Part I
Basic numerical techniques

1

Introduction

Summary In this chapter we summarise some important mathematical results which are frequently required to solve numerical and statistical problems arising in scientific work. First, we discuss some Taylor series expansions for functions of one variable; we continue with the notion of functions of two or more variables and of partial differentiation. Lastly, we outline the concept of a matrix and the basic matrix operations: addition, subtraction, multiplication and the derivation of an inverse.

1.1 The level of mathematics required

The level of mathematics necessary for understanding the techniques described in this book is quite modest. Yet, using these techniques we are able to solve many difficult numerical and statistical problems. We are able to solve non-linear equations in one or more unknowns, integrate and differentiate a given function (including empirical functions), smooth crude data, fit curves and interpolate. Some of the techniques we develop can be used to solve other complex problems. For example, the method of finite differences (chapter 6) is used extensively for solving differential equations.

Readers will be familiar with the mathematical concepts of differentiation (obtaining the slope of a curve) and integration (finding the area under a curve). There are certain other basic concepts and formulae which bear repeating and these are summarised in the remainder of this chapter.

1.2 The Taylor series expansion

This formula uses the value of a function $f(x)$ and its derivatives at a particular point x to produce the value of that function at a neighbouring point $x+h$.

■ $$f(x+h) = f(x) + \frac{h}{1!}f'(x) + \frac{h^2}{2!}f''(x) + \dots . \qquad (1.2.1)$$

Further reading: Conte [17] 16; Grossman and Turner [38] 407–27; Nielsen [73] 9; Sokolnikoff and Sokolnikoff [92] 30–42.

1.3 The exponential series

The following series for e^x is valid for all values of x:

■ $$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots . \qquad (1.3.1)$$

This formula can be derived using the Taylor series expansion (1.2.1) and recalling that the derivative of e^x is e^x.

Further reading: Conte [17] 16; Grossman and Turner [38] 428–34, 457–60.

1.4 The logarithmic series

The following series for the natural logarithm of $1+x$ is valid whenever $|x|<1$:

■ $$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \dots . \qquad (1.4.1)$$

This formula can be derived using the Taylor series expansion (1.2.1) and recalling that the derivative of $\ln(1+x)$ is $(1+x)^{-1}$.

Further reading: Conte [17] 16; Grossman and Turner [38] 428–34.

1.5 The binomial expansion

The following series for $(1+x)^n$ is valid for all n when $|x|<1$:

■ $$(1+x)^n = 1 + \binom{n}{1}x + \binom{n}{2}x^2 + \dots , \qquad (1.5.1)$$

where

■ $$\left.\begin{array}{l} \binom{n}{1} = \frac{n}{1}, \ \binom{n}{2} = \frac{n(n-1)}{1 \times 2}, \\ \binom{n}{3} = \frac{n(n-1)(n-2)}{1 \times 2 \times 3}, \text{ etc.} \end{array}\right\} \qquad (1.5.2)$$

The formula may be derived using the Taylor series expansion (1.2.1).

When n is a positive integer, the series is valid for all x and consists of a sum of $n+1$ terms. Furthermore

■ $$\binom{n}{r} = \frac{n!}{r!\,(n-r)!}. \qquad (1.5.3)$$

Numerical values of the binomial coefficient (1.5.3) are often displayed in a *Pascal triangle* (table 1.5.1).

Table 1.5.1. *Binomial coefficients* $\binom{n}{r}$ *for small integer values of n. Each entry is equal to the sum of the entry immediately above and the one above to the left (for example, 70 = 35 + 35)*

	r								
n	0	1	2	3	4	5	6	7	8
0	1								
1	1	1							
2	1	2	1						
3	1	3	3	1					
4	1	4	6	4	1				
5	1	5	10	10	5	1			
6	1	6	15	20	15	6	1		
7	1	7	21	35	35	21	7	1	
8	1	8	28	56	70	56	28	8	1

Example 1.5.1. Use (1.5.1) to find the square root of 1.03 correct to six decimal places.

$$
\begin{aligned}
(1.03)^{\frac{1}{2}} &= 1+\frac{\frac{1}{2}}{1}(0.03)+\frac{\frac{1}{2}(-\frac{1}{2})}{1\times 2}(0.03)^2+\frac{\frac{1}{2}(-\frac{1}{2})(-\frac{3}{2})}{1\times 2\times 3}(0.03)^3+\dots \\
&= 1+0.015-0.000\,1125+0.000\,0016875-\dots \\
&= 1.014\,889 \text{ (correct to six decimal places).}
\end{aligned}
$$

Further reading: Grossman and Turner [38] 32–42; Nielsen [73] 8.

