3

Special matrices

3.1 Introduction

In this chapter we show how one can modify the elimination method for the solution of $A\boldsymbol{x} = \boldsymbol{b}$ when the matrix A has certain special properties. In particular when $A \in \mathbb{R}^{n\times n}$ is symmetric and positive definite the amount of computational work can be halved. For matrices with a band structure, having nonzero elements only in positions close to the diagonal, the efficiency can be improved even more dramatically.

3.2 Symmetric positive definite matrices

Definition 3.1 *The matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ is said to be* **symmetric** *if $a_{ij} = a_{ji}$ for all i and j in the set $\{1, 2, \ldots, n\}$; i.e., if $A = A^{\mathrm{T}}$. The set of all symmetric matrices $A \in \mathbb{R}^{n\times n}$ will be denoted by $\mathbb{R}^{n\times n}_{\mathrm{sym}}$. A matrix $A \in \mathbb{R}^{n\times n}$ is called* **positive definite** *if*

$$\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x} > 0$$

for every vector $\boldsymbol{x} \in \mathbb{R}^n_* = \mathbb{R}^n \setminus \{\boldsymbol{0}\}$.

Example 3.1 *Consider the matrix $A \in \mathbb{R}^{2\times 2}$,*

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

and a vector $\boldsymbol{x} = (x_1, x_2)^{\mathrm{T}} \in \mathbb{R}^2_* = \mathbb{R}^2 \setminus \{\boldsymbol{0}\}$.

Clearly, $\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x} = ax_1^2 + (b + c)x_1x_2 + dx_2^2$. The quadratic form on the right-hand side is positive for all real numbers x_1, x_2 such that

https://doi.org/10.1017/CBO9780511801181.004 Published online by Cambridge University Press

$\boldsymbol{x} = (x_1, x_2)^{\mathrm{T}} \neq (0,0)^{\mathrm{T}} = \mathbf{0}$ if, and only if,

$$a > 0\,, \quad d > 0 \ \text{ and } \ (b+c)^2 < 4ad\,.$$

We see that if $A \in \mathbb{R}^{2\times 2}$ is positive definite, then the diagonal elements of A are positive. Further, noting that the third inequality can be rewritten as

$$(b-c)^2 < 4(ad-bc) = 4\,\det(A)\,,$$

we deduce that the determinant of a positive definite matrix $A \in \mathbb{R}^{2\times 2}$ is positive. This, of course, is still true in the special case when $A \in \mathbb{R}^{2\times 2}_{\mathrm{sym}}$, *i.e.*, when $b = c$. ◇

The next theorem extends the observations of the last example to any symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$.

Theorem 3.1 *Suppose that $n \geq 2$ and $A = (a_{ij}) \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ is positive definite; then:*

(i) *all the diagonal elements of A are positive, that is, $a_{ii} > 0$, for $i = 1, 2, \ldots, n$;*
(ii) *all the eigenvalues of A are real and positive, and the eigenvectors of A belong to $\mathbb{R}^n_*$;*
(iii) *the determinant of A is positive;*
(iv) *every submatrix B of A obtained by deleting any set of rows and the corresponding set of columns from A is symmetric and positive definite; in particular, every leading principal submatrix is positive definite;*
(v) *$a_{ij}^2 < a_{ii}a_{jj}$ for all i and j in $\{1, 2, \ldots, n\}$ such that $i \neq j$;*
(vi) *the element of A with largest absolute value lies on the diagonal;*
(vii) *if α is the largest of the diagonal elements of A, then*

$$|a_{ij}| \leq \alpha \qquad \forall\, i, j \in \{1, 2, \ldots, n\}\,.$$

Proof (i) Consider the vector $\boldsymbol{x} \in \mathbb{R}^n$ with only one nonzero element, in position $i \in \{1, 2, \ldots, n\}$. Since A is positive definite and $\boldsymbol{x} \in \mathbb{R}^n_*$, it follows that $x_i a_{ii} x_i = \boldsymbol{x}^{\mathrm{T}} A \boldsymbol{x} > 0$, and therefore $a_{ii} > 0$.

(ii) Suppose that $\lambda \in \mathbb{C}$ is an eigenvalue of A and let $\boldsymbol{x} \in \mathbb{C}^n_* = \mathbb{C}^n \backslash \{\mathbf{0}\}$ denote the associated eigenvector. Further, let $\bar{\boldsymbol{x}}$ denote the vector in $\mathbb{C}^n_*$ whose ith element is the complex conjugate of the ith element of

https://doi.org/10.1017/CBO9780511801181.004 Published online by Cambridge University Press

$\boldsymbol{x}$, $i = 1, 2, \ldots, n$. As $A\boldsymbol{x} = \lambda\boldsymbol{x}$, it follows that $\bar{\boldsymbol{x}}^{\mathrm{T}} A\boldsymbol{x} = \lambda(\bar{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{x})$, and therefore, using the symmetry of A,

$$\boldsymbol{x}^{\mathrm{T}} A\bar{\boldsymbol{x}} = \boldsymbol{x}^{\mathrm{T}} A^{\mathrm{T}}\bar{\boldsymbol{x}} = (\bar{\boldsymbol{x}}^{\mathrm{T}} A\boldsymbol{x})^{\mathrm{T}} = (\lambda(\bar{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{x}))^{\mathrm{T}} = \lambda(\boldsymbol{x}^{\mathrm{T}}\bar{\boldsymbol{x}})\,.$$

Complex conjugation then yields $\bar{\boldsymbol{x}}^{\mathrm{T}} A\boldsymbol{x} = \bar{\lambda}(\bar{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{x})$, and hence $\lambda(\bar{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{x}) = \bar{\lambda}(\bar{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{x})$. As $\boldsymbol{x} \neq \boldsymbol{0}$, it follows that $\lambda = \bar{\lambda}$; i.e., λ is a real number.

