeros of the characteristic polynomial.*

(iii) *If the eigenvalues λ_i and λ_j of A are distinct, then the corresponding eigenvectors $\boldsymbol{x}^{(i)}$ and $\boldsymbol{x}^{(j)}$ are orthogonal in $\mathbb{R}^n$, i.e.,*

$$\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{x}^{(j)} = 0 \qquad \text{if } \lambda_i \neq \lambda_j\,, \qquad i, j \in \{1, 2, \ldots, n\}\,.$$

(iv) *If λ_i is a root of multiplicity m of (5.2), then there is a linear subspace in $\mathbb{R}^n$ of dimension m, spanned by m mutually orthogonal eigenvectors associated with the eigenvalue λ_i.*

(v) *Suppose that each of the eigenvectors $\boldsymbol{x}^{(i)}$ of A is* **normalised**, *in other words, $\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{x}^{(i)} = 1$ for $i = 1, 2, \ldots, n$, and let X denote the square matrix whose columns are the normalised (orthogonal) eigenvectors; then, the matrix $\Lambda = X^{\mathrm{T}} A\, X$ is diagonal, and the diagonal elements of Λ are the eigenvalues of A.*

(vi) *Let $Q \in \mathbb{R}^{n\times n}$ be an orthogonal matrix and define $B \in \mathbb{R}^{n\times n}_{\rm sym}$ by $B = Q^{\mathrm{T}} A\, Q$; then, $\det(B - \lambda I) = \det(A - \lambda I)$ for each $\lambda \in \mathbb{R}$. The eigenvalues of B are the same as the eigenvalues of A, and the eigenvectors of B are the vectors $Q^{\mathrm{T}}\boldsymbol{x}^{(i)}$, $i = 1, 2, \ldots, n$.*

(vii) *Any vector $\boldsymbol{v} \in \mathbb{R}^n$ can be expressed as a linear combination of the (ortho)normalised eigenvectors $\boldsymbol{x}^{(i)}$, $i = 1, 2, \ldots, n$, of A, i.e.,*

$$\boldsymbol{v} = \sum_{i=1}^{n} \alpha_i \boldsymbol{x}^{(i)}, \qquad \alpha_i = \boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{v}\,.$$

(viii) *The trace of A, $\mathrm{Trace}(A) = \sum_{i=1}^{n} a_{ii}$, is equal to the sum of the eigenvalues of A.*

These properties should be familiar; proofs will be found in any standard text on linear algebra.[1]

[1] See, for example, T.S. Blyth and E.F. Robertson, *Basic Linear Algebra*, Springer Undergraduate Mathematics Series, Springer, 1998, A.G. Hamilton, *Linear Algebra*, Cambridge University Press, 1990, or R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, 1992.

5.2 The characteristic polynomial

Given that $A \in \mathbb{R}^{n\times n}$ and $n \leq 4$, it is quite easy to write down the characteristic polynomial $\det(A-\lambda I)$ by expanding the determinant, and then find the roots of this polynomial of degree n in order to determine the eigenvalues of A. If $n > 4$ there is no general closed formula for the roots of a polynomial in terms of its coefficients, and therefore we have to resort to a numerical technique. A further difficulty is that the roots may be very sensitive to small changes in the coefficients of the polynomial, and we find that the effect of rounding errors in the construction of the characteristic polynomial is usually catastrophic.

Example 5.1 *Consider, for example, the diagonal matrix of order* 16 *whose diagonal elements are* $j + \frac{1}{3}$, $j = 1, 2, \ldots, 16$*; the eigenvalues are, of course, just the diagonal elements. Constructing the characteristic polynomial, working with 10 significant digits throughout, gives the result*

$$\lambda^{16} - 141.3333333\lambda^{15} + 9193.333333\lambda^{14} - \cdots .$$

Using a standard numerical algorithm (such as Newton's method) for computing the roots of the polynomial and working with 10 significant digits gives the smallest root as 1.333333331, which is nearly correct to 10 significant digits. The three largest roots, however, are computed as, approximately, $15.5 \pm 1.3i$ and 16.7, which are very different from their true values $14.\dot{3}$, $15.\dot{3}$, $16.\dot{3}$, respectively, even though the matrix in this example is of quite modest size, and the eigenvalues are well spaced. Thus we conclude from this example that the numerical method which constructs the characteristic polynomial and finds its roots is completely unsatisfactory for general use, except for matrices of very small size. ◇

The fact that in general the roots of the characteristic polynomial cannot be given in closed form shows that any method must proceed by successive approximation. Although one cannot expect to produce the required eigenvalues exactly in a finite number of steps, we shall see that there exist rapidly convergent iterative methods for computing the eigenvalues and eigenvectors numerically.

5.3 Jacobi's method

This method uses a succession of orthogonal transformations to produce a sequence of matrices which approaches a diagonal matrix in the limit.

Each step in the process involves a matrix representing a plane rotation. We begin with a simple example.

Example 5.2 (The plane rotation matrix in $\mathbb{R}^2$) *Let us suppose that $\varphi \in [-\pi, \pi]$ and consider the matrix $R(\varphi) \in \mathbb{R}^{2\times 2}$ defined by*

$$R(\varphi) = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}.$$

For a vector $\boldsymbol{x} \in \mathbb{R}^2$, $R(\varphi)\boldsymbol{x}$ is the plane rotation of $\boldsymbol{x}$ around the origin by an angle φ (in the clockwise direction when $\varphi > 0$ and in the anticlockwise direction when $\varphi < 0$).

We note in passing that since $\cos(-\varphi) = \cos\varphi$, $\sin(-\varphi) = -\sin\varphi$ and $\cos^2\varphi + \sin^2\varphi = 1$, we have that

$$(R(\varphi))^{\mathrm{T}} = R(-\varphi) \qquad \text{and} \qquad R(\varphi)\,R(-\varphi) = I\,.$$

Hence $R(\varphi)$ is an orthogonal matrix; i.e.,

$$R(\varphi)R(\varphi)^{\mathrm{T}} = R(\varphi)^{\mathrm{T}}R(\varphi) = I\,,$$

where I is the 2×2 identity matrix.

The next definition extends the notion of plane rotation matrix to $\mathbb{R}^n$.

Definition 5.2 (The plane rotation matrix in $\mathbb{R}^n$) *Suppose that $n \geq 2$, $1 \leq p < q \leq n$ and $\varphi \in [-\pi, \pi]$. We consider the matrix $R^{(pq)}(\varphi) \in \mathbb{R}^{n\times n}$ whose elements are the same as those of the identity matrix $I \in \mathbb{R}^{n\times n}$, except for the four elements*

$$\begin{aligned} r_{pp} &= c\,, & r_{pq} &= s\,, \\ r_{qp} &= -s\,, & r_{qq} &= c\,, \end{aligned}$$

where $c = \cos\varphi$, $s = \sin\varphi$.

As in Example 5.2, it is a straightforward matter to show that

$$(R^{(pq)}(\varphi))^{\mathrm{T}} = R^{(pq)}(-\varphi)\,, \qquad R^{(pq)}(\varphi)\,R^{(pq)}(-\varphi) = I\,,$$

and that, therefore,

$$R^{(pq)}(\varphi)(R^{(pq)}(\varphi))^{\mathrm{T}} = (R^{(pq)}(\varphi))^{\mathrm{T}}R^{(pq)}(\varphi) = I\,.$$

Hence $R^{(pq)}(\varphi) \in \mathbb{R}^{n\times n}$ is an orthogonal matrix for any p, q such that $1 \leq p < q \leq n$, and any $\varphi \in [-\pi, \pi]$.

The basic result underlying Jacobi's method is encapsulated in the next theorem.

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

Theorem 5.2 *Suppose that $A \in \mathbb{R}^{n\times n}_{\rm sym}$. For each pair of integers (p,q) with $1 \le p < q \le n$, there exists $\varphi \in [-\pi/4, \pi/4]$ such that the (p,q) entry of the symmetric matrix $R^{(pq)}(\varphi)^{\rm T} A R^{(pq)}(\varphi)$ is equal to 0.*

Proof For the sake of notational simplicity, we shall write R instead of $R^{(pq)}(\varphi)$ throughout the proof, and abbreviate $c = \cos\varphi$ and $s = \sin\varphi$.

Consider the product $A' = AR$. Evidently the only difference between A' and A is in columns p and q; these columns of A' are linear combinations of the same two columns of A:

$$\left.\begin{array}{rcl} a'_{ip} &=& a_{ip}c - a_{iq}s \\ a'_{iq} &=& a_{ip}s + a_{iq}c \end{array}\right\}, \quad i = 1, 2, \ldots, n\,. \tag{5.3}$$

Multiplication of A' by $R^{\rm T}$ on the left gives a similar result, but affects rows p and q, rather than columns p and q. Writing $B = R^{\rm T}A'$ gives

$$\left.\begin{array}{rcl} b_{pj} &=& a'_{pj}c - a'_{qj}s \\ b_{qj} &=& a'_{pj}s + a'_{qj}c \end{array}\right\}, \quad j = 1, 2, \ldots, n\,. \tag{5.4}$$

Combining these equations shows that $B = R^{\rm T} A\, R$, where

$$\left.\begin{array}{rcl} b_{pp} &=& a_{pp}c^2 - 2a_{pq}sc + a_{qq}s^2, \\ b_{qq} &=& a_{pp}s^2 + 2a_{pq}sc + a_{qq}c^2, \\ b_{pq} &=& (a_{pp} - a_{qq})sc + a_{pq}(c^2 - s^2) = b_{qp}\,. \end{array}\right\} \tag{5.5}$$

The remaining elements of $B = R^{\rm T} A\, R$ in columns p and q are given by the expressions

$$\left.\begin{array}{rcl} b_{ip} &=& a_{ip}c - a_{iq}s \\ b_{iq} &=& a_{ip}s + a_{iq}c \end{array}\right\}, \quad i = 1, 2, \ldots, n\,, \quad i \ne p, q\,.$$

The matrix $B = R^{\rm T} A\, R$ is evidently symmetric, so the nondiagonal elements of B in rows p and q are also given by the same expressions.

Finally, we note that all the elements of B which do not lie either in row p or q or in column p or q are the same as the corresponding elements of A, that is,

$$b_{ij} = a_{ij}\,, \qquad \text{if } i \ne p, q \text{ and } j \ne p, q\,.$$

We see from (5.5) that in order to ensure that b_{pq}, the (p,q)-entry of the matrix $B = R^{\rm T} A\, R$, is equal to 0, it suffices to choose φ such that

$$\tan 2\varphi = \frac{2a_{pq}}{a_{qq} - a_{pp}}\,; \tag{5.6}$$

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

thus we select

$$\varphi = \frac{1}{2}\tan^{-1}\frac{2a_{pq}}{a_{qq}-a_{pp}} \in [-\pi/4, \pi/4]\,. \tag{5.7}$$

To see this, apply the trigonometric identities $c^2 - s^2 = \cos(2\varphi)$ and $sc = \frac{1}{2}\sin(2\varphi)$ to b_{pq} in (5.5), with $b_{pq} = 0$. That completes the proof.[1] □

We can avoid the trigonometric calculations involved in the formula (5.7) for φ by writing $t = s/c$, and seeing that t is required to satisfy

$$(a_{pp} - a_{qq})t + a_{pq}(1 - t^2) = 0\,. \tag{5.8}$$

If $a_{pq} = 0$, we can ensure that (5.8) holds by selecting $t = 0$ (which corresponds to choosing $\varphi = 0$). If $a_{pq} \neq 0$ and $a_{pp} = a_{qq}$, we put $t = 1$ (corresponding to $\varphi = \pi/4$). Finally, if $a_{pq} \neq 0$ and $a_{pp} \neq a_{qq}$, we solve the quadratic equation (5.8); there will be two distinct real roots, so we define t as the one that is smaller in absolute value. Having selected t, we then use the relation $\sec^2\varphi = 1 + \tan^2\varphi$ to calculate c by $c = 1/(1+t^2)^{1/2}$, and then s from $s = ct$.

Definition 5.3 (The classical Jacobi method) *Let $A \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ and define $A^{(0)} = A$. Given $k \geq 0$ and $A^{(k)} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$, the basic step of Jacobi's method computes $A^{(k+1)} \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ by first locating the largest in absolute value off-diagonal element $(A^{(k)})_{pq} = a^{(k)}_{pq}$ of the matrix $A^{(k)}$, and then setting $A^{(k+1)} = R^{(pq)}(\varphi_k)^{\mathrm{T}} A^{(k)} R^{(pq)}(\varphi_k)$ with φ_k chosen so as to reduce $(A^{(k+1)})_{pq}$ to zero. This process is then repeated until all the off-diagonal elements are smaller than a given positive tolerance ε.*

In order to show that as $k \to \infty$ the sequence of matrices $(A^{(k)})$ generated by successive steps of the classical Jacobi method converges to a diagonal matrix (whose diagonal entries are the eigenvalues of the original matrix A), we need the following result.

Lemma 5.1 *The sum of squares of the elements of a symmetric matrix is invariant under an orthogonal transformation: that is, if $A \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$*

[1] For future reference, note that a simple calculation based on (5.5) and (5.6) gives

$$b_{ii} - a_{ii} = \begin{cases} 0 & \text{if} \quad i \neq p, q\,, \\ -a_{pq}\tan\varphi & \text{if} \quad i = p\,, \\ a_{pq}\tan\varphi & \text{if} \quad i = q\,. \end{cases}$$

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

and $B = R^{\mathrm{T}} A\, R$ *where* $R \in \mathbb{R}^{n\times n}$ *is an orthogonal matrix, then*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2\,. \tag{5.9}$$

The quantity

$$\|A\|_F = \left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2}$$

is called the **Frobenius norm**[1] of $A \in \mathbb{R}^{n\times n}$. The Frobenius norm of $A \in \mathbb{R}^{n\times n}$ is the 2-norm of A, with A regarded as an element of a linear space of dimension n^2 over the field of real numbers; however, it is *not* a subordinate norm in the sense of Definition 2.10. In particular, the Frobenius norm on $\mathbb{R}^{n\times n}$ is not subordinate to the 2-norm on $\mathbb{R}^n$.

Now, one can express (5.9) equivalently by saying that the Frobenius norm of a symmetric matrix A is invariant under an orthogonal transformation: $\|R^{\mathrm{T}} AR\|_F = \|A\|_F$.