1.6 Partial differentiation

The slope of a function $f(x)$ in one dimension is the rate of increase of that function as x increases. A function of two variables x and y can be likened to a hill. The x-value may be compared with latitude, the y-value with longitude and the height at a particular point with the function value $f(x,y)$.

If we were to start walking at the bottom of the hill and head straight towards the summit, we might find the climbing rather difficult, because the slope is rather steep. Alternatively, we might zig-zag up the hill along paths which are less steep. Clearly, the slope at a given point (x,y) depends upon the direction in which we are heading.

In a mathematical context, two directions are important: the direction of constant y and increasing x (constant longitude and increasing latitude) and the direction of constant x and increasing y (constant latitude and increasing longitude). These two directions are at right angles.

The slope in the direction of constant y and increasing x is referred to as the partial derivative of $f(x,y)$ with respect to x and is denoted by $\partial f/\partial x$. It is obtained by treating y as a constant and differentiating $f(x,y)$ with respect to x in the usual way. Likewise, the slope in the direction of constant x and increasing y is referred to as the partial derivative of $f(x,y)$ with respect to y and is denoted by $\partial f/\partial y$. It is obtained by treating x as a constant and differentiating $f(x,y)$ with respect to y.

At the top of the hill, the slope in any direction is zero, including the direction (y constant, x increasing) and the direction (x constant, y increasing). At a maximum point of the function $f(x,y)$ (or minimum point, for that matter) $\partial f/\partial x$ and $\partial f/\partial y$ are both zero.

Example 1.6.1. Find the minimum value of

$$f(x,y)=x^2+y^2-x+y+xy-3.$$

We equate to zero the partial derivatives with respect to x and y:

$$
\begin{aligned}
2x-1+y&=0,\\
2y+1+x&=0.
\end{aligned}
$$

We solve these two simultaneous equations in two unknowns and find that the minimum is -4 at the point $(1,-1)$. This stationary point is a minimum point because the solution is unique and $f(x,y)\to\infty$ as both x and y tend to $\pm\infty$.

Further reading: Grossman and Turner [38] 435–8; Sokolnikoff and Sokolnikoff [92] 123–6, 160–6.

1.7 The two-dimensional Taylor series

The two-dimensional Taylor series takes the form

■
$$f(x+h,\, y+k) = f(x,y) + \left(h\frac{\partial}{\partial x} f(x,y) + k\frac{\partial}{\partial y} f(x,y)\right) + \frac{1}{2!}\left(h^2\frac{\partial^2}{\partial x^2} f(x,y) + 2hk\frac{\partial^2}{\partial x \partial y} f(x,y) + k^2\frac{\partial^2}{\partial y^2} f(x,y)\right) + \dots, \tag{1.7.1}$$

and it may be generalised to three or more variables.

Further reading: Sokolnikoff and Sokolnikoff [92] 155–7.

1.8 The concept of a matrix

A matrix is a rectangular array of elements. These elements are usually numbers. For example,

$$\begin{pmatrix} -2 \\ 3 \end{pmatrix}, \quad \begin{pmatrix} 1.3 & 1.2 \\ 1.7 & 0.6 \\ 0.2 & 1.3 \end{pmatrix}, \quad \begin{pmatrix} 4.3 & 0.0 \\ 1.9 & 0.3 \end{pmatrix}.$$

By the *dimension* of a matrix we mean the number of rows and columns in that matrix. Thus, a matrix of m rows and n columns is called $m \times n$ matrix. The above matrices are of dimension 2×1, 3×2 and 2×2 respectively.