The fact that the eigenvector associated with λ has real elements follows by noting that all elements of the singular matrix $A - \lambda I$ are real numbers. Therefore, the column vectors of $A - \lambda I$ are linearly dependent in $\mathbb{R}^n$. Hence there exist n real numbers $x_1, \ldots, x_n$ such that $(A - \lambda I)\boldsymbol{x} = \boldsymbol{0}$, where $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$.

Finally, as $A\boldsymbol{x} = \lambda\boldsymbol{x}$ with $\lambda \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^n_*$, we have that $\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x} = \lambda\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$. Since $\lambda = \boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x}/\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$ and A is positive definite, λ is the ratio of two positive real numbers and therefore also real and positive.

(iii) This follows from the fact that the determinant of A is equal to the product of its eigenvalues, and the previous result. Indeed, since A is symmetric, there exist an orthogonal matrix X and a diagonal matrix Λ, whose diagonal elements are the eigenvalues λ_i, $i = 1, 2, \ldots, n$, of A, such that $A = X^{\mathrm{T}}\Lambda X = X^{-1}\Lambda X$. By the Binet–Cauchy Theorem (see Chapter 2, end of Section 2.3),

$$\begin{aligned} \det(A) &= \det(X^{-1})\,\det(\Lambda)\,\det(X) \\ &= \frac{1}{\det(X)}\,\det(\Lambda)\,\det(X) \\ &= \det(\Lambda) = \lambda_1 \ldots \lambda_n > 0\,. \end{aligned}$$

(iv) Consider the vector $\boldsymbol{x} \in \mathbb{R}^n_*$ with zeros in the positions corresponding to the rows which have been deleted. Then,

$$\boldsymbol{x}^{\mathrm{T}} A\boldsymbol{x} = \boldsymbol{y}^{\mathrm{T}} B\boldsymbol{y}$$

where B is the submatrix of A containing the rows and columns which remain after deletion, and $\boldsymbol{y}$ is the vector consisting of the elements of $\boldsymbol{x}$ which were not deleted. Since the expression on the left is positive, the same is true of the expression on the right, for all vectors $\boldsymbol{y}$ except the zero vector. Therefore B is positive definite.

(v) By the previous result the 2×2 submatrix consisting of rows and columns r and s of A is positive definite, and its determinant is therefore positive.

(vi) This follows from the previous result, since it shows that $|a_{ij}|$ cannot exceed the greater of a_{ii} and a_{jj}.

(vii) This follows at once from the previous result. □

https://doi.org/10.1017/CBO9780511801181.004 Published online by Cambridge University Press

The converses of two of these results are also true:

(i) If all the eigenvalues of the symmetric matrix $A \in \mathbb{R}^{n\times n}$ are positive, then A is positive definite;
(ii) If the determinant of each leading principal submatrix of a matrix $A \in \mathbb{R}^{n\times n}$ is positive, then A is positive definite.

The proof of the second result is involved and will not be given here;[1] see, however, Example 3.1 for the case of $n = 2$. The proof of the first statement, on the other hand, is quite simple and proceeds as follows.

Since $A \in \mathbb{R}^{n\times n}$ is symmetric, it has a complete set of orthonormal eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ in $\mathbb{R}^n_*$, and the corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$ are all real. Given any vector $\boldsymbol{x} \in \mathbb{R}^n_*$, it can be expressed as

$$\boldsymbol{x} = \sum_{i=1}^{n} \alpha_i \boldsymbol{v}_i$$

where $\alpha_i \in \mathbb{R}$, $i = 1, 2, \ldots, n$, and $\alpha_1^2 + \cdots + \alpha_n^2 = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} > 0$. Since $A\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$, $i = 1, 2, \ldots, n$, it follows that

$$A\boldsymbol{x} = \sum_{i=1}^{n} \alpha_i \lambda_i \boldsymbol{v}_i \,.$$

As $\boldsymbol{v}_j^{\mathrm{T}}\boldsymbol{v}_i = 0$ for $i \neq j$ and $\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{v}_i = 1$, we deduce that

$$\begin{aligned} \boldsymbol{x}^{\mathrm{T}} A \boldsymbol{x} &= \sum_{i=1}^{n} \lambda_i \alpha_i^2 \\ &\geq \left(\min_{i=1}^{n} \lambda_i \right) \sum_{i=1}^{n} \alpha_i^2 > 0 \,, \end{aligned}$$

since $\min_{i=1}^{n} \lambda_i > 0$; therefore A is positive definite.

For a symmetric positive definite matrix A we can now obtain an LU factorisation $A = LU$ in which $U = L^{\mathrm{T}}$.