Proof of lemma The sum of squares of the elements of A is the same as the trace of A^2, for

$$\mathrm{Trace}(A^2) = \sum_{i=1}^{n}(A^2)_{ii} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}a_{ji} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2\,, \tag{5.10}$$

since A is symmetric. Analogously, as $B = R^{\mathrm{T}} AR$ is symmetric, we have that

$$\mathrm{Trace}(B^2) = \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}^2\,.$$

Thus, it remains to show that $\mathrm{Trace}(B^2) = \mathrm{Trace}(A^2)$. Now,

$$B^2 = (R^{\mathrm{T}} A\, R)(R^{\mathrm{T}} A\, R) = R^{\mathrm{T}} A^2 R\,, \tag{5.11}$$

since R is orthogonal. Hence B^2 is an orthogonal transformation of A^2 which, by virtue of Theorem 5.1 (vi), means that B^2 and A^2 have the same eigenvalues, and therefore the same trace, since the trace is the sum of the eigenvalues (see Theorem 5.1 (viii)). □

[1] Ferdinand Georg Frobenius (26 October 1849, Berlin-Charlottenburg, Prussia, Germany – 3 August 1917, Berlin, Germany), contributed to the theory of analytic functions, representation theory of groups, differential equation theory and the theory of elliptic functions.

Now we are ready to embark on the convergence analysis of the classical Jacobi method.

Theorem 5.3 *Suppose that $A \in \mathbb{R}^{n\times n}_{\text{sym}}$, $n \geq 2$. In the classical Jacobi method the off-diagonal entries in the sequence of matrices $(A^{(k)})$, generated from $A^{(0)} = A$ according to Definition 5.3, converge to 0 in the sense that*

$$\lim_{k\to\infty} \sum_{\substack{i,j=1 \\ i\neq j}}^{n} [(A^{(k)})_{ij}]^2 = 0\,. \tag{5.12}$$

Furthermore,

$$\lim_{k\to\infty} \sum_{i=1}^{n} [(A^{(k)})_{ii}]^2 = \text{Trace}(A^2)\,. \tag{5.13}$$

Proof Let a_{pq} be the off-diagonal element of A with largest absolute value, and let $B = (R^{(pq)}(\varphi))^{\mathrm{T}} A\, R^{(pq)}(\varphi)$, where φ is defined by (5.7). Then, letting $c = \cos\varphi$ and $s = \sin\varphi$, we have that

$$\begin{pmatrix} b_{pp} & b_{pq} \\ b_{qp} & b_{qq} \end{pmatrix} = \begin{pmatrix} c & s \\ -s & c \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{pmatrix} \begin{pmatrix} c & s \\ -s & c \end{pmatrix},$$

and Lemma 5.1 implies that

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

Writing

$$S(A) = \sum_{i,j=1}^{n} a_{ij}^2\,, \quad D(A) = \sum_{i=1}^{n} a_{ii}^2\,, \quad L(A) = \sum_{\substack{i,j=1 \\ i\neq j}}^{n} a_{ij}^2\,,$$

it follows that $S(A) = D(A) + L(A)$. Now $S(B) = S(A)$ by Lemma 5.1, and so $D(B) + L(B) = D(A) + L(A)$. The diagonal entries of B are the same as those of A, except the ones in rows p and q, $1 \leq p < q \leq n$. Further, as $b_{pq} = 0$, it follows that $b_{pp}^2 + b_{qq}^2 = a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2$. Therefore,

$$D(B) = D(A) + 2a_{pq}^2\,.$$

Consequently,

$$L(B) = L(A) - 2a_{pq}^2\,.$$

Now a_{pq} is the largest off-diagonal element of A; hence $L(A) \leq N a_{pq}^2$ where $N = n(n-1)$ is the number of off-diagonal elements, and therefore

$$L(B) \leq (1 - 2/N) L(A)\,. \tag{5.14}$$

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

On writing $A^{(0)} = A$, $A^{(1)} = B$, and generating subsequent members of the sequence $(A^{(k)})$ in a similar manner, as indicated in the algorithm in Definition 5.3, we deduce from (5.14) that

$$0 \leq L(A^{(k)}) \leq (1 - 2/N)^k L(A)\,, \qquad k = 1, 2, 3, \ldots\,, \tag{5.15}$$

where $N \geq 2$. Thus we conclude that $\lim_{k\to\infty} L(A^{(k)}) = 0$.

Now, (5.13) follows from (5.10) and (5.12) on noting that

$$\mathrm{Trace}(A^2) = S(A) = S(A^{(k)}) = D(A^{(k)}) + L(A^{(k)}) \qquad \forall\, k \geq 0\,,$$

and passing to the limit $k \to \infty$: $\mathrm{Trace}(A^2) = \lim_{k\to\infty} D(A^{(k)})$. □

According to Theorem 5.1 (viii) the trace of A^2 is the sum of the eigenvalues of A^2, and the eigenvalues of A^2 are the squares of the eigenvalues of A. Thus, we have shown that the sum of the squares of the diagonal elements in the sequence of matrices $(A^{(k)})$ generated by the classical Jacobi method converges to the sum of the squares of the eigenvalues of A. More work is required to show that for each $i = 1, 2, \ldots, n$ the sequence of diagonal elements $(a_{ii}^{(k)})$ converges to an eigenvalue of A as $k \to \infty$. We shall further discuss this question in the final paragraphs of Section 5.4. First, however, we describe another variant of Jacobi's method.

Definition 5.4 (The serial Jacobi method) *This version of Jacobi's method proceeds in a systematic order, using transformations $R^{(pq)}(\varphi)$ to reduce to zero the elements $(1,2)$, $(1,3)$, …, $(1,n)$, $(2,3)$, $(2,4)$, …, $(2,n)$, …, $(n-1,n)$ in this order. The complete step is then repeated iteratively.*

It is not difficult to prove that this method also converges. Both these variants of the Jacobi method converge quite rapidly; the rate of convergence is in practice much faster than is suggested by (5.15), and in fact it can be shown that convergence is ultimately quadratic.

It is time for an example!

Example 5.3 *Let us consider the 5×5 matrix*

$$A = \begin{pmatrix} 4 & 1 & 2 & 1 & 2 \\ 1 & 3 & 0 & -3 & 4 \\ 2 & 0 & 1 & 2 & 2 \\ 1 & -3 & 2 & 4 & 1 \\ 2 & 4 & 2 & 1 & 1 \end{pmatrix}. \tag{5.16}$$

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

The values of $D(A^{(k)})$ and $L(A^{(k)})$ after each iteration of the serial Jacobi method, with $A^{(0)} = A$, are shown in Table 5.1. The off-diagonal elements of the third iterate, $A^{(3)}$, are zero to 10 decimal digits. The diagonal elements of $A^{(3)}$, which give the eigenvalues, are

$$8.094,\ 1.690,\ -0.671,\ 7.170,\ -3.282\,.$$

Note that the eigenvalues do not appear in any particular order.

Table 5.1. *Convergence of the serial Jacobi iteration.*

k	$D(A^{(k)})$	$L(A^{(k)})$
0	43.000	88.00000000
1	126.309	4.69087885
2	130.981	0.01948855
3	131.000	0.00000000

This concludes the discussion about the use of Jacobi's method for computing the eigenvalues of a symmetric matrix A. 'Fine,' you might say, 'but how do we determine the *eigenvectors* of A?'

It turns out that by collecting the information accumulated in the course of the Jacobi iteration, it is fairly easy to calculate the eigenvectors of A. We begin by noting that if M is an orthogonal matrix such that $M^{\mathrm{T}}AM = D$, where D is diagonal, then the diagonal elements of D are the eigenvalues of A, and the columns of M are the corresponding eigenvectors of A.

In the course of the Jacobi iteration (be it classical or serial), we have constructed the plane rotations $R^{(p_j q_j)}(\varphi_j)$, $j = 1, 2, \ldots, k$. Thus, an approximation $M^{(k)}$ to the orthogonal matrix M can be obtained by considering the product of these rotation matrices: initially, we put $M^{(0)} = I$ and then we apply the column transformation $R^{(p_j q_j)}(\varphi_j)$ at each step $j = 1, 2, \ldots, k$. This corresponds to multiplying $M^{(j-1)}$ on the right by $R^{(p_j q_j)}(\varphi_j)$ for $j = 1, 2, \ldots, k$, and leads to the orthogonal matrix

$$M^{(k)} = R^{(p_1 q_1)}(\varphi_1) \ldots R^{(p_k q_k)}(\varphi_k)$$

which represents the required approximation to the orthogonal matrix M. The columns of $M^{(k)}$ will be the desired approximate eigenvectors

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

of A corresponding to the approximate eigenvalues which appear along the diagonal of $A^{(k)}$.

The Jacobi method usually converges in a reasonable number of iterations, and is a satisfactory method for small or moderate-sized matrices. However, there are many problems, particularly in the area of numerical solution of partial differential equations, which give rise to very large matrices that are sparse, with most of the elements being zero. A further consideration is that in many practical situations one does not need to compute all the eigenvalues. It is much more common to require a few of the largest eigenvalues and corresponding eigenvectors, or perhaps a few of the smallest. Jacobi's method is not suitable for such problems, as it always produces all the eigenvalues, and will not preserve the sparse structure of a matrix during the course of the iteration. For example, it is easy to see that if Jacobi's method is applied to a symmetric tridiagonal matrix, then at the end of one sweep all (but two) of the elements of the matrix will in general be nonzero and, although still symmetric, the transformed matrix is no longer tridiagonal. Later on in this chapter we shall consider numerical algorithms for computing selected eigenvalues of a matrix. Thus, as an overture to what will follow, we now outline a 'rough and ready' technique for locating the eigenvalues.

5.4 The Gerschgorin theorems

Gerschgorin's Theorem[1] provides a very simple way of determining a region that contains the eigenvalues of a matrix. It is very general, and does not assume that the matrix is symmetric; in fact we shall allow the elements of a square matrix of order n to be complex and write $A \in \mathbb{C}^{n\times n}$ to express this fact.

Definition 5.5 *Suppose that $n \geq 2$ and $A \in \mathbb{C}^{n\times n}$. The* **Gerschgorin discs** *D_i, $i = 1, 2, \ldots, n$, of the matrix A are defined as the closed circular regions*

$$D_i = \{z \in \mathbb{C}\colon |z - a_{ii}| \leq R_i\} \tag{5.17}$$

in the complex plane, where

$$R_i = \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \tag{5.18}$$

is the radius of D_i.

[1] After S.A. Gerschgorin; see the historical survey of Seiji Fujino and Joachim Fischer, Über S.A. Gerschgorin (1901–1933) [German: About S.A. Gershgorin (1901–1933)], *GAMM Mitt. Ges. Angew. Math. Mech.* **21**, no. 1, 15–19, 1998.

Theorem 5.4 (Gerschgorin's Theorem) *Let $n \geq 2$ and $A \in \mathbb{C}^{n\times n}$. All eigenvalues of the matrix A lie in the region $D = \bigcup_{i=1}^{n} D_i$, where D_i, $i = 1, 2, \ldots, n$, are the Gerschgorin discs of A defined by (5.17), (5.18).*

Proof Suppose that $\lambda \in \mathbb{C}$ and $\boldsymbol{x} \in \mathbb{C}^n \setminus \{\mathbf{0}\}$ are an eigenvalue and the corresponding eigenvector of A, so that

$$\sum_{j=1}^{n} a_{ij}x_j = \lambda x_i\,, \quad i = 1, 2, \ldots, n\,. \tag{5.19}$$

Suppose that x_k, with $k \in \{1, 2, \ldots, n\}$, is the component of $\boldsymbol{x}$ which has largest modulus, or one of those components if more than one have the same modulus. We note in passing that $x_k \neq 0$, given that $\boldsymbol{x} \neq \mathbf{0}$; also,

$$|x_j| \leq |x_k|\,, \quad j = 1, 2, \ldots, n\,. \tag{5.20}$$

This means that

$$\begin{aligned} |\lambda - a_{kk}|\,|x_k| &= |\lambda\, x_k - a_{kk}\, x_k| \\ &= \left|\sum_{j=1}^{n} a_{kj}x_j - a_{kk}x_k\right| \\ &= \left|\sum_{\substack{j=1\\ j\neq k}}^{n} a_{kj}x_j\right| \\ &\leq |x_k|R_k\,, \end{aligned} \tag{5.21}$$

which, on division by $|x_k|$, shows that λ lies in the Gerschgorin disc D_k of radius R_k centred at a_{kk}. Hence, $\lambda \in D = \bigcup_{i=1}^{n} D_i$. □

Theorem 5.5 (Gerschgorin's Second Theorem) *Let $n \geq 2$. Suppose that $1 \leq p \leq n-1$ and that the Gerschgorin discs of the matrix $A \in \mathbb{C}^{n\times n}$ can be divided into two disjoint subsets $D^{(p)}$ and $D^{(q)}$, containing p and $q = n - p$ discs respectively. Then, the union of the discs in $D^{(p)}$ contains p of the eigenvalues, and the union of the discs in $D^{(q)}$ contains $n - p$ eigenvalues. In particular, if one disc is disjoint from all the others, it contains exactly one eigenvalue, and if all the discs are disjoint then each disc contains exactly one eigenvalue.*

Proof We shall use a so-called *homotopy* (or continuation) argument.

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

For $0 \leq \varepsilon \leq 1$, we consider the matrix $B(\varepsilon) = (b_{ij}(\varepsilon)) \in \mathbb{C}^{n\times n}$, where

$$b_{ij}(\varepsilon) = \begin{cases} a_{ii} & \text{if } i = j\,, \\ \varepsilon a_{ij} & \text{if } i \neq j\,. \end{cases} \tag{5.22}$$

Then, $B(1) = A$, and $B(0)$ is the diagonal matrix whose diagonal elements coincide with those of A. Each of the eigenvalues of $B(0)$ is therefore the centre of one of the Gerschgorin discs of A; thus exactly p of the eigenvalues of $B(0)$ lie in the union of the discs in $D^{(p)}$. Now, the eigenvalues of $B(\varepsilon)$ are the zeros of its characteristic polynomial, which is a polynomial whose coefficients are continuous functions of ε; hence the zeros of this polynomial are also continuous functions of ε. Thus as ε increases from 0 to 1 the eigenvalues of $B(\varepsilon)$ move along continuous paths in the complex plane, and at the same time the radii of the Gerschgorin discs increase from 0 to the radii of the Gerschgorin discs of A. Since p of the eigenvalues lie in the union of the discs in $D^{(p)}$ when $\varepsilon = 0$, and these discs are disjoint from all of the discs in $D(q)$, these p eigenvalues must still lie in the union of the discs in $D^{(p)}$ when $\varepsilon = 1$, and the theorem is proved.

