chapter six

Principal components analysis

6.1 Definition of principal components

The technique of principal components analysis was first described by Karl Pearson (1901). He apparently believed that this was the correct solution to some of the problems that were of interest to biometricians at that time, although he did not propose a practical method of calculation for more than two or three variables. A description of practical computing methods came much later from Hotelling (1933). Even then, the calculations were extremely daunting for more than a few variables, because they had to be done by hand. It was not until computers became generally available that the technique achieved widespread use.

Principal components analysis is one of the simplest of the multivariate methods. The object of the analysis is to take p variables X_1, X_2, … , X_p, and find combinations of these to produce indices Z_1, Z_2, … , Z_p that are uncorrelated and in order of their importance in terms of the variation in the data. The lack of correlation means that the indices are measuring different dimensions of the data, and the ordering is such that $\text{Var}(Z_1) \geq \text{Var}(Z_2) \ \ldots \geq \text{Var}(Z_p)$, where $\text{Var}(Z_i)$ denotes the variance of Z_i. The Z indices are then the principal components. When doing a principal components analysis, there is always the hope that the variances of most of the indices will be so low as to be negligible. In that case, most of the variation in the full data set can be adequately described by the few Z variables with variances that are not negligible, and some degree of economy is then achieved.

Principal components analysis does not always work in the sense that a large number of original variables are reduced to a small number of transformed variables. Indeed, if the original variables are uncorrelated, then the analysis achieves nothing. The best results are obtained when the original variables are very highly correlated, positively or negatively. If that is the case, then it is quite conceivable that 20 or more original variables can be adequately represented by two or three principal components. If this desirable state of affairs does occur, then the important principal components will be of some interest as measures of the underlying dimensions in the data. It will also be of value to know that there is a good deal of redundancy in the original variables, with most of them measuring similar things.

Table 6.1 Correlations between the five body measurements of female sparrows calculated from the data of Table 1.1

	X_1	X_2	X_3	X_4	X_5
X_1, total length	1.000	—			
X_2, alar extent	0.735	1.000	—		
X_3, length of beak and head	0.662	0.674	1.000	—	
X_4, length of humerus	0.645	0.769	0.763	1.000	—
X_5, length of keel of sternum	0.605	0.529	0.526	0.607	1.000

Note: Only the lower part of the table is shown, because the correlation between X_i and X_j is the same as the correlation between X_j and X_i.

Before describing the calculations involved in a principal components analysis, it is of value to look briefly at the outcome of the analysis when it is applied to the data in Table 1.1 on five body measurements of 49 female sparrows. Details of the analysis are given in Example 6.1. In this case, the five measurements are quite highly correlated, as shown in Table 6.1. This is, therefore, good material for the analysis in question. It turns out that the first principal component has a variance of 3.62, whereas the other components all have variances that are much less than this (0.53, 0.39, 0.30, and 0.16). This means that the first principal component is by far the most important of the five components for representing the variation in the measurements of the 49 birds. The first component is calculated to be

$$Z_1 = 0.45\,X_1 + 0.46\,X_2 + 0.45\,X_3 + 0.47\,X_4 + 0.40\,X_5$$

where X_1–X_5 denote the measurements in Table 1.1 in order, after they have been standardized to have zero means and unit standard deviations.

Clearly, Z_1 is essentially just an average of the standardized body measurements, and it can be thought of as a simple index of size. The analysis given in Example 6.1 therefore leads to the conclusion that most of the differences between the 49 birds are a matter of size rather than shape.

6.2 *Procedure for a principal components analysis*

A principal components analysis starts with data on p variables for n individuals, as indicated in Table 6.2. The first principal component is then the linear combination of the variables $X_1, X_2, \ldots, X_p$:

$$Z_1 = a_{11}\,X_1 + a_{12}\,X_2 + a_{1p}\,X_p$$

Table 6.2 The form of data for a principal components analysis, with variables X_1 to X_p and observations on n cases

Case	X_1	X_2	...	X_p
1	x_{11}	x_{12}	...	x_{1p}
2	x_{21}	x_{22}	...	x_{2p}
n	x_{n1}	x_{n2}	...	x_{np}

that varies as much as possible for the individuals, subject to the condition that

$$a_{11}^2 + a_{12}^2 + ... + a_{1p}^2 = 1$$

Thus, Var(Z_1), the variance of Z_1, is as large as possible given this constraint on the constants a_{1j}. The constraint is introduced because if this is not done, then Var(Z_1) can be increased by simply increasing any one of the a_{1j} values.

The second principal component

$$Z_2 = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p$$

is chosen so that Var(Z_2) is as large as possible, subject to the constraint that

$$a_{21}^2 + a_{22}^2 + ... + a_{2p}^2 = 1$$

and also to the condition that Z_1 and Z_2 have zero correlation for the data. The third principal component

$$Z_3 = a_{31}X_1 + a_{32}X_2 + ... + a_{3p}X_p$$

is such that Var(Z_3) is as large as possible, subject to the constraint that

$$a_{31}^2 + a_{32}^2 + ... + a_{3p}^2 = 1$$

and also that Z_3 is uncorrelated with both Z_1 and Z_2. Further principal components are defined by continuing in the same way. If there are p variables, then there will be up to p principal components.

To use the results of a principal components analysis, it is not necessary to know how the equations for the principal components are derived. However, it is useful to understand the nature of the equations themselves. In fact, a principal components analysis involves finding the eigenvalues of the sample covariance matrix.

The calculation of the sample covariance matrix has been described in Sections 2.6 and 2.7. The covariance matrix is symmetric and has the form

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdot\cdot & c_{1p} \\ c_{21} & c_{22} & \cdot\cdot & c_{2p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ c_{p1} & c_{p2} & \cdot\cdot & c_{pp} \end{bmatrix}$$

where the diagonal element c_{ii} is the variance of X_i, and the off-diagonal terms $c_{ij} = c_{ji}$ are the covariance of variables X_i and X_j.

The variances of the principal components are the eigenvalues of the matrix **C**. There are p of these eigenvalues, some of which may be zero, but negative eigenvalues are not possible for a covariance matrix. Assuming that the eigenvalues are ordered as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, then λ_i corresponds to the ith principal component

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p$$

In particular, $\text{Var}(Z_i) = \lambda_i$ and the constants a_{i1}, a_{i2}, …, a_{ip} are the elements of the corresponding eigenvector, scaled so that

$$a_{i1}^2 + a_{i2}^2 + \ldots + a_{ip}^2 = 1$$

An important property of the eigenvalues is that they add up to the sum of the diagonal elements (the trace) of the matrix **C**. That is,

$$\lambda_1 + \lambda_2 + \ldots + \lambda_p = c_{11} + c_{22} + \ldots + c_{pp}$$

As c_{ii} is the variance of X_i and λ_i is the variance of Z_i, this means that the sum of the variances of the principal components is equal to the sum of the variances of the original variables. Therefore, in a sense, the principal components account for all the variation in the original data.

To avoid one or two variables having an undue influence on the principal components, it is usual to code the variables $X_1, X_2, \ldots, X_p$ to have means of zero and variances of one at the start of an analysis. The matrix **C** then takes the form

$$\mathbf{C} = \begin{bmatrix} 1 & c_{12} & \cdots & c_{1p} \\ c_{21} & 1 & \cdots & c_{2p} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ c_{p1} & c_{p2} & \cdots & 1 \end{bmatrix}$$

where $c_{ij} = c_{ji}$ is the correlation between X_i and X_j. In other words, the principal components analysis is carried out on the correlation matrix. In that case, the sum of the diagonal terms, and hence the sum of the eigenvalues, is equal to p, the number of X variables.

The steps in a principal components analysis are

1. Start by coding the variables $X_1, X_2, \ldots, X_p$, to have zero means and unit variances. This is usual, but is omitted in some cases where it is thought that the importance of variables is reflected in their variances.
2. Calculate the covariance matrix **C**. This is a correlation matrix if Step 1 has been done.
3. Find the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ and the corresponding eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$. The coefficients of the ith principal component are then the elements of $\mathbf{a}_i$, while λ_i is its variance.
4. Discard any components that only account for a small proportion of the variation in the data. For example, starting with 20 variables, it might be found that the first three components account for 90% of the total variance. On this basis, the other 17 components may reasonably be ignored.

Example 6.1: Body measurements of female sparrows

Some mention has already been made of what happens when a principal components analysis is carried out on the data on five body measurements of 49 female sparrows (Table 1.1). This example is now considered in more detail.

It is appropriate to begin with Step 1 of the four parts of the analysis that have just been described. Standardization of the measurements ensures that they all have equal weight in the analysis.

Table 6.3 The eigenvalues and eigenvectors of the correlation matrix for five measurements on 49 female sparrows

		Eigenvectors (coefficients for the principal components)				
Component	Eigenvalue	X_1	X_2	X_3	X_4	X_5
1	3.616	0.452	0.462	0.451	0.471	0.398
2	0.532	–0.051	0.300	0.325	0.185	–0.877
3	0.386	0.691	0.341	–0.455	–0.411	–0.179
4	0.302	–0.420	0.548	–0.606	0.388	0.069
5	0.165	0.374	–0.530	–0.343	0.652	–0.192

Note: The eigenvalues are the variances of the principal components. The eigenvectors give the coefficients of the standardized X variables used to calculate the principal components.

Omitting standardization would mean that the variables X_1 and X_2, which vary most over the 49 birds, would tend to dominate the principal components.

The covariance matrix for the standardized variables is the correlation matrix. This has already been given in lower triangular form in Table 6.1. The eigenvalues of this matrix are found to be 3.616, 0.532, 0.386, 0.302, and 0.164. These add to 5.000, the sum of the diagonal terms in the correlation matrix. The corresponding eigenvectors are shown in Table 6.3, standardized so that the sum of the squares of the coefficients is one for each of them. These eigenvectors then provide the coefficients of the principal components.

The eigenvalue for a principal component indicates the variance that it accounts for out of the total variances of 5.000. Thus, the first principal component accounts for (3.616/5.000)100% = 72.3% of the total variance. Similarly, the other principal components in order account for 10.6%, 7.7%, 6.0%, and 3.3%, respectively, of the total variance. Clearly, the first component is far more important than any of the others.

Another way of looking at the relative importance of principal components is in terms of their variance in comparison with the variance of the original variables. After standardization, the original variables all have variances of 1.0. The first principal component, therefore, has a variance of 3.616 original variables. However, the second principal component has a variance of only 0.532 of that of one of the original variables, while the other principal components account for even less variation. This confirms the importance of the first principal component in comparison with the others.

The first principal component is

$$Z_1 = 0.452X_1 + 0.462X_2 + 0.451X_3 + 0.471X_4 + 0.398X_5$$

where X_1 to X_5 are the standardized variables. The coefficients of the X variables are nearly equal, and this is clearly an index of the size of

the sparrows. It seems, therefore, that about 72.3% of the variation in the data is related to size differences among the sparrows.

The second principal component is

$$Z_2 = -0.051\,X_1 + 0.300\,X_2 + 0.325\,X_3 + 0.185\,X_4 - 0.877\,X_5$$

This is a contrast between variables X_2 (alar extent), X_3 (length of beak and head), and X_4 (length of humerus), on the one hand, and variable X_5 (length of the keel of the sternum), on the other. That is to say, Z_2 will be high if X_2, X_3, and X_4 are high but X_5 is low. On the other hand, Z_2 will be low if X_2, X_3, and X_4 are low but X_5 is high. Hence, Z_2 represents a shape difference between the sparrows. The low coefficient of X_1 (total length) means that the value of this variable does not affect Z_2 very much. The other principal components can be interpreted in a similar way. They therefore represent other aspects of shape differences.

The values of the principal components may be useful for further analyses. They are calculated in the obvious way from the standardized variables. Thus, for the first bird, the original variable values are $x_1 = 156$, $x_2 = 245$, $x_3 = 31.6$, $x_4 = 18.5$, and $x_5 = 20.5$. These standardize to $x_1 = (156 - 157.980)/3.654 = -0.542$, $x_2 = (245 - 241.327)/5.068 = 0.725$, $x_3 = (31.6 - 31.459)/0.795 = 0.177$, $x_4 = (18.5 - 18.469)/0.564 = 0.055$, and $x_5 = (20.5 - 20.827)/0.991 = -0.330$, where in each case the variable mean for the 49 birds has been subtracted, and a division has been made by the sample standard deviation for the 49 birds. The value of the first principal component for the first bird is therefore

$$\begin{aligned} Z_1 &= 0.452 \times (-0.542) + 0.462 \times 0.725 + 0.451 \times 0.177 + 0.471 \times 0.055 \\ &\quad + 0.398 \times (-0.330) \\ &= 0.064 \end{aligned}$$

The second principal component for the same bird is

$$\begin{aligned} Z_2 &= -0.051 \times (-0.542) + 0.300 \times 0.725 + 0.325 \times 0.177 + 0.185 \times 0.055 \\ &\quad - 0.877 \times (-0.330) \\ &= 0.602 \end{aligned}$$

The other principal components can be calculated in a similar way.

The birds being considered were picked up after a severe storm. The first 21 of them recovered, while the other 28 died. A question of some interest is, therefore, whether the survivors and nonsurvivors show any differences. It has been shown in Example 4.1 that there is no evidence of any differences in mean values. However, in Example 4.2, it has been shown that the survivors seem to have been less variable than the nonsurvivors. The situation will now be considered in terms of principal components.

Table 6.4 Comparison between survivors and nonsurvivors in terms of means and standard deviations of principal components

Principal component	Mean		Standard deviation	
	Survivors	Nonsurvivors	Survivors	Nonsurvivors
1	−0.100	0.075	1.506	2.176
2	0.004	−0.003	0.684	0.776
3	−0.140	0.105	0.522	0.677
4	0.073	−0.055	0.563	0.543
5	0.023	−0.017	0.411	0.408

The means and standard deviations of the five principal components are shown in Table 6.4 separately for survivors and nonsurvivors. None of the mean differences between survivors and nonsurvivors are significant from t-tests, and none of the standard deviation differences are significant on F-tests. However, Levene's test on deviations from medians (described in Chapter 4) just gives a significant difference between the variation of principal component 1 for survivors and nonsurvivors on a one-sided test at the 5% level. The assumption for the one-sided test is that, if anything, nonsurvivors were more variable than survivors. The variation is not significantly different for survivors and nonsurvivors with Levene's test on the other principal components. As principal component 1 measures overall size, it seems that stabilizing selection may have acted against very large and very small birds.

