Applied Multivariate Data Analysis

Second Edition

Brian S. Everitt
Institute of Psychiatry, King's College London, UK

and

Graham Dunn
School of Epidemiology and Health Sciences, University of Manchester, UK

John Wiley & Sons, Ltd

First published in Great Britain in 2001 by Arnold
This impression printed by Hodder Education,
a part of Hachette Livre UK,
338 Euston Road, London NW1 3BH

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN 978-0-4707-1117-0

8 9 10

Typeset in 10/12pt Times by Academic & Technical Typesetting. Bristol

3

Principal components analysis

3.1 Introduction

Principal components analysis is among the oldest and most widely used multivariate techniques. Originally introduced by Pearson (1901) and independently by Hotelling (1933), the basic idea of the method is to describe the variation of a set of multivariate data in terms of a set of uncorrelated variables, each of which is a particular linear combination of the original variables. The new variables are derived in decreasing order of importance so that, for example, the first principal component accounts for as much as possible of the variation in the original data. The second component is chosen to account for as much as possible in the remaining variation *subject* to being uncorrelated with the first component – and so on. The usual objective of this type of analysis is to see whether the first few components account for most of the variation in the original data. If so, they can be used to summarize the data with little loss of information. A reduction in dimensionality is thus achieved which might then be useful in simplifying later analysis.

Consider, for example, a set of data consisting of examination scores for several different subjects for each of a number of students. One question of interest might be how best to construct an informative index of examination performance. One obvious possibility would be the mean score for each student, although if the possible or observed range of examination scores varied from subject to subject, it might be more sensible to weight the scores in some way before calculating the average, or alternatively standardize the results for the separate examinations before attempting to combine them. Another possibility would be to use the first principal component derived from the observed examination results. This would give an index providing maximum discrimination between the students, with those examination scores that vary most within the sample of students being given the highest weight.

A further possible application for principal components analysis arises in the field of economics, where complex data – for example, prices, wage rates and

the cost of living – are often summarized by some kind of index number. When assessing changes in prices over time the economist will wish to allow for the fact that the prices of some commodities are more variable than others, or are considered more important than others; in each case the index will be weighted accordingly. In such examples, the first principal component can often satisfy the investigator's requirements (see Kendall, 1975).

But it is not always the first principal component that is of most interest to a researcher. A taxonomist, for example, when investigating variation in morphological measurements on animals, would be more likely to be interested in the second and subsequent components since these are likely to indicate shape; here the first component will often relate only to size.

Again the first principal component derived from, say, clinical psychiatric scores on patients may only provide an index of the severity of symptoms, and it is the remaining components that will give the psychiatrist information about the 'pattern' of symptoms.

In some applications, the principal components are an end in themselves and may be amenable to interpretation. More often they are obtained for use as input to another analysis. One example is provided by regression analysis; principal components may be useful when:

- there are too many explanatory variables relative to the number of observations;
- the explanatory variables are highly correlated.

Both situations lead to problems when applying regression techniques, problems which may be overcome by reducing the explanatory variables to a smaller number of principal components. Some applications of the technique are described in Rencher (1995).

3.2 Algebraic basics of principal components

The first principal component of the observations is that linear combination of the original variables whose sample variance is greatest among all possible such linear combinations. But how useful is this artificial variate constructed from the observed variables? To answer this question, we would first need to know the proportion of the total variance for which it accounted. If, for example, 80% of the variation in a multivariate data set involving six variables could be accounted for by a simple weighted average of the variable values, then almost all the variation could be expressed along a single continuum rather than in six-dimensional space. This would provide a highly parsimonious summary of the data that might be useful in later analysis.

The second principal component is defined as that linear combination of the original variables that accounts for a maximal proportion of the remaining variance subject to being uncorrelated with the first principal component. Subsequent components are defined similarly. The question now arises as to how the coefficients specifying the linear combination of the original variables

Box 3.1 Algebraic basis of principal components analysis

- The first principal component of the observations, y_1, is the linear combination

$$y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p$$

whose sample variance is greatest among all such linear combinations.

- Since the variance of y_1 could be increased without limit simply by increasing the coefficients $a_{11}, a_{12}, \ldots, a_{1p}$ (which we will write as the vector $\mathbf{a}_1$), a restriction must be placed on these coefficients. As we shall see later, a sensible constraint is to require that the sum of squares of the coefficients, $\mathbf{a}_1'\mathbf{a}_1$, should take the value one.
- The second principal component y_2 is the linear combination

$$y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p = \mathbf{a}_2'\mathbf{x}$$

which has the greatest variance subject to the following two conditions:

$$\mathbf{a}_2'\mathbf{a}_2 = 1,$$

$$\mathbf{a}_2'\mathbf{a}_1 = 0.$$

(The second condition above ensures that y_1 and y_2 are uncorrelated.)

- Similarly, the jth principal component is that linear combination $y_j = \mathbf{a}_j'\mathbf{x}$ which has greatest variance subject to the conditions

$$\mathbf{a}_j'\mathbf{a}_j = 1,$$

$$\mathbf{a}_j'\mathbf{a}_i = 0 \quad (i < j).$$

- To find the coefficients defining the first principal component we need to choose the elements of the vector $\mathbf{a}_1$ so as to maximize the variance of y_1 subject to the constraint $\mathbf{a}_1'\mathbf{a}_1 = 1$. The variance of y_1 is given by

$$\text{Var}(y_1) = \text{Var}(\mathbf{a}_1'\mathbf{x}) = \mathbf{a}_1'\mathbf{S}\mathbf{a}_1.$$

- To maximize a function of several variables subject to one or more constraints, the method of *Lagrange multipliers* is used. In this case this leads to the solution that $\mathbf{a}_1$ is the eigenvector of S corresponding to the largest eigenvalue – details are given in Morrison (1990) and Chatfield and Collins (1980).
- The other components are derived in similar fashion, with $\mathbf{a}_j$ being the eigenvector of S associated with the jth largest eigenvalue.
- If the eigenvalues of S are $\lambda_1, \lambda_2, \ldots, \lambda_p$, then since $\mathbf{a}_i'\mathbf{a}_i = 1$, the variance of the ith principal component is given by λ_i (see Exercise 3.3).
- The total variance of the p principal components will equal the total variance of the original variables so that

$$\sum_{i=1}^{p} \lambda_i = \text{trace}(\mathbf{S}).$$

- Consequently, the jth principal component accounts for a proportion P_j of the total variation on the original data, where

$$P_j = \frac{\lambda_j}{\text{trace}(\mathbf{S})}.$$

- The first p^* principal components, where $p^* < p$ account for P^* of the total variation in the original data, where

$$P^* = \frac{\sum_{i=1}^{p^*} \lambda_i}{\text{trace}(\mathbf{S})}.$$

defining each component are found. The algebra of the principal components is detailed in Box 3.1.

The derivation of principal components given in Box 3.1 is in terms of the eigenvalues and eigenvectors of the *covariance matrix*, **S**. In practice, however, it is far more usual to extract the components from the *correlation matrix*, **R**.

Box 3.2 Correlations and covariances of variables and components

- The covariance of the observed variables with the jth principal component are found as follows:

$$\begin{aligned}\text{Cov}(\mathbf{x}, y_j) &= \text{Cov}(\mathbf{x}, \mathbf{x}'\mathbf{a}_j) \\ &= E(\mathbf{x}\mathbf{x}')\mathbf{a}_j \\ &= \mathbf{S}\mathbf{a}_j \\ &= \lambda_j \mathbf{a}_j.\end{aligned}$$

- Consequently, the covariance of variable i with component j is given by

$$\text{Cov}(x_i, y_j) = \lambda_j a_{ji}.$$

- The correlation of variable i with component j is therefore

$$\begin{aligned}r_{x_i, y_j} &= \frac{\text{Cov}(x_i, y_j)}{\sigma_{x_i}\sigma_{y_i}} \\ &= \frac{\lambda_j a_{ji}}{s_{ii}^{1/2}\sqrt{\lambda_j}} \\ &= \frac{\sqrt{\lambda_j} a_{ji}}{s_{ii}^{1/2}}.\end{aligned}$$

- If the components are extracted from the correlation matrix rather than the covariance matrix, then

$$r_{x_i, y_j} = \sqrt{\lambda_j} a_{ji}.$$

The reasons are not difficult to identify. If we imagine a set of multivariate data where the variables $x_1, x_2, \ldots, x_p$ are of completely different types – for example, length, temperature, blood pressure and anxiety rating – then the structure of the principal components derived from the covariance matrix will depend upon the essentially arbitrary choice of units of measurement; changing lengths from centimetres to inches, say, will alter the derived components. Additionally, if there are large differences between the variances of the original variables, those whose variances are largest will tend to dominate the early components – an example illustrating this problem is given in Jolliffe (1986).