The same proof evidently still applies when the discs can be divided into any number of disjoint subsets. □

Example 5.4 *Consider the matrix*

$$A = \begin{pmatrix} 4.00 & 0.20 & -0.10 & 0.10 \\ 0.20 & -1.00 & -0.10 & 0.05 \\ -0.10 & -0.10 & 3.00 & 0.10 \\ 0.10 & 0.05 & 0.10 & -3.00 \end{pmatrix}. \tag{5.23}$$

Figure 5.1 shows, as solid circles, the Gerschgorin discs for this matrix; for instance, one of the discs has centre at 4.00 *and radius* 0.40. *The discs are clearly disjoint, so that each disc contains one eigenvalue of the matrix. The significance of the dotted circles will be explained in our next example.*

Example 5.5 *Let us consider the matrix A defined by (5.23), and then transform it into $B = KAK^{-1}$, where $K \in \mathbb{R}^{4\times 4}$ is the same as the identity matrix except that $k_{22} = \kappa > 0$.*

This transformation has the effect of multiplying the elements in row 2 by κ, and multiplying the elements in column 2 by $1/\kappa$; the diagonal element a_{22} thus remains unaltered. A small value of κ then means that the second disc of B is smaller than the second disc of A, but the other

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

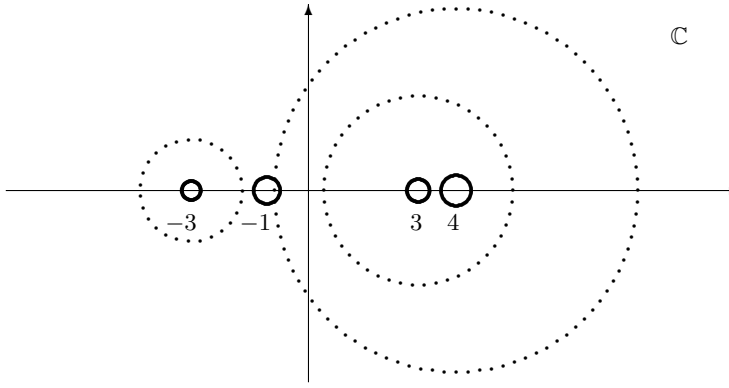

Fig. 5.1. Gerschgorin discs in the complex plane for the matrix A defined in (5.23) (solid circles) and for $B = KAK^{-1}$ (dotted circles). The numbers along the real axis denote the first coordinate of the centre point of each circle (the second coordinate being zero in each case).

discs grow larger. The dotted discs in Figure 5.1 are for the matrix B with $\kappa = 1/23$. For this value the other discs are still just disjoint from the disc centred at -1.00; the disc with centre at 4.00 almost touches the disc with centre at -1.00. The disc with centre -1.00 has radius 0.014, and is too small to be visible in the figure. The eigenvalue in this disc is -1.009 to three decimal digits. The same procedure can be used to reduce the size of each of the discs in turn. ◇

This idea is formalised in the next theorem.

Theorem 5.6 *Let $n \geq 2$, and suppose that in the matrix $A \in \mathbb{C}^{n\times n}$ all the off-diagonal elements are smaller in absolute value than ε, so that $|a_{ij}| < \varepsilon$, for all $i, j \in \{1, 2, \ldots, n\}$ with $i \neq j$. Suppose also that for a particular integer $r \in \{1, 2, \ldots, n\}$ the diagonal element a_{rr} is distant δ from all the other diagonal elements, so that $|a_{rr} - a_{ii}| > \delta$, for all i such that $i \neq r$. Then, provided that*

$$\varepsilon < \frac{\delta}{2(n-1)}\,, \tag{5.24}$$

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

there is an eigenvalue λ of A such that

$$|\lambda - a_{rr}| < 2(n-1)\varepsilon^2/\delta\,. \tag{5.25}$$

Proof We apply the **similarity transformation**

$$A \in \mathbb{C}^{n\times n} \mapsto A' = K\,A\,K^{-1} \in \mathbb{C}^{n\times n}\,,$$

where $K \in \mathbb{R}^{n\times n}$ is the same as the identity matrix, except that the diagonal element in row r is chosen to be $k_{rr} = \kappa > 0$. This has the effect of multiplying the off-diagonal elements of row r by κ, and the element in column r of row i, where $i \neq r$, by $1/\kappa$. The Gerschgorin disc from row r then has centre a_{rr} and radius not exceeding $\kappa(n-1)\varepsilon$, and the disc corresponding to row $i \neq r$ has centre a_{ii} and radius not exceeding $(n-2)\varepsilon + \varepsilon/\kappa$.

We now want to reduce the size of disc r by choosing a small value of κ, while keeping it disjoint from the rest. This is easily done by choosing $\kappa = 2\varepsilon/\delta$. The radius of disc r does not exceed $2(n-1)\varepsilon^2/\delta$, and the radius of disc $i \neq r$ does not exceed $(n-2)\varepsilon + \frac{1}{2}\delta$. The sum of these radii therefore satisfies

$$\begin{aligned} R_r + R_i &\leq 2(n-1)\varepsilon^2/\delta + (n-2)\varepsilon + \tfrac{1}{2}\delta \\ &< \varepsilon + (n-2)\varepsilon + \tfrac{1}{2}\delta \\ &< \delta\,, \end{aligned} \tag{5.26}$$

where we have used the given condition (5.24) twice. As the centres a_{rr} and a_{ii} of these discs are distant more than δ from each other, (5.26) shows that the two discs are disjoint, and the required result is proved. □

Theorem 5.6 is sufficient to show that for a matrix satisfying its hypotheses we can find a Gerschgorin disc whose radius is of order ε^2 provided that ε is sufficiently small. It also indicates that the spacing between the diagonal elements is important.

In particular, Theorem 5.6 applies to the matrix $A^{(k)}$ which results after k iterations of the Jacobi method. If at that stage all the off-diagonal elements have magnitude less than ε then there is one eigenvalue in each of the intervals $[a_{ii}^{(k)} - (n-1)\varepsilon, a_{ii}^{(k)} + (n-1)\varepsilon]$, provided that these intervals are disjoint; this follows from Theorem 5.5. If ε is sufficiently small compared with the distances between the diagonal elements of $A^{(k)}$, Theorem 5.6 may be used to give closer bounds on the eigenvalues.

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

We close this section with some comments on the convergence of the classical Jacobi iteration. According to the Cauchy–Schwarz inequality,

$$\left(\sum_{\substack{i,j=1 \\ i\neq j}}^{n} |a_{ij}^{(k)}| \right)^2 \leq \sum_{\substack{i,j=1 \\ i\neq j}}^{n} 1^2 \sum_{\substack{i,j=1 \\ i\neq j}}^{n} |a_{ij}^{(k)}|^2 = n(n-1) \sum_{i,j=1, i\neq j}^{n} |a_{ij}^{(k)}|^2 .$$

Therefore, also,

$$\left(\max_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}^{(k)}| \right)^2 \leq n(n-1) \sum_{\substack{i,j=1 \\ i\neq j}}^{n} |a_{ij}^{(k)}|^2 ,$$

so (5.12) implies that

$$\lim_{k\to\infty} \max_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}^{(k)}| = 0 .$$

In other words, the radii of the Gerschgorin discs for the matrices in the sequence $(A^{(k)})$ converge to 0 as $k \to \infty$. As $A^{(k)}$ and A have identical eigenvalues for all k, it follows from Theorems 5.4 and 5.5 that the set of limiting diagonal entries $\{\lim_{k\to\infty} a_{11}^{(k)}, \ldots, \lim_{k\to\infty} a_{nn}^{(k)}\}$ delivered by the Jacobi iteration is equal to the set of eigenvalues of A. This holds irrespective of the spacing between the diagonal entries.

5.5 Householder's method

The general method for finding the eigenvalues of a real symmetric matrix begins by applying an orthogonal transformation to reduce it to a tridiagonal matrix. This can be done in a finite number of steps by using Householder matrices.

Definition 5.6 *Given a vector* $\boldsymbol{v} \in \mathbb{R}^n_*$*, the corresponding* **Householder matrix** $H = H(\boldsymbol{v})$ *of order* n *is defined by*

$$H = I - \frac{2}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}} \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}},$$

where I *is the identity matrix of order* n*.*

Clearly, for any vector $\boldsymbol{x} \in \mathbb{R}^n$, we have

$$H\boldsymbol{x} = \boldsymbol{x} - 2\frac{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x}}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}}\boldsymbol{v} ,$$

and hence the vectors $H\boldsymbol{x}$, $\boldsymbol{x}$ and $\boldsymbol{v}$ are coplanar. In particular, if $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{0}$ then $H\boldsymbol{x} = \boldsymbol{x}$, and therefore the $(n-1)$-dimensional

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

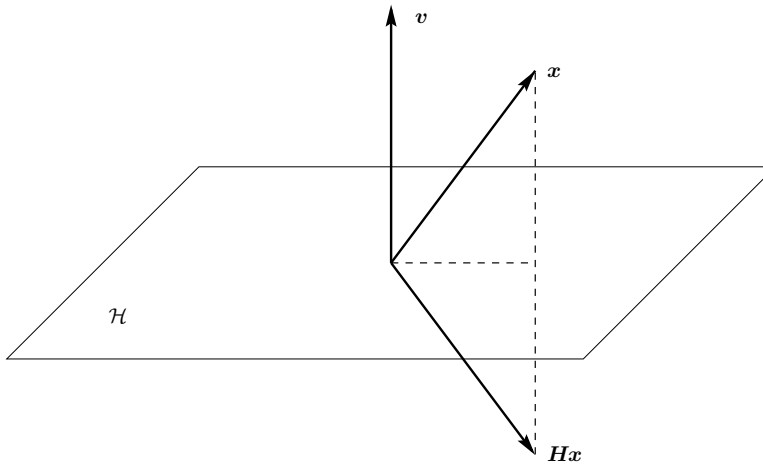

Fig. 5.2. Action of the Householder reflector $H\colon \boldsymbol{x} \mapsto H\boldsymbol{x}$, corresponding to $\boldsymbol{v} \in \mathbb{R}^n_*$, on a vector $\boldsymbol{x} \in \mathbb{R}^n$. $H\boldsymbol{x}$ is the reflection of $\boldsymbol{x}$ in the hyperplane $\mathcal{H}$ perpendicular to $\boldsymbol{v}$.

hyperplane $\mathcal{H}$ consisting of all vectors $\boldsymbol{x}$ that are perpendicular to $\boldsymbol{v}$ in $\mathbb{R}^n$ is invariant under the mapping $\boldsymbol{x} \mapsto H\boldsymbol{x}$. Finally, for any $\boldsymbol{x} \in \mathbb{R}^n$,

$$\boldsymbol{v}^{\mathrm{T}} H\boldsymbol{x} = -\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x}\,.$$

Hence, if the angle between $\boldsymbol{x}$ and $\boldsymbol{v}$ is denoted by φ, then the angle between $\boldsymbol{v}$ and $H\boldsymbol{x}$ is equal to $\pi+\varphi$. We conclude from these observations that the vector $H\boldsymbol{x}$ is the reflection of $\boldsymbol{x}$ in the hyperplane $\mathcal{H}$. For this reason, the mapping $\boldsymbol{x} \mapsto H\boldsymbol{x}$ is frequently referred to as **Householder reflector**, corresponding to the vector $\boldsymbol{v} \in \mathbb{R}^n_*$ (see Figure 5.2).

Lemma 5.2 *Every Householder matrix is symmetric and orthogonal.*

Proof As $I^{\mathrm{T}} = I$, $(\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}})^{\mathrm{T}} = (\boldsymbol{v}^{\mathrm{T}})^{\mathrm{T}}\boldsymbol{v}^{\mathrm{T}} = \boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$, and $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}$ is a (positive real) number, the symmetry of H follows. The orthogonality of H is a consequence of the identity

$$H^{\mathrm{T}}H = HH^{\mathrm{T}} = H^2 = I - \frac{4}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}}\,\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}} + \frac{4}{(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v})^2}(\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}})(\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}) = I\,,$$

since $(\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}})(\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}) = \boldsymbol{v}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v})\boldsymbol{v}^{\mathrm{T}} = (\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v})\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$ by the associativity of matrix multiplication. □

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

Lemma 5.3 *Let $1 \leq k < n$ and suppose that H_k is a $k \times k$ Householder matrix. Then, the matrix $H \in \mathbb{R}^{n\times n}$, written in partitioned form as*

$$H = \begin{pmatrix} I_{n-k} & 0 \\ 0^{\mathrm{T}} & H_k \end{pmatrix}$$

where I_{n-k} is the identity matrix of order $n-k$ and 0 is the $(n-k)\times k$ zero matrix, is also a Householder matrix.

The proof of this lemma is straightforward and is left as an exercise. (See Exercise 1.)

Lemma 5.4 *Given any vector $\boldsymbol{x} \in \mathbb{R}^n_*$, there exists a Householder matrix $H \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ such that all elements of the vector $H\boldsymbol{x}$ are zero, except the first; i.e., $H\boldsymbol{x}$ is a nonzero multiple of $\mathbf{e}_1$, the first column of the identity matrix.*

In geometrical terms this result can be rephrased by saying that for any vector $\boldsymbol{x} \in \mathbb{R}^n_*$ there exists an $(n-1)$-dimensional hyperplane $\mathcal{H}$ passing through the origin in $\mathbb{R}^n$ such that the reflection $H\boldsymbol{x}$ of $\boldsymbol{x}$ in $\mathcal{H}$ is equal to a nonzero multiple of $\mathbf{e}_1$. To find $\mathcal{H}$ it suffices to identify a vector $\boldsymbol{v} \in \mathbb{R}^n_*$ normal to $\mathcal{H}$. Since $\mathcal{H}$ is unaffected by rescaling $\boldsymbol{v}$ (see Definition 5.6), the length of $\boldsymbol{v}$ is immaterial. As noted in the discussion following Definition 5.6, the vectors $H\boldsymbol{x}$, $\boldsymbol{x}$ and $\boldsymbol{v}$ are coplanar. Therefore, we shall seek $\boldsymbol{v} \in \mathbb{R}^n_*$ as a suitable linear combination of $\boldsymbol{x}$ and $\boldsymbol{e}_1$.