Theorem 3.2 *Suppose that $n \geq 2$ and $A \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ is a positive definite matrix; then, there exists a lower triangular matrix $L \in \mathbb{R}^{n\times n}$ such that*

$$A = LL^{\mathrm{T}} \,.$$

This is known as the **Cholesky factorisation**[2] *of A.*

[1] For more details, see R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, 1992, Theorem 7.2.5.

[2] 'André-Louis Cholesky (1875–1918) was a French military officer involved in geodesy and surveying in Crete and North Africa just before World War I. He

Proof Since A is symmetric and positive definite, all the leading principal submatrices of A are positive definite, and hence by Theorem 2.2 the usual LU factorisation exists, with

$$A = L^{(1)}U^{(1)}\,,$$

$L^{(1)} \in \mathbb{R}^{n\times n}$ a unit lower triangular and $U^{(1)} \in \mathbb{R}^{n\times n}$ an upper triangular matrix. In this factorisation the product of the leading principal submatrices of $L^{(1)}$ and $U^{(1)}$ of order k is the leading principal submatrix of A of order k, $1 \le k \le n$. Since the determinant of this submatrix is positive and all the diagonal elements of $L^{(1)}$ are unity, it follows that

$$u_{11}^{(1)}u_{22}^{(1)}\dots u_{kk}^{(1)} > 0\,, \qquad k = 1, 2, \dots, n\,.$$

Thus all the diagonal elements of $U^{(1)}$ are positive. If we now define D to be the diagonal matrix with elements $d_{ii} = \sqrt{u_{ii}^{(1)}}$, $i = 1, 2, \dots, n$, we can write

$$A = L^{(1)}U^{(1)} = (L^{(1)}D)(D^{-1}U^{(1)}) = LU\,,$$

where now $l_{ii} = u_{ii} = \sqrt{u_{ii}^{(1)}}$. The symmetry of the matrix A shows that

$$LU = A = A^{\mathrm{T}} = U^{\mathrm{T}}L^{\mathrm{T}}\,,$$

so that

$$U(L^{\mathrm{T}})^{-1} = L^{-1}U^{\mathrm{T}}\,.$$

In this equality the left-hand side is upper triangular, and the right-hand side is lower triangular, and hence both sides must be diagonal. Therefore, $U = D^*L^{\mathrm{T}}$, where D^* is a diagonal matrix; but U and L^{T} have the same diagonal elements, so $D^* = I$ and $U = L^{\mathrm{T}}$.

The same argument shows that L and L^{T} are unique, except for the arbitrary choice of the signs of the square roots in the definition of the diagonal matrix D. If we make the natural choice, taking all the square roots to be positive, then the diagonal elements of L are positive, and the factorisation is unique. □

developed the method now named after him to compute solutions to the normal equations for some least squares data fitting problems arising in geodesy. His work was posthumously published on his behalf in 1924 by a fellow officer, Benoit, in the *Bulletin Géodésique*.' – Cleve Moler, *NA-Digest*, February 18, 1990, Volume 90, Issue 07, `http://www.netlib.org/na-digest-html/90/v90n07.html`

In practice we construct the elements of L directly, rather than forming $L^{(1)}$ and $U^{(1)}$ first. This is done in a similar way to the LU factorisation. Suppose that $i \le j$; we then require that

$$a_{ij} = \sum_{k=1}^{i} l_{ik} l_{jk}\,, \qquad 1 \le i \le j \le n\,. \tag{3.1}$$

Note that we have used the fact that $(L^{\mathrm{T}})_{kj} = l_{jk}$; the sum only extends up to $k = i$ since L is lower triangular. The same equation will also hold for $i > j$, since A is symmetric. For $i = j$, equation (3.1) gives

$$l_{11} = a_{11}^{1/2}\,, \qquad l_{ii} = \left\{ a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2 \right\}^{1/2}\,, \qquad 1 < i \le n\,. \tag{3.2}$$

As A is a positive definite matrix, $a_{11} > 0$ and therefore l_{11} is a positive real number. Further, as we have seen in the proof of the preceding theorem, $l_{ii} > 0$, $i = 2, 3, \ldots, n$. We find similarly that

$$l_{ji} = \frac{1}{l_{ii}} \left\{ a_{ij} - \sum_{k=1}^{i-1} l_{ik} l_{jk} \right\}\,, \qquad 1 \le i < j \le n\,. \tag{3.3}$$

These equations now enable us to calculate the elements of L in succession. For each $i \in \{1, 2, \ldots, n-1\}$, we first calculate l_{ii} from (3.2), and then calculate $l_{i+1\,i},\ l_{i+2\,i},\ \ldots, l_{ni}$ from (3.3). Finally, we compute l_{nn} using (3.2).

As, by hypothesis, the matrix $A \in \mathbb{R}^{n\times n}_{\mathrm{sym}}$ is positive definite, the required factorisation exists, so we can be sure that the divisor l_{ii} in (3.3), and the expression in the curly brackets in (3.2) whose square root is taken, will be positive. Thus, (3.2) implies that

$$l_{11}^2 = a_{11}\,, \qquad \max_{k=1}^{i-1} l_{ik}^2 \le a_{ii}\,, \qquad i = 2, 3, \ldots, n\,.$$

The elements of the factor L cannot therefore grow very large, and no pivoting is necessary.