The problem is overcome by extracting the components as the eigenvectors of **R**, which is equivalent to calculating the principal components from the original variables after each has been standardized to have unit variance. It should be noted, however, that there is rarely any simple correspondence between the components derived from **S**. And choosing to work with **R** rather than with **S** involves a definite but possibly arbitrary decision to make variables 'equally important'.

It is often of interest to determine the correlations or covariances between the original variables and the derived components; details of how this is done are given in Box 3.2. Note in particular that it is easy to express the coefficients of components derived from a correlation matrix as correlations of variables and components; the original coefficients are simply multiplied by the square root of the appropriate eigenvalue.

3.3 Rescaling principal components

If the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$, which define the principal components, are used to form a $p \times p$ matrix, $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_p]$, and the eigenvalues $\lambda_1, \ldots, \lambda_p$ are arranged in a diagonal matrix, $\mathbf{\Lambda}$, then it is easy to show that the covariance matrix of the original variables is given by

$$\mathbf{S} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}'. \tag{3.1}$$

(This assumes that the components have been extracted from **S** rather than from the correlation matrix, **R**.)

By rescaling the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_p$ so that the sum of squares of their elements is equal to the corresponding eigenvalue, λ_i, rather than unity, that is, calculating $\mathbf{a}_i^* = \lambda_i^{1/2}\mathbf{a}_i$, then (3.1) may be written more simply as

$$\mathbf{S} = \mathbf{A}^*(\mathbf{A}^*)', \tag{3.2}$$

where $\mathbf{A}^* = [\mathbf{a}_1^*, \ldots, \mathbf{a}_p^*]$.

The elements of $\mathbf{A}^*$ are such that the coefficients of the more important components are scaled up compared to those of the less important components, a scaling which is intuitively reasonable. The rescaled vectors have a number of other advantages, since their elements are analogous to *factor loadings*, as we shall see in Chapter 12. In the case of components arising from a correlation matrix, the rescaled coefficients give, as shown in Box 3.2, correlations between the components and the original variables. It is often these rescaled coefficients that are presented as the result of a principal components analysis.

3.4 Calculating principal component scores

Principal component scores for individual i with vector of variable values $\mathbf{x}_i$ are obtained from the equations

$$\begin{aligned} y_{i1} &= \mathbf{a}_1'(\mathbf{x}_i - \bar{\mathbf{x}}) \\ &\vdots \\ y_{ip} &= \mathbf{a}_p'(\mathbf{x}_i - \bar{\mathbf{x}}), \end{aligned} \tag{3.3}$$

where $\bar{\mathbf{x}}$ is the vector of mean values of the original variables. The transformed variables will have zero means and variances $\lambda_1, \ldots, \lambda_p$.

How the principal component scores might be used will be illustrated later.

3.5 Choosing the number of components

As described in Box 3.1, principal components analysis is seen to be a technique for transforming a set of observed variables into a new set of variables which are uncorrelated with one another. The variation in the original p variables is only *completely* accounted for by *all* p principal components. The usefulness of these transformed variables, however, stems from their property of accounting for the variance in decreasing proportions. So the question we need to ask is how many components are needed to provide an adequate summary of a given data set? A number of informal and more formal techniques are available. Here we shall concentrate on the former; examples of the use of formal inferential methods are given in Jolliffe (1986) and Rencher (1995).

The most common of the relatively *ad hoc* procedures that have been suggested are the following:

- Retain just enough components to explain some specified, large percentage of the total variation of the original variables. Values between 70% and 90% are usually suggested, although smaller values might be appropriate as p or n increases.
- Exclude those principal components whose eigenvalues are less than the average, $\sum_{i=1}^{p} \lambda_i / p$. Since $\sum_{i=1}^{p} \lambda_i = \text{trace}(\mathbf{S})$, the average eigenvalue is also the average variance of the original variables. This method then retains those components that account for more variance than the average for the variables. When the components are extracted from the correlation matrix, $\text{trace}(\mathbf{R}) = p$, and the average is therefore 1; components with eigenvalues less than 1 are therefore excluded. (This rule was originally suggested by Kaiser, 1958, but Jolliffe, 1972, proposes on the basis of a number of simulation studies that a more appropriate procedure would be to exclude components extracted from a correlation matrix whose associated eigenvalues are less than 0.7.)
- Cattell (1965) suggests examination of the plot of the λ_i against i, the so-called *scree diagram*. The number of components selected is the value of i corresponding to an 'elbow' in the curve, this point being considered to be where

'large' eigenvalues cease and 'small' eigenvalues begin. A modification described by Jolliffe (1986) is the *log-eigenvalue diagram* consisting of a plot of $\log(\lambda_i)$ against i.

Examples of the use of a number of these procedures will be given later.

3.6 Two simple examples of principal components analysis

3.6.1 Bivariate data with correlation coefficient r

Suppose we have just two variables, x_1 and x_2, measured on a sample of individuals, with sample correlation matrix given by

$$\mathbf{R} = \begin{pmatrix} 1.0 & r \\ r & 1.0 \end{pmatrix}. \tag{3.4}$$

The principal components of $\mathbf{R}$ are derived in Box 3.3.

Box 3.3 The principal components of bivariate data

- In order to find the principal components of a set of bivariate data with correlation r we need to find the eigenvalues and eigenvectors of the correlation matrix, $\mathbf{R}$.
- The eigenvalues are roots of the equation
$$|\mathbf{R} - \lambda\mathbf{I}| = 0.$$
- This leads to a quadratic equation in λ,
$$(1-\lambda)^2 - r^2 = 0,$$
giving eigenvalues $\lambda_1 = 1 + r$, $\lambda_2 = 1 - r$. Note that the sum of the eigenvalues is 2, equal to trace($\mathbf{R}$).
- The eigenvector corresponding to λ_1 is obtained by scaling the equation
$$\mathbf{R}\mathbf{a}_1 = \lambda_1\mathbf{a}_1.$$
- This leads to the equations
$$a_{11} + ra_{12} = (1+r)a_{11}, \qquad ra_{11} + a_{12} = (1+r)a_{12}.$$
- The two equations are identical and both reduce to $a_{11} = a_{12}$.
- If we now introduce the normalization constraint, $\mathbf{a}_1'\mathbf{a}_1 = 1$ we find that
$$a_{11} = a_{12} = \frac{1}{\sqrt{2}}.$$
- Similarly, we find the second eigenvector to be given by $a_{21} = 1/\sqrt{2}$ and $a_{22} = -1/\sqrt{2}$.
- The two principal components are then given by
$$y_1 = \frac{1}{\sqrt{2}}(x_1 + x_2), \qquad y_2 = \frac{1}{\sqrt{2}}(x_1 - x_2).$$

Notice that if $r < 0$ the order of the eigenvalues and hence of the principal components is reversed; if $r = 0$ the eigenvalues are both equal to 1 and any two solutions at right angles could be chosen to represent the two components. Two further points:

- There is an arbitrary sign in the choice of the elements of $\mathbf{a}_i$; it is customary to choose a_{i1} to be positive.
- The components do not depend on r, although the proportion of variance explained by each does change with r. As r tends to 1 the proportion of variance accounted for by y_1, namely $(1 + r)/2$, also tends to 1.

3.6.2 Head size in brothers

The data shown in Table 3.1 give the head length (in millimetres) for each of the first two adult sons in 25 families. The mean vector and variance–covariance matrix of the data are

$$\bar{\mathbf{x}} = \begin{pmatrix} 185.72 \\ 183.84 \end{pmatrix}, \qquad \mathbf{S} = \begin{pmatrix} 95.29 & 69.66 \\ 69.66 & 100.81 \end{pmatrix}.$$

The principal components of these data derived from their covariance matrix are:

PC1	PC2
0.693	0.721
0.721	−0.693

The observations are plotted in Figure 3.1, along with the axes corresponding to the principal components. The first of these passes through the mean of the data and has slope 0.721/0.693. The second also passes through the mean with slope

Table 3.1 Head lengths (mm) of first and second sons

First son	Second son	First son	Second son
191	179	195	201
181	185	183	188
176	171	208	192
189	190	197	189
188	197	192	187
179	186	183	174
174	186	190	195
188	187	163	161
195	183	186	173
181	182	175	165
192	185	174	178
176	176	197	200
190	187		

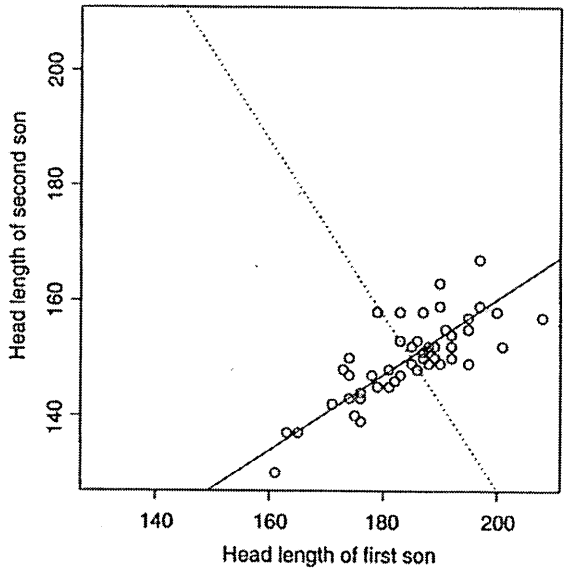

Figure 3.1 Head lengths of first and second sons, showing axes defined by the prinicipal components of the sample covariance matrix.