Proof of lemma We seek $H = I - [2/(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v})]\,\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$ with $\boldsymbol{v} = \boldsymbol{x} + c\mathbf{e}_1$, where c is a nonzero real number to be determined. Hence,

$$\begin{aligned} \boldsymbol{v}^{\mathrm{T}}\boldsymbol{x} &= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + c\beta\,, \\ \boldsymbol{v}^{\mathrm{T}}\boldsymbol{v} &= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + 2c\beta + c^2\,, \end{aligned}$$

where $\beta = \mathbf{e}_1^{\mathrm{T}}\boldsymbol{x}$ is the first entry of $\boldsymbol{x}$. A simple manipulation then shows that

$$H\boldsymbol{x} = \boldsymbol{x} - \frac{2}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}}\,\boldsymbol{v}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x}) = \frac{(c^2 - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})\boldsymbol{x} - 2c(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + c\beta)\mathbf{e}_1}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + 2c\beta + c^2}\,.$$

Thus, $H\boldsymbol{x}$ will be a multiple of $\mathbf{e}_1$ provided that we choose c so that $c^2 = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$. Also, to avoid division by 0, we need to ensure that $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + 2c\beta + c^2 \neq 0$. To do so, note that $c^2 \geq \beta^2$; therefore

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + 2c\beta + c^2 \geq (\beta + c)^2 \neq 0\,,$$

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

provided that $\beta + c \neq 0$, which can be ensured by selecting the appropriate sign for c, that is, by defining

$$c = \begin{cases} (\text{sign } \beta)\sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}} & \text{when } \beta \neq 0\,, \\ \sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}} & \text{when } \beta = 0\,. \end{cases}$$

With this choice of c, we have $H\boldsymbol{x} = -c\,\mathbf{e}_1$, as required. □

We now show how Householder matrices can be used to reduce a given matrix to tridiagonal form.

Theorem 5.7 *Given that $A \in \mathbb{R}^{n\times n}_{\text{sym}}$ and $n \geq 3$, there exists a matrix $Q_n \in \mathbb{R}^{n\times n}_{\text{sym}}$, a product of $n-2$ Householder matrices $H_{(n,k)} \in \mathbb{R}^{n\times n}_{\text{sym}}$, $k = 2, \ldots, n-1$, given by*

$$Q_n = H_{(n,n-1)}H_{(n,n-2)}\ldots H_{(n,2)}$$

such that $Q_n^{\mathrm{T}}AQ_n = T_n$ is tridiagonal; the matrix Q_n is orthogonal.

Proof The proof of the theorem will proceed by induction. Before embarking on this, we make some preparatory observations which highlight the key ideas in the proof.

Consider the matrix $A \in \mathbb{R}^{n\times n}_{\text{sym}}$, partitioned by its first row and column in the form

$$A = \begin{pmatrix} \alpha & \boldsymbol{b}^{\mathrm{T}} \\ \boldsymbol{b} & C \end{pmatrix},$$

where $\alpha \in \mathbb{R}$, $\boldsymbol{b} \in \mathbb{R}^{n-1}$ and $C \in \mathbb{R}^{(n-1)\times(n-1)}_{\text{sym}}$, and define

$$\mathcal{E}_1^n = \{\boldsymbol{v} \in \mathbb{R}^n\colon \boldsymbol{v} = (\lambda, 0, \ldots, 0)^{\mathrm{T}} \text{ for some } \lambda \in \mathbb{R}\}\,.$$

If $\boldsymbol{b}$ happens to belong to $\mathbb{R}^{n-1}_*$, then, by Lemma 5.4, there exists an $(n-1)\times(n-1)$ Householder matrix H_{n-1} such that each element of $H_{n-1}\boldsymbol{b}$, except the first, is equal to 0. If, on the other hand, $\boldsymbol{b} = \mathbf{0}$, then $H_{n-1}\boldsymbol{b} = \mathbf{0}$, trivially. Either way, $H_{n-1}\boldsymbol{b} \in \mathcal{E}_1^{n-1}$.

Let us extend the Householder matrix $H_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}_{\text{sym}}$, using Lemma 5.3 with $k = n-1$, to a Householder matrix $H_{(n,n-1)} \in \mathbb{R}^{n\times n}_{\text{sym}}$ by defining the $(1,1)$-entry of $H_{(n,n-1)}$ as 1 and choosing the remaining entries in the first row and first column of $H_{(n,n-1)}$ as 0. Then,

$$\begin{aligned} H_{(n,n-1)}^{\mathrm{T}}AH_{(n,n-1)} &= \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & H_{n-1}^{\mathrm{T}} \end{pmatrix}\begin{pmatrix} \alpha & \boldsymbol{b}^{\mathrm{T}} \\ \boldsymbol{b} & C \end{pmatrix}\begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & H_{n-1} \end{pmatrix} \\ &= \begin{pmatrix} \alpha & \boldsymbol{d}^{\mathrm{T}} \\ \boldsymbol{d} & D \end{pmatrix}, \end{aligned} \tag{5.27}$$

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

where

$$\boldsymbol{d} = H_{n-1}^{\mathrm{T}}\boldsymbol{b} = H_{n-1}\boldsymbol{b} \in \mathcal{E}_1^{n-1} \text{ and } D = H_{n-1}^{\mathrm{T}} C\, H_{n-1} \in \mathbb{R}_{\mathrm{sym}}^{(n-1)\times(n-1)}\,.$$

As $\boldsymbol{d} \in \mathcal{E}_1^{n-1}$, the first row and first column of $H_{(n,n-1)}^{\mathrm{T}} A H_{(n,n-1)}$ are of the desired form. It remains to transform the submatrix D to tridiagonal form. This will be achieved by proceeding inductively.

If $n = 3$, then the 3×3 matrix $H_{(n,n-1)}^{\mathrm{T}} A H_{(n,n-1)}$ is automatically tridiagonal since $\boldsymbol{d} \in \mathcal{E}_1^2$, and we complete the proof by taking $Q_3 = H_{(3,2)}$. We note in passing that if $\mathbf{f} \in \mathcal{E}_1^3$, then

$$Q_3^{\mathrm{T}}\mathbf{f} = H_{(3,2)}^{\mathrm{T}}\mathbf{f} = H_{(3,2)}\mathbf{f} \in \mathcal{E}_1^3\,,$$

as the $(1,1)$-entry of $H_{(3,2)}$ is 1 and the remaining entries in its first column are all 0.

Let us suppose that $n \geq 4$ and $A \in \mathbb{R}_{\mathrm{sym}}^{n\times n}$. Our inductive hypothesis is that the statement of the theorem has already been established for any real symmetric matrix of order $n-1$, *i.e.*, $D \in \mathbb{R}_{\mathrm{sym}}^{(n-1)\times(n-1)}$ can be transformed into tridiagonal form:

$$Q_{n-1}^{\mathrm{T}} D Q_{n-1} = T_{n-1},$$

where $Q_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$ is an orthogonal matrix that is a product of $n-3$ Householder matrices, each of size $(n-1)\times(n-1)$ –

$$Q_{n-1} = H_{(n-1,n-2)} \dots H_{(n-1,2)} \; -$$

and $Q_{n-1}^{\mathrm{T}}\mathbf{f} \in \mathcal{E}_1^{n-1}$ for any vector $\mathbf{f} \in \mathcal{E}_1^{n-1}$. This inductive hypothesis has already been verified above for 3×3 real symmetric matrices.

We now extend each of the $(n-1)\times(n-1)$ matrices $H_{(n-1,k)}$, for $k = 2, \dots, n-2$, to $n \times n$ Householder matrices $H_{(n,k)}$, $k = 2, \dots, n-2$, respectively, as in Lemma 5.3, and define

$$Q_n = H_{(n,n-1)} H_{(n,n-2)} \dots H_{(n,2)}\,.$$

Then, by (5.27),

$$\begin{aligned} Q_n^{\mathrm{T}} A Q_n &= \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & Q_{n-1}^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \alpha & \boldsymbol{d}^{\mathrm{T}} \\ \boldsymbol{d} & D \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & Q_{n-1} \end{pmatrix} \\ &= \begin{pmatrix} \alpha & \boldsymbol{d}^{\mathrm{T}} Q_{n-1} \\ Q_{n-1}^{\mathrm{T}}\boldsymbol{d} & T_{n-1} \end{pmatrix}. \end{aligned}$$

As $\boldsymbol{d} \in \mathcal{E}_1^{n-1}$, it follows from our inductive hypothesis that $Q_{n-1}^{\mathrm{T}}\boldsymbol{d}$ also belongs to $\mathcal{E}_1^{n-1}$, and therefore the last matrix is tridiagonal. As Q_n is a product of $n-2$ Householder matrices, each of size $n \times n$ and each

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

orthogonal, $Q_n \in \mathbb{R}^{n\times n}$ is itself orthogonal. Moreover, for any $\mathbf{f} \in \mathcal{E}_1^n$ we have $Q_n^{\mathrm{T}}\mathbf{f} \in \mathcal{E}_1^n$, since the $(1,1)$-entry of Q_n is 1 and the remaining entries in the first column of Q_n are 0. This concludes the inductive step, and completes the proof. □

The recursive transformation of a symmetric matrix to tridiagonal form outlined in the proof of Theorem 5.7 is called **Householder's method**. In implementing this method in practice it is important to carry out the transformations efficiently. Counting the arithmetic operations involved is straightforward but tedious, and shows that the complete reduction requires approximately $\frac{1}{3}n^3$ multiplications, for a moderately large value of n.

Example 5.6 *In order to illustrate Householder's method, we return to the matrix A defined in (5.16). The first stage uses the Householder matrix defined by the vector*

$$\boldsymbol{v} = (0.000,\ 4.162,\ 2.000,\ 1.000,\ 2.000)^{\mathrm{T}}\,. \tag{5.28}$$

The result of the transformation is the matrix

$$\begin{pmatrix} 4.000 & -3.162 & 0.000 & 0.000 & 0.000 \\ -3.162 & 5.300 & 1.232 & -0.332 & 0.284 \\ 0.000 & 1.232 & 1.653 & 3.312 & 0.275 \\ 0.000 & -0.332 & 3.312 & 5.149 & 1.123 \\ 0.000 & 0.284 & 0.275 & 1.123 & -3.102 \end{pmatrix}.$$

The leading element of the matrix is unchanged, and the first row and column have tridiagonal structure.

The second stage uses the Householder matrix with the vector

$$\boldsymbol{v} = (0.000,\ 0.000,\ 2.540,\ -0.332,\ 0.284)^{\mathrm{T}} \tag{5.29}$$

and gives the new matrix

$$\begin{pmatrix} 4.000 & -3.162 & 0.000 & 0.000 & 0.000 \\ -3.162 & 5.300 & -1.308 & 0.000 & 0.000 \\ 0.000 & -1.308 & 0.057 & -2.166 & 0.792 \\ 0.000 & 0.000 & -2.166 & 6.610 & 0.420 \\ 0.000 & 0.000 & 0.792 & 0.420 & -2.967 \end{pmatrix}.$$

This time the leading 2×2 minor is unaltered, and the first two rows and columns have tridiagonal structure.

The final stage uses the Householder matrix with vector

$$\boldsymbol{v} = (0.000,\ 0.000,\ 0.000,\ -4.471,\ 0.792)^{\mathrm{T}} \tag{5.30}$$

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

and gives the tridiagonal matrix

$$\begin{pmatrix} 4.000 & -3.162 & 0.000 & 0.000 & 0.000 \\ -3.162 & 5.300 & -1.308 & 0.000 & 0.000 \\ 0.000 & -1.308 & 0.057 & 2.306 & 0.000 \\ 0.000 & 0.000 & 2.306 & 5.208 & -3.411 \\ 0.000 & 0.000 & 0.000 & -3.411 & -1.565 \end{pmatrix}. \tag{5.31}$$

The numerical values are quoted here to three decimal digits, for simplicity. ◇

Having shown how to transform a symmetric matrix into tridiagonal form, we can now consider the problem of determining the eigenvalues of a tridiagonal matrix.

5.6 Eigenvalues of a tridiagonal matrix

Before developing a numerical algorithm for calculating the eigenvalues and the eigenvectors of a symmetric tridiagonal matrix, let us spend some time exploring the location of the eigenvalues. The main result of this section is the so-called Sturm sequence property,[1] stated in Theorem 5.9, which enables us to specify the number of eigenvalues of a symmetric tridiagonal matrix which exceed a given real number ϑ. The proof of the Sturm sequence property is based on Cauchy's Interlace Theorem which is of independent interest, and proving the latter is our first task.

To simplify the notation we now write the symmetric tridiagonal matrix in the form

$$T = \begin{pmatrix} a_1 & b_2 & & & & & & \\ b_2 & a_2 & b_3 & & & & & \\ & b_3 & a_3 & b_4 & & & & \\ & & \cdots & \cdots & \cdots & & & \\ & & & \cdots & \cdots & \cdots & & \\ & & & & \cdots & \cdots & \cdots & \\ & & & & & b_{n-1} & a_{n-1} & b_n \\ & & & & & & b_n & a_n \end{pmatrix}.$$

[1] Jacques Charles François Sturm (22 September 1803, Geneva, Helvetia (now Switzerland) – 18 December 1855, Paris, France). The results discussed here are based on Sturm's paper 'Mémoire sur la résolution des équations numériques', published in *Mémoires présentés par divers savants étrangers à l'Académie royale des sciences, section Sc. math. phys.*, **6**, 273–318, 1835, concerning the number of roots of a polynomial in an interval. In 1826 Sturm made the first accurate determination of the velocity of sound in water working with the Swiss engineer Daniel Colladon. In 1840 Sturm succeeded Poisson in the chair of mechanics in the Faculté des Sciences in Paris.