Figure 6.1 shows a plot of the values of the 49 birds for the first two principal components, which between them account for 82.9%

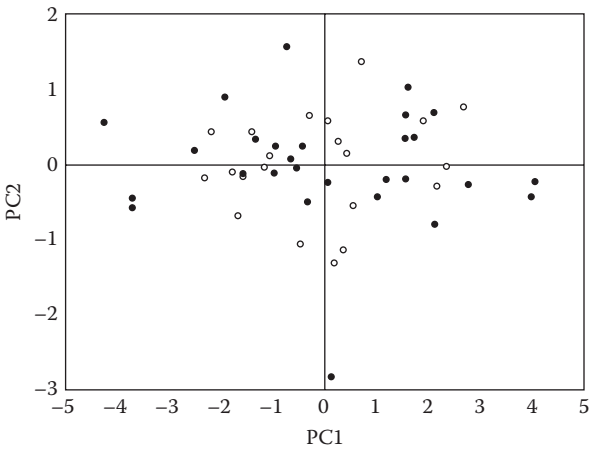

Figure 6.1 Plot of 49 female sparrows against values for the first two principal components, PC1 and PC2. ○ = survivor, ● = nonsurvivor.

of the variation in the data. The figure shows quite clearly how birds with extreme values for the first principal component failed to survive. Indeed, there is a suggestion that this was true for principal component 2 as well.

It is important to realize that some computer programs may give the principal components as shown with this example but with the signs of the coefficients of the body measurements reversed. For example, Z_2 might be shown as

$$Z_2 = 0.051X_1 - 0.300X_2 - 0.325X_3 - 0.185X_4 + 0.877X_5$$

This is not a mistake. The principal component is still measuring exactly the same aspect of the data, but in the opposite direction.

Example 6.2: Employment in European countries

As a second example of a principal components analysis, consider the data in Table 1.5 on the percentages of people employed in nine industry sectors in Europe. The correlation matrix for the nine variables is shown in Table 6.5. Overall, the values in this matrix are not particularly high, which indicates that several principal components will be required to account for the variation in the data.

The eigenvalues of the correlation matrix, with percentages of the total of 9.000 in parentheses, are 3.112 (34.6%), 1.809 (20.1%), 1.496 (16.6%), 1.063 (11.8%), 0.710 (7.9%), 0.311 (3.5%), 0.293 (3.3%), 0.204 (2.3%), and 0.000 (0.0%). The last eigenvalue is zero because the sum of the nine variables being analyzed is 100% before standardization. The principal component corresponding to this eigenvalue has the value zero for all the countries, and hence has a zero variance. If any linear

Table 6.5 The correlation matrix for percentages employed in nine industry groups in 30 countries in Europe in lower diagonal form, calculated from the data in Table 1.5

	AGR	MIN	MAN	PS	CON	SER	FIN	SPS	TC
AGR	1.000	—							
MIN	0.316	1.000	—						
MAN	−0.254	−0.672	1.000	—					
PS	−0.382	−0.387	0.388	1.000	—				
CON	−0.349	−0.129	−0.034	0.165	1.000	—			
SER	−0.605	−0.407	−0.033	0.155	0.473	1.000	—		
FIN	−0.176	−0.248	−0.274	0.094	−0.018	0.379	1.000	—	
SPS	−0.811	−0.316	0.050	0.238	0.072	0.388	0.166	1.000	—
TC	−0.487	0.045	0.243	0.105	−0.055	−0.085	−0.391	0.475	1.000

Note: The variables are the percentages employed in AGR, agriculture, forestry, and fishing; MIN, mining and quarrying; MAN, manufacturing; PS, power and water supplies; CON, construction; SER, services; FIN, finance; SPS, social and personal services; TC, transport and communications.

combination of the original variables in a principal components analysis is constant, then this must of necessity result in one of the eigenvalues being zero.

This example is not as straightforward as the previous one. The first principal component only accounts for about 35% of the variation in the data, and four components are needed to account for 83% of the variation. It is a matter of judgment as to how many components are important. It can be argued that only the first four should be considered, because these are the ones with eigenvalues greater than one. To some extent, the choice of the number of components that are important will depend on the use that is going to be made of them. For the present example, it will be assumed that a small number of indices are required to present the main aspects of differences between the countries, and for simplicity, only the first two components will be examined further. Between them, they account for about 55% of the variation in the original data.

The first component is

$$Z_1 = 0.51(\text{AGR}) + 0.37(\text{MIN}) - 0.25(\text{MAN}) - 0.31(\text{PS}) - 0.22(\text{CON}) \\ -0.38(\text{SER}) - 0.13(\text{FIN}) - 0.42(\text{SPS}) - 0.21(\text{TC})$$

where the abbreviations for variables are defined in Table 6.5. As the analysis has been done on the correlation matrix, the variables in this equation are the original percentages after they have each been standardized to have a mean of zero and a standard deviation of one. From the coefficients of Z_1, it can be seen that it is a contrast between the numbers engaged in agriculture, forestry and fishing (AGR) and mining and quarrying (MIN), and the numbers engaged in other occupations.

The second component is

$$Z_2 = -0.02(\text{AGR}) + 0.00(\text{MIN}) + 0.43(\text{MAN}) + 0.11(\text{PS}) - 0.24(\text{CON}) \\ - 0.41(\text{SER}) - 0.55(\text{FIN}) + 0.05(\text{SPS}) + 0.52(\text{TC})$$

which primarily contrasts the numbers in manufacturing (MAN) and transport and communications (TC) with the numbers in construction (CON), service industries (SER), and finance (FIN).

Figure 6.2 shows a plot of the 30 countries against their values for Z_1 and Z_2. The picture is certainly rather meaningful in terms of what is known about the countries. Most of the traditional Western democracies are grouped with slightly negative values for Z_1 and values for Z_2 between about plus and minus one. Gibraltar and Albania stand out as having rather distinct employment patterns, while the remaining countries lie in a band ranging from the former Yugoslavia ($Z_1 = -1.2$, $Z_2 = 2.2$) to Turkey ($Z_1 = 3.2$, $Z_2 = -0.3$).

As with the previous example, it is possible that some computer programs will produce the principal components shown here, but

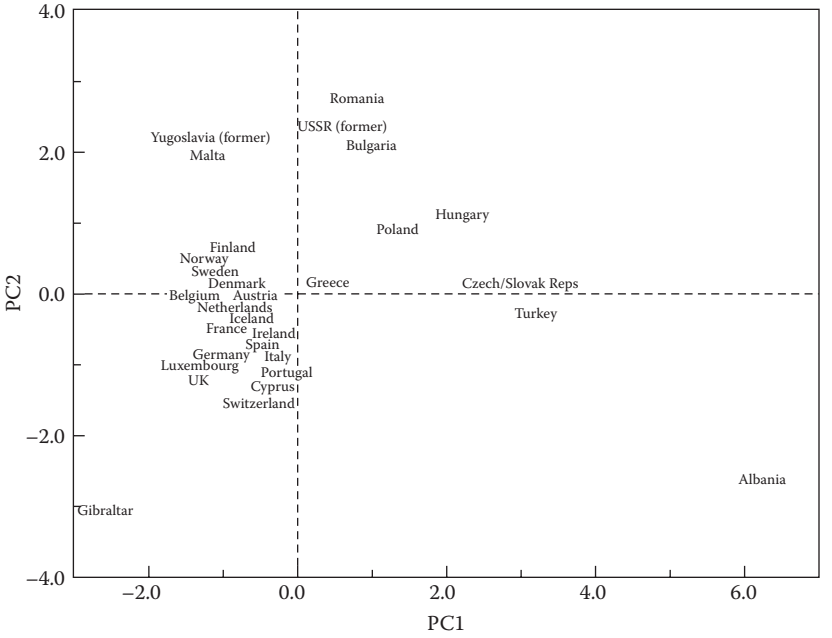

Figure 6.2 European countries plotted against the first two principal components for employment variables.

with the signs of the coefficients of the original variables reversed. The components still measure the same aspects of the data, but with the high and low values reversed.

6.3 Computer programs

The Appendix to this chapter provides the R code for carrying out a principal components analysis, and many standard statistical packages will carry out this analysis, because it is one of the most common types of multivariate analysis in use. When the analysis is not mentioned as an option in a package, it may still be possible to do the required calculations as a special type of factor analysis, as explained in Chapter 7. In that case, care will be needed to ensure that there is no confusion between the principal components and the factors, which are the principal components scaled to have unit variances.

This confusion can also occur with some programs that claim to be carrying out a principal component analysis. Instead of providing the values of the principal components (with variances equal to eigenvalues), they provide values of the principal components scaled to have variances of one.

6.4 Further reading

Principal components analysis is covered in almost all texts on multivariate analysis, and in greater detail by Jolliffe (1986) and Jackson (1991). Social scientists may also find the shorter monograph by Dunteman (1989) to be helpful.

Exercises

1. Table 6.6 shows six measurements on each of 25 pottery goblets excavated from prehistoric sites in Thailand, with Figure 6.3 illustrating the typical shape and the nature of the measurements. The main

Table 6.6 Measurements (in centimeters) taken on 25 prehistoric goblets from Thailand

Goblet	X_1	X_2	X_3	X_4	X_5	X_6
1	13	21	23	14	7	8
2	14	14	24	19	5	9
3	19	23	24	20	6	12
4	17	18	16	16	11	8
5	19	20	16	16	10	7
6	12	20	24	17	6	9
7	12	19	22	16	6	10
8	12	22	25	15	7	7
9	11	15	17	11	6	5
10	11	13	14	11	7	4
11	12	20	25	18	5	12
12	13	21	23	15	9	8
13	12	15	19	12	5	6
14	13	22	26	17	7	10
15	14	22	26	15	7	9
16	14	19	20	17	5	10
17	15	16	15	15	9	7
18	19	21	20	16	9	10
19	12	20	26	16	7	10
20	17	20	27	18	6	14
21	13	20	27	17	6	9
22	9	9	10	7	4	3
23	8	8	7	5	2	2
24	9	9	8	4	2	2
25	12	19	27	18	5	12

Note: The data were kindly provided by Professor C.F.W. Higham of the University of Otago, New Zealand. The variables are defined in Figure 6.3.

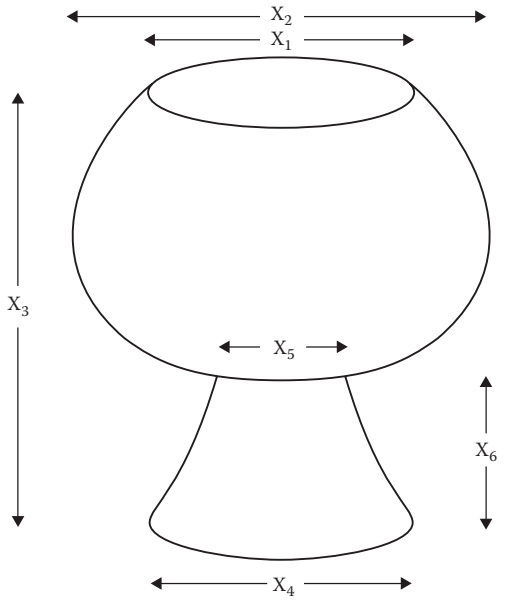

Figure 6.3 Measurements made on pottery goblets from Thailand.

question of interest for these data concerns similarities and differences between the goblets, with obvious questions being whether it is possible to display the data graphically to show how the goblets are related, and if so, whether there are any obvious groupings of similar goblets and any goblets that are particularly unusual. Carry out a principal components analysis and see whether the values of the principal components help to answer these questions.

One point that needs consideration with this exercise is the extent to which differences between goblets are due to shape differences rather than size differences. It may well be considered that two goblets that are almost the same shape but have very different sizes are really similar. The problem of separating size and shape differences has generated a considerable scientific literature, which will not be considered here. However, it can be noted that one way to remove the effects of size involves dividing the measurements for a goblet by the total height of the body of the goblet. Alternatively, the measurements on a goblet can be expressed as a proportion of the sum of all measurements on that goblet. These types of standardization of variables will ensure that the data values are similar for two goblets with the same shape but different sizes.

2. Table 6.7 shows estimates of the average protein consumption from different food sources for the inhabitants of 25 European countries

Table 6.7 Protein consumption (grams per person per day) in 25 European countries

Country	Red meat	White meat	Eggs	Milk	Fish	Cereals	Starchy foods	Pulses, nuts, and oilseeds	Fruit and vegetables	Total
Albania	10	1	1	9	0.0	42	1	6	2	72
Austria	9	14	4	20	2.0	28	4	1	4	86
Belgium	14	9	4	18	5.0	27	6	2	4	89
Bulgaria	8	6	2	8	1.0	57	1	4	4	91
Czechoslovakia	10	11	3	13	2.0	34	5	1	4	83
Denmark	11	11	4	25	10.0	22	5	1	2	91
East Germany	8	12	4	11	5.0	25	7	1	4	77
Finland	10	5	3	34	6.0	26	5	1	1	91
France	18	10	3	20	6.0	28	5	2	7	99
Greece	10	3	3	18	6.0	42	2	8	7	99
Hungary	5	12	3	10	0.0	40	4	5	4	83
Ireland	14	10	5	26	2.0	24	6	2	3	92
Italy	9	5	3	14	3.0	37	2	4	7	84
The Netherlands	10	14	4	23	3.0	22	4	2	4	86
Norway	9	5	3	23	10.0	23	5	2	3	83
Poland	7	10	3	19	3.0	36	6	2	7	93
Portugal	6	4	1	5	14.0	27	6	5	8	76
Romania	6	6	2	11	1.0	50	3	5	3	87
Spain	7	3	3	9	7.0	29	6	6	7	77
Sweden	10	8	4	25	8.0	20	4	1	2	82
Switzerland	13	10	3	24	2.0	26	3	2	5	88
United Kingdom	17	6	5	21	4.0	24	5	3	3	88
USSR	9	5	2	17	3.0	44	6	3	3	92
West Germany	11	13	4	19	3.0	19	5	2	4	80
Yugoslavia	4	5	1	10	1.0	56	3	6	3	89

as published by Weber (1973). Use principal components analysis to investigate the relationships between the countries on the basis of these variables.

