1

Introduction

1.1 Objects and Variables

Multivariate analysis deals with data containing observations on two or more variables, each measured on a set of objects. For example, we may have the set of examination marks achieved by certain students, or the cork deposit in various directions of a set of trees, or flower measurements for different species of iris (see Tables 1.2, 1.4, and 1.3, respectively). Each of these data has a set of "variables" (the examination marks, trunk thickness, and flower measurements) and a set of "objects" (the students, trees, and flowers). In general, if there are n objects, $o_1, \dots, o_n$ and p variables, $x_1, \dots, x_p$, the data contains np pieces of information. These may be conveniently arranged using an $(n \times p)$ "data matrix", in which each row corresponds to an object, and each column corresponds to a variable. For instance, three variables on five "objects" (students) are shown as a (5×3) data matrix in Table 1.1.

Note that all the variables need not be of the same type: in Table 1.1, x_1 is a "continuous" variable, x_2 is a discrete variable, and x_3 is a binary variable. Note also that attribute, characteristic, description, measurement, and response are synonyms for "variable", whereas individual, observation, plot, reading, item, and unit can be used in place of "object".

1.2 Some Multivariate Problems and Techniques

We may now illustrate various categories of multivariate technique.

1.2.1 Generalizations of Univariate Techniques

Most univariate questions are capable of at least one multivariate generalization. For instance, using Table 1.2, we may ask, as an example, "What is the appropriate underlying parent distribution of examination marks on various papers of a set of students?" "What are the summary statistics?" "Are the differences between average marks on different papers significant?", etc. These problems are direct generalizations of univariate problems, and their motivation is easy to grasp. See, for example, Chapters 2–7 and 13.

Multivariate Analysis, Second Edition. Kanti V. Mardia, John T. Kent, and Charles C. Taylor.

Table 1.1 Data matrix with five students as objects, where x_1 is age in years at entry to university, x_2 is marks out of 100 in an examination at the end of the first year, and x_3 is sex.

	Variables		
Objects	x_1	x_2	x_3
1	18.45	70	1
2	18.41	65	0
3	18.39	71	0
4	18.70	72	0
5	18.34	94	1

1 indicates male; 0 indicates female.

1.2.2 Dependence and Regression

The data in Table 1.2, which were collected at the University of Hull in the early 1970s, formed part of an investigation into the merits of open-book vs. closed-book examinations. Marks (out of 100) were given for 88 students on each of five subjects; these observations were sorted (almost) according to the average. Initially, we may enquire as to the degree of dependence between performance on different papers taken by the same students. It may be useful, for counseling or other purposes, to have some idea of how final degree marks ("dependent" variables) are affected by previous examination results or by other variables such as age and sex ("explanatory" variables). This presents the so-called regression problem, which is examined in Chapter 7.

1.2.3 Linear Combinations

Given examination marks on different topics (as in Table 1.2), the question arises of how to combine or average these marks in a suitable way. A straightforward method would use the simple arithmetic mean, but this procedure may not always be suitable. For instance, if the marks on some papers vary more than others, we may wish to weight them differently. This leads us to search for a linear combination (weighted sum) which is "optimal" in some sense. If all the examination papers fall in one group, then *principal component analysis* and *factor analysis* are two techniques that can help to answer such questions (see Chapters 9 and 10). In some situations, the papers may fall into more than one group – for instance, in Table 1.2, some examinations were "open book", while others were "closed book". In such situations, we may wish to investigate the use of linear combinations within each group separately. This leads to the method known as *canonical correlation analysis*, which is discussed in Chapter 11.

The idea of taking linear combinations is an important one in multivariate analysis, and we will return to it in Section 1.5.

Table 1.2 Marks in open- and closed-book examination out of 100.

Mechanics (C)	Vectors (C)	Algebra (O)	Analysis (O)	Statistics (O)
77	82	67	67	81
63	78	80	70	81
75	73	71	66	81
55	72	63	70	68
63	63	65	70	63
53	61	72	64	73
51	67	65	65	68
59	70	68	62	56
62	60	58	62	70
64	72	60	62	45
52	64	60	63	54
55	67	59	62	44
50	50	64	55	63
65	63	58	56	37
31	55	60	57	73
60	64	56	54	40
44	69	53	53	53
42	69	61	55	45
62	46	61	57	45
31	49	62	63	62
44	61	52	62	46
49	41	61	49	64
12	58	61	63	67
49	53	49	62	47
54	49	56	47	53
54	53	46	59	44
44	56	55	61	36
18	44	50	57	81
46	52	65	50	35
32	45	49	57	64
30	69	50	52	45
46	49	53	59	37
40	27	54	61	61
31	42	48	54	68
36	59	51	45	51
56	40	56	54	35

(Continued)

Table 1.2 (Continued)

Mechanics (C)	Vectors (C)	Algebra (O)	Analysis (O)	Statistics (O)
46	56	57	49	32
45	42	55	56	40
42	60	54	49	33
40	63	53	54	25
23	55	59	53	44
48	48	49	51	37
41	63	49	46	34
46	52	53	41	40
46	61	46	38	41
40	57	51	52	31
49	49	45	48	39
22	58	53	56	41
35	60	47	54	33
48	56	49	42	32
31	57	50	54	34
17	53	57	43	51
49	57	47	39	26
59	50	47	15	46
37	56	49	28	45
40	43	48	21	61
35	35	41	51	50
38	44	54	47	24
43	43	38	34	49
39	46	46	32	43
62	44	36	22	42
48	38	41	44	33
34	42	50	47	29
18	51	40	56	30
35	36	46	48	29
59	53	37	22	19
41	41	43	30	33
31	52	37	27	40
17	51	52	35	31
34	30	50	47	36
46	40	47	29	17
10	46	36	47	39

Table 1.2 (Continued)

Mechanics (C)	Vectors (C)	Algebra (O)	Analysis (O)	Statistics (O)
46	37	45	15	30
30	34	43	46	18
13	51	50	25	31
49	50	38	23	09
18	32	31	45	40
08	42	48	26	40
23	38	36	48	15
30	24	43	33	25
03	09	51	47	40
07	51	43	17	22
15	40	43	23	18
15	38	39	28	17
05	30	44	36	18
12	30	32	35	21
05	26	15	20	20
00	40	21	09	14

O indicates open book, and C indicates closed book.

1.2.4 Assignment and Dissection

Table 1.3 gives three (50×4) data matrices (or one (150×5) data matrix if the species is coded as a variable). In each matrix, the "objects" are 50 irises of species *Iris setosa, Iris versicolor*, and *Iris virginica*, respectively. The "variables" are

$$\begin{aligned} x_1 &= \text{sepal length}, & x_2 &= \text{sepal width}, \\ x_3 &= \text{petal length}, & x_4 &= \text{petal width}. \end{aligned}$$

The flowers of the first two iris species (*I. setosa* and *I. versicolor*) were taken from the same natural colony but the sample of the third iris species (*I. virginica*) is from a different colony; for more general details on the data, see Mardia (2023). If a new iris of unknown species has measurements $x_1 = 5.1,\ x_2 = 3.2,\ x_3 = 2.7$, and $x_4 = 0.7$, we may ask to which species it belongs. This presents the problem of *discriminant analysis*, which is discussed in Chapter 12. However, if we were presented with the 150 observations of Table 1.3 in an unclassified manner (say, before the three species were established), then the aim could have been to dissect the population into homogeneous groups. This problem is handled by *cluster analysis* (see Chapter 14).

1.2.5 Building Configurations

In some cases, the data consists not of an $(n \times p)$ data matrix but of $n(n-1)/2$ "distances" between all pairs of points. To get an intuitive feel for the structure of such data, a

Table 1.3 Measurements (in cm) on three types of irises.