$-0.693/0.721$. This example illustrates that principal components analysis is essentially just a rotation of the axes of the multivariate data scatter.

3.7 More complex examples of the application of principal components analysis

3.7.1 Olympic decathlon results

In Chapter 2 the results for the 34 competitors in the Olympic decathlon of 1988 were subjected to various graphical procedures. In this chapter principal components analysis will be applied to the data. But before undertaking the analysis we shall transform the data by taking negative values for the four running events. In this way the results of all ten events are scored in the same direction, with small values reflecting a poor performance and large values the reverse. In addition, we shall analyse the results on only the first 33 competitors, since the athlete who finished last in the competition might, from examining the plots in Chapter 2, justifiably be labelled an outlier. The results of a principal components analysis on the correlation matrix of the ten events (see Table 3.2) are shown in Table 3.3.

Notice first that the components as given are scaled so that the sums of squares of their elements are equal to one. To rescale them so that they represented correlations between variables and components, they would need to be multiplied by the square root of the corresponding eigenvalue. The coefficients defining the first component are all positive and it is clearly a measure of overall

Table 3.2 Correlation matrix for the 1988 Olympic decathlon results (first 33 competitors only)

	Event									
	1	2	3	4	5	6	7	8	9	10
1	1.000									
2	0.540	1.000								
3	0.208	0.142	1.000							
4	0.146	0.273	0.122	1.000						
5	0.606	0.515	−0.095	0.088	1.000					
6	0.638	0.478	0.296	0.307	0.546	1.000				
7	0.047	0.042	0.806	0.147	−0.142	0.110	1.000			
8	0.389	0.350	0.480	0.213	0.319	0.522	0.344	1.000		
9	0.065	0.182	0.598	0.116	−0.120	0.063	0.443	0.274	1.000	
10	0.261	0.396	−0.269	0.114	0.587	0.143	−0.402	0.031	−0.096	1.000

Note: The ten events are identified in Chapter 2.

performance (see later). This component has variance 3.42 and accounts for 34% of the total variance. The second component contrasts performance on the 'power' events such as shot, discus and javelin with the only real stamina event, the 1500 m. The second component has variance 2.61, so between them the first two components account for 60% of the total variance.

Only the first two components have eigenvalues greater than 1, so one of the *ad hoc* rules described previously would suggest retaining just these two to provide a parsimonious description of the data. (The scree diagram shown in Figure 3.2, however, suggests consideration of the first three components.)

The scores of the 33 competitors on the first two principal components can be used to produce a scatterplot of the data – see Figure 3.3. The plotted points are labelled by the finishing position of the competitors. The first component score largely ranks the athletes in finishing order, confirming its interpretation as an overall measure of performance. Further evidence for this interpretation can be obtained by looking at the relationship between the first principal component score and the number of points gained by each competitor (these points are allocated using a standard set of tables). Figure 3.4 shows first principal component score plotted against the points gained. The points lie almost on a straight line and the correlation coefficient takes the value 0.95.

3.7.2 Drug usage by American college students

The majority of adult and adolescent Americans regularly use psychoactive substances during an increasing proportion of their lifetime. Various forms of licit and illicit psychoactive substances use are prevalent, suggesting that patterns of psychoactive substance taking are a major part of the individual's behavioural repertory and have pervasive implications for the performance of other behaviours. In an investigation of these phenomena, Huba *et al.* (1981) collected data on drug usage rates for 1634 students in the seventh to ninth

Table 3.3 Principal components of Olympic decathlon results (first 33 competitors only)

	Component number									
Event	1	2	3	4	5	6	7	8	9	10
1	0.4159	0.1488	0.2675	0.0883	0.4423	0.0307	−0.2544	−0.6637	0.108	0.110
2	0.3941	0.1521	−0.1689	−0.2442	0.3689	0.0938	0.7505	0.1413	0.046	−0.056
3	0.2691	−0.4835	0.0985	−0.1078	−0.0098	−0.2300	−0.1107	0.0725	0.423	−0.651
4	0.2123	−0.0279	−0.8550	0.3879	−0.0019	−0.0745	−0.1351	−0.1554	−0.102	−0.119
5	0.3558	0.3522	0.1895	−0.0806	−0.1470	−0.3269	−0.1413	0.1468	−0.651	−0.337
6	0.4335	0.0696	0.1262	0.3823	0.0888	0.2105	−0.2725	0.6390	0.207	0.260
7	0.1758	−0.5033	0.0461	0.0256	0.0194	−0.6149	0.1440	0.0094	−0.167	0.535
8	0.3841	−0.1496	0.1369	0.1440	−0.7167	0.3478	0.2733	−0.2769	−0.018	0.066
9	0.1799	−0.3720	−0.1923	−0.6005	0.0956	0.4374	−0.3419	0.0585	−0.306	0.131
10	0.1701	0.4210	−0.2226	−0.4856	−0.3398	−0.3003	−0.1869	−0.0073	0.457	0.243
Variance	3.42	2.61	0.94	0.88	0.56	0.49	0.43	0.31	0.27	0.10

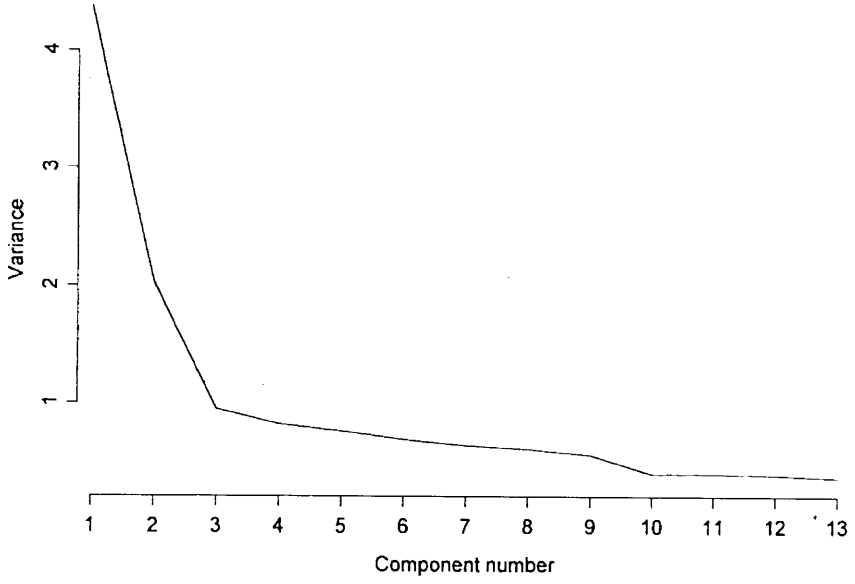

Figure 3.2 Scree diagram for the principal components of the Olympic decathlon results.

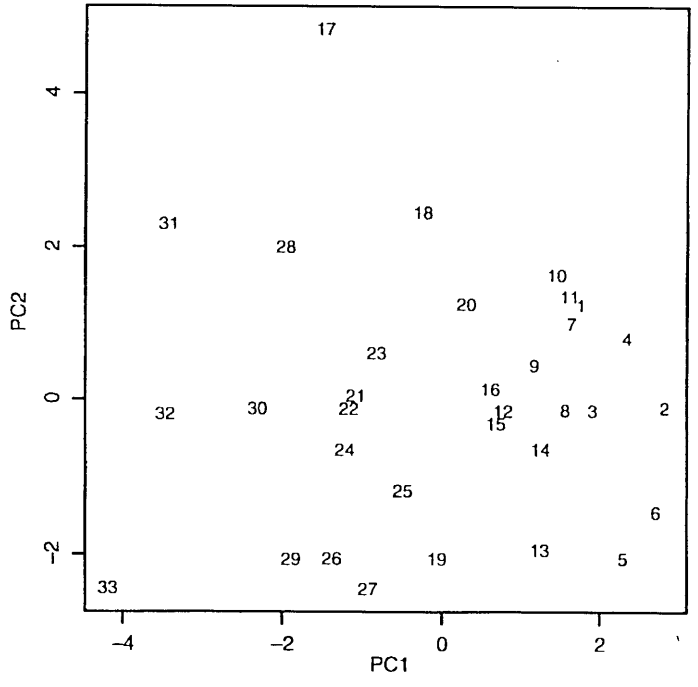

Figure 3.3 Scatterplot of the first two principal component scores for the Olympic decathlon results (first 33 competitors only).