References

Dunteman, G.H. (1989). *Principal Components Analysis*, Newbury Park, CA: Sage.
Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24: 417–41; 498–520.
Jackson, J.E. (1991). *A User's Guide to Principal Components*, New York: Wiley.
Jolliffe, I.T. (2002). *Principal Component Analysis*. 2nd Edn. New York: Springer.
Pearson, K. (1901). On lines and planes of closest fit to a system of points in space. *Philosophical Magazine* 2: 557–72.
Weber, A. (1973). *Agrarpolitik im Spannungsfeld der Internationalen Ernährungspolitik*. Kiel: Institut für Agrapolitik und Marktlehre.

Appendix: Principal Components Analysis (PCA) in R

The default R installation provides two computational methods for principal components analysis, performed by two functions, `princomp()` and `prcomp()`, which are loaded each time R is invoked. The former uses an algorithm that closely follows the procedure described in Section 6.2, based on the calculation of eigenvalues of the correlation matrix, or the covariance matrix if this is desired. In matrix algebra terminology, this technique is known as the *spectral decomposition of the covariance or correlation matrix*. On the other hand, `prcomp()` applies a method called the *singular value decomposition* (SVD) (Anton, 2013), a general procedure of matrix factorization, useful for handling matrices that are singular or nearly singular. The application of SVD to principal component analysis is supported by two facts: first, that the nonzero singular values of any real matrix **M** are the square roots of the nonzero eigenvalues of both $\mathbf{MM}^T$ and $\mathbf{M}^T\mathbf{M}$, the product of a matrix by its transpose, and vice versa; second, that the covariance or the correlation matrices can be expressed as the multiplication of these two matrices. In general, `princomp()` and `prcomp()` produce similar results. However, the implementation of the SVD algorithm is more accurate from a numerical point of view. Thus, `prcomp()` or any other R function for principal component analysis based on the SVD procedure is preferable.

R scripts with comments are provided in this book's website as computational aids for getting the results in Example 6.1, the analysis of the Bumpus sparrow data, and Example 6.2, the analysis of employment data in European countries. The basic command used in these examples is

```
pca.results<-prcomp(data, scale=TRUE,...)
```

The option `scale=TRUE` means that the principal components are computed on the correlation matrix. It is worth noticing that the object `pca.results` produced by `prcomp()` contains the singular values of the correlation or the covariance matrix. These values, labeled with the heading *Standard Deviation*, are revealed when the function `print(pca.results)` is executed. The eigenvalues are simply these standard deviations squared. In addition to the singular values, the eigenvectors, and the values for each principal component, two-dimensional graphical summaries of the PCA can be built by means of the function `plot()`, like the plots shown in Figures 6.1 and 6.2. It is also possible to produce a variation of the plot for the first two principal components through the command `biplot(pca.results)`. A biplot is a graphical summary of the PCA in which the first two principal components, plotted as points,

are simultaneously displayed with a projection of the variables in the two-dimensional reduced space, plotted as arrows. See Gower and Hand (1996) for further details.

Several R packages offer functions for principal component analysis in addition to `princomp` and `prcompr`. A list of some of these packages and functions is given here. The one to use can be chosen based on the descriptions provided for each package in the corresponding help files or manuals.

Package	Function for PCA	Reference
`stats`	`princomp()`	R documentation (R Core Team 2016)
`stats`	`prcomp()`	R documentation (R Core Team 2016)
`FactoMineR`	`PCA()`	Le et al. (2008)
`ade4`	`dudi.pca()`	Dray and Dufour (2007)
`vegan`	`rda()`	Oksanen et al. (2016)
`amap`	`acp()`	Lucas (2014)

References

Anton, H. (2013). *Elementary Linear Algebra.* 11th Edn. New York: Wiley.

Dray, S. and Dufour, A.B. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software* 22(4): 1–20.

Gower, J.C. and Hand, D.J. (1996). *Biplots. Monographs on Statistics and Applied Probability.* London: Chapman & Hall.

Le, S., Josse, J. and Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software* 25(1): 1–18.

Lucas, A. (2014). amap: Another Multidimensional Analysis Package. R package version 0.8-14. https://CRAN.R-project.org/package=amap

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., et al. (2016). vegan: Community Ecology Package. R package version 2.4-0. http://CRAN.R-project.org/package=vegan

R Core Team (2016). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing. https://www.r-project.org/

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

chapter seven

Factor analysis

7.1 The factor analysis model

Factor analysis has similar aims to principal components analysis. The basic idea is still that it may be possible to describe a set of p variables X_1, X_2, …, X_p in terms of a smaller number of indices or factors, and in the process get a better understanding of the relationship between these variables. There is, however, one important difference. Principal components analysis is not based on any particular statistical model, whereas factor analysis is based on a model.

Charles Spearman is credited with the early development of factor analysis. He studied the correlations between students' test scores of various types and noted that many observed correlations could be accounted for by a simple model (Spearman, 1904). For example, in one case, he obtained the matrix of correlations shown in Table 7.1 for how boys in a preparatory school scored on tests in Classics, French, English, mathematics, discrimination of pitch, and music. He noted that this matrix has the interesting property that any two rows are almost proportional if the diagonals are ignored. Thus, for rows Classics and English, there are the ratios

$$\frac{0.83}{0.67} \approx \frac{0.70}{0.64} \approx \frac{0.66}{0.54} \approx \frac{0.63}{0.51} \approx 1.2$$

Based on this observation, Spearman suggested that the six test scores could be described by the equation

$$X_i = a_i F + e_i$$

where:

- X_i is the ith score after it has been standardized to have a mean of zero and a standard deviation of one for all the boys
- a_i is a constant
- F is a factor value, which has a mean of zero and a standard deviation of one for all the boys
- e_i is the part of X_i that is specific to the ith test only

Table 7.1 Correlations between test scores for boys in a preparatory school

	Classics	French	English	Mathematics	Discrimination of pitch	Music
Classics	1.00	0.83	0.78	0.70	0.66	0.63
French	0.83	1.00	0.67	0.67	0.65	0.57
English	0.78	0.67	1.00	0.64	0.54	0.51
Mathematics	0.70	0.67	0.64	1.00	0.45	0.51
Discrimination of pitch	0.66	0.65	0.54	0.45	1.00	0.40
Music	0.63	0.57	0.51	0.51	0.40	1.00

Source: Data from Spearman, C., *Am. J. Psychol.*, 15, 201–93, 1904.

Spearman showed that a constant ratio between the rows of a correlation matrix follows as a consequence of these assumptions, and therefore, this is a plausible model for the data.

Apart from the constant correlation ratios, it also follows that the variance of X_i is given by

$$\begin{aligned} \mathrm{Var}(X_i) &= \mathrm{Var}(a_i F + e_i) \\ &= \mathrm{Var}(a_i F) + \mathrm{Var}(e_i) \\ &= a_i^2\ \mathrm{Var}(F) + \mathrm{Var}(e_i) \\ &= a_i^2 + \mathrm{Var}(e_i) \end{aligned}$$

because a_i is a constant, F and e_i are assumed to be independent, and the variance of F is assumed to be unity. Also, because $\mathrm{Var}(X_i) = 1$,

$$1 = a_i^2 + \mathrm{Var}(e_i)$$

Hence, the constant a_i, which is called the *factor loading*, is such that its square is the proportion of the variance of X_i that is accounted for by the factor.

On the basis of his work, Spearman formulated his two-factor theory of mental tests. According to this theory, each test result is made up of two parts: one that is common to all the tests (general intelligence) and another that is specific to the test in question. Later, this theory was modified to allow each test result to consist of a part due to several common factors plus a part specific to the test. This gives the general factor analysis model, which states that

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \ldots + a_{im}F_m + e_i$$

where:

X_i	is the ith test score with mean zero and unit variance
a_{i1} to a_{im}	are the factor loadings for the ith test
F_1 to F_m	are m uncorrelated common factors, each with mean zero and unit variance
e_i	is specific only to the ith test, is uncorrelated with any of the common factors, and has zero mean

With this model,

$$\begin{aligned} \mathrm{Var}(X_i) = 1 &= a_{i1}^2\ \mathrm{Var}(F_1) + a_{i2}^2\ \mathrm{Var}(F_2) + \ldots + a_{im}^2\ \mathrm{Var}(F_m) + \mathrm{Var}(e_i) \\ &= a_{i1}^2 + a_{i2}^2 + \ldots + a_{im}^2 + \mathrm{Var}(e_i) \end{aligned}$$

where:

$a_{i1}^2 + a_{i2}^2 + \ldots + a_{im}^2$ is called the *communality* of X_i (the part of its variance that is related to the common factors)

$\text{Var}(e_i)$ is called the *specificity* of X_i (the part of its variance that is unrelated to the common factors)

It can also be shown that the correlation between X_i and X_j is

$$r_{ij} = a_{i1}a_{j1} + a_{12}a_{j2} + \ldots + a_{im}a_{jm}$$

Hence, two test scores can only be highly correlated if they have high loadings on the same factors. Furthermore, $-1 \leq a_{jj} \leq +1$, as the communality cannot exceed one.

7.2 *Procedure for a factor analysis*

The data for a factor analysis have the same form as for a principal components analysis. That is, there are p variables with values for these n individuals, as shown in Table 6.2.

There are three stages to a factor analysis. To begin with, provisional factor loadings a_{ij} are determined. One approach starts with a principal components analysis and neglects the principal components after the first m, which are then taken to be the m factors. The factors found in this way are then uncorrelated with each other and are also uncorrelated with the specific factors. However, the specific factors are not uncorrelated with each other, which means that one of the assumptions of the factor analysis model does not hold. This may not matter much, providing that the communalities are high.

In whatever way the provisional factor loadings are determined, it is possible to show that they are not unique. If F_1, F_2, …, F_m are the provisional factors, then linear combinations of these of the form

$$F^*_1 = d_{11}F_1 + d_{12}F_2 + \ldots + d_{1m}F_m$$

$$F^*_2 = d_{21}F_1 + d_{22}F_2 + \ldots + d_{2m}F_m$$

$$\vdots$$

$$F^*_m = d_{m1}F_1 + d_{m2}F_2 + \ldots + d_{mm}F_m$$

can be constructed that are uncorrelated and explain the data just as well as the provisional factors. Indeed, there are an infinite number of

alternative solutions for the factor analysis model. This leads to the second stage in the analysis, which is called *factor rotation*. At this stage, the provisional factors are transformed to find new factors that are easier to interpret. To rotate or to transform in this context means essentially to choose the d_{ij} values in the equations in the above equations.

The last stage of an analysis involves calculating the factor scores. These are the values of the rotated factors F^*_1, F^*_2, … , F^*_m for each of the n individuals for which data are available.

Generally, the number of factors (m) is up to the user, although it may sometimes be suggested by the nature of the data. When a principal components analysis is used to find a provisional solution, a rough rule of thumb involves choosing m to be the number of eigenvalues greater than unity for the correlation matrix of the test scores. The logic here is the same as was explained in the previous chapter on principal components analysis. A factor associated with an eigenvalue of less than unity accounts for less variation in the data than one of the original test scores. In general, increasing m will increase the communalities of variables. However, communalities are not changed by factor rotation.

Factor rotation can be orthogonal or oblique. With orthogonal rotation, the new factors are uncorrelated, like the provisional factors. With oblique rotation, the new factors are correlated. Whichever type of rotation is used, it is desirable that the factor loadings for the new factors should be either close to zero or very different from zero. A near-zero a_{ij} means that X_i is not strongly related to the factor F_j. A large positive or negative value of a_{ij} means that X_i is determined by F_j to a large extent. If each test score is strongly related to some factors but not related to the others, then this makes the factors easier to identify than would otherwise be the case.

One method of orthogonal factor rotation that is often used is called *varimax rotation*. This is based on the assumption that the interpretability of factor j can be measured by the variance of the squares of its factor loadings, that is, the variance of a_{1j}^2, a_{2j}^2, …, a_{pj}^2. If this variance is large, then the a_{ij} values tend to be either close to zero or close to unity. Varimax rotation, therefore, maximizes the sum of these variances for all of the factors. Kaiser (1958) first suggested this approach. Later, he modified it slightly by normalizing the factor loadings before maximizing the variances of their squares, because this appears to give improved results. Varimax rotation can, therefore, be carried out with or without Kaiser normalization. Numerous other methods for orthogonal rotation have been proposed. However, varimax rotation seems to be a good standard approach.

Sometimes, factor analysts are prepared to give up the idea of the factors being uncorrelated so that the factor loadings should be as simple as possible. An oblique rotation may then give a better solution than an orthogonal one. Again, there are numerous methods available to do the oblique rotation.

A method for calculating the factor scores for individuals based on principal components is described in the section below. There are other methods available, so the one to be used will depend on the computer package or R code being used for an analysis.

7.3 Principal components factor analysis

It has been remarked above that one way to do a factor analysis is to begin with a principal components analysis and use the first few principal components as unrotated factors. This has the virtue of simplicity, although as the specific factors e_1, e_2, … , e_p are correlated, the factor analysis model is not quite correct. Sometimes, factor analysts do a principal components factor analysis first and then try other approaches afterward.