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	**Sepal width**	**Petal length**	**Petal width**	**Sepal length**	**Sepal width**	**Petal length**	**Petal width**	**Sepal length**	**Sepal width**	**Petal length**	**Petal width**
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5

Table 1.3 (Continued)

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	**Sepal width**	**Petal length**	**Petal width**	**Sepal length**	**Sepal width**	**Petal length**	**Petal width**	**Sepal length**	**Sepal width**	**Petal length**	**Petal width**
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

Source: Fisher (1936) / John Wiley & Sons.

configuration can be constructed of n points in a Euclidean space of low dimension (e.g. $p = 2$ or 3). Hopefully, the distances between the n points of this configuration will closely match the original distances. The problems of building and interpreting such configurations are studied in Chapter 15, on *multidimensional scaling*.

1.3 The Data Matrix

The general $(n \times p)$ data matrix with n objects and p variables can be written as follows:

$$\begin{array}{ll}
 & \overbrace{\begin{array}{ccccc} x_1 & \cdots & x_j & \cdots & x_p, \end{array}}^{\text{Variables}} \\
\text{Objects}\left\{\begin{array}{l} o_1 \\ \vdots \\ o_r \\ \vdots \\ o_n \end{array}\right. &
\begin{array}{ccccc} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{r1} & \cdots & x_{rj} & \cdots & x_{rp}. \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{array}
\end{array}$$

Table 1.1 shows one such data matrix (5×3) with five objects and three variables. In Table 1.3, there are three data matrices, each having 50 rows (objects) and 4 columns (variables). Note that these data matrices can be considered from two alternative points of view. If we compare two columns, then we are examining the relationship between variables. On the other hand, a comparison of two rows involves examining the relationship between different objects. For example, in Table 1.1, we may compare the first two columns to investigate whether there is a relationship between age at entry and marks obtained. Alternatively, looking at the first two rows will give a comparison between two students ("objects"), one male and one female.

The general $(n \times p)$ data matrix will be denoted $\boldsymbol{X}$ or $\boldsymbol{X}(n \times p)$. The element in row r and column j is x_{rj}. This denotes the observation of the jth variable on the rth object. We may write the matrix $\boldsymbol{X} = (x_{rj})$. The rows of $\boldsymbol{X}$ will be written $\boldsymbol{x}'_1, \boldsymbol{x}'_2, \dots, \boldsymbol{x}'_n$. Note that $\boldsymbol{x}_r$ denotes the rth row of $\boldsymbol{X}$ *written as a column*. The columns of $\boldsymbol{X}$ will be written with subscripts in parentheses as $\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)}, \dots, \boldsymbol{x}_{(p)}$; that is, we may write

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}'_1 \\ \vdots \\ \boldsymbol{x}'_n \end{bmatrix} = [\boldsymbol{x}_{(1)} \;\cdots\; \boldsymbol{x}_{(p)}],$$

where

$$\boldsymbol{x}_r = \begin{pmatrix} x_{r1} \\ \vdots \\ x_{rp} \end{pmatrix} \; r = 1, \dots, n \qquad \text{and} \qquad \boldsymbol{x}_{(j)} = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix} \; j = 1, \dots, p.$$

In general, we use square brackets for matrices and round brackets for vectors.

Note that on the one hand $\boldsymbol{x}_1$ is the p-vector denoting the p observations on the first *object*, while on the other hand $\boldsymbol{x}_{(1)}$ is the n-vector whose elements denote the observations on the first *variable*. In multivariate analysis, the rows $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ usually form a random sample, whereas the columns $\boldsymbol{x}_{(1)}, \dots, \boldsymbol{x}_{(p)}$ do not. This point is emphasized in the notation by the use of parentheses.

Clearly when n and p are even moderately large, the resulting np pieces of information may prove too numerous to handle individually. Various way of summarizing multivariate data are discussed in Sections 1.4, 1.7, and 1.8.

1.4 Summary Statistics

We give here the basic summary statistics and some standard notation.

1.4.1 The Mean Vector and Covariance Matrix

An obvious extension of the univariate notion of mean and variance leads to the following definitions. The sample mean of the jth variable is

$$\bar{x}_j = \frac{1}{n}\sum_{r=1}^{n} x_{rj}, \tag{1.1}$$

and the sample variance of the jth variable is

$$s_{jj} = \frac{1}{n}\sum_{r=1}^{n}(x_{rj} - \bar{x}_j)^2 = s_j^2, \quad \text{say,} \quad j = 1, \dots, p. \tag{1.2}$$

The sample covariance between the ith and jth variables is

$$s_{ij} = \frac{1}{n}\sum_{r=1}^{n}(x_{ri} - \bar{x}_i)(x_{rj} - \bar{x}_j). \tag{1.3}$$

The vector of means,

$$\bar{\boldsymbol{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix}, \tag{1.4}$$

is called the *sample mean vector*, or simply the "mean vector". It represents the center of gravity of the sample points $\boldsymbol{x}_r,\ r = 1, \dots, n$. The $p \times p$ matrix

$$\boldsymbol{S} = (s_{ij}), \tag{1.5}$$

with elements given by (1.2) and (1.3), is called the *sample covariance matrix*, or simply the "covariance matrix".

The above statistics may also be expressed in matrix notation. Corresponding to (1.1) and (1.4), we have

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{r=1}^{n}\boldsymbol{x}_r = \frac{1}{n}\boldsymbol{X}'\boldsymbol{1}, \tag{1.6}$$

where $\boldsymbol{1}$ is a column vector of n ones. Also,

$$s_{ij} = \frac{1}{n}\sum_{r=1}^{n}x_{ri}x_{rj} - \bar{x}_i\,\bar{x}_j,$$

so that

$$\boldsymbol{S} = \frac{1}{n}\sum_{r=1}^{n}(\boldsymbol{x}_r - \bar{\boldsymbol{x}})(\boldsymbol{x}_r - \bar{\boldsymbol{x}})' = \frac{1}{n}\sum_{r=1}^{n}\boldsymbol{x}_r\boldsymbol{x}_r' - \bar{\boldsymbol{x}}\,\bar{\boldsymbol{x}}'. \tag{1.7}$$

This may also be written as

$$\boldsymbol{S} = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X} - \bar{\boldsymbol{x}}\,\bar{\boldsymbol{x}}' = \frac{1}{n}(\boldsymbol{X}'\boldsymbol{X} - \frac{1}{n}\boldsymbol{X}'\boldsymbol{1}\boldsymbol{1}'\boldsymbol{X}),$$

using (1.6). Writing

$$\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}',$$

where $\boldsymbol{H}$ denotes the *centering matrix*, we find that

$$\boldsymbol{S} = \frac{1}{n}\ \boldsymbol{X}'\boldsymbol{H}\boldsymbol{X}, \tag{1.8}$$

which is a convenient matrix representation of the sample covariance matrix.

Since $\boldsymbol{H}$ is a symmetric idempotent matrix ($\boldsymbol{H} = \boldsymbol{H}', \boldsymbol{H} = \boldsymbol{H}^2$), it follows that for any p-vector $\boldsymbol{a}$,

$$\boldsymbol{a}'\boldsymbol{S}\boldsymbol{a} = \frac{1}{n}\ \boldsymbol{a}'\boldsymbol{X}\boldsymbol{H}'\boldsymbol{H}\boldsymbol{X}\boldsymbol{a} = \frac{1}{n}\ \boldsymbol{y}'\boldsymbol{y} \geq 0,$$

where $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{X}\boldsymbol{a}$. Hence, the covariance matrix $\boldsymbol{S}$ is positive semidefinite ($\boldsymbol{S} \geq 0$). For continuous data, we usually expect that $\boldsymbol{S}$ is not only positive semidefinite but positive definite if $n \geq p + 1$; see Table A.7 for more specific definitions.