The evaluation of l_{ii} from (3.2) requires $i - 1$ multiplications, $i - 1$ subtractions and one square root operation, a total of $2i - 1$ operations. The calculation of each l_{ij} from (3.3) also requires $2i - 1$ operations. The total number of operations required to construct L is therefore

$$\sum_{i=1}^{n} \sum_{j=i}^{n} (2i - 1) = \sum_{i=1}^{n} (2i - 1)(1 + n - i) = \tfrac{1}{6} n(n+1)(2n+1)\,.$$

https://doi.org/10.1017/CBO9780511801181.004 Published online by Cambridge University Press

For large n the number of operations required is approximately $\frac{1}{3}n^3$, which, as might be expected, is half the number given in Section 2.6 for the LU factorisation of a nonsymmetric matrix.

3.3 Tridiagonal and band matrices

As we shall see in the final chapters, in the numerical solution of boundary value problems for second-order differential equations one encounters a particular kind of matrix whose elements are mostly zeros, except for those along its main diagonal and the two adjacent diagonals. Matrices of this kind are referred to as tridiagonal. In order to motivate the definition of tridiagonal matrix stated in Definition 3.2 below, we begin with an example which is discussed in more detail in Chapter 13.

Example 3.2 *Consider the two-point boundary value problem*

$$-\frac{\mathrm{d}^2y}{\mathrm{d}x^2} + r(x)y = f(x), \qquad x \in (0,1),$$

$$y(0) = 0, \quad y(1) = 0.$$

where r and f are continuous functions of x defined on the interval $[0,1]$.

The numerical solution of the boundary value problem proceeds by selecting an integer $n \geq 4$, choosing a step size $h = 1/n$, and subdividing the interval $[0,1]$ by the points $x_k = kh$, $k = 0, 1, \ldots, n$. The numerical approximation to $y(x_k)$, the value of the analytical solution y at the point $x = x_k$, is denoted by Y_k. The values Y_k are obtained by solving the set of linear equations

$$-\frac{Y_{k+1} - 2Y_k + Y_{k-1}}{h^2} + r(x_k)Y_k = f(x_k)$$

for $k = 1, 2, \ldots, n-1$, together with the boundary conditions

$$Y_0 = 0, \quad Y_n = 0.$$

Equivalently,

$$\begin{aligned} a_kY_{k-1} + c_kY_k + b_kY_{k+1} &= d_k, \qquad k = 1, 2, \ldots, n-1, \\ Y_0 &= 0, \quad Y_n = 0, \end{aligned}$$

https://doi.org/10.1017/CBO9780511801181.004 Published online by Cambridge University Press

where

$$a_k = b_k = -1/h^2, \quad c_k = 2/h^2 + r(x_k)\,, \quad d_k = f(x_k)\,,$$

for $k = 1, 2, \ldots, n-1$.

Clearly, for $1 < k < n-1$, the kth equation in the linear system above involves only three of the $n-1$ unknowns: Y_{k-1}, Y_k and Y_{k+1}. ◇

The example motivates the following definition of a tridiagonal (or triple diagonal) matrix.

Definition 3.2 *Suppose that $n \geq 3$. A matrix $T = (t_{ij}) \in \mathbb{R}^{n\times n}$ is said to be* **tridiagonal** *if it has nonzero elements only on the main diagonal and the two adjacent diagonals; i.e.,*

$$t_{ij} = 0 \quad \text{if} \quad |i-j| \;>\; 1\,, \qquad i, j \in \{1, 2, \ldots, n\}\,.$$

Such matrices are also sometimes called **triple diagonal**.

It is easy to see that in the LU factorisation process of a tridiagonal matrix $T \in \mathbb{R}^{n\times n}$, without row interchanges, the unit lower triangular matrix $L \in \mathbb{R}^{n\times n}$ and the upper triangular matrix $U \in \mathbb{R}^{n\times n}$ each have only two elements in each row. Writing T in the compact notation

$$T = \begin{pmatrix} b_1 & c_1 & & & & \\ a_2 & b_2 & c_2 & & & \\ & a_3 & b_3 & c_3 & & \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ & & & & a_n & b_n \end{pmatrix}, \tag{3.4}$$

the factorisation may be written $T = LU$ where

$$L = \begin{pmatrix} 1 & & & & & \\ l_2 & 1 & & & & \\ & l_3 & 1 & & & \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ & & & l_n & 1 \end{pmatrix} \tag{3.5}$$

and

$$U = \begin{pmatrix} u_1 & v_1 & & & \\ & u_2 & v_2 & & \\ & & u_3 & v_3 & \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ & & & & u_n \end{pmatrix}, \tag{3.6}$$

https://doi.org/10.1017/CBO9780511801181.004 Published online by Cambridge University Press

with the convention that the missing elements in these matrices are all equal to zero. It is often convenient to define $a_1 = 0$ and $c_n = 0$. Multiplying L and U shows that $v_j = c_j$, and that the elements l_j and u_j can be calculated from

$$l_j = a_j/u_{j-1}\,, \quad u_j = b_j - l_j c_{j-1}\,, \qquad j = 2, 3, \ldots, n\,, \tag{3.7}$$

starting from $u_1 = b_1$.