The method for finding the unrotated factors is as follows. With p variables, there will be the same number of principal components. These are linear combinations of the original variables

$$\begin{aligned} Z_1 &= b_{11}X_1 + b_{12}X_2 + \ldots + b_{1p}X_p \\ Z_2 &= b_{21}X_1 + b_{22}X_2 + \ldots + b_{2p}X_p \\ &\vdots \\ Z_p &= b_{p1}X_1 + b_{p2}X_2 + \ldots + b_{pp}X_p \end{aligned} \tag{7.1}$$

where the b_{ij} values are given by the eigenvectors of the correlation matrix. This transformation from X values to Z values is orthogonal, so that the inverse relationship is simply

$$\begin{aligned} X_1 &= b_{11}Z_1 + b_{21}Z_2 + \ldots + b_{p1}Z_p \\ X_2 &= b_{12}Z_1 + b_{22}Z_2 + \ldots + b_{p2}Z_p \\ &\vdots \\ X_p &= b_{1p}Z_1 + b_{2p}Z_2 + \ldots + b_{pp}Z_p \end{aligned}$$

For a factor analysis, only m of the principal components are retained, so the last equations become

$$X_1 = b_{11}Z_1 + b_{21}Z_2 + \ldots + b_{m1}Z_m + e_1$$

$$X_2 = b_{12}Z_1 + b_{22}Z_2 + \ldots + b_{m2}Z_m + e_2$$

$$\vdots$$

$$X_p = b_{1p}Z_1 + b_{2p}Z_2 + \ldots + b_{mp}Z_m + e_p$$

where e_i is a linear combination of the principal components Z_{m+1} to Z_p. All that needs to be done now is to scale the principal components Z_1, Z_2, … , Z_m to have unit variances, as required for factors. To do this, Z_i must be divided by its standard deviation, which is $\sqrt{\lambda_i}$, the square root of the corresponding eigenvalue in the correlation matrix. The equations then become

$$X_1 = \sqrt{\lambda}_1 b_{11}F_1 + \sqrt{\lambda}_2 b_{21}F_2 + \ldots + \sqrt{\lambda}_m b_{m1}F_m + e_1$$

$$X_2 = \sqrt{\lambda}_1 b_{12}F_1 + \sqrt{\lambda}_2 b_{22}F_2 + \ldots + \sqrt{\lambda}_m b_{m2}F_m + e_2$$

$$\vdots$$

$$X_p = \sqrt{\lambda}_1 b_{1p}F_1 + \sqrt{\lambda}_2 b_{2p}F_2 + \ldots + \sqrt{\lambda}_m b_{mp}F_m + e_p$$

where $F_i = Z_i/\sqrt{\lambda_i}$. The unrotated factor model is then

$$X_1 = a_{11}F_1 + a_{12}F_2 + \ldots + a_{1m}F_m + e_1$$

$$X_2 = a_{21}F_1 + a_{22}F_2 + \ldots + a_{2m}F_m + e_2$$

$$\vdots$$

$$X_p = a_{p1}F_1 + a_{p2}F_2 + \ldots + a_{pm}F_m + e_p \qquad (7.2)$$

where $a_{ij} = \sqrt{\lambda_j} b_{ji}$.

After a varimax or other type of rotation, a new solution has the form

$$\begin{aligned} X_1 &= g_{11}F^*_1 + g_{12}F^*_2 + \ldots + g_{1m}F^*_m + e_1 \\ X_2 &= g_{21}F^*_1 + g_{22}F^*_2 + \ldots + g_{2m}F^*_m + e_2 \\ &\vdots \\ X_p &= g_{p1}F^*_1 + g_{p2}F^*_2 + \ldots + g_{pm}F^*_m + e_p \end{aligned} \tag{7.3}$$

where F^*_i represents the new ith factor.

The values of the ith unrotated factor are just the values of the ith principal component after these have been scaled to have a variance of one. The values of the rotated factors are more complicated to obtain, but it can be shown that these are given by the matrix equation

$$\mathbf{F^* = XG(G'G)^{-1}} \tag{7.4}$$

where:

- **F*** is an n × m matrix containing the values for the m rotated factors in its columns, with one row for each of the n original rows of data
- **X** is the n × p matrix of the original data for p variables and n observations, after coding the variables X_1 to X_p to have means of zero and variances of one
- **G** is the p × m matrix of rotated factor loadings given by Equation 7.3

7.4 *Using a factor analysis program to do principal components analysis*

Because many computer programs for factor analysis allow the option of using principal components as initial factors, it is possible to use the programs to do principal components analysis. All that has to be done is to extract the same number of factors as variables and not do any rotation. The factor loadings will then be as given by Equation 7.2, with m = p and $e_1 = e_2 = \ldots = e_p = 0$. The principal components are given by Equation 7.1, with $b_{ij} = a_{ji}/\sqrt{\lambda_i}$, where λ_i is the ith eigenvalue.

Example 7.1: Employment in European countries

In Example 6.2, a principal components analysis was carried out on the percentages of people employed in nine industry groups in 30 countries in Europe for the years 1989 to 1995 (Table 1.5). It is of some interest to continue the examination of these data using a factor analysis model.

The correlation matrix for the nine percentage variables is given in Table 6.5, and the eigenvalues and eigenvectors of this correlation matrix are shown in Table 7.2. There are four eigenvalues greater than unity, suggesting that four factors should be considered, which is what will be done here.

The eigenvectors in Table 7.2 give the coefficients of the X variables for Equation 7.1. These are changed into factor loadings for four factors using Equation 7.2, to give the model

$$X_1 = +\mathbf{0.90}\cdot F_1 - 0.03\cdot F_2 - 0.34\cdot F_3 + 0.02\cdot F_4 + e_1\,(0.93)$$

$$X_2 = +\mathbf{0.66}\cdot F_1 - 0.00\cdot F_2 + \mathbf{0.63}\cdot F_3 + 0.12\cdot F_4 + e_2\,(0.85)$$

$$X_3 = -0.43\cdot F_1 + \mathbf{0.58}\cdot F_2 - \mathbf{0.61}\cdot F_3 + 0.06\cdot F_4 + e_3\,(0.91)$$

$$X_4 = -\mathbf{0.56}\cdot F_1 + 0.15\cdot F_2 - 0.36\cdot F_3 + 0.02\cdot F_4 + \mathbf{e}_4\,(0.46)$$

$$X_5 = -0.39\cdot F_1 - 0.33\cdot F_2 + 0.09\cdot F_3 + \mathbf{0.81}\cdot F_4 + e_5\,(0.92)$$

$$X_6 = -\mathbf{0.67}\cdot F_1 - \mathbf{0.55}\cdot F_2 + 0.08\cdot F_3 + 0.17\cdot F_4 + e_6\,(0.79)$$

$$X_7 = -0.23\cdot F_1 - \mathbf{0.74}\cdot F_2 - 0.12\cdot F_3 - \mathbf{0.50}\cdot F_4 + e_7\,(0.87)$$

$$X_8 = -\mathbf{0.76}\cdot F_1 + 0.07\cdot F_2 + 0.44\cdot F_3 - 0.03\cdot F_4 + e_8\,(0.88)$$

$$X_9 = +0.36\cdot F_1 + \mathbf{0.69}\cdot F_2 + \mathbf{0.50}\cdot F_3 - 0.04\cdot F_4 + e_9\,(0.87)$$

Here, the values in parentheses are the communalities. For example, the communality for variable X_1 is $(0.90)^2+(-0.03)^2+(-0.34)^2+(0.02)^2=0.93$. The communalities are quite high for all variables except X_4 (power supplies). Most of the variance for the other eight variables is, therefore, accounted for by the four common factors.

Factor loadings that are 0.50 or more (ignoring the sign) are bold in these equations. These large and moderate loadings indicate how the variables are related to the factors. It can be seen that X_1 is almost entirely accounted for by factor 1 alone, X_2 is a mixture of factor 1 and factor 3, X_3 is accounted for by factor 1 and factor 2, and so on. An undesirable property of this choice of factors is that five of the nine X variables are related strongly to two of the factors. This suggests that a factor rotation may provide a simpler model for the data.

Table 7.2 Eigenvalues and eigenvectors for the European employment data of Table 1.5

	Eigenvectors								
Eigenvalue	X_1 AGR	X_2 MIN	X_3 MAN	X_4 PS	X_5 CON	X_6 SER	X_7 FIN	X_8 SPS	X_9 TC
3.111	0.512	0.375	−0.246	−0.315	−0.222	−0.382	−0.131	−0.428	−0.205
1.809	−0.024	−0.000	0.432	0.109	−0.242	−0.408	−0.553	0.055	0.516
1.495	−0.278	0.516	−0.503	−0.292	0.071	0.064	−0.096	0.360	0.413
1.063	0.016	0.113	0.058	0.023	0.783	0.169	−0.489	−0.317	−0.042
0.705	0.025	−0.345	0.231	−0.854	−0.064	0.269	−0.133	0.046	0.023
0.311	−0.045	0.203	−0.028	0.208	−0.503	0.674	−0.399	−0.167	−0.136
0.293	0.166	−0.212	−0.238	0.065	0.014	−0.165	−0.463	0.619	−0.492
0.203	0.539	−0.447	−0.431	0.157	0.030	0.203	−0.026	−0.045	0.504
0.000	−0.582	−0.419	−0.447	−0.030	−0.129	−0.245	−0.191	−0.410	−0.061

Note: The variables are the percentages employed in nine industry groups: AGR, agriculture, forestry, and fishing; MIN, mining and quarrying; MAN, manufacturing; PS, power and water supplies; CON, construction; SER, services; FIN, finance; SPS, social and personal services; TC, transport and communications.

A varimax rotation with Kaiser normalization was carried out. This produced the model

$$X_1 = +\mathbf{0.85} \cdot F_1 + 0.10 \cdot F_2 + 0.27 \cdot F_3 - 0.36 \cdot F_4 + e_1$$

$$X_2 = +0.11 \cdot F_1 + 0.30 \cdot F_2 + \mathbf{0.86} \cdot F_3 - 0.10 \cdot F_4 + e_2$$

$$X_3 = -0.03 \cdot F_1 + 0.32 \cdot F_2 - \mathbf{0.89} \cdot F_3 - 0.09 \cdot F_4 + \mathbf{e}_3$$

$$X_4 = -0.19 \cdot F_1 - 0.04 \cdot F_2 - \mathbf{0.64} \cdot F_3 + 0.14 \cdot F_4 + e_4$$

$$X_5 = -0.02 \cdot F_1 + 0.08 \cdot F_2 - 0.04 \cdot F_3 + \mathbf{0.95} \cdot F_4 + e_5$$

$$X_6 = -0.35 \cdot F_1 - 0.48 \cdot F_2 - 0.15 \cdot F_3 + \mathbf{0.65} \cdot F_4 + e_6$$

$$X_7 = -0.08 \cdot F_1 - \mathbf{0.93} \cdot F_2 + 0.00 \cdot F_3 - 0.01 \cdot F_4 + e_7$$

$$X_8 = -\mathbf{0.91} \cdot F_1 - 0.17 \cdot F_2 - 0.12 \cdot F_3 - 0.04 \cdot F_4 + e_8$$

$$X_9 = -\mathbf{0.73} \cdot F_1 + \mathbf{0.57} \cdot F_2 - 0.03 \cdot F_3 - 0.14 \cdot F_4 + e_9$$

The communalities are unchanged, and the factors are still uncorrelated. However, this is a slightly better solution than the previous one, as only X_9 is appreciably dependent on more than one factor.

At this stage, it is usual to try to put labels on factors. In the present case, this is not too difficult, based on the highest loadings only.

Factor 1 has a high positive loading for X_1 (agriculture, forestry, and fishing) and high negative loadings for X_8 (social and personal services) and X_9 (transport and communications). Therefore, it measures the extent to which people are employed in agriculture rather than services and communications. It can therefore be labeled "rural industries rather than social service and communication."

Factor 2 has high negative loadings for X_7 (finance) and a fairly high coefficient for X_9 (transport and communications). This can therefore be labeled "lack of finance industries."

Factor 3 has a high positive loading for X_2 (mining and quarrying), a high negative loading for X_3 (manufacturing), and a moderately high negative loading for X_4 (power supplies). This can therefore be labeled "mining rather than manufacturing."

Finally, factor 4 has a high positive loading for X_5 (construction) and a moderately high positive loading for X_6 (service industries). Therefore, "construction and service industries" seems to be a fair label in this case.

The **G** matrix of Equations 7.3 and 7.4 is given by the factor loadings shown in the second model in this example. For example, $g_{11} = 0.85$ and $g_{12} = 0.10$, to two decimal places. Using these loadings and carrying out the matrix calculations shown in Equation 7.4 provides the

Table 7.3 Rotated factor scores for 30 European countries

Country	Factor 1	Factor 2	Factor 3	Factor 4
Belgium	−0.97	−0.56	−0.10	−0.48
Denmark	−0.89	−0.47	−0.03	−0.67
France	−0.56	−0.78	−0.15	−0.25
Germany	0.05	−0.57	−0.47	0.58
Greece	0.48	0.19	−0.23	0.02
Ireland	0.28	−0.60	−0.36	0.03
Italy	0.25	−0.13	0.17	1.00
Luxembourg	−0.46	−0.36	0.02	0.92
Netherlands	−1.36	−1.56	−0.03	−2.09
Portugal	0.66	−0.45	−0.37	0.64
Spain	0.23	−0.11	−0.09	0.93
United Kingdom	−0.50	−1.14	−0.35	−0.04
Austria	0.18	0.05	−0.71	0.56
Finland	−0.78	−0.20	−0.21	−0.52
Iceland	−0.18	−0.04	−0.06	0.46
Norway	−1.36	−0.17	0.20	−0.42
Sweden	−1.20	−0.52	0.04	−0.74
Switzerland	0.12	−0.67	0.01	0.65
Albania	3.16	−1.82	1.76	−1.78
Bulgaria	0.47	1.56	−0.57	−0.65
Czech/Slovak Republics	−0.26	1.45	3.12	0.44
Hungary	−1.05	1.70	2.82	−0.15
Poland	0.97	0.71	−0.37	−0.42
Romania	1.11	1.73	−1.69	−0.81
USSR (Former)	0.08	2.09	−0.11	0.14
Yugoslavia (Former)	0.13	1.48	−1.70	0.17
Cyprus	0.46	−0.32	0.03	1.08
Gibraltar	−0.05	−1.05	0.08	3.26
Malta	−1.17	0.49	−0.79	−1.31
Turkey	2.15	0.07	0.15	−0.56

Note: Factor 1 is "rural industries rather than social service industries and communication," factor 2 is "lack of finance industries," factor 3 is "mining rather than manufacturing," and factor 4 is "construction industries."

values for the factor scores for each of the 30 countries in the original data set. These factor scores are shown in Table 7.3.