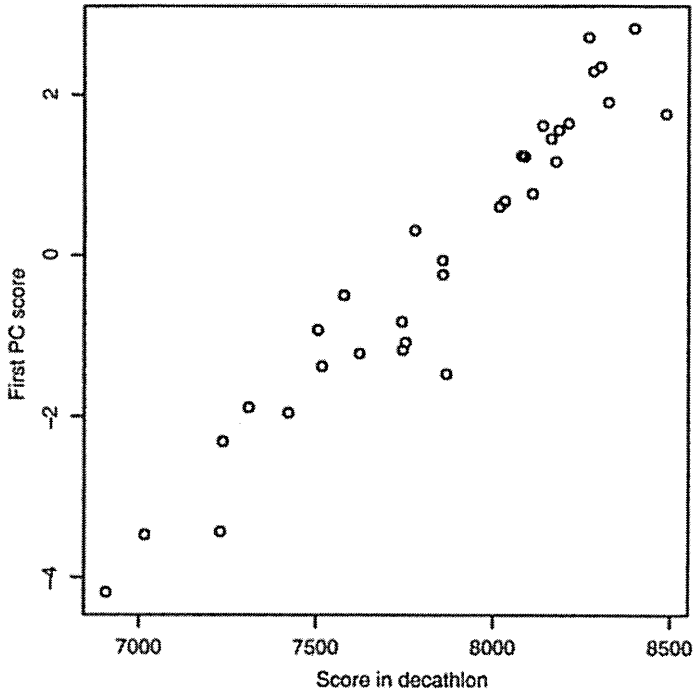

Figure 3.4 Scatterplot of the first principal component score for the Olympic decathlon results against competitor's score in the competition (first 33 competitors only).

grades in 11 schools in the greater metropolitan area of Los Angeles. Each participant completed a questionnaire about the number of times a particular substance had ever been used. The substances asked about were as follows:

1. cigarettes;
2. beer;
3. wine;
4. liquor;
5. cocaine;
6. tranquillizers;
7. drug store medications used to get high;
8. heroin and other opiates;
9. marijuana;
10. hashish;
11. inhalants (glue, gasoline, etc.);
12. hallucinogenics (LSD, mescaline, etc.);
13. amphetamine stimulants.

Responses were recorded on a five-point scale:

1. never tried;
2. only once;

Table 3.4 Correlation matrix for drug usage data

	Substance												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1												
2	0.447	1											
3	0.422	0.619	1										
4	0.435	0.604	0.583	1									
5	0.114	0.068	0.053	0.115	1								
6	0.203	0.146	0.139	0.258	0.349	1							
7	0.091	0.103	0.110	0.122	0.209	0.221	1						
8	0.082	0.063	0.066	0.097	0.321	0.355	0.201	1					
9	0.513	0.445	0.365	0.482	0.186	0.315	0.150	0.154	1				
10	0.304	0.318	0.240	0.368	0.303	0.377	0.163	0.219	0.534	1			
11	0.245	0.203	0.183	0.255	0.272	0.323	0.310	0.288	0.301	0.302	1		
12	0.101	0.088	0.074	0.139	0.279	0.367	0.232	0.320	0.204	0.368	0.340	1	
13	0.245	0.199	0.184	0.293	0.278	0.545	0.232	0.314	0.394	0.467	0.392	0.511	1

Note: See text for identification of the 13 substances.

3. a few times;
4. many times;
5. regularly.

The correlations between the usage rates of the 13 substances are shown in Table 3.4. The coefficients defining the 13 principal components of these correlations are shown in Table 3.5. These coefficients are scaled so that they represent correlations between observed variables and derived components. The eigenvalues corresponding to each component are also shown in Table 3.5. A scree diagram of the eigenvalues is shown in Figure 3.5. There is a relatively clear 'elbow' in the curve at the third eigenvalue.

The first three components account for 52% of the total variation in the data. The first component is clearly a measure of overall drug usage, as might be expected since all the correlations in Table 3.3 are positive. The second component contrasts 'legal' with 'illegal' substances (with the exception of marijuana, which has the same sign coefficient as the legal substances). Alternatively, this component might be seen as contrasting 'soft' and 'hard' drug usage. So after overall usage has been accounted for, the main source of variation is between individuals who consume the two different types of substance. The third component is essentially a contrast of drugstore and inhalant substance usage on the one hand with marijuana, hashish and amphetamine usage on the other.

Most users of principal components analysis search for an interpretation of the derived components which will allow them to be labelled. But this is not without its difficulties and critics; the following quotation from Marriott (1974) should act as a salutary warning about the dangers of over-interpretation.

> It must be emphasised that no mathematical method is, or could be, designed to give physically meaningful results. If a mathematical

Table 3.5 Principal components of drug usage correlations in Table 3.3

	Principal component number												
Substance	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0.58	0.40	−0.06	0.01	0.27	0.38	0.10	−0.02	0.46	−0.08	−0.09	−0.13	−0.12
2	0.60	0.57	0.13	0.09	−0.15	−0.12	−0.09	0.04	−0.04	−0.06	−0.05	0.38	−0.30
3	0.55	−0.56	0.13	0.09	−0.27	−0.13	−0.05	−0.06	0.07	−0.28	0.17	−0.09	0.34
4	0.66	0.46	0.05	0.06	−0.15	−0.13	0.00	−0.09	−0.15	0.40	−0.07	−0.30	−0.08
5	0.44	−0.41	0.05	0.53	0.38	−0.29	−0.24	−0.18	0.13	0.05	0.10	0.02	0.00
6	0.61	−0.37	−0.17	0.08	−0.11	−0.04	0.44	−0.38	−0.06	−0.06	−0.28	0.11	0.08
7	0.37	−0.27	0.71	0.30	0.22	−0.22	0.27	0.16	0.05	0.00	0.00	−0.02	−0.02
8	0.42	−0.45	−0.14	0.48	0.28	−0.32	0.13	0.39	−0.08	0.01	0.03	−0.04	−0.03
9	0.71	0.23	−0.23	−0.10	0.31	0.12	0.12	0.20	−0.10	0.23	0.13	0.24	0.28
10	0.69	−0.07	−0.35	−0.11	0.22	−0.20	−0.01	0.29	−0.26	−0.27	−0.16	−0.18	−0.07
11	0.58	−0.24	0.31	−0.18	0.07	0.42	−0.38	−0.28	−0.28	−0.05	−0.04	0.01	0.02
12	0.52	−0.47	−0.11	−0.26	−0.31	−0.12	−0.32	0.15	0.36	0.13	−0.17	0.07	0.10
13	0.69	−0.33	−0.23	−0.24	−0.18	−0.02	0.11	−0.14	0.04	−0.05	0.45	−0.06	−0.18
Variance	4.38	2.04	0.95	0.82	0.76	0.69	0.64	0.61	0.56	0.40	0.40	0.39	0.37

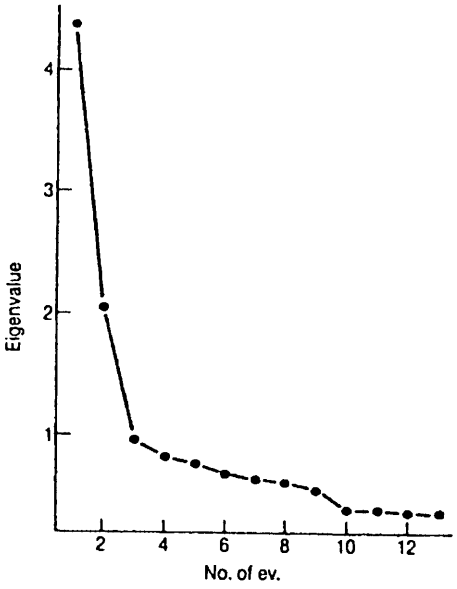

Figure 3.5 Scree diagram for the principal components of the drug usage data.

expression of this sort has an obvious physical meaning, it must be attributed to a lucky chance, or to the fact that the data have a strongly marked structure that shows up in analysis. Even in the latter case, quite small sampling fluctuations can upset the interpretation; for example, the first two principal components may appear in reverse order, or may become confused altogether. Reification then requires considerable skill and experience if it is to give a true picture of the physical meaning of the data.

3.8 Using principal components analysis to select a subset of variables

Although the first few principal component scores for each individual may provide a very useful summary of a set of multivariate data, *all* the original variables are needed in their computation. In many cases an investigator might be happier with determining a subset of the *original* variables which contains, in some sense, virtually all the information available in the complete set of these variables. In a series of papers, Jolliffe (1970; 1972; 1973), discusses a number of methods for selecting subsets of variables, several of which are based on principal components analysis.