As in one-dimensional statistics, it is often convenient to define the covariance matrix with a divisor of $n-1$ instead of n. Set

$$\boldsymbol{S}_u = \frac{1}{n-1}\boldsymbol{X}'\boldsymbol{H}\boldsymbol{X} = \frac{n}{n-1}\boldsymbol{S}. \tag{1.9}$$

If the data forms a random sample from a multivariate distribution, with finite second moments, then $\boldsymbol{S}_u$ is an *unbiased* estimate of the true covariance matrix (see Theorem 2.11).

The matrix

$$\boldsymbol{M} = \sum_{r=1}^{n}\boldsymbol{x}_r\boldsymbol{x}_r' = \boldsymbol{X}'\boldsymbol{X} \tag{1.10}$$

is called the *matrix of sums of squares and products* for obvious reasons. The matrix $n\boldsymbol{S}$ can appropriately be labeled as the matrix of *corrected* sums of squares and products.

The *sample correlation coefficient* between the ith and the jth variables is

$$r_{ij} = s_{ij}/(s_i s_j). \tag{1.11}$$

Unlike s_{ij}, the correlation coefficient is invariant under both changes of scale and origin of the ith and jth variables. Clearly, $|r_{ij}| \leq 1$. The matrix

$$\boldsymbol{R} = (r_{ij}) \tag{1.12}$$

with $r_{ii} = 1$ is called the *sample correlation matrix*. Note that $\boldsymbol{R} \geq 0$ (i.e. $\boldsymbol{R}$ is positive semidefinite). If $\boldsymbol{R} = \boldsymbol{I}$, we say that the variables are uncorrelated. If $\boldsymbol{D} = \operatorname{diag}(s_i)$, then

$$\boldsymbol{R} = \boldsymbol{D}^{-1}\boldsymbol{S}\boldsymbol{D}^{-1}, \quad \boldsymbol{S} = \boldsymbol{D}\boldsymbol{R}\boldsymbol{D}. \tag{1.13}$$

Example 1.4.1 Table 1.4 gives a (28×4) data matrix (Rao, 1948) related to weights of bark deposits of 28 trees in the four directions north (N), east (E), south (S), and west (W).

It is found that

$$\bar{x}_1 = 50.536, \qquad \bar{x}_2 = 46.179, \qquad \bar{x}_3 = 49.679, \qquad \bar{x}_4 = 45.179.$$

These means suggest *prima-facie* differences with more deposits in the N–S directions than in the E–W directions. The covariance and correlation matrices are

$$\boldsymbol{S} = \begin{array}{c} \\ \text{N} \\ \text{E} \\ \text{S} \\ \text{W} \end{array} \begin{array}{c} \begin{array}{cccc} \text{N} & \text{E} & \text{S} & \text{W} \end{array} \\ \begin{bmatrix} 280.034 & 215.761 & 278.136 & 218.190 \\ & 212.075 & 220.879 & 165.254 \\ & & 337.504 & 250.272 \\ & & & 217.932 \end{bmatrix} \end{array},$$

$$\boldsymbol{R} = \begin{array}{c} \\ \text{N} \\ \text{E} \\ \text{S} \\ \text{W} \end{array} \begin{array}{c} \begin{array}{cccc} \text{N} & \text{E} & \text{S} & \text{W} \end{array} \\ \begin{bmatrix} 1 & 0.885 & 0.905 & 0.883 \\ & 1 & 0.826 & 0.769 \\ & & 1 & 0.923 \\ & & & 1 \end{bmatrix} \end{array}.$$

Since these matrices are symmetric, only the upper half needs to be shown.

Table 1.4 Weights of cork deposits (in centigrams) for 28 trees in the four directions.

N	E	S	W	N	E	S	W
72	66	76	77	91	79	100	75
60	53	66	63	56	68	47	50
56	57	64	58	79	65	70	61
41	29	36	38	81	80	68	58
32	32	35	36	78	55	67	60
30	35	34	26	46	38	37	38
39	39	31	27	39	35	34	37
42	43	31	25	32	30	30	32
37	40	31	25	60	50	67	54
33	29	27	36	35	37	48	39
32	30	34	28	39	36	39	31
63	45	74	63	50	34	37	40
54	46	60	52	43	37	39	50
47	51	52	43	48	54	57	43

Source: Rao (1948) / Oxford University Press.

Comparing the diagonal terms of $\boldsymbol{S}$, we note that the sample variance is largest in the south direction. Furthermore, the matrix $\boldsymbol{R}$ does not seem to have a "circular" pattern; e.g. the correlation between N and S is relatively high, while the correlation between the other pair of opposite cardinal points W and E is lowest. □

1.4.2 Measures of Multivariate Scatter

The matrix $\boldsymbol{S}$ is one possible multivariate generalization of the univariate notion of variance, measuring scatter about the mean. However, sometimes it is convenient to have a *single* number to measure multivariate scatter. Two common such measures are

1. the *generalized variance*, $|\boldsymbol{S}|$ (the determinant of $\boldsymbol{S}$) and
2. the *total variation*, $\operatorname{tr} \boldsymbol{S}$ (the sum of the diagonal entries of $\boldsymbol{S}$).

A motivation for these measures is given in Section 1.5.3. For both measures, large values indicate a high degree of scatter about $\bar{\boldsymbol{x}}$, and low values represent concentration about $\bar{\boldsymbol{x}}$. However, each measure reflects different aspects of the variability in the data. The generalized variance plays an important role in maximum-likelihood estimation (Chapter 6), and the total variation is a useful concept in principal component analysis (Chapter 9).

Example 1.4.2 (Gnanadesikan and Gupta, 1970) An experimental subject spoke 10 different words seven times each, and five speech measurements were taken on each utterance.

For each word, we have five variables and seven observations. The (5 × 5) covariance matrix was calculated for each word, and the 10 generalized variances were as follows:

$$2.9,\quad 1.3,\quad 641.6,\quad 26\,828.8,\quad 262\,404.3,$$

$$169.2,\quad 3106.8,\quad 617\,671.2\quad 6.7\quad 3.0.$$

Ordering these generalized variances, we find that for this speaker the second word has the least variation and the eighth word has most variation. A general point of interest for identification is to find which word has least variation for a particular speaker. □

1.5 Linear Combinations

Taking linear combinations of the variables is one of the most important tools of multivariate analysis. A few suitably chosen combinations may provide more information than a multiplicity of the original variables, often because the dimension of the data is reduced. Linear transformations can also simplify the structure of the covariance matrix, making interpretation of the data more straightforward.

Consider a linear combination

$$y_r = a_1 x_{r1} + \cdots + a_p x_{rp}, \quad r = 1, \ldots, n, \tag{1.14}$$

where $a_1, \ldots, a_p$ are given. From (1.6), the mean $\bar{y}$ of the y_r is given by

$$\bar{y} = \frac{1}{n}\boldsymbol{a}' \sum_{r=1}^{n} \boldsymbol{x}_r = \boldsymbol{a}'\bar{\boldsymbol{x}}, \tag{1.15}$$

and the variance is given by

$$s_y^2 = \frac{1}{n}\sum_{r=1}^{n}(y_r - \bar{y})^2 = \frac{1}{n}\sum_{r=1}^{n}\boldsymbol{a}'(\boldsymbol{x}_r - \bar{\boldsymbol{x}})(\boldsymbol{x}_r - \bar{\boldsymbol{x}})'\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{S}\boldsymbol{a}, \tag{1.16}$$

where we have used (1.7).

In general, we may be interested in a q-dimensional linear transformation,

$$\boldsymbol{y}_r = \boldsymbol{A}\boldsymbol{x}_r + \boldsymbol{b}, \quad r = 1, \ldots, n, \tag{1.17}$$

which may be written as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{A}' + \boldsymbol{1}\boldsymbol{b}'$, where $\boldsymbol{A}$ is a $(q \times p)$ matrix, and $\boldsymbol{b}$ is a q-vector. Usually, $q \leq p$.