For two matrices to be *equal* they must have the same number of elements arranged in exactly the same pattern and have the same elements in the same places. Clearly

$$\begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix} \neq \begin{pmatrix} 1 & 2 & 0 \\ 3 & 1 & 4 \end{pmatrix}.$$

If two matrices **A** and **B** have the same dimension, the *sum* of the two matrices is a matrix **C**, obtained by adding the corresponding elements of **A** and **B**. Thus, if

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ -3 & 4 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} -3 & 2 \\ 4 & 1 \end{pmatrix},$$

$$\mathbf{C} = \mathbf{A} + \mathbf{B} = \begin{pmatrix} -2 & 4 \\ 1 & 5 \end{pmatrix}.$$

In a similar manner,

$$\mathbf{D} = \mathbf{A} - \mathbf{B} = \begin{pmatrix} 4 & 0 \\ -7 & 3 \end{pmatrix}.$$

The sum and difference are not defined when **A** and **B** are not of the same dimension.

A matrix with all its elements zero is called a *zero matrix* and is usually denoted by **0**.

When a matrix is multiplied by a constant, the result is a matrix each of whose elements is equal to the corresponding element of the original matrix multiplied by the constant. Thus,

$$-3\begin{pmatrix}1 & 2\\ 3 & 4\end{pmatrix}=\begin{pmatrix}-3 & -6\\ -9 & -12\end{pmatrix}=\begin{pmatrix}-1 & -2\\ -3 & -4\end{pmatrix}3.$$

The *product of two matrices* is denoted by writing two matrices alongside each other *without* a multiplication sign. For example, **AB**. The product **AB** only exists if the number of columns in **A** is equal to the number of rows in **B**. The resulting product matrix **AB** has the same number of rows as **A** and the same number of columns as **B**. The element in the *r*th row and the *s*th column of **AB** is obtained by multiplying each element in the *r*th row of **A** by the corresponding element in the *s*th column of **B** and adding the products. Thus, the element in the second row and first column of the product

$$\text{second row}\rightarrow\begin{pmatrix}2 & 3 & 5\\ 1 & 9 & 2\\ 4 & 7 & 6\\ 14 & 12 & 13\end{pmatrix}\begin{pmatrix}-1 & -5\\ 12 & 11\\ 8 & 0\end{pmatrix}$$
$$\uparrow \text{first column}$$

is $1\times(-1)+9\times 12+2\times 8=123$.

In this case the first matrix is of dimension 4×3 and the second of 3×2. The product matrix is of dimension 4×2, and it takes the form

$$\begin{pmatrix}74 & 23\\ 123 & 94\\ 128 & 57\\ 234 & 62\end{pmatrix}.$$

This method of defining a matrix product may seem complicated, but it is extremely useful, because we often need sums of the products of pairs of numbers. Examples are given at the end of this section. Although the product **AB** may be defined, **BA** may not; the above example shows this. Furthermore, even when **AB** and **BA** are both defined, the two products are not in general equal.

A square matrix ($n\times n$ say) with ones down the *principal diagonal* and zeros elsewhere is called the *unit matrix* and is usually denoted by **I**. For example,

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

It is easy to verify that whenever the product is defined

$$\mathbf{AI}=\mathbf{IA}=\mathbf{A}.$$

A square matrix $\mathbf{A}$ is said to have an *inverse* $\mathbf{A}^{-1}$ if there exists $\mathbf{A}^{-1}$ such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}.$$

A square matrix without an inverse is said to be *singular.* Methods for determining the inverse of a matrix are given in section 1.10.

A matrix of dimension $1 \times n$ is often referred to as a *row vector* or *vector.* A matrix of dimension $n \times 1$ may be referred to as a *column vector* or *vector.* The above rules of addition, subtraction and multiplication still apply.

A square matrix with zeros everywhere except on the principal diagonal is called a *diagonal matrix.* If all the elements on the diagonal are non-zero, the matrix has an inverse, and the inverse is also diagonal. The inverse is obtained by taking the reciprocals of the diagonal elements. For example,

$$\begin{pmatrix} -2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix}^{-1} = \begin{pmatrix} -\frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{4} \end{pmatrix}.$$

The *transpose matrix* $\mathbf{A}'$ of a matrix $\mathbf{A}$ is obtained by interchanging rows and columns. Thus,

$$\mathbf{A} = \begin{pmatrix} 1 & 5 & 9 \\ 2 & 6 & 10 \\ 3 & 7 & 11 \\ 4 & 8 & 12 \end{pmatrix} \text{has transpose } \mathbf{A}' = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \end{pmatrix}.$$