One such method is to use one or other of the criteria discussed in Section 3.5 to decide on the number of components which account for a substantial proportion of the variation in the original variables – this number is taken to indicate

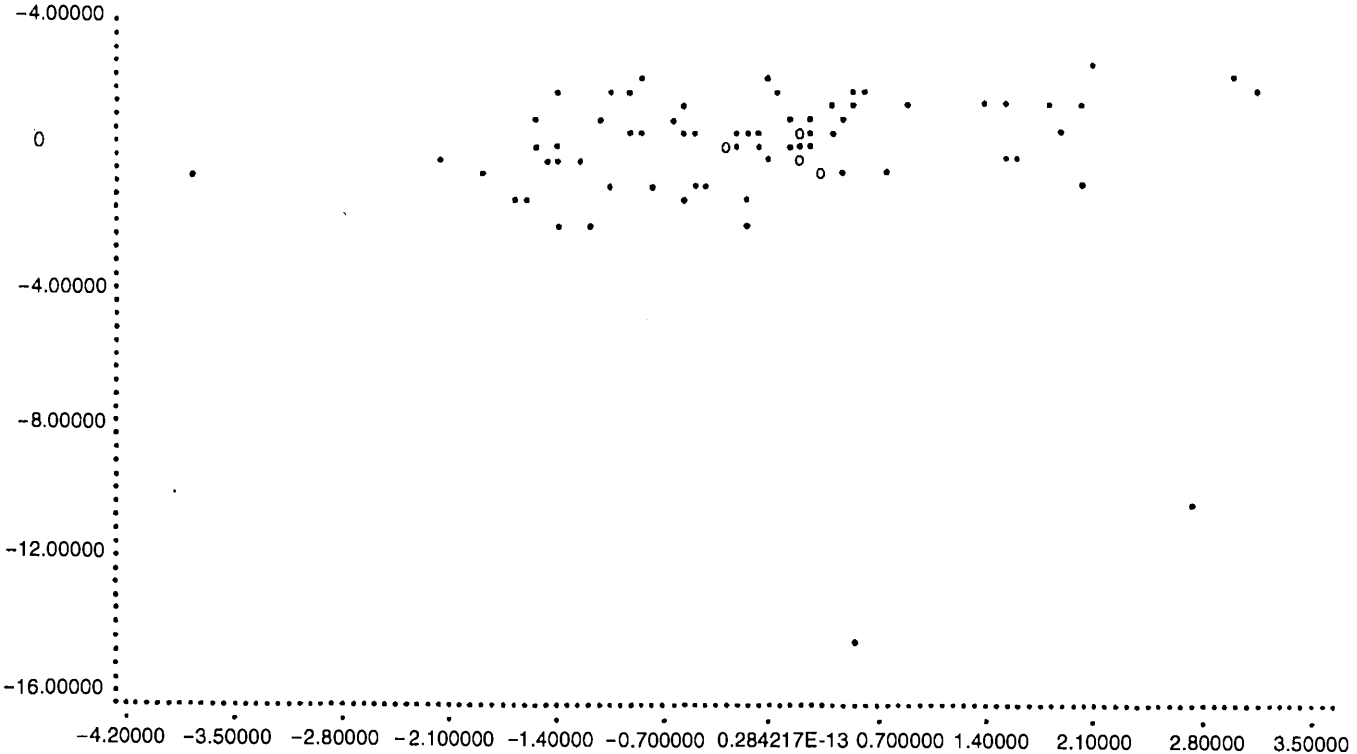

Figure 3.6 Scatterplot produced from the last two components of a set of five-dimensional data.

the effective dimensionality of the data and will be the size of the subset of original variables to be retained. Each variable is selected, one associated with each component, as the one not already chosen which has the greatest absolute coefficient value on the component. To illustrate this procedure we will use the drug usage data analysed in the previous section. If we now use the eigenvalue criterion suggested by Jolliffe, that is, retain components with eigenvalues greater than 0.7, five components are kept. Examining the coefficients defining these components (see Table 3.5), we find we are led to the following five variables:

1. marijuana usage;
2. beer usage;
3. drugstore usage;
4. cocaine usage;
5. hallucinogenics usage.

3.9 Using the last few principal components

Although, when using principal components analysis in practice, primary interest is most commonly centred on the components with largest variance, there may be occasions where examination of the last few components might prove helpful. These components may, for example, be of assistance in identifying outliers which are adding insignificant dimensions or obscuring singularities in the data. A set of five-dimensional data was constructed to contain two such outliers, and Figure 3.6 shows the two-dimensional plot of the data using principal components 4 and 5. This diagram clearly identifies the two outliers; repeating the analysis with these observations removed results in a solution in which only four principal components are needed to account for the variation in the data, demonstrating for the remaining observations the existence of a linear dependency. It is of interest to compare Figure 3.6 with Figure 3.7, which gives a plot of these data in the space of the first two principal components – this shows no evidence of any outlying observations.

3.10 The biplot

A *biplot* is a graphical representation of the information in an $n \times p$ data matrix. The 'bi' is a reflection that the technique allows the display of both information about the variables as indicated by their variances and covariances *and* the relationships between individuals as indicated by particular measures of inter-individual distance. The technique is described in Gabriel (1981) and in full detail in Gower and Hand (1996). Here only an example of its use will be given, using data on the athletics records of 55 countries given in Table 3.6. The biplot diagram for these data is shown in Figure 3.8. In this diagram the lengths of the arrows reflect the variances of the corresponding variable, and

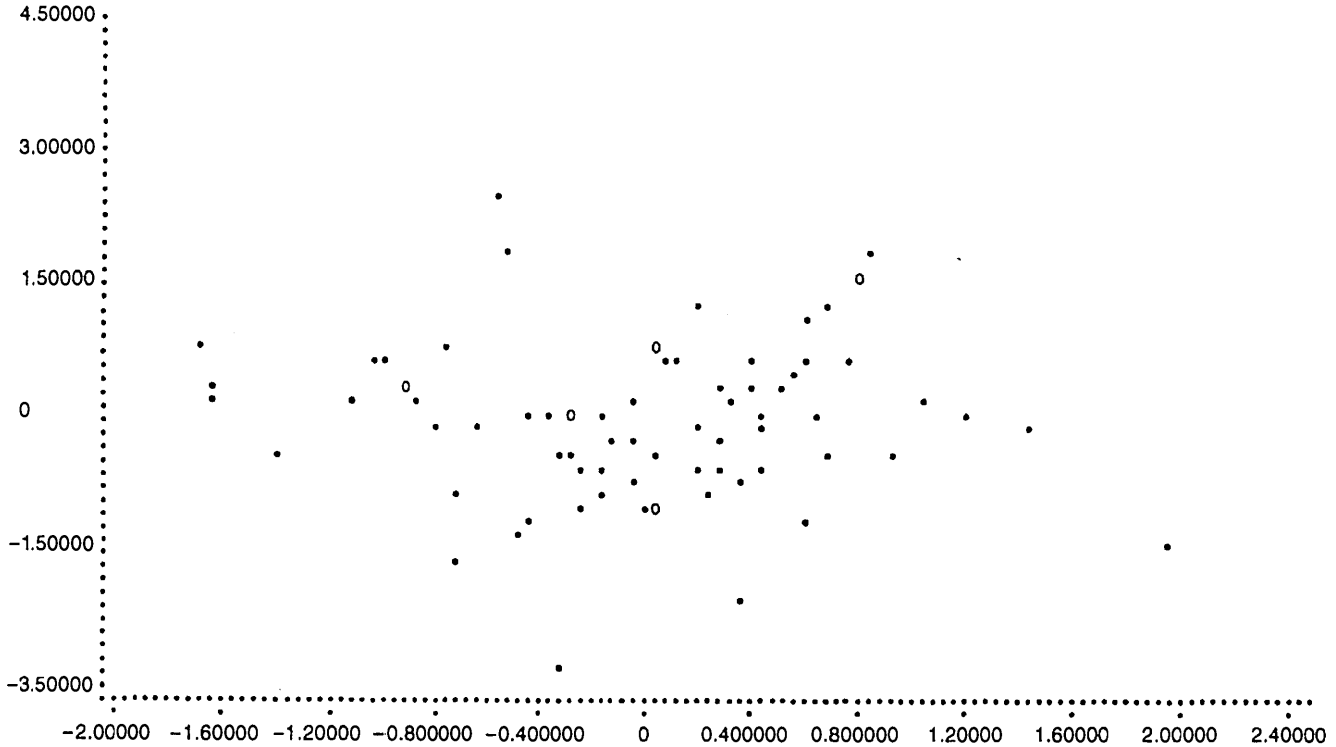

Figure 3.7 Scatterplot of the data used in producing Figure 3.5 now plotted in the space of the first two components.