The mean vector and covariance matrix of the new objects $\boldsymbol{y}_r$ are given by

$$\bar{\boldsymbol{y}} = \boldsymbol{A}\bar{\boldsymbol{x}} + \boldsymbol{b}, \tag{1.18}$$

$$\boldsymbol{S}_y = \frac{1}{n}\sum_{r=1}^{n}(\boldsymbol{y}_r - \bar{\boldsymbol{y}})(\boldsymbol{y}_r - \bar{\boldsymbol{y}})' = \boldsymbol{A}\boldsymbol{S}\boldsymbol{A}'. \tag{1.19}$$

If $\boldsymbol{A}$ is nonsingular (so, in particular, $q = p$), then

$$\boldsymbol{S} = \boldsymbol{A}^{-1}\boldsymbol{S}_y(\boldsymbol{A}')^{-1}. \tag{1.20}$$

Here are several important examples of linear transformations that are used later in the book. For simplicity, all of the transformations are centered to have mean $\boldsymbol{0}$.

1.5.1 The Scaling Transformation

Let $\boldsymbol{y}_r = \boldsymbol{D}^{-1}(\boldsymbol{x}_r - \overline{\boldsymbol{x}}),\ r = 1, \dots, n,$ where $\boldsymbol{D} = \text{diag}(s_i)$. This transformation scales each variable to have unit variance and thus eliminates the arbitrariness in the choice of scale. For example, if $\boldsymbol{x}_{(1)}$ measures lengths, then $\boldsymbol{y}_{(1)}$ will be the same whether $\boldsymbol{x}_{(1)}$ is measured in inches or meters. Note that $\boldsymbol{S}_y = \boldsymbol{R}$.

1.5.2 Mahalanobis Transformation

If $\boldsymbol{S} > 0$, then $\boldsymbol{S}^{-1}$ has a unique symmetric positive-definite square root $\boldsymbol{S}^{-1/2}$ (see (A.66)). The Mahalanobis transformation is defined by

$$\boldsymbol{z}_r = \boldsymbol{S}^{-1/2}(\boldsymbol{x}_r - \overline{\boldsymbol{x}}), \quad r = 1, \dots, n. \tag{1.21}$$

Then, $\boldsymbol{S}_z = \boldsymbol{I}$, so that this transformation eliminates the correlation between the variables and standardizes the variance of each variable.

1.5.3 Principal Component Transformation

By the spectral decomposition theorem, the covariance matrix $\boldsymbol{S}$ may be written in the form

$$\boldsymbol{S} = \boldsymbol{G}\boldsymbol{L}\boldsymbol{G}', \tag{1.22}$$

where $\boldsymbol{G}$ is an orthogonal matrix, and $\boldsymbol{L}$ is a diagonal matrix of the eigenvalues of $\boldsymbol{S}$, $l_1 \geq l_2 \geq \cdots \geq l_p \geq 0$. The principal component transformation is defined by the *rotation*

$$\boldsymbol{w}_r = \boldsymbol{G}'(\boldsymbol{x}_r - \overline{\boldsymbol{x}}), \quad r = 1, \dots, n. \tag{1.23}$$

Since $\boldsymbol{S}_w = \boldsymbol{G}'\boldsymbol{S}\boldsymbol{G} = \boldsymbol{L}$ is diagonal, the columns of $\boldsymbol{W}$, called principal components, represent *uncorrelated* linear combinations of the variables. In practice, one hopes to summarize most of the variability in the data using only the principal components with the highest variances and then reducing the dimension.

Since the principal components are uncorrelated with variances $l_1, \dots, l_p$, it seems natural to define the "overall" spread of the data by some symmetric monotonically increasing function of $l_1, \dots, l_p$, such as $\prod l_i$ and $\sum l_i$. From Section A.6, $|\boldsymbol{S}| = |\boldsymbol{L}| = \prod l_i$ and $\text{tr}\,\boldsymbol{S} = \text{tr}\,\boldsymbol{L} = \sum l_i$. Thus, the rotation to principal components provides a motivation for the measures of multivariate scatter discussed in Section 1.4.2.

Note that alternative versions of the transformations in Sections 1.5.1–1.5.3 can be defined using $\boldsymbol{S}_u$ instead of $\boldsymbol{S}$.

Example 1.5.1 *A transformation of the cork data* If in the cork data of Table 1.4, the aim is to investigate whether the bark deposits are uniformly spread, our interest would be in linear combinations such as

$$y_1 = \text{N} + \text{S} - \text{E} - \text{W}, \qquad y_2 = \text{N} - \text{S}, \qquad y_3 = \text{E} - \text{W}.$$

Here,

$$\boldsymbol{A} = \begin{bmatrix} 1 & -1 & 1 & -1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}.$$

From Example 1.4.1 and Equation (1.18), we find that the mean vector of the transformed variables is

$$\bar{\boldsymbol{y}}' = (8.857,\ 0.857,\ 1.000),$$

and the covariance matrix is

$$\begin{bmatrix} 124.12 & -20.27 & -25.96 \\ & 61.27 & 26.96 \\ & & 99.50 \end{bmatrix}.$$

Obviously, the mean of y_1 is higher than that of y_2 and y_3, indicating possibly more cork deposit along the north–south axis than along the east–west axis. However, $\text{var}(y_1)$ is larger. If we standardize the variables so that the sum of squares of the coefficients is unity, by letting

$$z_1 = (\text{N} + \text{S} - \text{E} - \text{W})/2, \qquad z_2 = (\text{N} - \text{S})/\sqrt{2}, \qquad z_3 = (\text{E} - \text{W})/\sqrt{2},$$

the variances are more similar:

$$\text{var}(z_1) = 31.03, \qquad \text{var}(z_2) = 30.63, \qquad \text{var}(z_3) = 49.75.$$

□

1.6 Geometrical Ideas

In Section 1.3, we mentioned the two alternative perspectives that can be used to view the data matrix. On the one hand, we may be interested in comparing the *columns* of the data matrix, that is the *variables*. This leads to the techniques known as *R-techniques*, so-called because the correlation matrix $\boldsymbol{R}$ plays an important role in this approach. R-techniques are important in principal component analysis, factor analysis, and canonical correlation analysis.

Alternatively, we may compare the *rows* of the data matrix, that is the different *objects*. This leads to techniques such as discriminant analysis, cluster analysis, and multidimensional scaling, which are known as *Q-techniques*.

These two approaches correspond to different geometric ways of representing the $(n \times p)$ data matrix. First, the columns can be viewed as p points in an n-dimensional space, which we call *R-space* or *object space*. For example, the correlation matrix has a natural interpretation in the object space of the centered matrix $\boldsymbol{Y} = \boldsymbol{HX}$. The correlation r_{ij} is just the cosine of the angle θ_{ij} subtended at the origin between the two corresponding columns,

$$\cos\theta_{ij} = \frac{\|\boldsymbol{y}'_{(i)}\|\ \|\boldsymbol{y}_{(j)}\|}{\|\boldsymbol{y}_{(i)}\|\ \|\boldsymbol{y}_{(j)}\|} = \frac{s_{ij}}{s_i s_j} = \boldsymbol{r}_{ij}.$$

Note that the correlation coefficients are a measure of similarity because their values are large when variables are close to one another.