Let us suppose that our aim is to solve the system of linear equations $T\boldsymbol{x} = \boldsymbol{r}$, where the matrix $T \in \mathbb{R}^{n\times n}$ is tridiagonal and nonsingular, and $\boldsymbol{r} \in \mathbb{R}^n$. Having calculated the elements of the matrices L and U in the LU factorisation $T = LU$ using (3.7), the forward and backsubstitution are then also very simple. Letting $\boldsymbol{y} = U\boldsymbol{x}$, the equation $L\boldsymbol{y} = \boldsymbol{r}$ gives

$$y_1 = r_1\,, \tag{3.8}$$

$$y_j = r_j - l_j y_{j-1}\,, \qquad j = 2, 3, \ldots, n\,, \tag{3.9}$$

and finally from $U\boldsymbol{x} = \boldsymbol{y}$ we get

$$x_n = y_n/u_n\,, \tag{3.10}$$

$$x_j = (y_j - v_j x_{j+1})/u_j\,, \qquad j = n-1, n-2, \ldots, 1\,. \tag{3.11}$$

The LU factorisation of a tridiagonal matrix requires approximately $3n$ operations. The forward and backsubstitution together involve approximately $5n$ operations. Thus, the whole solution process requires approximately $8n$ operations. The total amount of work is therefore far less than for a full matrix, being of order n for large n, compared with $\frac{2}{3}n^3$ for a full matrix. The method we have described is a minor variation on what is often known as the *Thomas algorithm*.[1]

So far we have assumed that pivoting was not necessary; clearly any interchange of rows will destroy the tridiagonal structure of T. However, it is easy to see that the only interchanges required will be between two adjacent rows.

Theorem 3.3 *Suppose that $n \geq 3$ and $T \in \mathbb{R}^{n\times n}$ is a tridiagonal matrix; then, there exists a permutation matrix $P \in \mathbb{R}^{n\times n}$ such that*

$$PA = L^{(1)}U^{(1)} \tag{3.12}$$

[1] After Llewellyn H. Thomas, a distinguished physicist, who in the 1950s held positions at Columbia University and at IBM's Watson Research Laboratory. He is probably best known in connection with the Thomas–Fermi electron gas model. The terminology 'Thomas algorithm' comes from David Young. Thomas, L.H., *Elliptic Problems in Linear Difference Equations over a Network*, Watson Sci. Comput. Lab. Rept, Columbia University, New York, 1949. See *NA-Digest* V.96, 09, `http://www.netlib.org/cgi-bin/mfs/02/96/v96n09.html`

where $L^{(1)} \in \mathbb{R}^{n\times n}$ is unit lower triangular with at most two nonzero elements in each row, and $U^{(1)} \in \mathbb{R}^{n\times n}$ is upper triangular with at most three nonzero elements in each row.

The proof of this theorem is left as an exercise (see Exercise 6). It shows that the effect of pivoting is at worst to lead to an additional superdiagonal in the upper triangular factor.

In an important class of problems it is also easy to show that pivoting is unnecessary. We have shown this to be true for a symmetric positive definite matrix, and we can now show that it is also true for a tridiagonal matrix which is strictly diagonally dominant.

Definition 3.3 *A matrix $A \in \mathbb{R}^{n\times n}$ is said to be* **diagonally dominant** *if*

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|\,, \qquad i = 1, 2, \ldots, n\,;$$

A is said to be **strictly diagonally dominant** *if strict inequality holds for each i.*

Theorem 3.4 *Suppose that $n \geq 3$, $T \in \mathbb{R}^{n\times n}$ is tridiagonal, as in (3.4), and*

$$|b_j| > |a_j| + |c_j|\,, \qquad j = 1, 2, \ldots, n \tag{3.13}$$

(with the convention $a_1 = 0$, $c_n = 0$); then T is nonsingular, and it can be written, without pivoting, in the form $T = LU$ where $L \in \mathbb{R}^{n\times n}$ is unit lower triangular and $U \in \mathbb{R}^{n\times n}$ is upper triangular. The condition (3.13) ensures that the matrix T is strictly diagonally dominant.

Proof We first show by induction that $|u_j| > |c_j|$ for all $j = 1, 2, \ldots, n$. This inequality trivially holds for $j = 1$ since

$$|u_1| = |b_1| > |a_1| + |c_1| = |c_1|\,.$$

Now let $j \in \{2, \ldots, n\}$ and adopt the inductive hypothesis:

$$\texttt{Hyp}_{j-1}\colon \qquad |u_{k-1}| > |c_{k-1}| \qquad \forall\, k \in \{2, \ldots, j\}\,.$$

(As we have already seen, $\texttt{Hyp}_1$ is true.) Then, from (3.7) we see that

$$\begin{aligned} |u_j| &\geq \left|\, |b_j| - |a_j| \left|\frac{c_{j-1}}{u_{j-1}}\right| \,\right| \\ &\geq |\,|b_j| - |a_j|\,| > |c_j| \end{aligned} \tag{3.14}$$

https://doi.org/10.1017/CBO9780511801181.004 Published online by Cambridge University Press

by the condition of strict diagonal dominance (3.13), which then shows that $\mathtt{Hyp}_j$ holds. That completes the inductive step.

We have thus proved that $|u_j| > |c_j|$ for all $j = 1, 2, \ldots, n$. In particular, we deduce that $u_j \neq 0$ for all $j \in \{1, 2, \ldots, n\}$; hence the LU factorisation $T = LU$ defined by (3.7) exists. Further,

$$\det(T) = \det(L)\det(U) = \det(U) = u_1 u_2 \ldots u_n \neq 0\,,$$

so T is nonsingular.