Table 3.6 Athletics records for 55 countries

	Event							
	1	2	3	4	5	6	7	8
Argentina	10.39	20.81	46.84	1.81	3.70	14.04	29.36	137.72
Australia	10.31	20.06	44.84	1.74	3.57	13.28	27.66	128.30
Austria	10.44	20.81	46.82	1.79	3.60	13.26	27.72	135.90
Belgium	10.34	20.68	45.04	1.73	3.60	13.22	27.45	129.95
Bermuda	10.28	20.58	45.91	1.80	3.75	14.68	30.55	146.62
Brazil	10.22	20.43	45.21	1.73	3.66	13.62	28.62	133.13
Burma	10.64	21.52	48.30	1.80	3.85	14.45	30.28	139.95
Canada	10.17	20.22	45.68	1.76	3.63	13.55	28.09	130.15
Chile	10.34	20.4	46.20	1.79	3.71	13.61	29.30	134.03
China	10.51	21.04	47.30	1.81	3.73	13.90	29.13	133.53
Columbia	10.43	21.05	46.10	1.82	3.74	13.49	27.88	131.35
Cook Is	12.18	23.2	52.94	2.02	4.24	16.70	35.38	164.70
Costa Rica	10.94	21.9	48.66	1.87	3.84	14.03	28.81	136.58
Czech	10.35	20.65	45.64	1.76	3.58	13.42	28.19	134.32
Denmark	10.56	20.52	45.89	1.78	3.61	13.50	28.11	130.78
Dom Rep	10.14	20.65	46.80	1.82	3.82	14.91	31.45	154.12
Finland	10.43	20.69	45.49	1.74	3.61	13.27	27.52	130.87
France	10.11	20.38	45.28	1.73	3.57	13.34	27.97	132.30
GDR	10.12	20.33	44.87	1.73	3.56	13.17	27.42	129.92
FRG	10.16	20.37	44.50	1.73	3.53	13.21	27.61	132.23
GB	10.11	20.21	44.93	1.70	3.51	13.01	27.51	129.13
Greece	10.22	20.71	46.56	1.78	3.64	14.59	28.45	134.60
Guatemala	10.98	21.82	48.40	1.89	3.80	14.16	30.11	139.33
Hungary	10.26	20.62	46.02	1.77	3.62	13.49	28.44	132.58
India	10.60	21.42	45.73	1.76	3.73	13.77	28.81	131.98
Indonesia	10.59	21.49	47.80	1.84	3.92	14.73	30.79	148.83
Ireland	10.61	20.96	46.30	1.79	3.56	13.32	27.81	132.35
Israel	10.71	21.00	47.80	1.77	3.72	13.66	28.93	137.55
Italy	10.01	19.72	45.26	1.73	3.60	13.23	27.52	131.08
Japan	10.34	20.71	45.86	1.79	3.64	13.41	27.72	128.63
Kenya	10.46	20.66	44.92	1.73	3.55	13.10	27.80	129.75
Korea (North)	10.91	21.94	47.30	1.85	3.77	14.13	29.67	130.87
Korea (South)	10.34	20.89	46.90	1.79	3.77	13.96	29.23	136.25
Luxemburg	10.34	20.77	47.40	1.82	3.67	13.64	29.08	141.27
Malaysia	10.40	20.92	46.30	1.82	3.80	14.64	31.01	154.10
Mauritus	11.19	22.45	47.70	1.88	3.83	15.06	31.77	152.23
Mexico	10.42	21.30	46.10	1.80	3.65	13.46	27.95	129.20
Netherlands	10.52	29.95	45.10	1.74	3.62	13.36	27.61	129.02
New Zealand	10.51	20.88	46.10	1.74	3.54	13.21	27.70	128.98
Norway	10.55	21.16	46.71	1.76	3.62	13.34	27.69	131.48
Papua New Guinea	10.96	21.78	47.90	1.90	4.01	14.72	31.36	148.22
Philippines	10.78	21.64	46.24	1.81	3.83	14.74	30.64	145.27
Poland	10.16	20.24	45.36	1.76	3.60	13.29	27.89	131.58
Portugal	10.53	21.17	46.70	1.79	3.62	13.13	27.38	128.65
Rumania	10.41	20.98	45.87	1.76	3.64	13.25	27.67	132.50
Singapore	10.38	21.28	47.40	1.88	3.89	15.11	31.32	157.77
Spain	10.42	20.77	45.98	1.76	3.55	13.31	27.73	131.57
Sweden	10.25	20.61	45.63	1.77	3.61	13.29	27.94	130.63
Switzerland	10.37	20.45	45.78	1.78	3.55	13.22	27.91	131.20

Table 3.6 Continued

	Event							
	1	2	3	4	5	6	7	8
Tapei	10.59	21.29	46.80	1.79	3.77	14.07	20.07	139.27
Thailand	10.39	21.09	47.91	1.83	3.84	15.23	32.56	149.90
Turkey	10.71	21.43	47.60	1.79	3.67	13.56	28.58	131.50
USA	9.93	19.75	43.86	1.73	3.53	13.20	27.43	128.22
USSR	10.07	20.00	44.60	1.75	3.59	13.20	27.53	130.55
W Samoa	10.82	21.86	49.00	2.02	4.24	16.28	34.91	161.83

Events: (1) 100 m (s), (2) 200 m (s), (3) 400 m (s), (4) 800 m (min), (5) 1500 m (min), (6) 5000 m (min), (7) 10 000 m (min), (8) marathon (min).

the angles between them indicate the size of their correlations, small angles corresponding to high correlations. The relative positions of the points corresponding to countries indicate the similarities and differences of their athletics records. The positions of these points relative to the lines representing the variables reflect a country's times on the various events. The diagram clearly

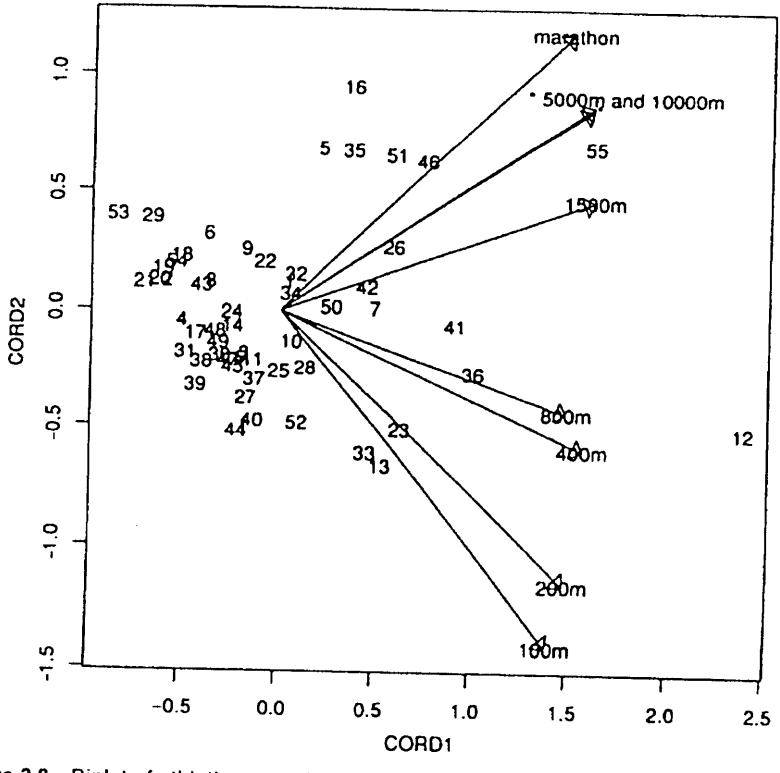

Figure 3.8 Biplot of athletics records for 55 countries.

picks out the Cook Islands (12) and Western Samoa (55) as the countries with the poorest athletics performances. In addition, Bermuda (5), Malaysia (35), Singapore (46), Thailand (51) and the Dominican Republic (16) appear to form a small 'cluster' of countries whose athletes are particularly poor at the marathon. Italy (29) and the USA (53) are seen to have the best overall athletics records.

3.11 Geometrical interpretation of principal components analysis

In geometrical terms it is easy to show that the first principal component defines the line of best fit (in the least-squares sense) to the p-dimensional observations in the sample. These observations may therefore be represented in one dimension by taking their projection onto this line, that is, finding their first principal component score. If the observations happen to be collinear in p dimensions this representation would account completely for the variation in the data and the sample covariance matrix would have only one non-zero eigenvalue. In practice, of course, such collinearity is extremely unlikely, and an improved representation would be given by projecting the p-dimensional observations onto the space of the best fit, this being defined by the first two principal components. Similarly, the first p^* components give the best fit in p^* dimensions. If the observations fit exactly into a space of p^* dimensions, it would be indicated by the presence of $p - p^*$ zero eigenvalues of the covariance matrix. This would imply the presence of $p - p^*$ linear relationships between the variables. Such constraints are sometimes referred to as *structural relationships*.