From studying the factor scores, it can be seen that the values for factor 1 emphasize the importance of rural industries rather than

Free ebooks ==> www.Ebook777.com

services and communications in Albania and Turkey. The values for factor 2 indicate that Bulgaria, Hungary, Romania, and the former USSR had few people employed in finance, but the Netherlands and Albania had large numbers employed in this area. The values for factor 3 contrast Albania and the Czech/Slovak Republics, with high levels of mining rather than manufacturing, with Romania and Yugoslavia, where the reverse is true. Finally, the values for factor 4 contrast Gibraltar, with high numbers in construction and service industries, with the Netherlands and Albania, where this is far from being the case.

It would be possible and reasonable to continue the analysis of this set of data by trying models with fewer factors and different methods of factor extraction. However, sufficient has been said already to indicate the general approach, so the example will be left at this point.

It should be kept in mind by anyone attempting to reproduce this analysis that different statistical packages may give the eigenvectors shown in Table 7.2 except that all the coefficients have their signs reversed. A sign reversal may also occur through a factor rotation, so that the loadings for a rotated factor are the opposite of what is shown above. Sign reversals like this just reverse the interpretation of the factor concerned. For example, if the loadings for the rotated factor 1 were the opposite of those shown above, then it would be interpreted as social and personal services, and transport and communications rather than rural industries.

7.5 Options in analyses

Computer programs for factor analysis, including different R codes, may allow a number of different options for the analysis, which is likely to be rather confusing for the novice in this area. Typically, there might be four or five methods for the initial extraction of factors and about the same number of methods for rotating these factors (including no rotation). This gives in the order of 20 different types of factor analysis that can be carried out, with results that will differ to some extent at least.

There is also the question of the number of factors to extract. Some packages may make an automatic choice, which may or may not be acceptable. The possibility of trying different numbers of factors therefore increases the choices for an analysis even more.

On the whole, it is probably best to avoid using too many options when first practicing factor analysis. The use of principal components as initial factors with varimax rotation, as used in the example in this chapter, is a reasonable start with any set of data. The maximum likelihood method for extracting factors is a good approach in principle, and might also be tried if this is available.

www.Ebook777.com

7.6 The value of factor analysis

Factor analysis is something of an art, and it is certainly not as objective as many statistical methods. For this reason, some statisticians are skeptical about its value. For example, Chatfield and Collins (1986) list six problems with factor analysis and conclude that "factor analysis should not be used in most practical situations." Similarly, Seber (2004) notes as a result of simulation studies that even if the postulated factor model is correct, then the chance of recovering it using available methods is not high.

On the other hand, factor analysis is widely used to analyze data and, no doubt, will continue to be widely used in the future. The reason for this is that users find the results useful for gaining insight into the structure of multivariate data. Therefore, if it is thought of as a purely descriptive tool, with limitations that are understood, it must take its place as one of the important multivariate methods. What should be avoided is carrying out a factor analysis on a single small sample that cannot be replicated and then assuming that the factors obtained must represent underlying variables that exist in the real world.

7.7 Discussion and further reading

Factor analysis is discussed in many texts on multivariate analysis, although, as noted in the previous section, the topic is sometimes not presented enthusiastically (Chatfield and Collins, 1986; Seber, 2004). Recent texts are generally more positive. For example, Rencher (2002) discusses at length the validity of factor analysis and why it often fails to work. He notes that there are many sets of data for which factor analysis should not be used, but others for which the method is useful.

Factor analysis as discussed in this chapter is often referred to as *exploratory factor analysis*, because it starts with no assumptions about the number of factors that exist or the nature of these factors. In this respect, it differs from what is called *confirmatory factor analysis*, which requires the number of factors and the factor structure to be specified in advance. In this way, confirmatory factor analysis can be used to test theories about the structure of the data.

Confirmatory factor analysis is more complicated to carry out than exploratory factor analysis. The details are described by Bernstein (1988, chapter 7), and Tabachnick and Fidell (2013). Confirmatory factor analysis is a special case of structural equation modeling, which is covered in Chapter 14 of the latter book.

Exercise

Using Example 7.1 as a model, carry out a factor analysis of the data in Table 6.7 on protein consumption from 10 different food sources for the

inhabitants of 25 European countries. Identify the important factors underlying the observed variables and examine the relationships between the countries with respect to these factors.

References

Bernstein, I.H. (1988). *Applied Multivariate Analysis*. Berlin: Springer.

Chatfield, C. and Collins, A.J. (1986). *Introduction to Multivariate Analysis*. London: Chapman and Hall.

Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23: 187–200.

Rencher, A.C. (2002). *Methods of Multivariate Statistics*. 2nd Edn. New York: Wiley.

Seber, G.A.F. (2004). *Multivariate Observations*. New York: Wiley.

Spearman, C. (1904). "General intelligence", objectively determined and measured. *American Journal of Psychology* 15: 201–93.

Tabachnick, B.G. and Fidell, L.S. (2013). *Using Multivariate Statistics*. 6th Edn. Boston, MA: Pearson.

Appendix: Factor Analysis in R

In its default package `stats`, R offers the function `factanal()` as a maximum likelihood (ML) method for extracting factors, a topic noted briefly in Section 7.5. Thus, ML factor analysis is also considered the default factor analysis in R. However, in Section 7.5, it was also emphasized that there are different approaches in factor analysis, each approach associated with a particular algorithm. Psychometric researchers have been the most interested in applying the range of existing algorithms for factor analysis. This explains why the R package psych, created and maintained by Revelle (2016a), has been considered as the main tool for psychometric applications of several multivariate methods, including factor analysis.

The `psych` package offers the `fa` function, from which the user may choose one of five methods of factor analysis (minimum residual, principal axis, weighted least squares, generalized least squares, and maximum likelihood factor analysis). Nevertheless, none of these options follows exactly the algorithm described in Section 7.3, whereby a principal components analysis (PCA) is used to produce initial factors, followed by a varimax rotation and the calculation of factor scores, which are also known as Bartlett scores, using Equation 7.4. It is not difficult to execute most of the steps of this PCA factor analysis with the set of R functions already considered in previous chapters (e.g., `prcomp`, matrix multiplication, and matrix inversion). The particular step completing this algorithm, varimax rotation with Kaiser normalization, can be performed with the `varimax()` function implemented in the stats package. However, this way of doing a PCA factor analysis can be avoided with `principal()`, another function in the `psych` package. Although this function is thought just to be doing a PCA, its output is organized in such a way that the component loadings are more suitable for a typical factor analysis, showing the best m factors. The developer of `psych` argues that the presence of `principal()` in his package as a choice for factor analysis, in addition to the algorithms executed by the `fa()` function, is because "psychologists typically use PCA in a manner similar to factor analysis and thus the principal function produces output that is perhaps more understandable than that produced by `princomp` in the stats package" (Revelle, 2016b). The command required to replicate the factor analysis described in Chapter 7 is then

```
principal(data, nfactors=4, rotate="varimax").
```

At this book's website, the reader will find two R scripts written to carry out the factor analysis performed for Example 7.1. One script does the calculations in the fastest way via the `principal()` function. The second script makes use of functions `prcomp()` and `varimax()`. It has been written for instructive purposes, so that the reader can follow in detail the application of Equations 7.1 through 7.4 in this chapter.

References

Revelle, W. (2016a). *PSYCH: Procedures for Personality and Psychological Research*. Evanston, IL: Northwestern University. http://CRAN.R-project.org/package=psych. Version = 1.6.6.

Revelle, W. (2016b). *An Overview of the psych Package: Vignette of psych Procedures for Psychological, Psychometric, and Personality Research*. https://cran.fhcrc.org/web/packages/psych/

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

chapter eight

Discriminant function analysis

8.1 The problem of separating groups

The problem that is addressed with discriminant function analysis is the extent to which it is possible to separate two or more groups of individuals, given measurements for these individuals on several variables. For example, with the data in Table 1.1 on five body measurements of 21 surviving and 28 nonsurviving sparrows, it is interesting to consider whether it is possible to use the body measurements to separate survivors and nonsurvivors. Also, for the data shown in Table 1.2 on four dimensions of Egyptian skulls for samples from five time periods, it is reasonable to consider whether the measurements can be used to assign skulls to different time periods.

In the general case, there will be m random samples from different groups, with sizes $n_1, n_2, \ldots n_m$, and values will be available for p variables $X_1, X_2, \ldots, X_p$ for each sample member. Thus, the data for a discriminant function analysis takes the form shown in Table 8.1. The data for a discriminant function analysis do not need to be standardized to have zero means and unit variances prior to the start of the analysis, as is usual with principal components and factor analysis. This is because the outcome of a discriminant function analysis is not affected in any important way by the scaling of individual variables.

8.2 Discrimination using Mahalanobis distances

One approach to discrimination is based on Mahalanobis distances, as defined in Section 5.3. The mean vectors for the m samples can be regarded as estimates of the true mean vectors for the groups. The Mahalanobis distances from the individual cases to the group centers can then be calculated, and each individual can be allocated to the group to which it is closest. This may or may not be the group that the individual actually came from, so the percentage of correct allocations is an indication of how well groups can be separated using the available variables.

This procedure is more precisely defined as follows. Let

$$\bar{\mathbf{x}}_i' = (\bar{x}_{1i}, \bar{x}_{2i}, \ldots, \bar{x}_{pi})'$$

Table 8.1 The form of data for a discriminant function analysis with m groups with possibly different sizes, and p variables measured on each individual case

Case	X_1	X_2	…	X_p	Group
1	x_{111}	x_{112}	…	x_{11p}	1
2	x_{211}	x_{212}	…	x_{21p}	1
⋮	⋮	⋮	⋮	⋮	⋮
n_1	$x_{n_1 11}$	$x_{n_1 12}$	…	$x_{n_1 1p}$	1
1	x_{121}	x_{122}	…	x_{12p}	2
2	x_{221}	x_{222}	…	x_{22p}	2
⋮	⋮	⋮	⋮	⋮	⋮
n_2	$x_{n_2 21}$	$x_{n_2 22}$	…	$x_{n_2 2p}$	2
1	x_{1m1}	x_{1m2}	…	x_{1mp}	m
2	x_{2m1}	x_{2m2}	…	x_{2mp}	m
⋮	⋮	⋮	⋮	⋮	⋮
n_m	$x_{n_m m1}$	$x_{n_m m2}$	…	$x_{n_m mp}$	m

denote the vector of mean values for the sample from the ith group, let $\mathbf{C}_i$ denote the covariance matrix for the same sample, and let $\mathbf{C}$ denote the pooled sample covariance matrix, where these vectors and matrices are calculated as explained in Section 2.7. Then, the Mahalanobis distance from an observation $\mathbf{x}' = (x_1, x_2, \ldots, x_p)'$ to the center of group i is estimated to be

$$\mathbf{D_i^2} = (\mathbf{x} - \bar{\mathbf{x}}_\mathbf{i})'\mathbf{C}^{-1}(\mathbf{x} - \bar{\mathbf{x}}_\mathbf{i})$$

$$\sum_{r=1}^{p}\sum_{s=1}^{p}\left(x_r - \bar{x}_{ri}\right)c^{rs}\left(x_s - \bar{x}_{si}\right) \tag{8.1}$$

where c^{rs} is the element in the rth row and the sth column of $\mathbf{C}^{-1}$. The observation $\mathbf{x}$ is then allocated to the group for which D_i^2 has the smallest value.

8.3 Canonical discriminant functions

It is sometimes useful to be able to determine functions of the variables $X_1, X_2, \ldots, X_p$ that in some sense separate the m groups as much as is possible. The simplest approach then involves taking a linear combination of the X variables

$$Z = a_1X_1 + a_2X_2 + \ldots + a_pX_p$$

Table 8.2 An analysis of variance on the Z index

Source of variation	Degrees of freedom	Mean square	F-ratio
Between groups	m – 1	M_B	M_B/M_W
Within groups	N – m	M_W	
	N – 1		

for this purpose. Groups can be well separated using values of Z if the mean value of this variable changes considerably from group to group, with the values within a group being fairly constant.

One way to determine the coefficients $a_1, a_2, \ldots, a_p$ in the index involves choosing these so as to maximize the F-ratio for a one-way analysis of variance. Thus, if there are a total of N individuals in all the groups, an analysis of variance on Z values takes the form shown in Table 8.2. Hence, a suitable function for separating the groups can be defined as the linear combination for which the F-ratio M_B/M_w is as large as possible, as first suggested by Fisher (1936).

When this approach is used, it turns out that it may be possible to determine several linear combinations for separating groups. In general, the number available, e.g., s, is the smaller of p and m – 1. The linear combinations are referred to as canonical discriminant functions.

The first function

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p$$

gives the maximum possible F-ratio for a one-way analysis of variance for the variation within and between groups. If there is more than one function, then the second one

$$Z_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$

gives the maximum possible F-ratio on a one-way analysis of variance, subject to the condition that there is no correlation between Z_1 and Z_2 within groups. Further functions are defined in the same way. Thus, the ith canonical discriminant function

$$Z_i = = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p$$

is the linear combination for which the F-ratio on an analysis of variance is maximized, subject to Z_i being uncorrelated with Z_1, Z_2, … , and Z_{i-1} within groups.