Second, the n rows may be taken as n points in p-dimensional *Q-space* or *variable space*. A natural way to compare two rows $\boldsymbol{x}_r$ and $\boldsymbol{x}_s$ is to look at the Euclidean distance between them:

$$\|\boldsymbol{x}_r - \boldsymbol{x}_s\|^2 = (\boldsymbol{x}_r - \boldsymbol{x}_s)'(\boldsymbol{x}_r - \boldsymbol{x}_s).$$

An alternative procedure is to transform the data by one of the transformations of Sections 1.5.1 or 1.5.2 and then look at the Euclidean distance between the transformed rows. Such distances play a role in cluster analysis. The most important of these distances is the *Mahalanobis distance* D_{rs}, given by

$$D_{rs}^2 = \|\boldsymbol{z}_r - \boldsymbol{z}_s\|^2 = (\boldsymbol{x}_r - \boldsymbol{x}_s)'\boldsymbol{S}^{-1}(\boldsymbol{x}_r - \boldsymbol{x}_s). \tag{1.24}$$

The Mahalanobis distance underlies Hotelling's T^2 test and the theory of discriminant analysis. Note that the Mahalanobis distance can alternatively be defined using $\boldsymbol{S}_u$ instead of $\boldsymbol{S}$.

1.7 Graphical Representation

1.7.1 Univariate Scatters

For $p = 1$ and $p = 2$, we can draw and interpret scatter plots for the data, but for $p = 3$ the difficulties of drawing such plots can be appreciated. For $p > 3$, the task becomes hopeless, although computer facilities exist, which allow one to examine the projection of a multivariate scatter onto any given plane; see Wickham et al. (2015) and the references therein. However, the need for graphical representations when $p > 3$ is greater than for the univariate case since the relationships cannot be understood by looking at a data matrix. A simple starting point is to look into univariate plots for the p variables side by side. For the cork data of Table 1.4, such a representation is given in Figure 1.1, which indicates that the distributions are somewhat skew. In this case, the variables are measured in the same units, and therefore, a direct comparison of the plots is possible. In general, we should standardize the variables before using such a plot.

Figure 1.1 does not give any idea of relationships between the variables. However, one way to exhibit the interrelationships is to plot all the observations consecutively along the x-axis, representing different variables by different symbols. For the cork data, a plot of this type is given in Figure 1.2. It shows the very noticeable tree differences as well as differences in pattern that are associated with the given ordering of the trees. (That is, the experimenters appear to have chosen groups of small- and large-trunked trees alternately.) We also observe that the 15th tree may be an outlier. These features due to ordering

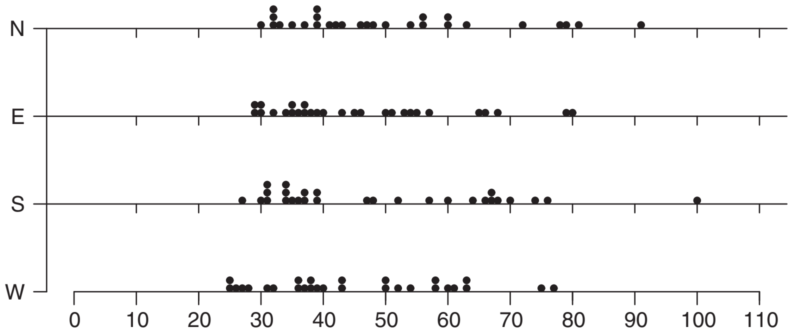

Figure 1.1 Univariate representation of the cork data of Table 1.4.

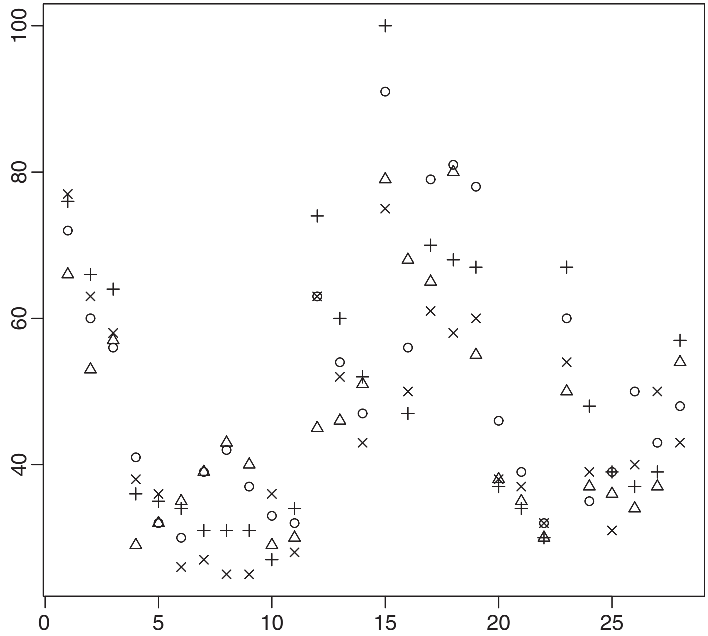

Figure 1.2 Consecutive univariate representation. Source: Adapted from Pearson (1956). ○ = N, △ = E, + = S, × = W.

would remain if the Mahalanobis residuals defined by (1.21) were plotted. Of course, the Mahalanobis transformation removes the main effect and adjusts every variance to unity and every covariance to zero.

1.7.2 Bivariate Scatters

Another way of understanding the interrelationships is to look into all $p(p-1)/2$ bivariate scatter diagrams of the data. For the cork data, with four variables, there are six such diagrams, and these could be presented as a *matrix* of simple scatterplots. Since the information is repeated in the upper and lower diagonals, there are various possibilities for these panels. One such example is shown in Figure 1.3 for the cork data.

Another method of graphing four variables in two dimensions is as follows. First, draw a scatter plot for two variables, say (N,E). Then, indicate the values of S and W of a point by plotting the values in two directions from the point (N,E). Figure 1.4 gives such a plot where W is plotted along the negative x-axis, and S is plotted along the negative y-axis using each point (N,E) as the origin. The markedly linear scatter of the points shows a strong dependence between N and E. A similar dependence between S and W is shown by the similar lengths of the two "legs" of each point. Dependence between N and S is reflected by the S legs being longer for large N, and a similar pattern is observed between N and W, E and

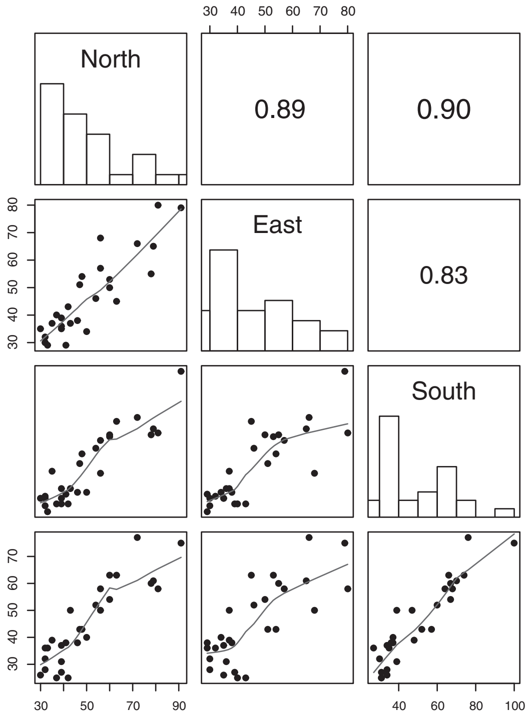

Figure 1.3 Matrix of scatterplots for the cork data. The upper diagonal shows the correlations (with text size proportional to the correlation), the lower diagonal shows a scatterplot (with a fitted smooth line), and the diagonal panels show the histograms.