3.12 Projection pursuit

Principal components analysis is often used to obtain a two- or three-dimensional representation of multivariate data, so that the data may be examined graphically and possibly interesting structure detected visually. Used in this way, the method is an example of a class of exploratory techniques for multivariate data usually labelled *projection pursuit*. According to Jones and Sibson (1987):

> The term projection pursuit was first used by Friedman and Tukey (1974) to name a technique for the exploratory analysis of reasonably large and reasonably multivariate data sets; projection pursuit reveals structure in data by offering selected low-dimensional orthogonal projections of it for inspection.

In essence, projection pursuit methods seek a one- or two-dimensional projection of the data that maximizes some measure of 'interestingness'. Principal components is, for example, a projection pursuit technique in which the index of interestingness is the proportion of total variance accounted for by the projected data. It relies for its success on the tendency for large variation to

be also interestingly structured variation, a connection which is not necessary and which often fails to hold in practice.

Many other indices of 'interestingness' have been suggested. Details are given in Jones and Sibson (1987) and Ripley (1996), but as a relatively simple example of the approach, let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be a p-dimensional data set of size n and consider only one-dimensional projection pursuit. Then the projection is specified by a p-dimensional vector $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$. The projected data in this case are one-dimensional data points $z_1, z_2, \ldots, z_n$, where $z_j = \boldsymbol{\alpha}'\mathbf{x}_j$ for

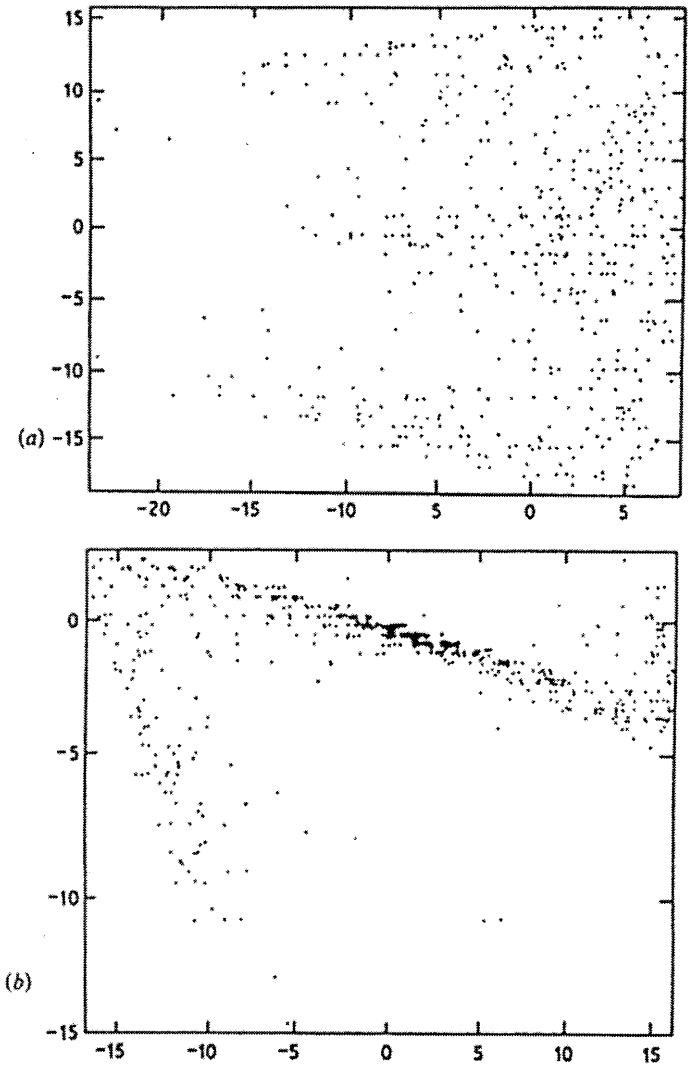

Figure 3.9 Two-dimensional projections of high-energy physics data: (a) principal components; (b) projection pursuit. (Reproduced from Friedman and Tukey (1974) with permission of the Institute of Electrical and Electronic Engineers, Inc.)

$j = 1, \ldots, n$. This projection is then compared with an 'uninteresting' projection, for example, the standard normal distribution, using a distance measure suitable for two functions f and g, for example,

$$\int [f(z) - g(z)]^2 \, dz. \tag{3.5}$$

Once an index is chosen, a projection is selected by numerically maximizing the index over the choice of projection. For details, see Ripley (1996), Hall (1989) and Sun (1991).

An example comparing a principal components projection and one obtained from a more complex projection pursuit procedure is provided by Friedman and Tukey (1974). The data consist of 500 seven-dimensional observations taken in a particle physics experiment. Full details are available in the original paper. The scatterplots in Figure 3.9 show projections of the data onto the first two principal components and onto a plane found by projection pursuit. The latter shows structure not apparent in the principal components plot.

3.13 Summary

Principal components looks for a few linear combinations of the original variables that can be used to summarize a data set, losing in the process as little information as possible. The derived variables might be used in a variety of ways, in particular for simplifying later analyses and providing informative plots of the data. The method consists of transforming a set of correlated variables to a new set of variables which are uncorrelated. Consequently, it should be noted that if the original variables are themselves almost uncorrelated there is little point in carrying out a principal components analysis, since it will merely find components which are close to the original variables but arranged in decreasing order of variance.

As a method for obtaining a low-dimensional view of multivariate data, principal components analysis is an example of a projection pursuit technique. But the more complex of these methods may lead to views of the data which are more informative than those given by principal components.

Exercises

3.1 Suppose that $\mathbf{x}' = [x_1, x_2]$ is such that $x_2 = 1 - x_1$ and $x_1 = 1$ with probability p and $x_1 = 0$ with probability $q = 1 - p$. Find the covariance matrix of $\mathbf{x}$ and its eigenvalues and eigenvectors.

3.2 The eigenvectors of a covariance matrix, $\mathbf{S}$, scaled so that their sums of squares are equal to the corresponding eigenvalue, are $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_p$. Show that

$$\mathbf{S} = \mathbf{c}_1\mathbf{c}_1' + \mathbf{c}_2\mathbf{c}_2' + \ldots + \mathbf{c}_p\mathbf{c}_p'.$$

3.3 If the eigenvalues of $\mathbf{S}$ are $\lambda_1, \lambda_2, \ldots, \lambda_p$ show that if the coefficients defining the principal components are scaled so that $\mathbf{a}_i'\mathbf{a}_i = 1$ then the variance of the ith principal component is λ_i.

3.4 If two variables, X and Y, have covariance matrix $\mathbf{S}$ given by

$$\mathbf{S} = \begin{pmatrix} a & c \\ c & b \end{pmatrix},$$

show that if $c \neq 0$ then the first principal component is

$$\sqrt{\frac{c^2}{c^2 + (V_1 - a)^2}}X + \frac{c}{|c|}\sqrt{\frac{(V_1 - a)^2}{c^2 + (V_1 - a)^2}}Y,$$

where V_1 is the variance explained by the first principal component. What is the value of V_1?

3.5 Find the principal components of the following correlation matrix and compare how the one- and two-component solutions reproduce the matrix:

$$\mathbf{R} = \begin{pmatrix} 1.0000 & & \\ 0.6579 & 1.0000 & \\ 0.0034 & -0.0738 & 1.0000 \end{pmatrix}.$$

3.6 MacDonnell (1902) obtained measurements on seven physical characteristics for each of 3000 criminals. The seven variables measured were (1) head length, (2) head breadth, (3) face breadth, (4) left finger length, (5) left forearm length, (6) left foot length, (7) height. The corresponding correlation matrix is

$$\mathbf{R} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{array}{c} \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{array} \\ \begin{pmatrix} 1.000 & & & & & & \\ 0.402 & 1.000 & & & & & \\ 0.396 & 0.618 & 1.000 & & & & \\ 0.301 & 0.150 & 0.321 & 1.000 & & & \\ 0.305 & 0.135 & 0.289 & 0.846 & 1.000 & & \\ 0.339 & 0.206 & 0.363 & 0.759 & 0.797 & 1.000 & \\ 0.340 & 0.183 & 0.345 & 0.661 & 0.800 & 0.736 & 1.000 \end{pmatrix} \end{array}.$$

Find the principal components of these data and interpret the results.

3.7 Find the principal components of the air pollution data set described in the previous chapter and use the derived variables to produce a scatterplot matrix of the cities. What conclusions about the data can you draw from the analysis?