The determinants of the successive principal minors of a matrix of this form can easily be calculated by recurrence. Defining $p_r(\lambda)$ to be the determinant of the leading principal minor of order r of $T - \lambda I$, we see that

$$\begin{aligned} p_1(\lambda) &= a_1 - \lambda\,, \\ p_2(\lambda) &= (a_2 - \lambda)(a_1 - \lambda) - b_2^2\,. \end{aligned}$$

Expanding $p_r(\lambda)$ in terms of the elements of the last row, and then in terms of the last column, we obtain the relation

$$p_r(\lambda) = (a_r - \lambda)p_{r-1}(\lambda) - b_r^2 p_{r-2}(\lambda)\,, \qquad r = 2, 3, \ldots, n\,,$$

with the convention that

$$p_0(\lambda) \equiv 1\,.$$

In the rest of this section we shall assume that all the off-diagonal elements b_i are nonzero. For suppose that $b_k = 0$ for some k in the set $\{2, 3, \ldots, n\}$; then, the eigenvalues of the matrix T comprise the eigenvalues of the matrix consisting of the first $k-1$ rows and columns, together with the eigenvalues of the matrix consisting of the last $n-k+1$ rows and columns. These two problems become separated and can be treated independently; if several of the off-diagonal elements are zero, the matrix can be partitioned into a number of smaller matrices which can then be dealt with independently.

Theorem 5.8 (Cauchy's Interlace Theorem) *Let $n \geq 3$. The roots of p_r separate those of p_{r+1}, for $r = 1, 2, \ldots, n-1$; i.e., between two consecutive roots of p_{r+1} there is exactly one root of the polynomial p_r, $r = 1, 2, \ldots, n-1$.*

Proof The proof is by induction. It is trivial to show that the property holds for $r = 1$: the two roots

$$\tfrac{1}{2}\left[a_1 + a_2 \pm \sqrt{(a_1 - a_2)^2 + 4b_2^2}\right]$$

of p_2 are separated by a_1, the only root of the linear polynomial p_1.

Suppose that the statement is true when $r = i - 1$, $2 \leq i \leq n - 1$, so that the roots of p_{i-1} separate those of p_i. On denoting by α and β two consecutive roots of p_i, the inductive hypothesis implies that p_{i-1} has exactly one root between α and β, which means that $p_{i-1}(\alpha)$ and

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

$p_{i-1}(\beta)$ have opposite signs. Now,

$$p_{i+1}(\lambda) = (a_{i+1} - \lambda)p_i(\lambda) - b_{i+1}^2 p_{i-1}(\lambda)\,,$$

so that, as α and β are roots of p_i, it follows that $p_{i+1}(\alpha)$ and $p_{i+1}(\beta)$ also have opposite signs. Hence p_{i+1} has at least one root between α and β. Choosing α and β to be each pair of consecutive roots of p_i in turn we have therefore located $i-1$ roots of p_{i+1}.

Next choose α to be the algebraically smallest root of p_i. It is easy to see that each of the polynomials $p_1, p_2, \ldots, p_n$ tends to $+\infty$ as $\lambda \to -\infty$. By the inductive hypothesis, p_{i-1} has no roots smaller than α, so $p_{i-1}(\alpha)$ is positive; hence from the recurrence relation $p_{i+1}(\alpha)$ is negative, and therefore p_{i+1} must have a root smaller than α. A similar argument shows that p_{i+1} has a root greater than the largest root of p_i, so that we have located all the $i+1$ roots of p_{i+1}. There is exactly one root of p_i between each pair of consecutive roots of p_{i+1}, and the interlacing property follows. □

We have shown in particular that all the roots of each p_r are distinct. Moreover $p_i(\lambda)$ and $p_{i-1}(\lambda)$ cannot both vanish for the same λ, for if this were to happen the recurrence relation would show that this value of λ is a root of p_r for all values of $r \in \{0, 1, \ldots, n\}$; but p_0 evidently never vanishes.

Theorem 5.9 (The Sturm sequence property) *Let us suppose that $\vartheta \in \mathbb{R}$ and consider the sequence $p_i(\vartheta)$, $i = 0, 1, \ldots, n$. The number of agreements in sign between consecutive members of the sequence is the same as the number of eigenvalues of the matrix T which are strictly greater than ϑ.*

Proof Given that $\lambda \in \mathbb{R}$ and $1 \leq j \leq n$, we write $s_j(\lambda)$ for the number of agreements in sign in the sequence

$$p_0(\lambda)\,, p_1(\lambda)\,, \ldots, p_j(\lambda)\,,$$

and $g_j(\lambda)$ for the number of roots of the polynomial p_j which are strictly greater than λ.

It is trivial to see that $s_1(\vartheta) = g_1(\vartheta)$. The proof now proceeds by induction. Let us suppose that $2 \leq k \leq n$ and adopt the inductive hypothesis that $s_{k-1}(\vartheta) = g_{k-1}(\vartheta)$; we shall prove that $s_k(\vartheta) = g_k(\vartheta)$.

Under our hypothesis, either $s_k(\vartheta) = s_{k-1}(\vartheta) + 1$, if $p_k(\vartheta)$ and $p_{k-1}(\vartheta)$ have the same sign, or $s_k(\vartheta) = s_{k-1}(\vartheta)$ if they have opposite sign. Suppose that ϑ lies in the interval between the two consecutive roots α and

β of p_{k-1}. Then, there is exactly one root of p_k between α and β; denote this root by φ. As we saw in the proof of the previous theorem $p_k(\lambda)$ is positive when λ is large and negative, and the sign of $p_k(\lambda)$ is determined by the number of roots of p_k which are less than λ. Hence if $\vartheta < \varphi$ both p_k and p_{k-1} have the same number of roots less than ϑ, so that $p_k(\vartheta)$ and $p_{k-1}(\vartheta)$ have the same sign, and $s_k(\vartheta) = s_{k-1}(\vartheta) + 1$. Also if p_k and p_{k-1} have the same number of roots less than ϑ, then p_k must have one more root which is greater than ϑ; this means that $g_k(\vartheta) = g_{k-1}(\vartheta) + 1$. Hence $s_k(\vartheta) = g_k(\vartheta)$. A similar argument shows that $s_k(\vartheta) = g_k(\vartheta)$ in the alternative situation where $\vartheta > \varphi$. It is also a simple matter to modify the argument slightly for the cases where ϑ is less than the smallest root of p_{k-1}, or greater than the largest root of p_{k-1}, and so the inductive step is complete. □

The theorem and proof do not allow for any of the members of the sequence being zero, in which case the sign becomes undefined. A more careful analysis is tedious but not difficult; it shows that the theorem still holds if we adopt the convention that when $p_j(\vartheta)$ is zero it is given the same sign as $p_{j-1}(\vartheta)$. As we have already seen, two consecutive members of the sequence cannot both be zero.

Our next example will illustrate the application of the Sturm sequence property.

Example 5.7 *Determine the second largest eigenvalue of the matrix*

$$A = \begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & -1 & 2 & 0 \\ 0 & 2 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}. \tag{5.32}$$

If the eigenvalues are λ_j, where $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4$, we wish to find λ_2. Now, it is easy to see from Theorem 5.5 that all the eigenvalues lie in the interval $[-4, 4]$. We take the midpoint of this interval, and evaluate the Sturm sequence with $\vartheta = 0$, giving

$$p_0(0) = 1\,,\quad p_1(0) = 3\,,\quad p_2(0) = -4\,,\quad p_3(0) = -16\,,\quad p_4(0) = -12\,.$$

In this sequence there are three agreements of sign:

$$(1, 3)\,,\quad (-4, -16)\quad \text{and}\quad (-16, -12)\,.$$

Hence $s_4(0) = 3$, and the matrix has three eigenvalues greater than 0; this means that λ_2 must lie in the right-hand half of the interval

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

$[-4, 4]$, that is, in $[0, 4]$. We construct the Sturm sequence for $\vartheta = 2$, the midpoint of the interval, giving

$$p_0(2) = 1\,,\ \ p_1(2) = 1\,,\ \ p_2(2) = -4\,,\ \ p_3(2) = -0\,,\ \ p_4(2) = 4\,.$$

Notice that here $p_3(2)$ is zero, and is given the negative sign to agree with $p_2(2)$. The number of agreements in sign here is two, so two of the eigenvalues are greater than 2, and λ_2 must lie in $[2, 4]$, the right-hand half of the interval $[0, 4]$. For $\vartheta = 3$ we obtain the sequence

$$1\,,\ \ +0\,,\ \ -1\,,\ \ 2\,,\ \ -3\,,$$

with only one agreement of sign, so this time λ_2 must lie in the left-hand half $[2, 3]$ of the interval $[2, 4]$, and we repeat the process, taking $\vartheta = \frac{5}{2}$, the midpoint of $[2, 3]$. This time the sequence is

$$1\,,\ \ \frac{1}{2}\,,\ \ -\frac{11}{4}\,,\ \ \frac{17}{8}\,,\ \ -\frac{7}{16}\,,$$

with one agreement in sign, showing that $\lambda_2 < 2.5$.

The process of bisection can be repeated as many times as required to locate the eigenvalue to a given accuracy. After 13 stages we find that $\lambda_2 = 2.450$ correct to three decimal digits. ◇

This method is very similar to the usual bisection process for finding a solution of $f(x) = 0$, beginning with an interval $[a, b]$ such that $f(a)$ and $f(b)$ have opposite signs. A great advantage of the Sturm sequence method is that it not only determines the eigenvalue, but also indicates which eigenvalue it is. If we used the Jacobi method of Section 5.3 we would have to determine *all* the eigenvalues, sort them into order, and then choose the second largest eigenvalue as λ_2.

The Sturm sequence method will also determine how many eigenvalues of a matrix lie in a given interval (α, β); all that we need is to construct the Sturm sequences $(p_j(\alpha))_{j=0,1,\ldots,n}$ and $(p_j(\beta))_{j=0,1,\ldots,n}$; then, the required number of eigenvalues is $s_n(\alpha) - s_n(\beta)$.

It is very important to calculate the sequence $p_j(\vartheta)$ directly from the recurrence relation. For instance, in Example 5.7, with $\vartheta = 2.445$ we obtain

$$\begin{aligned}
p_0(2.445) &= 1\,,\\
p_1(2.445) &= 3 - 2.445 = 0.555\,,\\
p_2(2.445) &= (-1 - 2.445) \times 0.555 - 1 \times 1 = -2.9120\,,\\
p_3(2.445) &= (1 - 2.445) \times -2.9120 - 4 \times 0.555 = 1.9878\,,\\
p_4(2.445) &= (1 - 2.445) \times 1.9878 - 1 \times -2.9120 = 0.0396\,.
\end{aligned}$$

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

The alternative, to construct explicit forms for the polynomials $p_j(\lambda)$, $j = 0, 1, \ldots, n$, and *then* evaluate $p_j(\vartheta)$ by inserting the value of $\lambda = \vartheta$ into each of the polynomials $p_j(\lambda)$, will lead to the construction of the explicit form of the characteristic polynomial of the matrix, which is $p_n(\lambda)$, and we have already seen that this is affected disastrously by rounding errors. The calculation by direct use of the recurrence relation is perfectly satisfactory.

Example 5.8 *As a second example, we return to the matrix A in (5.16), which has been transformed to the tridiagonal form (5.31), to determine the largest eigenvalue.*

Table 5.2. *Bisection process for the largest eigenvalue. In the table k denotes the iteration number, ϑ_k the kth iterate approximating the unknown eigenvalue λ_1, and $s_4(\vartheta_k)$ signifies the number of sign agreements in the Sturm sequence $p_0(\vartheta_k), \ldots, p_4(\vartheta_k)$.*

k	ϑ_k	$s_4(\vartheta_k)$
1	0.000	3
2	5.463	2
3	8.194	0
4	6.829	2
5	7.511	1
6	7.853	1
7	8.024	1
8	8.109	0
9	8.066	1
10	8.088	1
11	8.098	0
12	8.093	1
13	8.096	0
14	8.094	0
15	8.094	1

Table 5.2 shows the result of the bisection process, using the Sturm sequence. The ∞-norm of the tridiagonal matrix is 10.926, so the process begins with the interval $[-10.926, 10.926]$.[1] The largest eigenvalue

[1] To explain this choice, let us note that if $\lambda \in \mathbb{C}$ is an eigenvalue of $A \in \mathbb{C}^{n\times n}$ and $\boldsymbol{x} \in \mathbb{C}^n \setminus \{\boldsymbol{0}\}$ is the corresponding eigenvector, then $|\lambda|\,\|\boldsymbol{x}\| = \|\lambda\boldsymbol{x}\| = \|A\boldsymbol{x}\| \leq \|A\|\,\|\boldsymbol{x}\|$; *i.e.*, $|\lambda| \leq \|A\|$, for any subordinate matrix norm $\|\cdot\|$ and any eigenvalue λ of A.

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

is 8.094, to three decimal digits, agreeing with the result of Jacobi's method, in Section 5.3. This table also shows how some savings are possible when all the eigenvalues are required. We see from the table that use of $\vartheta = 7.511$ gives 1 agreement in sign, while $\vartheta = 6.829$ gives 2 agreements in sign. The bisection process for the second largest eigenvalue can therefore begin with the interval $[6.829,\ 7.511]$. ◇

The method of bisection may appear rather crude, but it has the great advantage of guaranteed success, and is very little affected by rounding errors. Moreover, the amount of work involved is not large. If we have calculated the squares of the off-diagonal entries, b_r^2, of the matrix T in advance, each computation of all members of the sequence requires about $2n$ multiplications. If the bisection process is continued for 40 stages, the eigenvalue will be determined to about nine significant digits, and if we require to calculate m of the eigenvalues to this accuracy, we shall need about $80mn$ multiplications. If m is a good deal smaller than n, the order of the matrix, this is likely to be a great deal smaller than the work involved in the process of reduction to tridiagonal form, which, as we have seen, is about $\frac{1}{3}n^3$ multiplications. In most practical problems it is the initial Householder reduction to tridiagonal form which accounts for most of the computational work.

5.7 The QR algorithm

In this section we discuss briefly the QR algorithm, an alternative method for determining the eigenvalues of a tridiagonal matrix. In principle it could be applied to a full matrix, but it is more efficient to use the Householder method to reduce the matrix to tridiagonal form first. The basis of the method is the QR factorisation of the matrix which we have already encountered in Chapter 2, in the solution of least squares problems. In contrast with Section 2.9, however, where we were concerned with the solution of least squares problems for rectangular matrices $A \in \mathbb{R}^{m\times n}$, here the focus is on eigenvalue problems for symmetric tridiagonal matrices $A \in \mathbb{R}^{n\times n}$; we shall therefore revisit the derivation of the QR factorisation by adopting a slightly different approach from the one proposed in Section 2.9.