Finding the coefficients of the canonical discriminant functions turns out to be an eigenvalue problem. The within-sample matrix of sums of squares and cross products, **W**, and the total sample matrix of sums of squares and cross products matrix, **T**, are calculated as described in Section 4.7. From these, the between-groups matrix

$$\mathbf{B} = \mathbf{T} - \mathbf{W}$$

can be determined. Next, the eigenvalues and eigenvectors of the matrix $\mathbf{W}^{-1}\mathbf{B}$ have to be found. If the eigenvalues are $\lambda_1 > \lambda_2 > \ldots > \lambda_s$, then λ_i is the ratio of the between-group sum of squares to the within-group sum of squares for the ith linear combination, Z_i, while the elements of the corresponding eigenvector $\mathbf{a}_i' = (a_{i1}, a_{i2}, \ldots, a_{ip})$ are the coefficients of the X variables for this index.

The canonical discriminant functions $Z_1, Z_2, \ldots, Z_s$ are linear combinations of the original variables chosen in such a way that Z_1 reflects group differences as much as possible, Z_2 captures as much as possible of the group differences not displayed by Z_1, Z_3 captures as much as possible of the group differences not displayed by Z_1 and Z_2, and so on. The hope is that the first few functions are sufficient to account for almost all the important group differences. In particular, if only the first one or two functions are needed for this purpose, then a simple graphical representation of the relationship between the various groups is possible by plotting the values of these functions for the sample individuals.

8.4 Tests of significance

Several tests of significance are useful in conjunction with a discriminant function analysis. In particular, the T^2-test of Section 4.3 can be used to test for a significant difference between the mean values for any pair of groups, while one of the tests described in Section 4.7 can be used to test for overall significant differences between the means for the m groups.

In addition, a test is sometimes used for testing whether the mean of the discriminant function Z_j differs significantly from group to group. This is based on the individual eigenvalues of the matrix $\mathbf{W}^{-1}\mathbf{B}$. For example, sometimes, the statistic

$$\varphi_j^2 = \left\{N - 1 - (p + m)/2\right\} \log_e (1 + \lambda_j)$$

is used, where N is the total number of observations in all groups. This statistic is then tested against the chi-squared distribution with $p + m - 2j$ degrees of freedom (df), and a significantly large value is considered to provide evidence that the population mean values of Z_j vary from group

Free ebooks ==> www.Ebook777.com

to group. Alternatively, the sum $\varphi_j^2 + \varphi_{j+1}^2 + \ldots + \varphi_s^2$ is sometimes used for testing for group differences related to discriminant functions Z_j to Z_s. This is tested against the chi-squared distribution, with the df being the sum of those associated with the component terms. Other tests of a similar nature are also used.

Unfortunately, these tests are suspect to some extent, because the jth discriminant function in the population may not appear as the jth discriminant function in the sample due to sampling errors. For example, the estimated first discriminant function (corresponding to the largest eigenvalue for the sample matrix $\mathbf{W}^{-1}\mathbf{B}$) may in reality correspond to the second discriminant function for the population being sampled. Simulations indicate that this can upset the chi-squared tests described in the previous paragraph quite seriously. Therefore, it seems that the tests should not be relied on to decide how many of the obtained discriminant functions represent real group differences. See Harris (2001) for an extended discussion of the difficulties surrounding these tests and alternative ways to examine the nature of group differences.

One useful type of test that is valid, at least for large samples, involves calculating the Mahalanobis distance from each of the observations to the mean vector for the group containing the observation, as discussed in Section 5.3. These distances should follow approximately chi-squared distributions with p df. Hence, if an observation is very significantly far from the center of its group in comparison with the chi-squared distribution, then this brings into question whether the observation really came from the group in question.

8.5 Assumptions

The methods discussed so far in this chapter are based on two assumptions. First, for all the methods, the population within-group covariance matrix should be the same for all groups. Second, for tests of significance, the data should be multivariate normally distributed within groups.

In general, it seems that multivariate analyses that assume normality may be upset quite badly if this assumption is not correct. This contrasts with the situation with univariate analyses such as regression and analysis of variance, which are generally quite robust to this assumption. However, a failure of one or both assumptions does not necessarily mean that a discriminant function analysis is a waste of time. For example, it may well turn out that excellent discrimination is possible on data from nonnormal distributions, although it may not then be simple to establish the statistical significance of the group differences. Furthermore, discrimination methods that do not require the equality of population covariance matrices are available, as discussed in Section 8.12.

www.Ebook777.com

Example 8.1: Comparison of samples of Egyptian skulls

This example concerns the comparison of the values for four measurements on male Egyptian skulls for five samples ranging in age from the early predynastic period (circa 4000 BC) to the Roman period (circa AD 150). The data are shown in Table 1.2, and it has already been established that the mean values differ significantly from sample to sample (Example 4.3), with the differences tending to increase with the time difference between samples (Example 5.3).

The within-sample and total sample matrices of sums of squares and cross products are calculated as described in Section 4.7. They are found to be

$$\mathbf{W} = \begin{bmatrix} 3061.67 & 5.33 & 11.47 & 291.30 \\ 5.33 & 3405.27 & 754.00 & 412.53 \\ 11.47 & 754.00 & 3505.97 & 164.33 \\ 291.30 & 412.53 & 164.33 & 1472.13 \end{bmatrix}$$

and

$$\mathbf{T} = \begin{bmatrix} 3563.89 & -222.81 & -615.16 & 426.73 \\ -222.81 & 3635.17 & 1046.28 & 346.47 \\ -615.16 & 1046.28 & 4309.27 & -16.40 \\ 426.73 & 346.47 & -16.40 & 1533.33 \end{bmatrix}$$

The between-sample matrix is therefore

$$\mathbf{B} = \mathrm{T} - \mathbf{W} = \begin{bmatrix} 502.83 & -228.15 & -626.63 & 135.43 \\ -228.15 & 229.91 & 292.28 & -66.07 \\ -626.63 & 292.28 & 803.30 & -180.73 \\ 135.43 & -66.07 & -180.73 & 61.30 \end{bmatrix}$$

The eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ are found to be $\lambda_1 = 0.437$, $\lambda_2 = 0.035$, $\lambda_3 = 0.015$, and $\lambda_4 = 0.002$, and the corresponding canonical discriminant functions are

$$\begin{aligned} Z_1 &= -0.0107X_1 + 0.0040X_2 + 0.0119X_3 - 0.0068X_4 \\ Z_2 &= \ 0.0031X_1 + 0.0168X_2 - 0.0046X3 - 0.0022X_4 \\ Z_3 &= -0.0068X_1 + 0.0010X_2 + 0.0000X_3 + 0.0247X_4 \end{aligned} \qquad (8.2)$$

and

$$Z_4 = 0.0126X_1 - 0.0001X_2 + 0.0112X_3 + 0.0054X_4$$

Because λ_1 is much larger than the other eigenvalues, it is apparent that most of the sample differences are described by Z_1 alone.

The X variables in Equation 8.2 are the values as shown in Table 1.2 without standardization. The nature of the variables is illustrated in Figure 1.1, from which it can be seen that large values of Z_1 correspond to skulls that are tall but narrow, with long jaws and short nasal heights.

The Z_1 values for individual skulls are calculated in the obvious way. For example, the first skull in the early predynastic sample has $X_1 = 131$ mm, $X_2 = 138$ mm, $X_3 = 89$ mm, and $X_4 = 49$ mm. Therefore, for this skull

$$Z_1 = -0.0107 \times 131 + 0.0040 \times 138 + 0.0119 \times 89 - 0.0068 \times 49 = -0.124$$

The means and standard deviations found for the Z_1 values for the five samples are shown in Table 8.3. It can be seen that the mean of Z_1 has become lower over time, indicating a trend toward shorter, broader skulls with short jaws but relatively large nasal heights. This is, however, very much an average change. If the 150 skulls are allocated to the samples to which they are closest according to the Mahalanobis distance function of Equation 8.1, then only 51 of them (34%) are allocated to the samples to which they really belong

Table 8.3 Means and standard deviations for the discriminant function Z_1 with five samples of Egyptian skulls

Sample	Mean	Standard deviation
Early predynastic	–0.029	0.097
Late predynastic	–0.043	0.071
12th and 13th Dynasties	–0.099	0.075
Ptolemaic	–0.143	0.080
Roman	–0.167	0.095

Table 8.4 Results obtained when 150 Egyptian skulls are allocated to the group for which they have the minimum Mahalanobis distance

Source group	Number allocated to group					Total
	1	2	3	4	5	
1	12	8	4	4	2	30
2	10	8	5	4	3	30
3	4	4	15	2	5	30
4	3	3	7	5	12	30
5	2	4	4	9	11	30

(Table 8.4). Thus, although this discriminant function analysis has been successful in pinpointing the changes in skull dimensions over time, it has not produced a satisfactory method for aging individual skulls.

Example 8.2: Discriminating between groups of European countries

The data shown in Table 1.5 on the percentages employed in nine industry groups in 30 European countries have already been examined by principal components analysis and by factor analysis (Examples 6.2 and 7.1). Here, they will be considered from the point of view of the extent to which it is possible to discriminate between groups of countries on the basis of employment patterns. In particular, four natural groups existed in the period when the data were collected. These were (1) the European Union (EU) countries of Belgium, Denmark, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, Spain, and the United Kingdom; (2) the European Free Trade Area (EFTA) countries of Austria, Finland, Iceland, Norway, Sweden, and Switzerland; (3) the Eastern European countries of Albania, Bulgaria, the Czech/Slovak Republics, Hungary, Poland, Romania, the former USSR, and the former Yugoslavia; and (4) the other countries of Cyprus, Gibraltar, Malta, and Turkey. These four groups can be used as a basis for a discriminant function analysis. Wilks' lambda test (Section 4.7) gives a very highly significant result ($p < 0.001$), so there is very clear evidence, overall, that these groups are meaningful.

Apart from rounding errors, the percentages in the nine industry groups add to 100% for each of the 30 countries. This means that any one of the nine percentage variables can be expressed as 100 minus the remaining variables. It is, therefore, necessary to omit one of the variables from the analysis to carry out the analysis. The last variable, the percentage employed in transport and communications, has therefore been omitted for the analysis that will now be described.

The number of canonical variables is three in this example, this being the minimum of the number of variables ($p = 8$) and the number of groups minus one ($m - 1 = 3$). These canonical variables are found to be

$$Z_1 = 0.427\ \text{AGR} + 0.295\ \text{MIN} + 0.359\ \text{MAN} + 0.339\ \text{PS} + 0.222\ \text{CON} + 0.688\ \text{SER} + 0.464\ \text{FIN} + 0.514\ \text{SPS}$$

$$Z_2 = 0.674\ \text{AGR} + 0.579\ \text{MIN} + 0.550\ \text{MAN} + 1.576\ \text{PS} + 0.682\ \text{CON} + 0.658\ \text{SER} + 0.349\ \text{FIN} + 0.682\ \text{SPS}$$

and

$$Z_3 = 0.732\ \text{AGR} + 0.889\ \text{MIN} + 0.873\ \text{MAN} + 0.410\ \text{PS} + 0.524\ \text{CON} +$$
$$0.895\ \text{SER} + 0.714\ \text{FIN} + 0.764\ \text{SPS}$$

Different computer programs are likely to output these canonical variables with all the signs reversed for the coefficients of one or more of the variables. Also, it may be desirable to reverse the signs that are output. Indeed, with this example, the output from the computer program had negative coefficients for all the variables with Z_1 and Z_2. The signs were therefore all reversed to make the coefficients positive. It is important to note that it is the original percentages employed that are to be used in these equations, rather than these percentages after they have been standardized to have zero means and unit variances.

The eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ corresponding to the three canonical variables are $\lambda_1 = 5.349$, $\lambda_2 = 0.570$, and $\lambda_3 = 0.202$. The first canonical variable is therefore clearly the most important.

Because all the coefficients are positive for all three canonical variables, it is difficult to interpret what exactly they mean in terms of the original variables. It is helpful in this respect to consider instead the correlations between the original variables and the canonical variables, as shown in Table 8.5. This table includes the original variable TC (transport and communications), because the correlations for this variable are easily calculated once the values of Z_1 to Z_3 are known for all the European countries.

It can be seen that the first canonical variable has correlations above 0.5 for SER (services), FIN (finance), and SPS (social and personal services), and a correlation of −0.5 or less for AGR (agriculture, forestry, and fisheries), and MIN (mining). This canonical variable therefore represents service types of industry rather than traditional industries. There are no really large positive or negative correlations

Table 8.5 Correlations between the original percentages in different employment groups and the three canonical variates

Group	Z_1	Z_2	Z_3
AGR	−0.50	0.37	0.09
MIN	−0.62	0.03	0.20
MAN	−0.02	−0.20	0.12
PS	0.17	0.18	−0.23
CON	0.14	0.26	−0.34
SER	0.82	−0.01	0.08
FIN	0.61	−0.36	−0.09
SPS	0.56	−0.19	−0.28
TC	−0.22	−0.47	−0.41

between the second canonical variate and the original variables. However, considering the largest correlations, it seems to represent agriculture and construction, with an absence of transport, communications, and financial services. Finally, the third canonical variable also shows no large correlations, but represents, if anything, an absence of transport, communication, and construction.

Plots of the countries against their values for the canonical variables are shown in Figure 8.1. The plot of the second canonical variable against the first one shows a clear distinction between the eastern countries on the left-hand side and the other groups on the left. There is no clear separation between the EU and EFTA countries, with Malta and Cyprus being in the same cluster. Turkey and Gibraltar from the "other" group of countries appear in the top at the right-hand side. It can be clearly seen how most separation occurs with the horizontal values for the first canonical variate. Based on the interpretation of the canonical variables given in the previous paragraph, it appears that in the eastern countries, there is an emphasis on traditional industries rather than service industries, whereas the opposite tends to be true for the other countries. Similarly, Turkey and Gibraltar stand out because of the emphasis on agriculture and construction rather than transport, communications, and financial services. For Gibraltar, there are apparently none engaged in agriculture, but a very high percentage in construction.