3.8 The data in Table 3.7 give measurements on five meterological variables over an 11-year period. The variables are:

1. rainfall in November and December (mm);
2. average July temperature (°C);

Table 3.7 Meteorological data

Year	1	2	3	4	5
1920–21	87.9	19.6	1.0	1661	28.37
1921–22	89.9	15.2	90.1	968	23.77
1922–23	153.0	19.7	56.6	1353	26.04
1923–24	132.1	17.0	91.0	1293	25.74
1924–25	88.8	18.3	93.7	1153	26.68
1925–26	220.9	17.8	106.9	1286	24.29
1926–27	117.7	17.8	65.5	1104	28.00
1927–28	109.0	18.3	41.8	1574	28.37
1928–29	156.1	17.8	57.4	1222	24.96
1929–30	181.5	16.8	140.6	902	21.66
1930–31	181.4	17.0	74.3	1150	24.37

3. rainfall in July (mm);
4. radiation in July (curies);
5. average harvest yield (quintals per hectare).

Carry out a principal components analysis of both the covariance matrix and the correlation matrix of the data and compare the results. Which set of components leads to to most meaningful interpretation?

2.7 The labelled scatterplot of the birth death rates is shown in Figure C.2. It does appear to support the interpretation of the bivariate density estimate of the data given in the text.

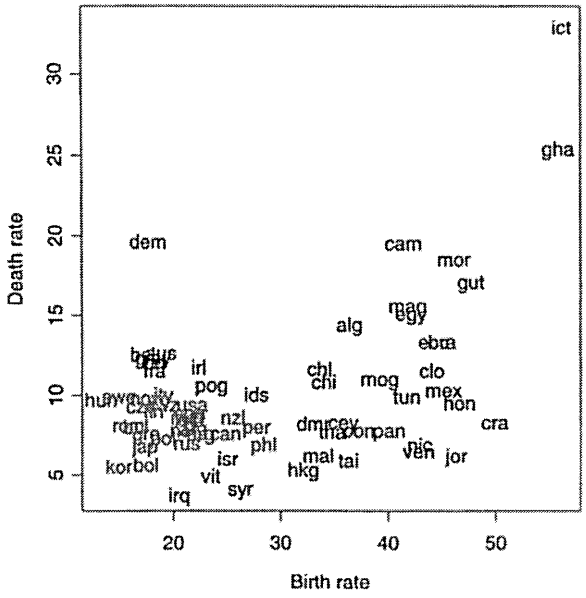

Figure C.2 Scatterplot of birth and death rate data showing points labelled by country.

Chapter 3

3.1 We first need to find the expected value of $\mathbf{x}$:

$$E(\mathbf{x}) = \begin{pmatrix} E(x_1) \\ E(x_2) \end{pmatrix} = \begin{pmatrix} p \\ q \end{pmatrix}.$$

The covariance matrix of $\mathbf{x}$ is defined to be

$$E(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))'.$$

This leads to the following:

$$\begin{aligned} &= E\begin{pmatrix} x_1 - p \\ x_2 - q \end{pmatrix}(x_1 - p \quad x_2 - q) \\ &= E\begin{pmatrix} (x_1 - p)^2 & (x_1 - p)(x_2 - q) \\ (x_2 - q)(x_1 - p) & (x_2 - q)^2 \end{pmatrix} \\ &= \begin{pmatrix} p(1-p) & -pq \\ -pq & q(1-p) \end{pmatrix}. \end{aligned}$$

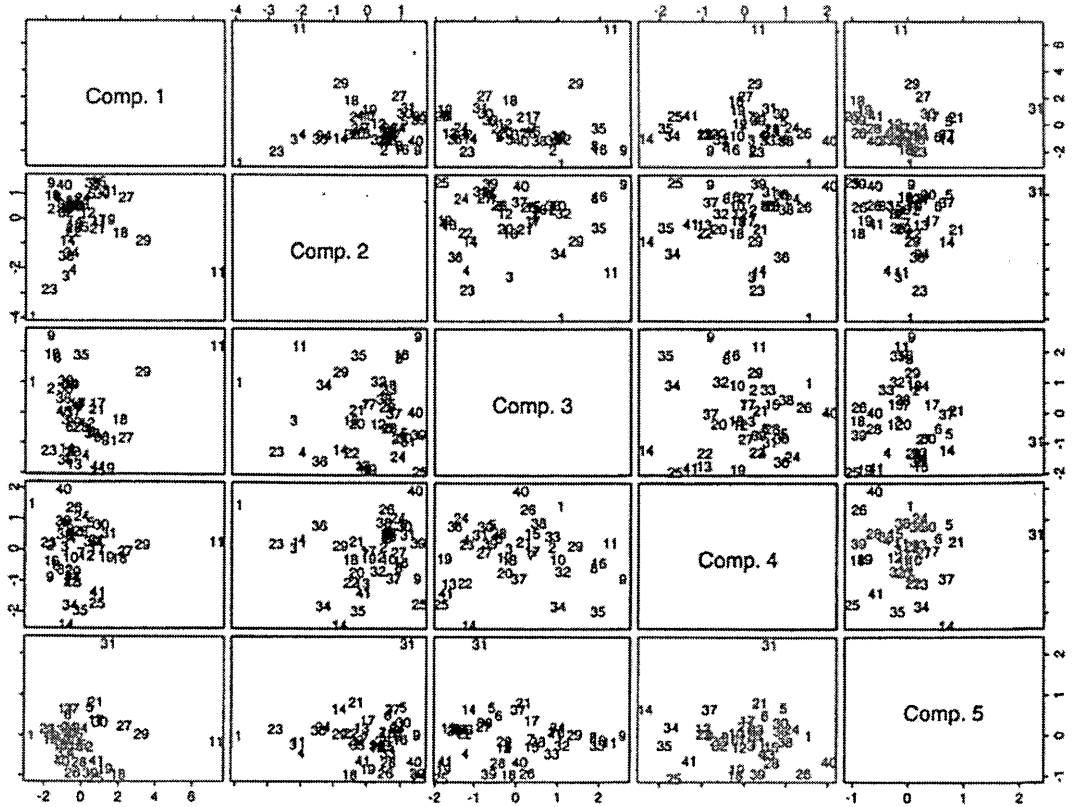

Figure C.3 Scatterplot matrix of principal component scores of the air pollution data.

3.7 The principal components of the correlation matrix of the air pollution data are shown in Table C.1 and the scatterplot matrix for the first five principal component scores is given in Figure C.3. (The cities are labelled by their number as given in Table 2.7.) Clearly Chicago (11), Philadelphia (29) and perhaps Providence (31) are outliers.

Table C.1 Principal components of the correlation matrix for the air pollution data (Exercise 3.7)

Var	PC1	PC2	PC3	PC4	PC5	PC6	PC7
SO_2	0.490	0.100	0.100	0.404	0.730	−0.183	0.150
Temp	−0.315	0.100	0.677	−0.185	0.162	−0.611	0.100
Manuf	0.541	−0.226	0.267	0.100	0.164	0.180	−0.745
Pop	0.488	−0.282	0.345	−0.113	−0.349	0.100	0.649
Wind	0.250	0.100	−0.311	−0.862	0.268	−0.150	0.100
Precip	0.100	0.626	0.492	0.184	0.161	0.554	0.100
Days	0.260	0.678	−0.110	0.110	−0.440	−0.505	0.100
Proportion of variance	0.39	0.606	0.805	0.932	0.9820	0.996	1.00

Chapter 4

4.1 The two-dimensional correspondence analysis solution for the data on eye colour and hair colour is shown in Figure C.4.

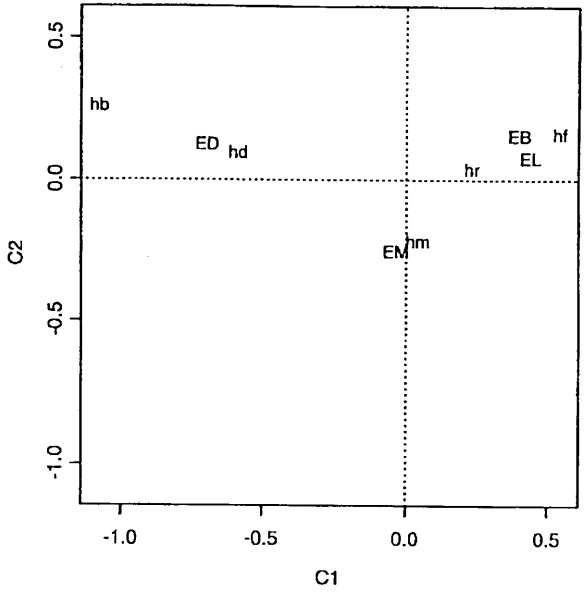

Figure C.4 Two-dimensional correspondence analysis solution for eye colour/hair colour data.