5.7.1 The QR factorisation revisited

Suppose that $n \geq 3$ and $A \in \mathbb{R}^{n\times n}$ is a symmetric tridiagonal matrix. We first show how to construct an orthogonal matrix $Q \in \mathbb{R}^{n\times n}$ and

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

an upper triangular matrix $R \in \mathbb{R}^{n\times n}$ such that $A = QR$; the problem is similar to the LU factorisation used in solving systems of linear equations, but here we have an orthogonal matrix Q instead of a lower triangular matrix L.

We construct the matrix Q as a product of plane rotation matrices $R^{p\,p+1}(\varphi) \in \mathbb{R}^{n\times n}$ (see Definition 5.2), with a suitably chosen φ. In order to explain what is meant here by 'suitably chosen', we note that in the product

$$B = R^{p\,p+1}(\varphi)A \tag{5.33}$$

the element $b_{p+1\,p}$ is easily found to be

$$b_{p+1\,p} = -s\,a_{pp} + c\,a_{p+1\,p}\,,$$

where $s = \sin\varphi$ and $c = \cos\varphi$. We can make $b_{p+1\,p} = 0$ by choosing

$$s = \frac{a_{p+1\,p}}{\rho}\,, \quad c = \frac{a_{pp}}{\rho}\,, \quad \rho = (a_{pp}^2 + a_{p+1\,p}^2)^{1/2}\,. \tag{5.34}$$

We note in passing that

$$\begin{aligned} b_{pp} &= ca_{pp} + sa_{p+1\,p}\,, \\ b_{p\,p+1} &= ca_{p\,p+1} + sa_{p+1\,p+1}\,, \\ b_{p+1\,p+1} &= -sa_{p\,p+1} + ca_{p+1\,p+1}\,. \end{aligned}$$

The remaining elements of B are the same as those of A.

To summarise the important points, upon multiplying the symmetric tridiagonal matrix A on the left by $R^{p\,p+1}(\varphi)$, where $c = \cos\varphi$ and $s = \sin\varphi$ in $R^{p\,p+1}(\varphi)$ are chosen as indicated in (5.34), we obtain a tridiagonal matrix $B = (b_{ij}) \in \mathbb{R}^{n\times n}$ such that $b_{p+1\,p} = 0$.

After this brief preparation, we embark on the description of the QR factorisation. Let us suppose that we successively multiply A on the left by the $n-1$ plane rotation matrices,

$$Q_1 = R^{12}(\varphi_1)\,, \quad Q_2 = R^{23}(\varphi_2)\,, \quad \ldots\,, \quad Q_{n-1} = R^{n-1\,n}(\varphi_{n-1})\,,$$

with $\varphi_1, \varphi_2, \ldots, \varphi_{n-1}$ selected according to (5.34); more precisely,

> for $p = 1, 2, \ldots, n-1$,
> φ_p is chosen so as to set the $(p+1, p)$-entry of $Q_p \ldots Q_1 A$ to zero.

Given that the elements below the diagonal of the matrix

$$Q_{p-1} \ldots Q_1 A\,, \qquad 2 \le p \le n-1\,,$$

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

which are already equal to zero, remain zero upon multiplication by the next rotation matrix Q_p in the sequence, we deduce that, after successive multiplications of A on the left by $Q_1, Q_2, \ldots, Q_{n-1}$, the matrix

$$Q_{n-1}\, Q_{n-2} \ldots Q_1\, A = R\,, \tag{5.35}$$

is upper triangular. In fact, since A is tridiagonal, R is tridiagonal and upper triangular; consequently, R is **bidiagonal** in the sense that $R_{ij} = 0$ if $i \neq j, j-1$.

As the matrices $Q_p = R^{p\,p+1}(\varphi_p)$, $p = 1, 2, \ldots, n-1$, are orthogonal, and therefore $Q_p^{\mathrm{T}} Q_p = I$, on multiplying (5.35) on the left by $Q_1^{\mathrm{T}}\, Q_2^{\mathrm{T}} \ldots Q_{n-1}^{\mathrm{T}}$, we find that

$$A = Q\, R\,,$$

where

$$Q = Q_1^{\mathrm{T}}\, Q_2^{\mathrm{T}} \ldots Q_{n-1}^{\mathrm{T}}$$

is an orthogonal matrix (as it is a product of orthogonal matrices). The next subsection describes the QR algorithm, based on the QR factorisation, for the numerical solution of the eigenvalue problem (5.1) where the matrix $A \in \mathbb{R}^{n\times n}$ is symmetric and tridiagonal.

5.7.2 The definition of the QR algorithm

Suppose that $A \in \mathbb{R}^{n\times n}$ is symmetric and tridiagonal. The QR algorithm defines a sequence of symmetric tridiagonal matrices $A^{(k)} \in \mathbb{R}^{n\times n}$, $k = 0, 1, 2, \ldots$, starting with $A^{(0)} = A$, as follows.

Suppose that $k \geq 0$. The kth step of the QR algorithm takes the symmetric tridiagonal matrix $A^{(k)}$ and chooses a **shift** $\mu_k \in \mathbb{R}$ (the choice of μ_k will be discussed below), then forming the QR factorisation

$$A^{(k)} - \mu_k\, I = Q^{(k)}\, R^{(k)}\,.$$

We then multiply $Q^{(k)}$ and $R^{(k)}$ in the reverse order, and construct the new matrix $A^{(k+1)}$ defined by

$$A^{(k+1)} = R^{(k)}\, Q^{(k)} + \mu_k\, I\,.$$

Recalling that the matrix $Q^{(k)}$ is orthogonal, it is a simple matter to see that $A^{(k+1)} = Q^{(k)T} A^{(k)} Q^{(k)}$, so that $A^{(k+1)}$ and $A^{(k)}$ have the same eigenvalues. As $A^{(0)} = A$, all matrices in the sequence $(A^{(k)})$ have the same eigenvalues as A itself. It is also easy to show that each of the matrices $A^{(k)}$ is symmetric and tridiagonal. (See Exercise 7.)

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

The choice of the shift parameter μ_k is very important; if correctly chosen the sequence of matrices $A^{(k)}$ converges very rapidly to a matrix in which one of the off-diagonal elements is zero. If this element is in the first or last row, we have thereby identified one of the eigenvalues; if it is one of the intermediate elements, we can split the matrix into two separate matrices of lower order. In either case we can repeat the iterative process with smaller matrices, until all the eigenvalues are found.

The usual simple choice of the shift parameter in the kth step is

$$\mu_k = a_{nn}^{(k)},$$

the last diagonal element of the matrix $A^{(k)}$. In general, after a few steps of the iteration the element at position $(n, n-1)$ will become negligibly small. One of the eigenvalues of the resulting matrix is then the last diagonal element, and we continue the process with the matrix of order $n-1$ obtained by removing the last row and column. There are special circumstances where this choice of shift is unsatisfactory, and other situations where another choice is more efficient, but we shall not discuss the details any further. The proof of the convergence of this method is long and technical; details will be found in the books cited in the Notes at the end of the chapter.

The method does not determine the eigenvalues in any particular order, so if we require only a small number of the largest eigenvalues, for example, the Sturm sequence method is preferable. The usual recommendation is that the QR algorithm should be used on a matrix of order n if more than about $\frac{1}{4}n$ of the eigenvalues are required.

Example 5.9 *We apply the QR algorithm to the tridiagonal matrix (5.31).*

After one step of the iteration the matrix $A^{(1)} = R^{(0)}\,Q^{(0)} + \mu_0\,I$, with $\mu_0 = a_{55}^{(0)} = a_{55}$, is

$$A^{(1)} = \begin{pmatrix} 7.034 & -2.271 & 0 & 0 & 0 \\ -2.271 & 2.707 & -0.744 & 0 & 0 \\ 0 & -0.744 & 5.804 & 3.202 & 0 \\ 0 & 0 & 3.202 & -0.464 & 1.419 \\ 0 & 0 & 0 & 1.419 & -2.082 \end{pmatrix}.$$

In successive iterations $k = 1, 2, 3, 4, 5$, the element $a_{54}^{(k)}$ has the values 1.419, -1.262, 0.965, -0.223, 0.002, and after the next iteration $a_{54}^{(6)}$

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

vanishes to 10 decimal digits. The element $a_{55}^{(6)}$ is -3.282, which is therefore an eigenvalue.

We then remove the last row and column, and continue the process on the resulting 4×4 matrix. After just one iteration the element at position $(4,3)$ vanishes to 7 decimal digits, giving the eigenvalue -0.671. We remove the last row and column and continue with the resulting 3×3 matrix. After one iteration of the resulting 3×3 matrix the element at position $(3,2)$ is 0.0005, and another iteration gives the accurate eigenvalue 1.690. We are now left with a 2×2 matrix, and the calculation of the last two eigenvalues is trivial. The number of iterations required to isolate each eigenvalue reduces as the algorithm reduces the size of the matrix; this sort of behaviour is typical.

The numerical values agree with those obtained by Jacobi's method, and the bisection method. ◇

5.8 Inverse iteration for the eigenvectors

We saw in Section 5.3 that Jacobi's method can also, if required, produce the eigenvectors of the matrix, but the use of Householder's algorithm, in conjunction with the Sturm sequence method or the QR algorithm, only gives the eigenvalues. Suppose that $A \in \mathbb{R}^{n \times n}$ is a symmetric matrix, and assume that we have a good approximation $\vartheta \in \mathbb{R}$ to the required eigenvalue $\lambda \in \mathbb{R}$ of A, and some approximation $\boldsymbol{v}^{(0)} \in \mathbb{R}^n_*$, $\|\boldsymbol{v}^{(0)}\|_2 = 1$, to the associated eigenvector $\boldsymbol{v} \in \mathbb{R}^n_*$, $\|\boldsymbol{v}\|_2 = 1$. It is implicitly assumed that $\vartheta \neq \lambda$ and that ϑ is not an eigenvalue of A, so that the matrix $A - \vartheta I$ is nonsingular. The method of **inverse iteration** defines the sequence of vectors $\boldsymbol{v}^{(k)}$, $k = 0, 1, \ldots$, as follows: given $\boldsymbol{v}^{(k)} \in \mathbb{R}^n_*$, find $\boldsymbol{w}^{(k)} \in \mathbb{R}^n_*$ and then $\boldsymbol{v}^{(k+1)} \in \mathbb{R}^n_*$ from

$$\begin{aligned} (A - \vartheta I)\boldsymbol{w}^{(k)} &= \boldsymbol{v}^{(k)}\,, \\ \boldsymbol{v}^{(k+1)} &= c_k \boldsymbol{w}^{(k)}\,, \end{aligned} \tag{5.36}$$

where $c_k = 1/\sqrt{\boldsymbol{w}^{(k)\mathrm{T}}\boldsymbol{w}^{(k)}} = 1/\|\boldsymbol{w}^{(k)}\|_2$. Hence, we conclude that $\|\boldsymbol{v}^{(k)}\|_2 = 1$, $k = 0, 1, 2, \ldots$.

Theorem 5.10 *Suppose that $A \in \mathbb{R}^{n \times n}_{\mathrm{sym}}$. The sequence of vectors $(\boldsymbol{v}^{(k)})$ in $\mathbb{R}^n_*$ defined in the process of inverse iteration (5.36) converges to the normalised eigenvector $\boldsymbol{v} \in \mathbb{R}^n_*$ corresponding to the eigenvalue $\lambda \in \mathbb{R}$ which is closest to $\vartheta \in \mathbb{R}$, provided that λ is a simple eigenvalue and the initial vector $\boldsymbol{v}^{(0)} \in \mathbb{R}^n_*$ is not orthogonal to the vector $\boldsymbol{v}$.*

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

Proof According to Theorem 5.1 (vii), the vector $\boldsymbol{v}^{(0)}$ can be expressed as a linear combination of the (ortho)normalised eigenvectors $\boldsymbol{x}^{(j)}$ in $\mathbb{R}^n_*$, $j = 1, 2, \ldots, n$, of the matrix A in the form

$$\boldsymbol{v}^{(0)} = \sum_{j=1}^{n} \alpha_j \boldsymbol{x}^{(j)}\,, \quad \alpha_j = \boldsymbol{v}^{(0)\mathrm{T}} \boldsymbol{x}^{(j)}. \tag{5.37}$$

Let $\lambda_s \in \mathbb{R}$ denote the eigenvalue of A which is closest to $\vartheta \in \mathbb{R}$. We shall prove that the sequence $(\boldsymbol{v}^{(k)})$ converges, as $k \to \infty$, to the eigenvector $\boldsymbol{v} = \boldsymbol{x}^{(s)} \in \mathbb{R}^n_*$ associated with λ_s, provided that $\alpha_s = \boldsymbol{v}^{(0)\mathrm{T}} \boldsymbol{x}^{(s)} \neq 0$. On expanding

$$\boldsymbol{w}^{(0)} = \sum_{j=1}^{n} \beta_j \boldsymbol{x}^{(j)}\,,$$

inserting this expansion into the first line of (5.36) with $k = 0$ and comparing the resulting left-hand side with the expansion (5.37) of $\boldsymbol{v}^{(0)}$ on the right, we find that $(\lambda_j - \vartheta)\beta_j = \alpha_j$. Our hypothesis that $\alpha_s \neq 0$ implies that $\lambda_s \neq \vartheta$. Further, as λ_s is the eigenvalue closest to ϑ, it then follows that $\lambda_j - \vartheta \neq 0$ for all $j \in \{1, 2, \ldots, n\}$. Hence,

$$\boldsymbol{v}^{(1)} = c_0 \boldsymbol{w}^{(0)} = c_0 \sum_{j=1}^{n} \frac{\alpha_j}{\lambda_j - \vartheta} \boldsymbol{x}^{(j)}\,.$$

Repeating this argument for $k = 1, 2, \ldots, m-1$ gives

$$\boldsymbol{v}^{(m)} = c_{m-1} \ldots c_0 \sum_{j=1}^{n} \frac{\alpha_j}{(\lambda_j - \vartheta)^m} \boldsymbol{x}^{(j)}\,. \tag{5.38}$$

Now $\boldsymbol{v}^{(m)\mathrm{T}} \boldsymbol{v}^{(m)} = 1$, and therefore,

$$c_{m-1} \ldots c_0 = \left[\sum_{j=1}^{n} \frac{\alpha_j^2}{(\lambda_j - \vartheta)^{2m}} \right]^{-1/2}. \tag{5.39}$$

Substituting (5.39) into (5.38), we obtain

$$\boldsymbol{v}^{(m)} = \frac{\sum_{j=1}^{n} \frac{\alpha_j}{(\lambda_j - \vartheta)^m} \boldsymbol{x}^{(j)}}{\left[\sum_{j=1}^{n} \frac{\alpha_j^2}{(\lambda_j - \vartheta)^{2m}} \right]^{1/2}} = \frac{\boldsymbol{x}_s + \sum_{j \neq s} \left(\frac{\alpha_j}{\alpha_s} \right) \left(\frac{\lambda_s - \vartheta}{\lambda_j - \vartheta} \right)^m \boldsymbol{x}^{(j)}}{\left[1 + \sum_{j \neq s} \left(\frac{\alpha_j}{\alpha_s} \right)^2 \left(\frac{\lambda_s - \vartheta}{\lambda_j - \vartheta} \right)^{2m} \right]^{1/2}}.$$

Since

$$\left| \frac{\lambda_s - \vartheta}{\lambda_j - \vartheta} \right| < 1 \qquad \forall\, j \in \{1, 2, \ldots, n\} \setminus \{s\}\,,$$

we find that $\lim_{m \to \infty} \boldsymbol{v}^{(m)} = \boldsymbol{x}_s = \boldsymbol{v}$; that completes the proof. □

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

If the estimate ϑ is within rounding error of λ_s and the eigenvalues are well spaced, the convergence of the sequence $(\boldsymbol{v}^{(k)})$ will be extremely rapid: usually a couple of iterations will be sufficient.