A square matrix $\mathbf{A}$ is *symmetric* if $\mathbf{A}' = \mathbf{A}$. For example,

$$\begin{pmatrix} 2 & 5 & 6 \\ 5 & -3 & -7 \\ 6 & -7 & 4 \end{pmatrix}.$$

Example 1.8.1. One important use of matrix algebra is in the study of systems of linear equations. To see why, let us consider the matrix product

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 0 & 2 \\ 3 & -1 & -2 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix}.$$

The result is a 3×1 matrix or column vector

$$\begin{pmatrix} x + 2y + 3z \\ 4x \quad + 2z \\ 3x - y - 2z \end{pmatrix}.$$

If we equate this vector to the column vector

$$\begin{pmatrix} 5 \\ 8 \\ 0 \end{pmatrix},$$

we are in effect expressing in matrix notation the three simultaneous equations

$$\left.\begin{aligned} x + 2y + 3z &= 5, \\ 4x \qquad + 2z &= 8, \\ 3x - \quad y - 2z &= 0. \end{aligned}\right\}$$

Let us write

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 0 & 2 \\ 3 & -1 & -2 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 5 \\ 8 \\ 0 \end{pmatrix}.$$

A and **C** are known and **X** unknown. To solve the three simultaneous equations, we need to find **X** such that

$$\mathbf{AX} = \mathbf{C}.$$

If **A** has an inverse, we can premultiply both sides by $\mathbf{A}^{-1}$ and obtain

$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{C}.$$

We now have the solution to the three simultaneous linear equations. The reader should note that inverting a matrix is usually an inefficient way of obtaining the solution to a system of equations, but the short-hand matrix notation is very convenient for mathematical purposes.

Example 1.8.2. The cost of a new house depends upon the type of structure and the price of land, labour and materials. House *A* requires 10 000 units of land, 1000 units of labour and 2000 units of materials. House *B*, on the other hand, requires 7500 units of land, 1500 units of labour and 1500 units of materials. These components may be summarised in a 3 x 2 matrix.

House *A*	House *B*	
10 000	7 500	Land
1 000	1 500	Labour
2 000	1 500	Materials

The prices of the components depend upon the locality, and the unit prices (in dollars) in each of four localities are summarised in the following 4 x 3 matrix:

Land	Labour	Materials	
3	4	2	Locality 1
2	4	3	Locality 2
2	3	4	Locality 3
1	3	5	Locality 4

The family man does not want to know all this detail. He merely wants to know the total cost of each type of dwelling in each of the four localities. The total cost of House *A* in locality 2 may be calculated as follows:

$$2 \times 10\,000 + 4 \times 1000 + 3 \times 2000 = 30\,000.$$

Total costs for both types of house in all four localities may be summarised in a 4×2 matrix, and it is soon apparent that this matrix is the product of the above two matrices. Thus,

$$\begin{matrix} & \text{House } A & \text{House } B \\ \text{Locality 1} \\ \text{Locality 2} \\ \text{Locality 3} \\ \text{Locality 4} \end{matrix} \begin{pmatrix} 38\,000 & 31\,500 \\ 30\,000 & 25\,500 \\ 31\,000 & 25\,500 \\ 23\,000 & 19\,500 \end{pmatrix} = \begin{pmatrix} 3 & 4 & 2 \\ 2 & 4 & 3 \\ 2 & 3 & 4 \\ 1 & 3 & 5 \end{pmatrix} \begin{pmatrix} 10\,000 & 7\,500 \\ 1\,000 & 1\,500 \\ 2\,000 & 1\,500 \end{pmatrix}.$$

The first matrix on the right-hand side gives the unit cost of materials, etc. When it is multiplied by the second matrix (the number of units required), we obtain the total cost matrix.

Further reading: Conte [17] 144–6, 148–50, 152–4; Grossman and Turner [38] 105–35; Hartree [41] 152; Sokolnikoff and Sokolnikoff [92] 114–20.