The plot of the third canonical variable against the first one shows no real vertical separation of the EU, EFTA, and Other groups of countries, although there are some obvious patterns, such as the Scandinavian countries appearing close together.

The discriminant function analysis has been successful in this example in separating the eastern countries from the others, with less success in separating the other groups. The separation is perhaps clearer than what was obtained using principal components, as shown in Figure 6.2.

8.6 *Allowing for prior probabilities of group membership*

Computer programs often allow many options for a discriminant function analysis. One situation is that the probability of membership is inherently different for different groups. For example, if there are two groups, it might be known that most individuals fall into group 1, while very few fall into group 2. In that case, if an individual is to be allocated to a group, it makes sense to bias the allocation procedure in favor of group 1. Thus, the process of allocating an individual to the group from which it has the smallest Mahalanobis distance should be modified. To allow for this, some computer programs permit prior probabilities of group membership to be taken into account in the analysis.

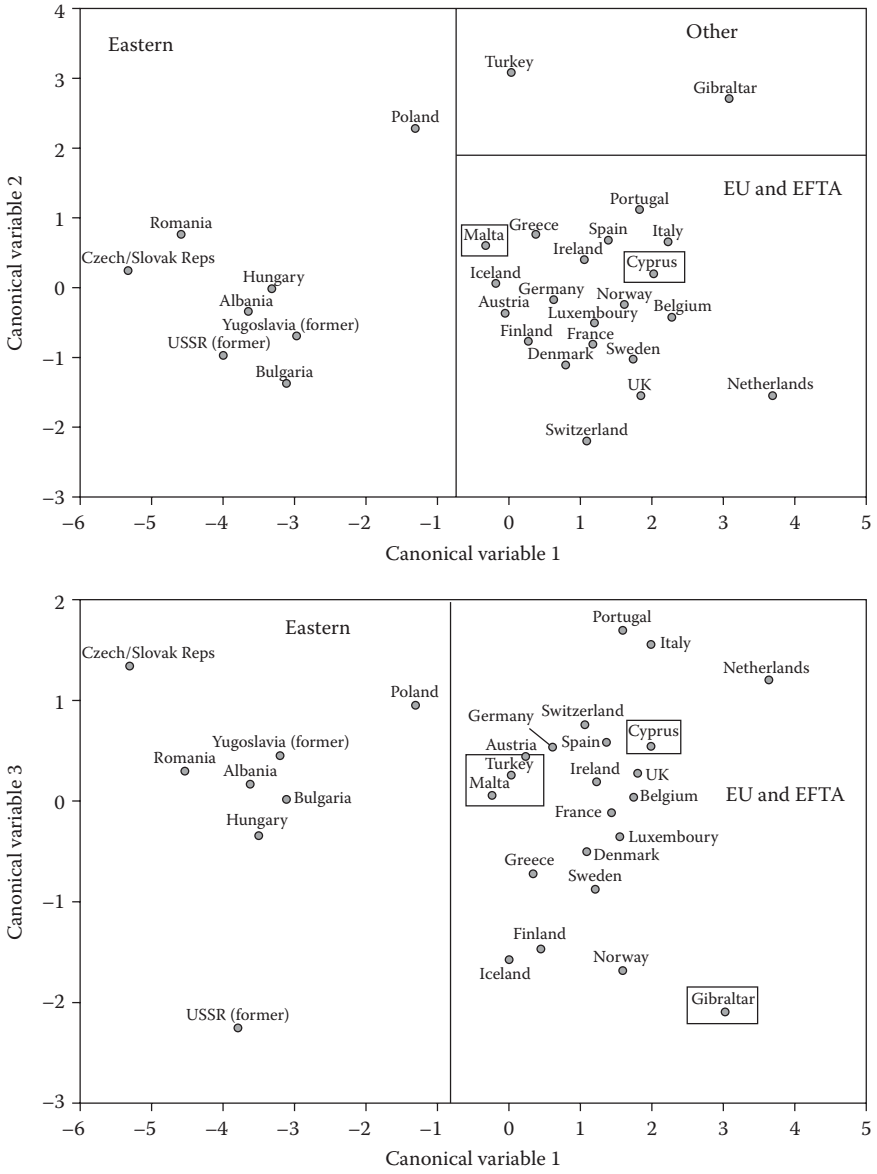

Figure 8.1 Plot of 30 European countries against their values for the first three canonical discriminant functions. Small boxes indicate countries in the other category that are not separated from the EU and EFTA groups.

8.7 Stepwise discriminant function analysis

Another possible modification of the basic analysis involves carrying it out in a stepwise manner. In this case, variables are added to the discriminant functions one by one until it is found that adding extra variables does not give significantly better discrimination. There are many different criteria that can be used for deciding which variables to include in the analysis and which to miss out.

A problem with stepwise discriminant function analysis is the bias that the procedure introduces into significance tests. Given enough variables, it is almost certain that some combination of them will produce significant discriminant functions by chance alone. If a stepwise analysis is carried out, then it is advisable to check its validity by rerunning it several times with a random allocation of individuals to groups to see how significant are the results obtained. For example, with the Egyptian skull data, the 150 skulls could be allocated completely at random to five groups of 30, the allocation being made a number of times, and a discriminant function analysis run on each random set of data. Some idea could then be gained of the probability of getting significant results through chance alone.

This type of randomization analysis to verify a discriminant function analysis is unnecessary for a standard nonstepwise analysis provided there is no reason to suspect the assumptions behind the analysis. It could, however, be informative in cases where the data are clearly not normally distributed within groups, or where the within-group covariance matrix is not the same for each group. For example, Manly (2007, Example 12.4) shows a situation in which the results of a standard discriminant function analysis are clearly suspect by comparison with the results of a randomization analysis.

8.8 Jackknife classification of individuals

Using an allocation matrix such as that shown in Table 8.4 must tend to have a bias in favor of allocating individuals to the group that they really come from. After all, the group means are determined from the observations in that group. It is, therefore, not surprising to find that an observation is closest to the center of the group where the observation helped to determine that center.

To overcome this bias, some computer programs carry out what is called a *jackknife classification* of observations. This involves allocating each individual to its closest group without using that individual to help determine a group center. In this way, any bias in the allocation is avoided. In practice, there is often not a great deal of difference between the straightforward classification and the jackknife classification, with the jackknife classification usually giving a slightly smaller number of correct allocations.

8.9 Assigning ungrouped individuals to groups

Some computer programs allow the input of data values for a number of individuals for which the true group is not known. It is then possible to assign these individuals to the group that they are closest to, in the Mahalanobis distance sense, on the assumption that they have to come from one of the m groups that are sampled. Obviously, in these cases, it will not be known whether the assignment is correct. However, the error in the allocation of individuals from known groups is an indication of how accurate the assignment process is likely to be. For example, the results shown in Table 8.4 indicate that allocating Egyptian skulls to different time periods using skull dimensions is liable to result in many errors.

8.10 Logistic regression

A rather different approach to discrimination between two groups involves making use of logistic regression. To explain how this is done, the more usual use of logistic regression will first be briefly reviewed.

The general framework for logistic regression is that there are m groups to be compared, with group i consisting of n_i items, of which Y_i exhibit a positive response (a success), and $n_i - Y_i$ exhibit a negative response (a failure). The assumption is then made that the probability of a success for an item in group i is given by

$$\pi_i = \frac{\exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}\right)} \tag{8.3}$$

where x_{ij} is the value of some variable X_j that is the same for all items in the group. In this way, the variables X_1 to X_p are allowed to influence the probability of a success, which is assumed to be the same for all items in the group, irrespective of the successes or failures of the other items in that or any other group. Similarly, the probability of a failure is $1 - \pi_i$ for all items in the ith group. It is permissible for some or all of the groups to contain just one item. Indeed, some computer programs only allow for this to be the case.

There need be no concern about arbitrarily choosing what to call a success and what to call a failure. It is easy to show that reversing these designations in the data simply results in all the β values and their estimates changing sign, and consequently changing π_i into $1 - \pi_i$.

The function that is used to relate the probability of a success to the X variables is called a *logistic function*. Unlike the standard multiple regression function, the logistic function forces estimated probabilities to lie within the range zero to one. It is for this reason that logistic regression is more sensible than linear regression as a means of modeling probabilities.

There are numerous computer programs available for fitting Equation 8.3 to data, that is, for estimating the values of β_0 to β_p, including R codes, as discussed in the Appendix to this chapter.

In the context of discrimination with two samples, three different types of situations have to be considered:

1. The data consist of a single random sample taken from a population of items that is itself divided into two parts. The application of logistic regression is then straightforward, and the fitted Equation 8.3 can be used to give an estimate of the probability of an item being in one part of the population (i.e., being a success) as a function of the values that the item possesses for variables X_1 to X_p. In addition, the distribution of success probabilities for the sampled items is an estimate of the distribution of these probabilities for the full population.
2. Separate sampling is used, whereby a random sample of size n_1 is taken from the population of items of one type (the successes), and an independent random sample of size n_2 is taken from the population of items of the second type (the failures). Logistic regression can still be used. However, the estimated probability of a success obtained from the estimated function must be interpreted in terms of the sampling scheme and the sample sizes used.
3. Groups of items are chosen to have particular values for the variables X_1 to X_p such that these variable values change from group to group. The number of successes in each group is then observed. In this case, the estimated logistic regression equation gives the probability of a success for an item conditional on the values that the item possesses for X_1 to X_p. The estimated function is, therefore, the same as for Situation (1), but the sample distribution of probabilities of a success is in no way an estimate of the distribution that would be found in the combined population of items that are successes or failures.

The following examples illustrate the differences between Situations (1) and (2), which are the ones that most commonly occur. Situation (3) is really just a standard logistic regression situation, and will not be considered further here.

Example 8.3: Storm survival of female sparrows (reconsidered)

The data in Table 1.1 consist of values for five morphological variables for 49 female sparrows taken in a moribund condition to Hermon Bumpus' laboratory at Brown University, Rhode Island, after a severe storm in 1898. The first 21 birds recovered and the remaining 28 died, and there is some interest in knowing whether it is possible to discriminate between these two groups on the basis of the five measurements. It has already been shown that there are no significant

differences between the mean values of the variables for survivors and nonsurvivors (Example 4.1), although the nonsurvivors may have been more variable (Example 4.2). A principal components analysis has also confirmed the test results (Example 6.1).

This is a situation of Type (1) if the assumption is made that the sampled birds were randomly selected from the population of female sparrows in some area close to Bumpus' laboratory. Actually, the assumption of random sampling is questionable, because it is not clear exactly how the birds were collected. Nevertheless, the assumption will be made for this example.

The logistic regression option in many standard computer packages can be used to fit the model

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_5 x_{i5})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_5 x_{i5})}$$

where:

X_1 = total length (mm)
X_2 = alar extent (mm)
X_3 = length of beak and head (mm)
X_4 = length of the humerus (mm)
X_5 = length of the sternum (mm)
π_i denotes the probability of the ith bird recovering from the storm

A chi-squared test for whether the variables account significantly for the difference between survivors and nonsurvivors gives the value 2.85 with five df, which is not at all significantly large when compared with chi-squared tables. There is, therefore, no evidence from this analysis that the survival status was related to the morphological variables. Estimated values for β_0 to β_5 are shown in Table 8.6,

Table 8.6 Estimates of the constant term and the coefficients of X variables when a logistic regression model is fitted to data on the survival of 49 female sparrows

Variable	β Estimate	Standard error	Chi-squared	p-Value
Constant	13.582	15.865	—	—
Total length	–0.163	0.140	1.36	0.244
Alar extent	–0.028	0.106	0.07	0.794
Length beak and head	–0.084	0.629	0.02	0.894
Length humerus	1.062	1.023	1.08	0.299
Length keel of sternum	0.072	0.417	0.03	0.864

Note: The chi-squared value is (estimated β value/standard error)^2. The p-value is the probability of a value this large from the chi-squared distribution with one degree of freedom. A small p-value (say, less than 0.05) provides evidence that the true value of the β parameter concerned is not equal to zero.

together with estimated standard errors and chi-squared statistics for testing whether the individual estimates differ significantly from zero. Again, there is no evidence of any significant effects.

Example 8.4: Comparison of two samples of Egyptian skulls

As an example of separate sampling, in which the sample size in the two groups being compared is not necessarily related in any way to the respective population sizes, consider the comparison between the first and last samples of Egyptian skulls for which data are provided in Table 1.2. The first sample consists of 30 male skulls from burials in the area of Thebes during the early predynastic period (circa 4000 BC) in Egypt, and the last sample consists of 30 male skulls from burials in the same area during the Roman period (circa AD 150). For each skull, measurements are available for X_1 = maximum breadth, X_2 = basibregmatic height, X_3 = basialveolar length, and X_4 = nasal height, all in millimeters (Figure 1.1). For the purpose of this example, it will be assumed that the two samples were effectively randomly chosen from their respective populations, although there is no way of knowing how realistic this is.

Obviously, the equal sample sizes in no way indicate that the population sizes in the two periods were equal. The sizes are, in fact, completely arbitrary, because many more skulls have been measured from both periods, and an unknown number of skulls have either not survived intact or not been found. Therefore, if the two samples are lumped together and treated as a sample of size 60 for the estimation of a logistic regression equation, then it is clear that the estimated probability of a skull with certain dimensions being from the early predynastic period may not really be estimating the true probability at all.

In fact, it is difficult to define precisely what is meant by the true probability in this example, because the population is not at all clear. A working definition is that the probability of a skull with specified dimensions being from the predynastic period is equal to the proportion of all skulls with the given dimensions that are from the predynastic period in a hypothetical population of all male skulls from either the predynastic or the Roman period that might have been recovered by archaeologists in the Thebes region.