The formula (3.7) and the inequalities $|u_j| > |c_j|$, $j = 1, 2, \ldots, n$, now imply that

$$\begin{aligned} |u_j| &\leq |b_j| + |l_j|\,|c_{j-1}| \\ &= |b_j| + |a_j c_{j-1}|/|u_{j-1}| \\ &\leq |b_j| + |a_j|\,, \qquad j = 1, 2, \ldots, n\,, \end{aligned} \tag{3.15}$$

so the elements u_j cannot grow large, and rounding errors are kept under control without pivoting. □

It is easy to see that the same result holds under the weaker assumption that the matrix is diagonally dominant, but not necessarily strictly diagonally dominant, provided that we also require that all the elements c_j, $j = 1, 2, \ldots, n-1$, are nonzero (see Exercise 5).

Note also that the matrix constructed in Example 3.2 satisfies this condition, provided that the function r is nonnegative; this often holds in practical boundary value problems.

If the matrix $T \in \mathbb{R}^{n\times n}$ is symmetric and positive definite, as well as tridiagonal, it can be factorised in the form $T = LL^{\mathrm{T}}$, where $L \in \mathbb{R}^{n\times n}$ is lower triangular with nonzero elements only on and immediately below the diagonal. If we use the notation $d_i = l_{ii}$, $e_i = l_{i\,i-1}$ we easily find from (3.2) and (3.3) that the elements can be calculated in succession from the following formulae:

$$\begin{aligned} d_1 &= b_1^{1/2}\,, \\ e_i &= c_{i-1}/d_{i-1}\,, \quad d_i = \left(b_i - e_i^2\right)^{1/2}\,, \qquad i = 2, 3, \ldots, n\,. \end{aligned}$$

This calculation involves about $4n$ operations. Including also the work required by the forward and backsubstitution stages, the complete solution of $T\boldsymbol{x} = \boldsymbol{b}$ will be found to involve about $10n$ operations. For the tridiagonal matrix the Cholesky factorisation method thus requires more work for the complete solution than the Thomas algorithm; in this case there is no particular advantage in exploiting the symmetry of the matrix in this way.

https://doi.org/10.1017/CBO9780511801181.004 Published online by Cambridge University Press

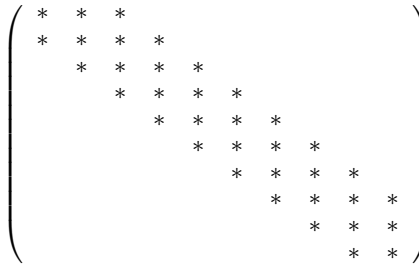

Fig. 3.1. The asterisks indicate the 36 nonzero elements in this 10×10 Band(1,2) matrix.

More generally, a system of equations may often involve a matrix of band type.

Definition 3.4 $B \in \mathbb{R}^{n \times n}$ *is a* **band matrix** *if there exist nonnegative integers* $p < n$ *and* $q < n$ *such that* $b_{ij} = 0$ *for all* $i, j \in \{1, 2, \ldots, n\}$ *such that* $p < i - j$ *or* $q < j - i$. *The band is of width* $p + q + 1$, *with* p *elements to the left of the diagonal and* q *elements to the right of the diagonal, in each row. Such a matrix is said to be* Band(p, q).

Thus, for example, a tridiagonal matrix is Band(1,1), and an $n \times n$ lower triangular matrix is Band$(n - 1,0)$.

An example of a Band(1,2) matrix $A \in \mathbb{R}^{10 \times 10}$ is shown in Figure 3.1, where each nonzero element in the matrix is identified by an asterisk. In addition to its main diagonal, the matrix has nonzero elements on its lower subdiagonal and two of its superdiagonals.

It is easy to see that, provided that no interchanges are necessary, such a band matrix can be written in the form $B = LU$, where L is Band$(p,0)$ and U is Band$(0,q)$ (see Exercise 7). It is also fairly simple to count the operations required in this calculation; the result is approximately proportional to $np(p + 2q)$ when n is moderately large. The most common situation has $q = p$, and then the number of operations is approximately proportional to np^2. As in the tridiagonal case, this is much smaller than n^3 when p and q are fairly small compared with n.

3.4 Monotone matrices

If a positive real number a is increased by $\varepsilon > 0$ to $a + \varepsilon$, then its reciprocal a^{-1} decreases to $(a + \varepsilon)^{-1}$. It is not usually true, however,

https://doi.org/10.1017/CBO9780511801181.004 Published online by Cambridge University Press

that if we increase some or all of the elements of a nonsingular matrix $A \in \mathbb{R}^{n\times n}$, then the elements of the inverse $A^{-1} \in \mathbb{R}^{n\times n}$ will decrease. This useful property holds for the class of monotone matrices defined below.

The discussion in this section is not related to Gaussian elimination and LU factorisation, but it is of relevance in the iterative solution of systems of linear equations with monotone matrices which arise in the course of numerical approximation of boundary value problems for certain ordinary and partial differential equations.