The proof of Theorem 5.10 breaks down if $\alpha_s = 0$, *i.e.*, when the initial vector $\boldsymbol{v}^{(0)}$ is exactly orthogonal to the required eigenvector. However, this does not mean that the iteration (5.36) will also break down; for the effect of rounding error will almost always introduce a small multiple of the vector $\boldsymbol{x}^{(s)}$ into the expansion of $\boldsymbol{v}^{(0)}$ in terms of the $\boldsymbol{x}^{(j)}$ with $j = 1, 2, \ldots, n$, and the required eigenvector will then be obtained in a small number of iterations. This is a useful property of the method, since in practice it is not possible to check whether or not $\boldsymbol{v}^{(0)}$ is orthogonal to $\boldsymbol{v}$, given that the eigenvector $\boldsymbol{v}$ is unknown.

There will also be a problem if there is a multiple eigenvalue, or two eigenvalues are very close together: in the first case $|\lambda_s - \vartheta|/|\lambda_j - \vartheta| = 1$ for some $j \neq s$, and the proof of Theorem 5.10 breaks down; in the second case $|\lambda_s - \vartheta|/|\lambda_j - \vartheta| \approx 1$ for some $j \neq s$, leading to very slow convergence.

The computation of $\boldsymbol{w}^{(k)}$ from (5.36) requires the solution of a system of linear equations whose matrix is $A - \vartheta I$. This matrix will usually be nearly singular – in fact, our objective in choosing ϑ was to make $A - \vartheta I$ exactly singular. In general the solution of such a system is extremely dangerous, because of the effect of rounding errors; in this case, however, the effect of rounding error will be to introduce a multiple of the dominant eigenvector, and this is exactly what is required. An analysis of the effect of rounding errors will confirm this fact, but would take too long here.[1]

There are two ways in which we can implement the inverse iteration process. One obvious possibility would be to use the original matrix $A \in \mathbb{R}^{n\times n}$, as implied in (5.36). An alternative is to replace A in this equation by the tridiagonal matrix $T \in \mathbb{R}^{n\times n}$ supplied by Householder's method. The calculation is then very much quicker, but produces the eigenvector of T; to obtain the corresponding eigenvector of A we must then apply to this vector the sequence of Householder transformations which were used in the original reduction to tridiagonal form. It is easy to show that this is the most efficient method.

[1] For further details, we refer to Sec. 4.3 in B. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980, and Section 7.6.1 in G.H. Golub and C.F. Van Loan, *Matrix Computations*, Third Edition, Johns Hopkins University Press, Baltimore, 1996.

Inverse iteration with the original matrix $A \in \mathbb{R}^{n\times n}$ requires the LU decomposition of A, followed by one or more forward and backsubstitution operations. As we saw in Section 2.6, the LU decomposition requires approximately $\frac{1}{3}n^3$ multiplications. The same process with the tridiagonal matrix T, using the Thomas algorithm, involves only a small multiple of n multiplications.

Having found an eigenvector of the tridiagonal matrix $T \in \mathbb{R}^{n\times n}$, so that

$$T\boldsymbol{v} = \lambda \boldsymbol{v}\,,$$

we use the fact that $Q^{\mathrm{T}}AQ = T$ to write

$$AQ\boldsymbol{v} = \lambda Q\boldsymbol{v}\,,$$

so that the vector $Q\boldsymbol{v}$ is an eigenvector of A. Using Theorem 5.7, this means that the required eigenvector of A is

$$H_{(n,n-1)} \dots H_{(n,2)}\boldsymbol{v}\,,$$

where the matrices $H_{(n,j)} \in \mathbb{R}^{n\times n}$, $j = 2, \dots, n-1$, are Householder matrices. To multiply a vector $\boldsymbol{x}$ by a Householder matrix $H = H(\boldsymbol{u})$ we write

$$H\boldsymbol{x} = (I - \alpha \boldsymbol{u}\boldsymbol{u}^{\mathrm{T}})\boldsymbol{x} = \boldsymbol{x} - \alpha(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x})\boldsymbol{u}\,.$$

Assuming that $\alpha = 2/(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u})$ is known, this requires the calculation of the scalar product $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}$, and then subtracting a multiple of the vector $\boldsymbol{u}$ from the vector $\boldsymbol{x}$. This evidently involves $2n$ multiplications. Hence the calculation of $Q\boldsymbol{v}$ requires only $2n(n-2)$ multiplications, and the work involved in the whole process is proportional to n^2, instead of n^3. In fact the total is less than $2n(n-2)$, since a more careful count can use the fact that many of the elements in the vector $\boldsymbol{u}$ are known to be zero.

Example 5.10 *Returning to the tridiagonal matrix (5.31), the QR algorithm has given an accurate eigenvalue which is* 8.094 *to three decimal digits. Beginning the inverse iteration (5.36) with a randomly chosen vector* $\boldsymbol{v}^{(0)} \in \mathbb{R}^5_*$*, we find that*

$$\boldsymbol{v}^{(1)} = (-0.0249, -0.0574, -0.3164, 0.4256, 0.8455)^{\mathrm{T}}.$$

Successive iterations make no change in this vector, as might be expected, since the eigenvalue used was accurate to within rounding error.

This is therefore the eigenvector of the tridiagonal matrix (5.31), to

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

within rounding error. To obtain the eigenvector of the original matrix (5.16) we multiply $\boldsymbol{v}^{(1)}$ *in succession by the three Householder matrices defined by the vectors (5.30), (5.29) and (5.28). The result is the eigenvector*

$$\boldsymbol{v} = (-0.0249,\ -0.5952,\ -0.1920,\ -0.2885,\ 0.7246)^{\mathrm{T}}.$$

Using this vector and the accurately calculated eigenvalue, we can check the result, and find that the elements of $A\boldsymbol{v} - \lambda\boldsymbol{v}$ *are of the same order as rounding error.*

5.9 The Rayleigh quotient

In this section we develop a simple technique based on the concept of Rayleigh quotient,[1] for obtaining an accurate approximation to an eigenvalue of a symmetric matrix when a reasonably accurate approximation to the associated eigenvector is already available.

Definition 5.7 *Given a vector* $\boldsymbol{x} \in \mathbb{R}^n_*$ *and a matrix* $A \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$*, the associated* **Rayleigh quotient** $R(\boldsymbol{x})$ *is defined as the real number*

$$R(\boldsymbol{x}) = \frac{\boldsymbol{x}^{\mathrm{T}} A\, \boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}\,. \tag{5.40}$$

Clearly, if $\boldsymbol{x} \in \mathbb{R}^n_*$ is an eigenvector corresponding to an eigenvalue $\lambda \in \mathbb{R}$ of a matrix $A \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$, then $R(\boldsymbol{x}) = \lambda$. More generally, if $\boldsymbol{x}$ is any nonzero vector in $\mathbb{R}^n$, then a number of further properties of the Rayleigh quotient are immediate deductions from the expansion of $\boldsymbol{x}$ in terms of the eigenvectors of A.

Theorem 5.11 *Suppose that the matrix* $A \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ *has the eigenvalues* $\lambda_j \in \mathbb{R}$*,* $j = 1, 2, \ldots, n$*, and the corresponding normalised eigenvectors* $\boldsymbol{x}^{(j)} \in \mathbb{R}^n_*$*,* $j = 1, 2, \ldots, n$*. If the vector* $\boldsymbol{x}$ *is expressed in terms of the*

[1] John William Strutt, Lord Rayleigh (12 November 1842, Langford Grove (near Maldon), Essex, England – 30 June 1919, Terling Place, Witham, Essex, England). In 1879 Rayleigh wrote a paper on travelling waves which set the foundation for the modern theory of solitons. His theory of scattering (1871) was the first correct explanation of why the sky is blue: the intensity of light scattered from small particles is inversely proportional to the fourth power of the wavelength; for this reason, the intensity of the short-wavelength blue component dominates in the scattered light reaching our eyes. From 1879 to 1884 Rayleigh was the second Cavendish Professor of Physics at Cambridge, succeeding Maxwell, and he was awarded the Nobel prize in 1904 for the discovery of the gas argon.

eigenvectors $\boldsymbol{x}^{(j)}$, $j = 1, 2, \ldots, n$, *as*

$$\boldsymbol{x} = \sum_{j=1}^{n} \alpha_j \, \boldsymbol{x}^{(j)} \,, \tag{5.41}$$

then

$$R(\boldsymbol{x}) = \frac{\sum_{j=1}^{n} \lambda_j \, \alpha_j^2}{\sum_{j=1}^{n} \alpha_j^2} \,. \tag{5.42}$$

On noting that $\boldsymbol{x}^{(i)\mathrm{T}} \boldsymbol{x}^{(j)}$ is equal to 1 when $i = j$ and to 0 otherwise, (5.42) follows trivially by inserting (5.41) into (5.40).

Theorem 5.12 *Let* $A \in \mathbb{R}^{n \times n}_{\mathrm{sym}}$. *For any vector* $\boldsymbol{x} \in \mathbb{R}^n_*$,

$$\lambda_{\min} \leq R(\boldsymbol{x}) \leq \lambda_{\max} \,, \tag{5.43}$$

where $\lambda_{min} \in \mathbb{R}$ *and* $\lambda_{max} \in \mathbb{R}$ *are respectively the least and greatest of the eigenvalues of* A. *These bounds are attained when* $\boldsymbol{x}$ *is the corresponding eigenvector.*

Proof The inequalities follow immediately from (5.42) by noting that $\lambda_{\min} \leq \lambda_j \leq \lambda_{\max}$, $j = 1, 2, \ldots, n$. □

Theorem 5.13 *Suppose that* $\boldsymbol{x} \in \mathbb{R}^n_*$ *is a normalised vector, that is,* $\|\boldsymbol{x}\|_2 = 1$. *Assume, further, that* $\boldsymbol{x}^{(k)} \in \mathbb{R}^n_*$ *is the* k*th normalised eigenvector of* $A \in \mathbb{R}^{n \times n}$, *and that*

$$\|\boldsymbol{x} - \boldsymbol{x}^{(k)}\|_2 = \mathcal{O}(\varepsilon)$$

for a small $\varepsilon \in \mathbb{R}$. *Then,*

$$R(\boldsymbol{x}) = \lambda_k + \mathcal{O}(\varepsilon^2) \,.$$

Proof It follows from (5.41) that $\boldsymbol{x}^{\mathrm{T}} \boldsymbol{x}^{(k)} = \alpha_k$, and therefore,

$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{x}^{(k)}\|_2^2 &= (\boldsymbol{x} - \boldsymbol{x}^{(k)})^{\mathrm{T}} (\boldsymbol{x} - \boldsymbol{x}^{(k)}) \\ &= \|\boldsymbol{x}\|_2^2 - 2\boldsymbol{x}^{\mathrm{T}} \boldsymbol{x}^{(k)} + \|\boldsymbol{x}^{(k)}\|_2^2 \\ &= 2(1 - \alpha_k). \end{aligned}$$

Hence, $\alpha_k = 1 + \mathcal{O}(\varepsilon^2)$. Further,

$$1 = \|\boldsymbol{x}\|_2^2 = \sum_{j=1}^{n} \alpha_j^2$$

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

$$\begin{aligned} &= \alpha_k^2 + \sum_{j\neq k} \alpha_j^2 \\ &= 1 + \mathcal{O}(\varepsilon^2) + \sum_{j\neq k} \alpha_j^2. \end{aligned}$$

Consequently, $\alpha_j = \mathcal{O}(\varepsilon)$ for all $j \neq k$. The result then follows from (5.42) which (with $\sum_{j=1}^n \alpha_j^2 = \|\boldsymbol{x}\|_2^2 = 1$) yields that

$$\begin{aligned} R(\boldsymbol{x}) &= \lambda_k \alpha_k^2 + \sum_{j\neq k} \lambda_j \alpha_j^2 \\ &= \lambda_k + \mathcal{O}(\varepsilon^2)\,. \end{aligned}$$

□

This important result means that if we have a fairly close approximation $\boldsymbol{x}$ to an eigenvector of A, then the Rayleigh quotient $R(\boldsymbol{x})$ gives very easily a much more accurate approximation to the corresponding eigenvalue.

5.10 Perturbation analysis

It is often necessary to have an estimate of how much the eigenvalues and eigenvectors of a matrix are affected by changes in the elements. Such perturbations may arise, for example, when the matrix elements are obtained by physical measurements which are inexact, or they might result from finite difference approximations of a differential equation, as will be seen in Chapter 13. The last two theorems in this chapter address some of these questions. We begin with the following preliminary result.