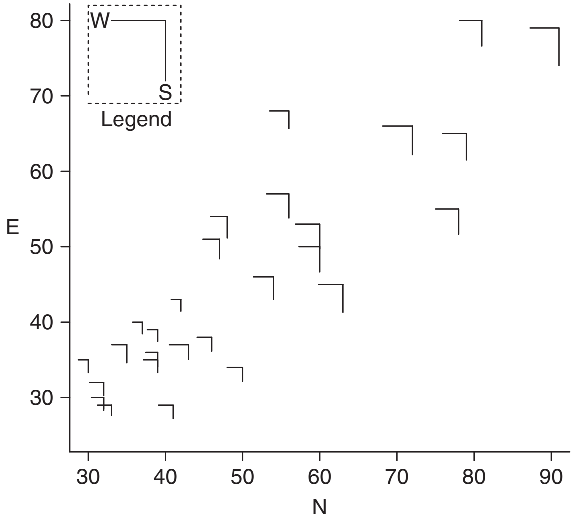

Figure 1.4 A glyph representation of the cork data of Table 1.4.

W, and E and S. This method of *glyphs* (Anderson, 1960) can be extended to several variables, but it does not help in understanding the multivariate complex as a whole. A related method due to Chernoff (1973) uses similar ideas to represent the observations on each object by a human face; see Mardia (2023) for an example using the iris data.

1.7.3 Harmonic Curves

Consider the rth data point $\boldsymbol{x}'_r = (x_{r1}, \dots, x_{rp})$, $r = 1, \dots, n$. An interesting method (Andrews, 1972; Ball and Hall, 1970) involves plotting the curve

$$f_{\boldsymbol{x}_r}(t) = \frac{x_{r1}}{\sqrt{2}} + x_{r2}\sin t + x_{r3}\cos t + \cdots + \begin{cases} x_{rp}\sin[pt/2] & p \text{ even} \\ x_{rp}\cos[(p-1)t/2] & p \text{ odd} \end{cases} \tag{1.25}$$

for each data point $\boldsymbol{x}_r$, $r = 1, \dots, n$, over the interval $-\pi \le t \le \pi$. Thus, there will be n harmonic curves drawn in two dimensions. Two data points are compared by visually studying the curves over $[-\pi, \pi]$. Note that the square of the L_2 distance

$$\int_{-\pi}^{\pi} [f_{\boldsymbol{x}}(t) - f_{\boldsymbol{y}}(t)]^2 \, \mathrm{d}t$$

between two curves $f_{\boldsymbol{x}}(t)$ and $f_{\boldsymbol{y}}(t)$ simplifies to

$$\pi \|\boldsymbol{x} - \boldsymbol{y}\|^2,$$

which is proportional to the square of the Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{y}$.

Some practical hints for the plotting of harmonic curves with a large number of objects are given in Gnanadesikan (1977). Here, we use a small simulated data set to illustrate their use. In this data set, there are $n = 20$ observations with $p = 11$ variables. The means and standard deviations of the first 15 (group 1) observations and the remaining 5 (group 2) observations are given in Table 1.5. It can be noted that these differ mainly in the mean values of the last three variables.

The harmonic curves are plotted in Figure 1.5, for the original data, the data with each value increased by 10, the standardization $(\boldsymbol{x}_{(j)} - \bar{x}_j)/s_j$, $j = 1, \dots, p$, where $\bar{x}_j$ is the mean of the jth column $\boldsymbol{x}_{(j)}$, and a random permutation of the columns. Given that harmonic curves depend on the order in which the variables are written down, as well as the location and

Table 1.5 Sample means $\bar{\boldsymbol{x}}$ and standard deviations (s_j) for the two groups of data, with $n_1 = 15$ and $n_2 = 5$ observations.

	vars	1	2	3	4	5	6	7	8	9	10	11
Group 1	Mean	0.4	0.4	−0.1	−0.3	0.2	−0.4	0.0	0.2	0.1	0.0	0.1
	SD	0.9	1.2	1.0	0.6	1.2	1.3	1.1	0.8	0.8	1.1	1.3
Group 2	Mean	0.7	0.0	0.1	0.2	0.1	−0.4	−0.2	0.5	5.9	5.0	5.1
	SD	1.2	0.6	0.9	1.3	0.8	0.5	1.2	1.4	1.0	1.0	0.6

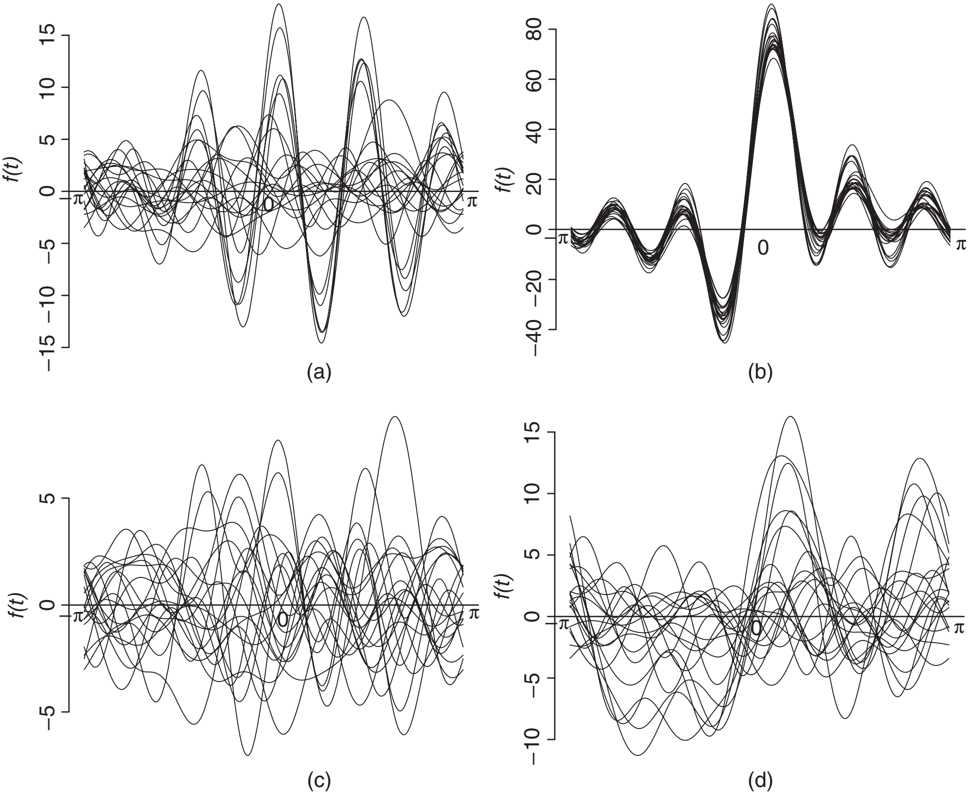

Figure 1.5 Harmonic curves for simulated data with $n = 10$ observations and $p = 11$ variables. (A summary of the two groups is given in Table 1.5.) (a) Original data; (b) data translated in all variables by 10; (c) each variable is standardized (by subtracting the mean and dividing by the standard deviation for that variable); (d) a random permutation $(7, 9, 3, 1, 10, 11, 6, 8, 5, 2, 4)$ of the variables.

scaling of the data, their interpretation requires some care. In the first plot, the second group of five observations gives curves that are away from the main group, but this is much less clear in the other plots.

One of the main uses of harmonic plots is to identify clusters in high-dimensional data. Further methods for cluster analysis are described in Chapter 14.