Definition 3.5 *The nonsingular matrix $A \in \mathbb{R}^{n\times n}$ is said to be* **monotone** *if all the elements of the inverse A^{-1} are nonnegative.*

Example 3.3 *Suppose that a and d are positive real numbers, and b and c are nonnegative real numbers such that $ad > bc$. Then,*

$$A = \begin{pmatrix} a & -b \\ -c & d \end{pmatrix}$$

is a monotone matrix. This is easily seen by considering the inverse of the matrix A,

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & b \\ c & a \end{pmatrix},$$

and noting that all elements of A^{-1} are nonnegative.

Next we introduce the concept of ordering in $\mathbb{R}^n$ and $\mathbb{R}^{n\times n}$.

Definition 3.6 *For vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$ we use the notation*

$$\boldsymbol{x} \succeq \boldsymbol{y}$$

to mean that

$$x_i \geq y_i\,, \qquad i = 1, 2, \ldots, n\,.$$

In the same way, for matrices A and B in $\mathbb{R}^{n\times n}$ we write

$$A \succeq B$$

to mean that

$$a_{ij} \geq b_{ij}\,, \qquad i, j = 1, 2, \ldots, n\,.$$

The sign $\succeq$ is read 'succeeds or is equal to' or, simply, 'is greater than or equal to'.

https://doi.org/10.1017/CBO9780511801181.004 Published online by Cambridge University Press

Note that, given two arbitrary matrices A and B in $\mathbb{R}^{n\times n}$, in general none of $A \succeq B$, $A = B$ and $B \succeq A$ will be true. Therefore the relation $\succeq$ is a partial, rather than a total, ordering on $\mathbb{R}^{n\times n}$; the same is true of the ordering $\succeq$ on $\mathbb{R}^n$.

Theorem 3.5 *(i) Suppose that the nonsingular matrix $A \in \mathbb{R}^{n\times n}$ is monotone, $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^n$, and the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$ are the solutions of*

$$A\boldsymbol{x} = \boldsymbol{b}\,, \quad A\boldsymbol{y} = \boldsymbol{c}\,,$$

respectively. If $\boldsymbol{b} \succeq \boldsymbol{c}$, then $\boldsymbol{x} \succeq \boldsymbol{y}$.

(ii) Suppose that A and B are nonsingular matrices in $\mathbb{R}^{n\times n}$ and that both are monotone. If $A \succeq B$, then $B^{-1} \succeq A^{-1}$.

Proof (i) Since the elements of A^{-1} are nonnegative and

$$\boldsymbol{x} - \boldsymbol{y} \;=\; A^{-1}(\boldsymbol{b} - \boldsymbol{c})\,,$$

the result follows from the fact that all elements of the vector $A^{-1}(\boldsymbol{b}-\boldsymbol{c})$ appearing on the right-hand side of this equality are nonnegative.

(ii) Since $A \succeq B$ and all the elements of B^{-1} are nonnegative, it follows that

$$B^{-1}A \;\succeq\; B^{-1}B \;=\; I\,.$$

In the same way, since all the elements of A^{-1} are nonnegative, it follows that

$$B^{-1} \;= B^{-1}A\,A^{-1} \succeq A^{-1},$$

as required. □

The following theorem will be useful in Chapter 13.

Theorem 3.6 *Suppose that $n \geq 3$ and $T \in \mathbb{R}^{n\times n}$ is a tridiagonal matrix of the form (3.4) with the properties*

$$a_i < 0\,, \quad i = 2, 3, \ldots, n\,, \qquad c_i < 0\,, \quad i = 1, 2, \ldots, n-1\,,$$

and

$$a_i + b_i + c_i \geq 0\,, \qquad i = 1, 2, \ldots, n\,,$$

where we have followed the convention that $a_1 = 0$, $c_n = 0$; then, the matrix T is monotone.

https://doi.org/10.1017/CBO9780511801181.004 Published online by Cambridge University Press

Proof Let $k \in \{1, 2, \ldots, n\}$. Column k of the inverse T^{-1} is the solution of the linear system $T\boldsymbol{c}^{(k)} = \boldsymbol{e}^{(k)}$, where $\boldsymbol{e}^{(k)}$ is column k of the identity matrix of size n, having a single nonzero element, 1, in row k. By applying the Thomas algorithm to this linear system, it is easy to deduce by induction from (3.7) that $l_j \leq 0$, $u_j \geq 0$ and $v_j \leq 0$ for all j; the argument is very similar to the proof of Theorem 3.4. It then follows from (3.8) and (3.9) that, in the notation of the Thomas algorithm, the vectors $\boldsymbol{y}$ and $\boldsymbol{x}$ have nonnegative elements. Hence column k of the inverse T^{-1} has nonnegative elements. Since the same is true for each $k \in \{1, 2, \ldots, n\}$, it follows that T is monotone. □

3.5 Notes

Symmetric systems of linear algebraic equations arise in the numerical solution of self-adjoint boundary value problems for differential equations with real-valued coefficients.

For further details on the Cholesky factorisation, the reader may consult any of the books listed in the Notes at the end of Chapter 2, particularly Chapter 10 of N.J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

Classical iterative methods for the solution of systems of linear equations with monotone matrices are discussed, for example, in

- Richard S. Varga, *Matrix Iterative Analysis,* Prentice–Hall, Englewood Cliffs, NJ, 1962.

A more recent reference on iterative algorithms for linear systems is

- Owe Axelson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, 1996.