It can be shown (Seber, 2004, p. 312) that if a logistic regression is carried out on a lumped sample to estimate Equation 8.3, then the modified equation

$$\pi_i = \frac{\exp\left(\beta_0 - \log_e\left\{\left(n_1P_2\right)/\left(n_2P_1\right)\right\} + \beta_1x_{i1} + \beta_2x_{i2} + \ldots + \beta_px_{ip}\right)}{1 + \exp\left(\beta_0 - \log_e\left\{\left(n_1P_2\right)/\left(n_2P_1\right)\right\} + \beta_1x_{i1} + \beta_2x_{i2} + \ldots + \beta_px_{ip}\right)} \quad (8.4)$$

is what really gives the probability that an item with the specified X values is a success. Here, Equation 8.4 differs from Equation 8.3 because of the term $\log_e\{(n_1P_2)/(n_2P_1)\}$ in the numerator and the

denominator, where P_1 is the proportion of items in the full population of successes and failures that are successes, and $P_2 = 1 - P_1$ is the proportion of the population that are failures. This, then, means that estimating the probability of an item with the specified X values being a success requires that P_1 and P_2 are either known or can somehow be estimated separately from the sample data to adjust the estimated logistic regression equation for the fact that the sample sizes n_1 and n_2 are not proportional to the population frequencies of successes and failures. In the example being considered, this requires that estimates of the relative frequencies of predynastic and Roman skulls in the Thebes area must be known to be able to estimate the probability that a skull is predynastic, given the values that it possesses for the variables X_1 to X_4.

Logistic regression was applied to the lumped data from the 60 predynastic and Roman skulls, with a predynastic skull being treated as a success. The resulting chi-squared test for the extent to which success is related to the X variables is 27.13 with four df. This is significantly large at the 0.1% level, giving very strong evidence of a relationship. The estimates of the constant term and the coefficients of the X variables are shown in Table 8.7. It can be seen that the estimate of β_1 is significantly different from zero at about the 1% level, and β_3 is significantly different from zero at the 2% level. Hence, X_1 and X_3 appear to be the important variables for discriminating between the two types of skull.

The fitted function can be used to discriminate between the two groups of skulls by assigning values for P_1 and $P_2 = 1 - P_1$ in Equation 8.4. As already noted, it is desirable that these should correspond to the population proportions of predynastic and Roman skulls. However, this is not possible, because these proportions are not known. In practice, therefore, arbitrary values must be assigned. For example, suppose that P_1 and P_2 are both set equal to 0.5. Then, $\log_e\{(n_1P_2)/(n_2P_1)\} = \log_e(1) = 0$, because $n_1 = n_2$, and Equations 8.3 and 8.4 become identical. The logistic function, therefore, estimates the

Table 8.7 Estimates of the constant term and the coefficients of X variables when a logistic regression model is fitted to data on 30 predynastic and 30 Roman period male Egyptian skulls

Variable	**β** Estimate	Standard error	Chi-squared	p-Value
Constant	−6.732	13.081	—	—
Maximum breadth	−0.202	0.075	7.13	0.008
Basibregmatic height	0.129	0.079	2.66	0.103
Basialveolar length	0.177	0.073	5.84	0.016
Nasal height	−0.008	0.104	0.01	0.939

Note: See Table 8.6 for an explanation of the columns.

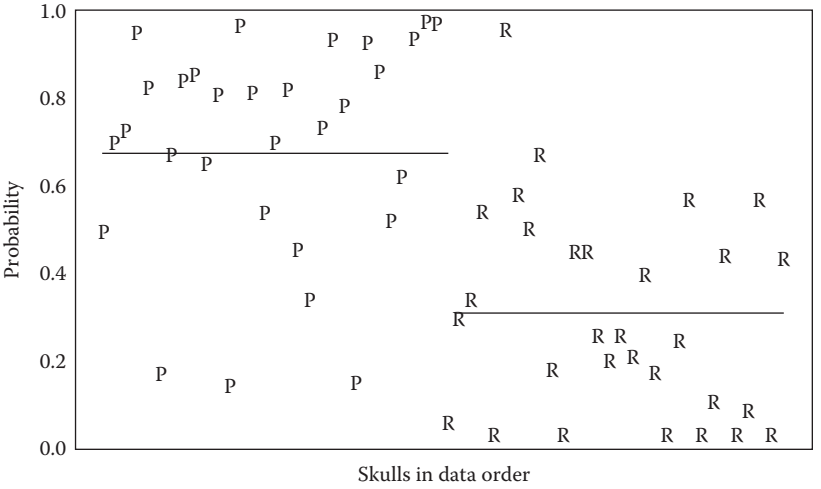

Figure 8.2 Values from the fitted logistic regression function plotted for 30 predynastic (P) and 30 Roman (R) skulls. The horizontal lines indicate the average group probabilities.

probability of a skull being predynastic in a population with equal frequencies of predynastic and Roman skulls.

The extent to which the logistic equation is effective for discrimination is indicated in Figure 8.2, which shows the estimated values of π_i for the 60 sample skulls. There is a distinct difference in the distributions of values for the two samples, with the mean for predynastic skulls being about 0.7 and the mean for Roman skulls being about 0.3. However, there is also a considerable overlap between the distributions. As a result, if the sample skulls are classified as being predynastic when the logistic equation gives a value greater than 0.5 or as Roman when the equation gives a value of less than 0.5, then six predynastic skulls are misclassified as being Roman, and seven Roman skulls are misclassified as being predynastic.

8.11 Computer programs

Major statistical packages, including R (see the Appendix to this chapter), generally have a discriminant function option that applies the methods described in Sections 8.2 through 8.5, based on the assumption of normally distributed data. Because the details of the order of calculations, the way the output is given, and the terminology vary considerably, manuals may have to be studied carefully to determine precisely what is done by any individual program. Logistic regression is also fairly widely available. In some programs, there is the restriction that all items are assumed

to have different values for X variables. However, it is more common for groups of items with common X values to be permitted.

8.12 Discussion and further reading

The assumption that samples are from multivariate distributions with the same covariance matrix, which is required for the use of the methods described in Sections 8.2 through 8.5, can be relaxed. If the samples being compared are assumed to come from multivariate normal distributions with different covariance matrices, then a method called *quadratic discriminant function analysis* can be applied. This option is also available in many computer packages. See Seber (2004, p. 297) for more information about this method and a discussion of its performance relative to the more standard linear discriminant function analysis.

Discrimination using logistic regression has been described in Section 8.10 in terms of the comparison of two groups. More detailed treatments of this method are provided by Hosmer et al. (2013) and Collett (2002). The method can also be generalized for discrimination between more than two groups if necessary, under several names, including *polytomous regression*. See Hosmer et al. (2013, Chapter 8) for more details. This type of analysis is a standard option in many computer packages.

Exercises

1. Consider the data in Table 4.5 for nine mandible measurements on samples from five different canine groups. Carry out a discriminant function analysis to see how well it is possible to separate the groups using the measurements.
2. Still considering the data in Table 4.5, investigate each canine group separately to see whether logistic regression shows a significant difference between males and females for the measurements. Note that in view of the small sample sizes available for each group, it is unreasonable to expect to fit a logistic function involving all nine variables with good estimates of parameters. Therefore, consideration should be given to fitting functions using only a subset of the variables.

References

Collett, D. (2002). *Modelling Binary Data*. 2nd Edn. Boca Raton, FL: Chapman and Hall/CRC.

Fisher, R.A. (1936). The utilization of multiple measurements in taxonomic problems. *Annals of Eugenics* 7: 179–88.

Harris, R.J. (2001). *A Primer on Multivariate Statistics*. 2nd Edn. New York: Psychology.

Hosmer, D.W., Lemeshow, S., and Sturdivant, R.X. (2013). *Applied Logistic Regression*. 3rd Edn. New York: Wiley.

Manly, B.F.J. (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. 3rd Edn. Boca Raton, FL: Chapman and Hall/CRC.

Seber, G.A.F. (2004). *Multivariate Observations*. New York: Wiley-Interscience.

Appendix: Discriminant Function Analysis in R

A.1 Canonical discriminant analysis in R

From a computational point of view, the method of canonical discriminant functions encompasses the eigenvalue analysis of $\mathbf{W}^{-1}\mathbf{B}$ (with **W** and **B** defined in Section 8.3), a task that can be carried out with the R-functions `eigen()` (described in the Appendix for Chapter 2) and `manova` (described in the Appendix for Chapter 4). This strategy is illustrated for Example 8.1 (the comparison of samples of Egyptian skulls) with an R script that can be downloaded from this book's website.

An alternative to using R programming is provided by the function `lda()` (linear discriminant analysis) included in the package `MASS` (Venables and Ripley, 2002). This uses two different methods of variable specification: either

```
discan.object<-lda(group.f~X1+X2+...,...)
```

or

```
discan.object<-lda(Xmat, group.f,...)
```

In the first method, `group.f` is a grouping factor and the variables `X1`, `X2`, … are the discriminator variables, reminding us that discriminant analysis involves continuous independent variables and a categorical dependent variable (i.e., the `group.f` label). The second option assumes that `Xmat` is a matrix or data frame whose columns are the discriminators. It is possible to specify probabilities of group membership in `lda()` with the `prior` option.

It is important to notice that the canonical coefficients produced by `eigen()` and `lda()` differ from those shown in Equation 8.2, because R scales eigenvectors in several ways. For example, `eigen()` forces each eigenvector to have a unit norm. These variations are of no concern, as one set of coefficients can be computed from another set with a suitable linear transformation. In the case of `lda`, the canonical coefficients are normalized so that the within-groups covariance matrix **W** is spherical (i.e., **W** is a multiple of the identity matrix **I**). Actually, the `print` method of `lda()` does not produce eigenvalues. Instead, it produces singular values, which are the ratio of the between- and within-group standard deviations of the linear discriminant variables. In addition, the output of `lda()` includes the proportion of trace, which is the proportion accounted by each eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$ with respect to the sum of all the eigenvalues.

This proportion can be interpreted equivalently as the proportion of the between-group variance present on each discriminant axis.

When `lda` is executed, the user may choose to produce a classification table similar to Table 8.4 in Example 8.1 through the function `predict`, using

```
table(group.f, predict(discan.object)$class)
```

The function `predict()` accepts an additional parameter with the name of a data frame containing new data to be assigned to a particular group based on the canonical discriminant functions. It is worth noting that `predict()` is not necessary when considering jackknifing classification of individuals, which is a procedure that can be produced in `lda()` with the option `CV=TRUE`. Here, `CV` stands for *cross validation*, which is another name given in multivariate analysis for the jackknife classification method. An example of this command is

```
discan.object.cv<-lda(group.f~X1+X2+..., CV=TRUE,...)
```

Here, the content of the object `discan.object.cv` (a list) is different from that generated by `lda()` with `CV=FALSE` (the default). Now, `discan.object.cv` includes the vector `class`, and it is not difficult to build a table with original membership of individuals and their jackknife classification based on the discriminant analysis. See an R code exemplifying this procedure in the book's website.

The function `lda()` also includes a `plot` method, which is useful for displaying one, two, or more linear discriminant functions using

```
plot(discan.object, dimen,...)
```

The resulting plot depends on the parameter `dimen`, the number of dimensions chosen. See the R documentation for further details.

An alternative and helpful R function for plot visualization in canonical discriminant analysis is offered by `candisc()`, which is present in the package with the same name (Friendly and Fox, 2016). The function `candisc` uses a multivariate linear model like that produced by the `lm()` function. It generates its own scaling of eigenvectors, and its corresponding plot method permits the display of centroids or means of discriminant scores for each group. R codes containing `lda` and `candisc` functions are available in the book's website, as computational aids for the discriminant analyses described in Examples 8.1 and 8.2.

A.2 Discriminant analysis based on logistic regression in R

Logistic regression is available in R as one option of the function `glm()` for fitting generalized linear models (Hilbe, 2009). A typical logistic regression analysis is written as

```
model.logistic <-glm
    (Y ~X1 +X2 +..., family=binomial(link="logit"),...)
```

where:

`Y`	is a binary response variable
`X1, X2, …`	are explanatory variables

To use logistic regression for discriminant analysis of two samples, it is only necessary to code one sample as 1 and the other as 0 and assign these binary values to a new variable `Y`. The parameter estimates of the logistic regression can be obtained with the `summary()` function

```
summary(model.logistic)
```

In addition, chi-squared tests for model comparison are accessible through the command `anova()`. Thus,

```
anova(model.1, model.2, test="Chisq")
```

evaluates whether the fit of `model.2` improves over the fit of `model.1`, assuming that the variables in `model.1` is a subset of the variables in `model.2`. The chi-squared tests indicated in the Examples for Section 8.10 can be carried out in that way, using R: `model.1` is the intercept-only model (i.e., no variables) and `model.2` includes all the discriminators of interest. R codes are provided in the book's website exemplifying the use of `glm` and `anova` in Examples 8.3 and 8.4.

The extensions of linear discriminant analysis that are described in Sections 8.7 and 8.12 are available for the R user. The function `stepclass` located in the package `klaR` (Weihs et al., 2005) is what R offers for those interested in stepwise discriminant function analysis, while the function `qda()` in the package `MASS` (Venables and Ripley, 2002) allows the separation of groups using quadratic discriminant analysis. Finally, polytomous regression, the extension of two-group logistic regression for more than two groups, is accessible through the function `multinom()` from the package `nnet` (Venables and Ripley, 2002).

References

Friendly, M. and Fox, J. (2016). candisc: Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis. R package version 0.7-0. http://CRAN.R-project.org/package=candisc

Hilbe, J.M. (2009). *Logistic Regression Models*. Boca Raton, FL: Chapman and Hall/CRC.

Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics*. 4th Edn. New York: Springer.

Weihs, C., Ligges, U., Luebke, K., and Raabe, N. (2005). klaR analyzing German business cycles. In Baier, D., Decker, R., and Schmidt-Thieme, L. (eds). *Data Analysis and Decision Support*, pp. 335–43. Berlin: Springer.