1.7.4 Parallel Coordinates Plot

In Figure 1.1, each observation is plotted once on each row of the plot. A parallel coordinates plot simply connects the points that belong to the same observation. For the cork data, this gives the plot shown in Figure 1.6. For these data, all the variables are measured in the same units, and so, it is appropriate to retain the scale for each axis. However, in general, each variable should be rescaled so that (for example) the range occupies $[0, 1]$. Such plots can then reveal clusters of observations and correlations between variables, provided that

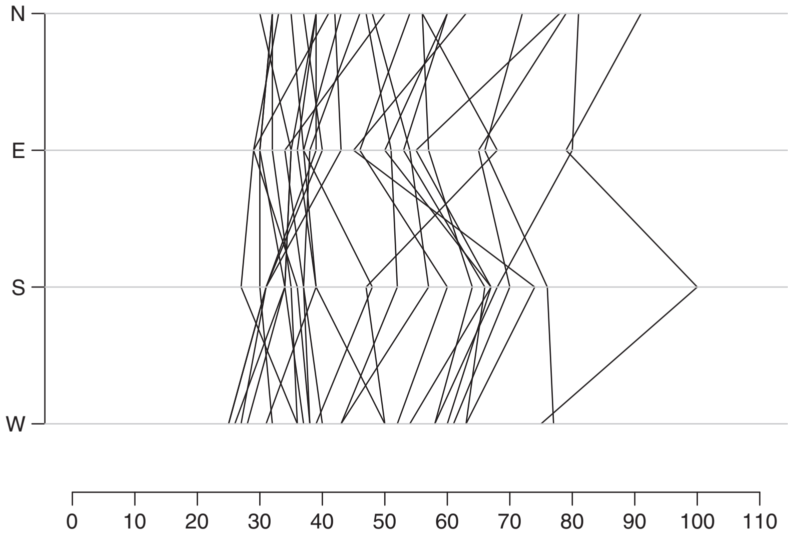

Figure 1.6 Parallel coordinates plot for the cork data.

the variables are appropriately (re-)ordered; see Wegman (1990) for some examples and extensions to this graphical tool.

1.8 Measures of Multivariate Skewness and Kurtosis

In Section 1.4, we have given a few basic summary statistics based on the first- and the second-order moments. In general, a kth-order central moment for the variables $i_1, \cdots, i_s$ is

$$M^{(j_1,\ldots,j_s)}_{i_1,\ldots,i_s} = \frac{1}{n}\sum_{r=1}^{n}\prod_{t=1}^{s}(x_{ri_t} - \bar{x}_{i_t})^{j_t},$$

where $j_1 + \cdots + j_s = k,\ j_t \neq 0,\ t = 1, \ldots, s$. As with mean and variance, we would like extensions to the multivariate case of summary measures such as $b_1 = m_3^2/s^6$ and $b_2 = m_4/s^4$, the univariate measures of skewness-squared, and kurtosis. Here,

$$m_k = \frac{1}{n}\sum_{r=1}^{n}(x_r - \bar{x})^k$$

is the (univariate) kth sample moment about the mean.

Using the invariant functions

$$g_{rs} = (\boldsymbol{x}_r - \bar{\boldsymbol{x}})'\boldsymbol{S}^{-1}(\boldsymbol{x}_s - \bar{\boldsymbol{x}}),$$

Mardia (1970b) has defined multivariate measures of skewness and kurtosis by

$$b_{1,p} = \frac{1}{n^2}\sum_{r,s=1}^{n} g_{rs}^3 \qquad (1.26)$$

and

$$b_{2,p} = \frac{1}{n}\sum_{r=1}^{n} g_{rr}^2. \tag{1.27}$$

The following properties are worth noting:

(1) $b_{1,p}$ depends only on the moments up to third order, whereas $b_{2,p}$ depends only on the moments up to fourth order excluding the third-order moments.

(2) These measures are invariant under *affine* transformations

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}.$$

(Similar properties hold for b_1 and b_2 under changes of scale and origin.)

(3) For $p = 1,\ b_{1,p} = b_1$ and $b_{2,p} = b_2$.

(4) Let D_r be the Mahalanobis distance between $\boldsymbol{x}_r$ and $\overline{\boldsymbol{x}}$, and let

$$\cos\alpha_{rs} = (\boldsymbol{x}_r - \overline{\boldsymbol{x}})'\boldsymbol{S}^{-1}(\boldsymbol{x}_s - \overline{\boldsymbol{x}})/D_rD_s$$

denote the cosine of the Mahalanobis angle between $(\boldsymbol{x}_r - \overline{\boldsymbol{x}})$ and $(\boldsymbol{x}_s - \overline{\boldsymbol{x}})$. Then, Equations (1.26) and (1.27) reduce to

$$b_{1,p} = \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n}(D_rD_s\cos\alpha_{rs})^3 \tag{1.28}$$

and

$$b_{2,p} = \frac{1}{n}\sum_{r=1}^{n} D_r^4. \tag{1.29}$$

Example 1.8.1 For the iris data shown in Table 1.2, the dimension p is equal to 4, and we compute $b_{1,4}$ and $b_{2,4}$ for each species and the pooled data. The results are shown in Table 1.6. It can be seen that the species have similar skewness and kurtosis values. □

Although $b_{1,p}$ cannot be negative, we note from (1.28) that if the data points are uniformly distributed on a p-dimensional hypersphere, then $b_{1,p}$ will be small (for example, typically less than 1 for $2 \le p \le 10$) since $\sum\cos\alpha_{rs} \simeq 0$ and D_r and D_s are approximately equal. We should expect $b_{1,p} \gg 0$ if there is a departure from spherical symmetry.

The statistic $b_{2,p}$ will pick up extreme behavior in the Mahalanobis distances of objects from the sample mean. The use of $b_{1,p}$ and $b_{2,p}$ to detect departure from multivariate normality is described in Chapter 6.

Table 1.6 Measures of multivariate skewness $b_{1,p}$ and kurtosis $b_{2,p}$ for each of the iris species ($n = 50, p = 4$) and the pooled data.

	I. setosa	*I. versicolor*	*I. virginica*	Pooled
Skewness	3.079	3.022	3.152	2.697
Kurtosis	26.54	22.88	24.30	23.74

Exercises and Complements

1.4.1 Under the transformation

$$\boldsymbol{y}_r = \boldsymbol{A}\boldsymbol{x}_r + \boldsymbol{b}, \quad r = 1, \dots, n$$

show that

$$\text{(i) } \bar{\boldsymbol{y}} = \boldsymbol{A}\bar{\boldsymbol{x}} + \boldsymbol{b}, \qquad \text{(ii) } \boldsymbol{S}_y = \boldsymbol{A}\boldsymbol{S}_x\boldsymbol{A}',$$

where $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}_x$ are the mean vector and the covariance matrix for the $\boldsymbol{x}_r$.

1.4.2 Let

$$\boldsymbol{S}(\boldsymbol{a}) = \frac{1}{n}\sum_{r=1}^{n}(\boldsymbol{x}_r - \boldsymbol{a})(\boldsymbol{x}_r - \boldsymbol{a})'$$

be the covariance matrix about $\boldsymbol{x} = \boldsymbol{a}$. Show that

$$\boldsymbol{S}(\boldsymbol{a}) = \boldsymbol{S} + (\bar{\boldsymbol{x}} - \boldsymbol{a})(\bar{\boldsymbol{x}} - \boldsymbol{a})'.$$

(i) Using (A.20) or otherwise, show that

$$|\boldsymbol{S}(\boldsymbol{a})| = |\boldsymbol{S}|\{1 + (\bar{\boldsymbol{x}} - \boldsymbol{a})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{a})\},$$

and consequently,

$$\min_{\boldsymbol{a}}|\boldsymbol{S}(\boldsymbol{a})| = |\boldsymbol{S}|.$$

(ii) Show that $\min_{\boldsymbol{a}} \operatorname{tr} \boldsymbol{S}(\boldsymbol{a}) = \operatorname{tr} \boldsymbol{S}$.

(iii) Note that $|\boldsymbol{S}| = \prod l_i$ and $\operatorname{tr} \boldsymbol{S} = \sum l_i$ are monotonically increasing symmetric functions of the eigenvalues $l_1, \dots l_p$ of $\boldsymbol{S}$. Use this observation in constructing other measures of multivariate scatter.

(iv) Using the inequality $g \le a$, where a and g are the arithmetic and geometric means of a set of positive numbers, show that $|\boldsymbol{S}| \le (n^{-1}\operatorname{tr} \boldsymbol{S})^n$.