Theorem 5.14 *Let $M \in \mathbb{R}^{n\times n}_{\text{sym}}$, with eigenvalues λ_i and corresponding orthonormal eigenvectors $\boldsymbol{v}_i, i = 1, 2, \ldots, n$, and suppose that $\boldsymbol{u} \neq \boldsymbol{0}$ and $\boldsymbol{w}$ are vectors in $\mathbb{R}^n$ and μ is a real number such that*

$$(M - \mu I)\boldsymbol{u} = \boldsymbol{w}\,. \tag{5.44}$$

Then, at least one eigenvalue λ_j of M satisfies

$$|\lambda_j - \mu| \leq \|\boldsymbol{w}\|_2 / \|\boldsymbol{u}\|_2\,.$$

Proof If μ is equal to one of the eigenvalues the proof is trivial, so we shall assume that $\mu \neq \lambda_k$, $k = 1, 2, \ldots, n$. We write the vectors $\boldsymbol{u}$ and

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

$\boldsymbol{w}$ as linear combinations of the eigenvectors of M, so that

$$\boldsymbol{u} = \sum_{k=1}^{n} \alpha_k \boldsymbol{v}_k \,, \qquad \boldsymbol{w} = \sum_{k=1}^{n} \beta_k \boldsymbol{v}_k \,.$$

Substituting in (5.44), we may equate coefficients of the linearly independent vectors $\boldsymbol{v}_k$, $k = 1, 2, \ldots, n$, to deduce that

$$(\lambda_k - \mu)\alpha_k = \beta_k \,, \qquad k = 1, 2, \ldots, n \,.$$

Now suppose that λ_j is the eigenvalue which is closest to μ; this means that

$$|\lambda_j - \mu| \leq |\lambda_k - \mu| \,, \qquad k = 1, 2, \ldots, n \,.$$

Since the eigenvectors $\boldsymbol{v}_i$, $i = 1, 2, \ldots, n$, are orthonormal in $\mathbb{R}^n$, we have

$$\sum_{k=1}^{n} \alpha_k^2 = \|\boldsymbol{u}\|_2^2 \,, \qquad \sum_{k=1}^{n} \beta_k^2 = \|\boldsymbol{w}\|_2^2 \,.$$

Hence

$$\sum_{k=1}^{n} \frac{\beta_k^2}{(\lambda_k - \mu)^2} = \|\boldsymbol{u}\|_2^2 \,,$$

which gives

$$\|\boldsymbol{w}\|_2^2 = \sum_{k=1}^{n} \beta_k^2 \geq (\lambda_j - \mu)^2 \sum_{k=1}^{n} \frac{\beta_k^2}{(\lambda_k - \mu)^2} = (\lambda_j - \mu)^2 \|\boldsymbol{u}\|_2^2 \,,$$

as required. □

We shall now use this result to show that in the case of a symmetric matrix A, small symmetric perturbations of A lead to small changes in the eigenvalues of A.

Theorem 5.15 (Bauer–Fike Theorem (symmetric case)) *Suppose that $A, E \in \mathbb{R}^{n\times n}_{\text{sym}}$ and $B = A - E$. Assume, further, that the eigenvalues of A are denoted by $\lambda_j, j = 1, 2, \ldots, n$, and μ is an eigenvalue of B. Then, at least one eigenvalue λ_j of A satisfies*

$$|\lambda_j - \mu| \leq \|E\|_2 \,.$$

Proof This is a straightforward consequence of the previous theorem. Suppose that $\boldsymbol{u}$ is the normalised eigenvector of B corresponding to the eigenvalue μ, so that $B\boldsymbol{u} = \mu\boldsymbol{u}$. Then,

$$(A - \mu I)\boldsymbol{u} = (B + E - \mu I)\boldsymbol{u} = E\boldsymbol{u} \,.$$

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

It then follows from Theorem 5.14 that there is an eigenvalue λ_j of A such that

$$|\lambda_j - \mu| \le \|E\boldsymbol{u}\|_2 \le \|E\|_2 \, \|\boldsymbol{u}\|_2 = \|E\|_2 \,,$$

as required. □

Example 5.11 *Consider the* 3×3 *Hilbert matrix*

$$A = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}$$

and its perturbation

$$B = \begin{pmatrix} 1.0000 & 0.5000 & 0.3333 \\ 0.5000 & 0.3333 & 0.2500 \\ 0.3333 & 0.2500 & 0.2000 \end{pmatrix}$$

which results by rounding each entry of A *to four decimal digits.*

In this case, $E = A - B$ and $\|E\|_2 = 3.3 \times 10^{-5}$. Let μ be an eigenvalue of B; then, according Theorem 5.15, at least one of the eigenvalues λ_1, λ_2, λ_3 of the matrix A satisfies the inequality

$$|\lambda_j - \mu| \le 3.3 \times 10^{-5} \,. \tag{5.45}$$

Indeed, the true eigenvalues of A and B are, respectively,

$$\lambda_1 = 0.002687338072 \,, \quad \lambda_2 = 0.1223270673 \,, \quad \lambda_3 = 1.408318925 \,,$$

and

$$\mu_1 = 0.002664493933 \,, \quad \mu_2 = 0.1223414532 \,, \quad \mu_3 = 1.408294053 \,.$$

Therefore,

$$\lambda_1 - \mu_1 = 2.29 \times 10^{-5}, \quad \lambda_2 - \mu_2 = -1.44 \times 10^{-5}, \quad \lambda_3 - \mu_3 = 2.49 \times 10^{-5},$$

which is in agreement with (5.45). ◇

5.11 Notes

Theorem 5.15 is a special case of the following general result, known as the Bauer–Fike Theorem.[1]

[1] F.L. Bauer and C.T. Fike, Norms and exclusion theorems, *Num. Math.* **2**, 137–141, 1960.

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

Theorem 5.16 *Assume that $A \in \mathbb{C}^{n\times n}$ is diagonalisable; i.e., there exists a nonsingular matrix $X \in \mathbb{C}^{n\times n}$ such that $X^{-1}AX = \Lambda$, where Λ is a diagonal matrix whose diagonal entries λ_j, $j = 1, \ldots, n$, are the eigenvalues of A. Suppose further that $E \in \mathbb{C}^{n\times n}$, $B = A - E$, and μ is an eigenvalue of B. Then, at least one eigenvalue λ_j of A satisfies*

$$|\lambda_j - \mu| \leq \kappa_2(X)\|E\|_2\,,$$

where $\kappa_2(X) = \|X\|_2\,\|X^{-1}\|_2$ is the condition number of the matrix X in the matrix 2-norm $\|\cdot\|_2$ on $\mathbb{C}^{n\times n}$.

In the special case when A, $E \in \mathbb{R}^{n\times n}_{\text{sym}}$, the matrix X can be chosen to be orthogonal; *i.e.*, $X^{-1} = X^{\mathrm{T}}$. Therefore, $\|X\|_2 = \|X^{-1}\|_2 = 1$, and hence $\kappa_2(X) = 1$, in accordance with the inequality stated in Theorem 5.15. Theorems 5.15 and 5.16 estimate how far the eigenvalues of A are perturbed by changes in the elements of A. The question as to how large the changes in the eigenvectors may be is more difficult; it is discussed in detail in

- J.H. Wilkinson, *The Algebraic Eigenvalue Problem,* Clarendon Press, Oxford University Press, New York, 1988.

Chapter 8 of Wilkinson's book outlines the convergence proof of the QR iteration, while the convergence of Jacobi's method is covered in Chapter 5 of that book. For further details, see also Chapter 9 of

- B. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

Exercises

5.1 Give a proof of Lemma 5.3.

5.2 Use Householder matrices to transform the matrix

$$A = \begin{pmatrix} 2 & 1 & 2 & 2 \\ 1 & -7 & 6 & 5 \\ 2 & 6 & 2 & -5 \\ 2 & 5 & -5 & 1 \end{pmatrix}$$

to tridiagonal form.

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

5.3 Use Sturm sequences to show that no eigenvalue of the matrix

$$A = \begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & 2 & -2 & 0 \\ 0 & -2 & 4 & \alpha \\ 0 & 0 & \alpha & 1 \end{pmatrix}$$

lies in the interval $(0, 1)$ if $5\alpha^2 > 8$, and that exactly one eigenvalue of A lies in this interval if $5\alpha^2 < 8$.

5.4 Given any two nonzero vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$, construct a Householder matrix H such that $H\boldsymbol{x}$ is a scalar multiple of $\boldsymbol{y}$; note that if $H\boldsymbol{x} = c\boldsymbol{y}$, then $c^2 = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}/\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}$. Is the matrix unique?

5.5 Suppose that the matrix $D \in \mathbb{R}^{n\times n}$ is diagonal with distinct diagonal elements $d_{11}, \ldots, d_{nn}$. Let $A \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$, with $|a_{ij}| \leq 1$ for all $i, j \in \{1, 2, \ldots, n\}$, and assume that $\varepsilon \in \mathbb{R}$ is so small that ε^2 can be neglected, and that the matrix $D + \varepsilon A$ has eigenvalue $\lambda + \varepsilon\mu$ and corresponding eigenvector $\mathbf{e} + \varepsilon\mathbf{u}$. Show that $\lambda = d_{jj}$ for some $j \in \{1, 2, \ldots, n\}$ and that $\mu = a_{jj}$. Write down the elements of $\mathbf{e}$, and show that

$$u_i = -\frac{a_{ij}}{d_{ii} - \lambda}\,, \qquad i \neq j\,.$$

Explain why the requirement that eigenvectors should be normalised implies that $u_j = 0$.

5.6 With the same notation as in Exercise 5, suppose now that $d_{11} = d_{22} = \cdots = d_{kk}$, that d_{kk}, $d_{k+1,k+1}$, $\ldots, d_{nn}$ are distinct, and that ε^3 can be neglected. Writing the matrices and the eigenvector in partitioned form, so that

$$\begin{pmatrix} d_{11}I_k + \varepsilon A_1 & \varepsilon A_2 \\ \varepsilon A_2^{\mathrm{T}} & D_{n-k} + \varepsilon A_3 \end{pmatrix} \begin{pmatrix} \mathbf{e} + \varepsilon\boldsymbol{u} + \varepsilon^2\boldsymbol{x} \\ \mathbf{f} + \varepsilon\boldsymbol{v} + \varepsilon^2\boldsymbol{y} \end{pmatrix}$$

$$= (\lambda + \varepsilon\mu + \varepsilon^2\nu) \begin{pmatrix} \mathbf{e} + \varepsilon\boldsymbol{u} + \varepsilon^2\boldsymbol{x} \\ \mathbf{f} + \varepsilon\boldsymbol{v} + \varepsilon^2\boldsymbol{y} \end{pmatrix},$$

show that $\lambda = d_{11}$, $\mathbf{f} = \mathbf{0}$, and that μ is an eigenvalue of A_1 with corresponding eigenvector $\mathbf{e}$. Show how $\boldsymbol{v}$ is obtained from the solution of $(D_{n-k} - \lambda I)\boldsymbol{v} = -A_2^{\mathrm{T}}\mathbf{e}$, and that

$$(A_1 - \mu)\boldsymbol{u} = \nu\mathbf{e} - A_2\mathbf{v}\,.$$

Explain how the vector $\boldsymbol{u}$ can be obtained in terms of the eigenvectors and eigenvalues of the matrix A_1, assuming that these eigenvalues are distinct.

5.7 Suppose that $A \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ is tridiagonal, that $A - \mu I = QR$ and $B = RQ + \mu I$, where $\mu \in \mathbb{R}$, $Q \in \mathbb{R}^{n\times n}$ is a product of plane rotations and $R \in \mathbb{R}^{n\times n}$ is upper triangular and tridiagonal. Show that B can be written as an orthogonal transformation of A, and that B is symmetric. Show also that the only nonzero elements in the matrix B which are below the diagonal lie immediately below the diagonal; deduce that B is tridiagonal.

5.8 Perform one step of the QR algorithm, using the shift $\mu = a_{nn}$, for the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Show that the QR algorithm does not converge for this matrix. (This is a special case in which a different shift must be used.)

5.9 Perform one step of the QR algorithm, using the shift $\mu = a_{nn}$, for the matrix

$$A = \begin{pmatrix} 13 & 4 \\ 4 & 10 \end{pmatrix}.$$

5.10 Carry out two steps of inverse iteration for the matrix

$$A = \begin{pmatrix} 2 & 2 \\ 2 & 5 \end{pmatrix},$$

using the eigenvalue estimate $\vartheta = 5$ and the initial vector

$$\boldsymbol{v}^{(0)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Verify that the elements of the vector $\boldsymbol{v}^{(2)}$ agree with those of the true eigenvector with an accuracy of about 5%. Evaluate the Rayleigh quotient using the vector $\boldsymbol{v}^{(2)}$, and verify that the result agrees with the true eigenvalue to about 1 in 3000.

5.11 An eigenvalue and eigenvector of the matrix A may be evaluated by solving the system of nonlinear equations

$$\begin{aligned} (A - \lambda I)\boldsymbol{x} &= 0\,, \\ \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} &= 1 \end{aligned}$$

for the unknowns λ and $\boldsymbol{x}$. Using Newton's method, starting

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press

from estimates $\lambda^{(0)}$ and $\boldsymbol{x}^{(0)}$, show that the next iteration is determined by

$$\begin{aligned} A\,\boldsymbol{\delta x} - \delta\lambda\,\boldsymbol{x}^{(0)} &= -(A - \lambda^{(0)}I)\boldsymbol{x}^{(0)}\,, \\ -\boldsymbol{x}^{(0)\mathrm{T}}\boldsymbol{\delta x} &= \tfrac{1}{2}(\boldsymbol{x}^{(0)\mathrm{T}}\boldsymbol{x}^{(0)} - 1) \end{aligned}$$

and $\boldsymbol{x}^{(1)} = \boldsymbol{x}^{(0)} + \boldsymbol{\delta x}, \lambda^{(1)} = \lambda^{(0)} + \delta\lambda$. Comment on the difference between this method and the method of inverse iteration in Section 5.8.

5.12 Suppose that $A \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ and that Jacobi's method has produced an orthogonal matrix R and a symmetric matrix B such that $B = R^{\mathrm{T}}AR$. Suppose also that $|b_{ij}| < \varepsilon$ for all $i \neq j$. Show that, for each $j = 1, 2, \ldots, n$, there is at least one eigenvalue λ of A such that

$$|\lambda - b_{jj}| < \varepsilon\sqrt{n}\,.$$

5.13 Suppose that $A \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ and that the Householder reduction and QR algorithm have produced an orthogonal matrix Q and a tridiagonal matrix T such that $T = Q^{\mathrm{T}}AQ$. Suppose also that $|t_{n,n-1}| < \varepsilon$. Show that there is at least one eigenvalue λ of A such that

$$|\lambda - t_{nn}| < \varepsilon\,.$$

https://doi.org/10.1017/CBO9780511801181.006 Published online by Cambridge University Press