1.9 Determinants and cofactors

Every square matrix **A** has a number associated with it known as its *determinant*. This may be written as det **A** or |**A**|. In the case of a 1×1 matrix, the determinant is merely the numerical value of the single element. For a 2×2 matrix, the determinant is equal to the product of the two diagonal elements minus the product of the two off-diagonal elements. Thus, for example,

$$\begin{vmatrix} 2 & 1 \\ 3 & 4 \end{vmatrix} = 2 \times 4 - 3 \times 1 = 5.$$

For matrices of higher dimension, we choose a row or column. (Any row or column will do.) We then multiply each element in that row or column by a number known as its cofactor and add the products. The sum is the determinant of the matrix.

To find the *cofactor* of the element in the ith row and jth column of an $n \times n$ matrix **A**, we delete the entire ith row and the entire jth column from that matrix to produce an $(n-1) \times (n-1)$ matrix and evaluate the determinant of this smaller matrix. This is the cofactor we require when $i+j$ is even. When $i+j$ is odd, the cofactor is obtained by changing the sign of this determinant.

Sometimes the complete *matrix of cofactors* of a matrix **A** is required. We then replace each element of **A** by its cofactor.

Example 1.9.1. The matrix of cofactors of the 2×2 matrix

$$\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 3 & 4 \end{pmatrix} \text{ is } \begin{pmatrix} 4 & -3 \\ -1 & 2 \end{pmatrix}.$$

A rule for calculating the determinant of a 2×2 matrix was given above. This rule is really only a special case of the more general rule for an $n \times n$ matrix.

Let us now calculate the determinant of **A** using the general rule.

By the first row: det **A** = $2\times 4+1\times(-3)=5$;
by the second row: det **A** = $3\times(-1)+4\times 2=5$;
by the first column: det **A** = $2\times 4+3\times(-1)=5$;
by the second column: det **A** = $1\times(-3)+4\times 2=5$.

The same answer is obtained whichever row or column is used.

Example 1.9.2. The matrix of cofactors of the 3 x 3 matrix

$$\mathbf{A}=\begin{pmatrix}-2 & 5 & 3\\ 4 & 7 & 1\\ 6 & 9 & 8\end{pmatrix} \text{ is}$$

$$\begin{pmatrix} \begin{vmatrix}7 & 1\\ 9 & 8\end{vmatrix} & -\begin{vmatrix}4 & 1\\ 6 & 8\end{vmatrix} & \begin{vmatrix}4 & 7\\ 6 & 9\end{vmatrix}\\ -\begin{vmatrix}5 & 3\\ 9 & 8\end{vmatrix} & \begin{vmatrix}-2 & 3\\ 6 & 8\end{vmatrix} & -\begin{vmatrix}-2 & 5\\ 6 & 9\end{vmatrix}\\ \begin{vmatrix}5 & 3\\ 7 & 1\end{vmatrix} & -\begin{vmatrix}-2 & 3\\ 4 & 1\end{vmatrix} & \begin{vmatrix}-2 & 5\\ 4 & 7\end{vmatrix} \end{pmatrix} = \begin{pmatrix}47 & -26 & -6\\ -13 & -34 & 48\\ -16 & 14 & -34\end{pmatrix}.$$

The determinant of **A** may be calculated using any row or column. For example,

by the second column: det **A** = $5\times(-26)+7\times(-34)+9\times 14=-242$;
by the third row: det **A** = $6\times(-16)+9\times 14+8\times(-34)=-242$.

The reader should verify that the same answer is obtained whichever row or column is used.

Note that it is not necessary to compute the complete matrix of cofactors to find the determinant of a matrix.

Further reading: Conte [17] 146–8; Grossman and Turner [38] 145–52; Scheid [89] 347–50; Sokolnikoff and Sokolnikoff [92] 102–9.

1.10 Matrix inversion

Sophisticated programs for matrix inversion are available on all computer systems and programs are also available on many desk machines. The reader is advised to make use of these programs. There are occasions, however, when it is convenient to invert a small matrix by hand. We now outline two methods by which the inverse of a non-singular square matrix may be obtained. Most general matrix inversion programs are similar to the first method we describe. The second method is useful for hand computation with small matrices, but it is inefficient (in terms of the number of operations required) for large matrices.

The matrix representation of a system of linear equations was described in example 1.8.1. Let us now look at a simple elimination procedure for solving a pair of simultaneous linear equations and deduce a method for inverting a