1.4.3 (a) Fisher (1947) gives data relating to the body weight in kilograms (x_1) and the heart weight in grams (x_2) of 144 cats. For the 47 female cats, the sums and sums of squares and products are given by

$$\boldsymbol{X}_1'\boldsymbol{1} = \begin{pmatrix} 110.9 \\ 432.5 \end{pmatrix} \text{ and } \boldsymbol{X}_1'\boldsymbol{X}_1 = \begin{bmatrix} 265.13 & 1029.62 \\ 1029.62 & 4064.71 \end{bmatrix}.$$

Show that the mean vector and covariance matrix are $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ given in Example 5.2.3.

(b) For the 97 male cats from Fisher's data, the statistics are

$$\boldsymbol{X}_2'\boldsymbol{1} = \begin{pmatrix} 281.3 \\ 1098.3 \end{pmatrix} \text{ and } \boldsymbol{X}_2'\boldsymbol{X}_2 = \begin{bmatrix} 836.75 & 3275.55 \\ 3275.55 & 13056.17 \end{bmatrix}.$$

Find the mean vector and covariance matrix for the sample of male cats.

(c) Regarding the 144 male and female cats as a single sample, calculate the mean vector and covariance matrix.

(d) Calculate the correlation coefficient for each of (a), (b), and (c).

1.5.1 Let $\boldsymbol{M} = \boldsymbol{X}'\boldsymbol{X}$, where $\boldsymbol{X}$ is a data matrix. Show that

$$m_{ii} = \boldsymbol{x}'_{(i)}\boldsymbol{x}_{(i)} = n(s_{ii} + \bar{x}_i^2), \qquad m_{ij} = \boldsymbol{x}'_{(i)}\boldsymbol{x}_{(j)} = n(s_{ij} + \bar{x}_i\bar{x}_j).$$

1.5.2 Show that the scaling transformation in Section 1.5.1 can be written as

$$\boldsymbol{Y} = \boldsymbol{HXD}^{-1}, \qquad \boldsymbol{Y}' = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_n].$$

Use the fact that

$$\boldsymbol{Y}'\boldsymbol{1} = \boldsymbol{0}, \qquad \boldsymbol{Y}'\boldsymbol{HY} = \boldsymbol{D}^{-1}\boldsymbol{X}'\boldsymbol{HXD}^{-1}$$

to show that

$$\bar{\boldsymbol{y}} = \boldsymbol{0}, \qquad \boldsymbol{S}_y = \boldsymbol{R}.$$

1.5.3 Show that the Mahalanobis transformation in Section 1.5.2 can be written as

$$\boldsymbol{Z} = \boldsymbol{HXS}^{-1/2}, \qquad \boldsymbol{Z}' = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_n].$$

Hence, following the method of Exercise 1.5.2, show that $\bar{\boldsymbol{z}} = \boldsymbol{0}$, and $\boldsymbol{S}_z = \boldsymbol{I}$.

1.5.4 For the mark data in Table 1.1, show that

$$\begin{aligned}
\bar{\boldsymbol{x}}' &= (18.458, 74.400, 0.400), \\
\boldsymbol{S} &= \begin{bmatrix} 0.0159 & -0.4352 & -0.0252 \\ & 101.8400 & 3.0400 \\ & & 0.2400 \end{bmatrix}, \\
\boldsymbol{S}^{-1} &= \begin{bmatrix} 76.7016 & 0.1405 & 6.2742 \\ & 0.0160 & -0.1885 \\ & & 7.2132 \end{bmatrix}, \\
\boldsymbol{S}^{-1/2} &= \begin{bmatrix} 8.7405 & 0.0205 & 0.5521 \\ & 0.1013 & -0.0732 \\ & & 2.6274 \end{bmatrix}.
\end{aligned}$$

1.6.1 If

$$D_{rs}^2 = (\boldsymbol{x}_r - \boldsymbol{x}_s)'\boldsymbol{S}^{-1}(\boldsymbol{x}_r - \boldsymbol{x}_s),$$

we may write

$$D_{rs}^2 = q_{rr} + q_{ss} - 2q_{rs},$$

where

$$q_{rs} = \boldsymbol{x}'_r\boldsymbol{S}^{-1}\boldsymbol{x}_s.$$

Writing

$$g_{rs} = (\boldsymbol{x}_r - \bar{\boldsymbol{x}})'\boldsymbol{S}^{-1}(\boldsymbol{x}_s - \bar{\boldsymbol{x}}),$$

we see that

$$g_{rs} = q_{rs} + \bar{\boldsymbol{x}}'\boldsymbol{S}^{-1}\bar{\boldsymbol{x}} - \boldsymbol{x}'_r\boldsymbol{S}^{-1}\bar{\boldsymbol{x}} - \boldsymbol{x}'_s\boldsymbol{S}^{-1}\bar{\boldsymbol{x}}.$$

Therefore,

$$D_{rs}^2 = g_{rr} + g_{ss} - 2g_{rs}.$$

(i) Show that

$$\sum_r g_{rs} = 0 \quad \text{and} \quad \sum_r g_{rr} = \sum_r \operatorname{tr} \boldsymbol{S}^{-1}(\boldsymbol{x}_r - \overline{\boldsymbol{x}})(\boldsymbol{x}_r - \overline{\boldsymbol{x}})' = n \operatorname{tr} \boldsymbol{I}_p = np.$$

(ii) Therefore, show that

$$\sum_r D_{rs}^2 = np + ng_{ss} \qquad \text{or} \qquad g_{ss} = \frac{1}{n}\sum_r D_{rs}^2 - p$$

and

$$\sum_{r,s=1}^{n} D_{rs}^2 = 2n^2 p.$$

1.8.1 (Mardia, 1970b) Let $u_{pq} = M_{pq}/s_1^p s_2^q$, where

$$M_{pq} = \frac{1}{n}\sum_{r=1}^{n}(x_{r1} - \overline{x}_1)^p (x_{r2} - \overline{x}_2)^q.$$

Show that

$$\begin{aligned} b_{1,2} \ &= \ (1-r^2)^{-3}[u_{30}^2 + u_{03}^2 + 3(1+2r^2)(u_{12}^2 + u_{21}^2) - 2r^3 u_{30}u_{03} \\ &\quad + 6r\{u_{30}(ru_{12} - u_{21}) + u_{03}(ru_{21} - u_{12}) - (2+r^2)u_{12}u_{21}\}] \end{aligned}$$

and

$$b_{2,2} = (1-r^2)^{-2}[u_{40} + u_{04} + 2u_{22} + 4r(ru_{22} - u_{13} - u_{31})].$$

Hence, for $s_1 = s_2 = 1,\ r = 0$, show that

$$b_{1,2} = M_{30}^2 + M_{03}^2 + 3M_{12}^2 + 3M_{21}^2$$

and

$$b_{2,2} = M_{40} + M_{04} + 2M_{22}.$$

Thus, $b_{1,2}$ accumulates the effects of M_{21}, M_{12}, M_{03}, and M_{30}, while $b_{2,2}$ accumulates the effects of M_{22}, M_{04}, and M_{40}.

1.8.2 Carry out a simulation experiment to see if you can determine something about the distribution of $b_{1,p}$ and $b_{2,p}$ for the case that the data come from a normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{I}$. Specifically, for a variety of n (with fixed $p = 2$) and a variety of p with fixed n generate, say, 100 samples of size n from $N_p(\mathbf{0}, \boldsymbol{I})$, and for each sample compute $b_{1,p}$ and $b_{2,p}$. Compute the sample mean (and variance) of these quantities and plot them against n and p as appropriate. You may also consider a histogram of a larger sample (maybe 10 000) for $n = 100$ (say) and $p = 5$ (say) to investigate the shape of the two distributions.