In particular, Chapter 6 of Axelson's book considers the relevance of monotone matrices in the context of iterative solution of systems of linear equations.

Theorem 3.6 is a slight variation on the following general result.

Theorem 3.7 *A sufficient condition for $A \in \mathbb{R}^{n\times n}$ to be a monotone matrix is that A is an **M-matrix**, that is, (a) $a_{ij} \leq 0$ for all $i, j \in \{1, 2, \ldots, n\}$ such that $i \neq j$, and (b) there exists a vector $\boldsymbol{g} \in \mathbb{R}^n$ with positive elements such that all elements of $A\boldsymbol{g} \in \mathbb{R}^n$ are positive.*

https://doi.org/10.1017/CBO9780511801181.004 Published online by Cambridge University Press

Exercises

3.1 Find the Cholesky factorisation of the matrix

$$A = \begin{pmatrix} 4 & 6 & 2 \\ 6 & 10 & 3 \\ 2 & 3 & 5 \end{pmatrix}.$$

3.2 Use the method of Cholesky factorisation to solve the system of equations

$$\begin{aligned} x_1 - 2x_2 + 2x_3 &= 4\,, \\ -2x_1 + 5x_2 - 3x_3 &= -7\,, \\ 2x_1 - 3x_2 + 6x_3 &= 10\,. \end{aligned}$$

3.3 Let $n \geq 3$. The $n \times n$ tridiagonal matrix T has the diagonal elements

$$T_{ii} = 2\,, \quad i = 1, 2, \ldots, n\,,$$

and the off-diagonal elements

$$T_{i\,i+1} = T_{i+1\,i} = -1\,, \quad i = 1, 2, \ldots, n-1\,.$$

In the factorisation $T = LU$, where $L \in \mathbb{R}^{n\times n}$ is unit lower triangular and $U \in \mathbb{R}^{n\times n}$ is upper triangular, show that

$$L_{i+1\,i} = -i/(i+1)\,, \qquad i = 1, 2, \ldots, n\,,$$

and find expressions for the elements of U. What is the determinant of T?

3.4 Let $n \geq 3$ and $1 \leq k \leq n$. Define the vector $\boldsymbol{v}^{(k)} \in \mathbb{R}^n$ with elements given by

$$v_i^{(k)} = \begin{cases} i(n+1-k)\,, & i = 1, \ldots, k\,, \\ k(n+1-i)\,, & i = k+1, \ldots, n\,. \end{cases}$$

Evaluate M_{kj}, the inner product of the vector $\boldsymbol{v}^{(k)}$ with column j of the matrix T defined in Exercise 3. (The inner product $\langle \boldsymbol{v}, \boldsymbol{w} \rangle$ of two vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ in $\mathbb{R}^n$ is defined as the real number $\boldsymbol{v}^{\mathrm{T}}\boldsymbol{w}$.) Hence give expressions for the elements of the inverse matrix T^{-1}, and verify that this inverse is symmetric. Find the ∞-norm of the inverse, $\|T^{-1}\|_\infty$, and show that the condition number of T is

$$\kappa_\infty(T) = \frac{1}{2}(n+1)^2\,, \quad n \text{ odd}\,.$$

What is the condition number $\kappa_\infty(T)$ when n is even?

https://doi.org/10.1017/CBO9780511801181.004 Published online by Cambridge University Press

3.5 Given that $n \geq 3$, in the notation of Theorem 3.4 suppose that

$$|b_j| \geq |a_j| + |c_j|\,, \quad j = 1, 2, \ldots, n\,,$$

and

$$|c_j| > 0\,, \quad j = 1, 2, \ldots, n-1\,,$$

with the convention that $a_1 = 0$ and $c_n = 0$. Show that the factorisation $T = LU$ exists without pivoting, and can be constructed by the Thomas algorithm. Give an example of a matrix T which satisfies these conditions, except that $c_k = 0$ for some $k \in \{1, 2, \ldots, n-1\}$ and such that T is singular and cannot be written in the form $T = LU$ without pivoting.

3.6 Let $n \geq 3$ and suppose that the matrix $T \in \mathbb{R}^{n\times n}$ is tridiagonal. Show that there exists a permutation matrix $P \in \mathbb{R}^{n\times n}$ such that

$$PA = L^{(1)}U^{(1)}$$

where $L^{(1)} \in \mathbb{R}^{n\times n}$ is unit lower triangular with at most two nonzero elements in each row, and $U^{(1)} \in \mathbb{R}^{n\times n}$ is upper triangular with at most three nonzero elements in each row.

3.7 Suppose that the matrix B is Band(p,q), and that there exists a factorisation $B = LU$ without row interchanges. Show that L is Band(p,0) and U is Band(0,q).

3.8 Suppose that $n \geq 4$, that the matrix $A \in \mathbb{R}^{n\times n}$ is Band(3,3), and has the LU factorisation $A = LU$, so that $L \in \mathbb{R}^{n\times n}$ is Band(3,0) and $U \in \mathbb{R}^{n\times n}$ is Band(0,3). Suppose also that $a_{i+2,i} = 0$, $a_{i,i+2} = 0$ for $i = 1, 2, \ldots, n-2$. By considering u_{24} and l_{42}, or otherwise, show that in general the elements $l_{i+2,i}$ and $u_{i,i+2}$ are not zero.