3
Principal Components Analysis

3.1 Introduction

The basic aim of principal components analysis is to describe the variation in a set of correlated variables, $x_1, x_2, \ldots, x_q$, in terms of a new set of uncorrelated variables, $y_1, y_2, \ldots, y_q$, each of which is a linear combination of the x variables. The new variables are derived in decreasing order of "importance" in the sense that y_1 accounts for as much of the variation in the original data amongst all linear combinations of $x_1, x_2, \ldots, x_q$. Then y_2 is chosen to account for as much as possible of the remaining variation, subject to being uncorrelated with y_1, and so on. The new variables defined by this process, $y_1, y_2, \ldots, y_q$, are the principal components.

The general hope of principal components analysis is that the first few components will account for a substantial proportion of the variation in the original variables, $x_1, x_2, \ldots, x_q$, and can, consequently, be used to provide a convenient lower-dimensional summary of these variables that might prove useful for a variety of reasons. Consider, for example, a set of data consisting of examination scores for several different subjects for each of a number of students. One question of interest might be how best to construct an informative index of overall examination performance. One obvious possibility would be the mean score for each student, although if the possible or observed range of examination scores varied from subject to subject, it might be more sensible to weight the scores in some way before calculating the average, or alternatively standardize the results for the separate examinations before attempting to combine them. In this way it might be possible to spread the students out further and so obtain a better ranking. The same result could often be achieved by applying principal components to the observed examination results and using the student's scores on the first principal component to provide a measure of examination success that maximally discriminated between them.

A further possible application for principal components analysis arises in the field of economics, where complex data are often summarized by some kind of index number, for example, indices of prices, wage rates, cost of living, and so on. When assessing changes in prices over time, the economist will wish to allow for the fact that prices of some commodities are more variable than others, or that the prices of some of the commodities are considered more important than others; in each case

the index will need to be weighted accordingly. In such examples, the first principal component can often satisfy the investigators requirements.

But it is not always the first principal component that is of most interest to a researcher. A taxonomist, for example, when investigating variation in morphological measurements on animals for which all the pairwise correlations are likely to be positive, will often be more concerned with the second and subsequent components since these might provide a convenient description of aspects of an animal's "shape"; the latter will often be of more interest to the researcher than aspects of an animal's "size" which here, because of the positive correlations, will be reflected in the first principal component. For essentially the same reasons, the first principal component derived from say clinical psychiatric scores on patients may only provide an index of the severity of symptoms, and it is the remaining components that will give the psychiatrist important information about the "pattern" of symptoms.

In some applications, the principal components may be an end in themselves and might be amenable to interpretation in a similar fashion as the factors in an *exploratory factor analysis* (see Chapter 4). More often they are obtained for use as a means of constructing an informative graphical representation of the data (see later in the chapter), or as input to some other analysis. One example of the latter is provided by regression analysis. Principal components may be useful here when:

- There are too many explanatory variables relative to the number of observations.
- The explanatory variables are highly correlated.

Both situations lead to problems when applying regression techniques, problems that may be overcome by replacing the original explanatory variables with the first few principal component variables derived from them. An example will be given later and other applications of the technique are described in Rencher (1995).

A further example of when the results from a principal components analysis may be useful in the application of *multivariate analysis of variance* (see Chapter 7) is when there are too many original variables to ensure that the technique can be used with reasonable power. In such cases the first few principal components might be used to provide a smaller number of variables for analysis.

3.2 Algebraic Basics of Principal Components

The first principal component of the observations is that linear combination of the original variables whose sample variance is greatest amongst all possible such linear combinations. The second principal component is defined as that linear combination of the original variables that accounts for a maximal proportion of the remaining variance subject to being uncorrelated with the first principal component. Subsequent components are defined similarly. The question now arises as to how the coefficients specifying the linear combinations of the original variables defining each component are found. The algebra of *sample* principal components is summarized in Display 3.1.

Display 3.1
Algebraic Basis of Principal Components Analysis

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

$$y_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1q}x_q$$

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

$$y_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2q}x_q$$

  i.e., $y_2 = \mathbf{a}_2'\mathbf{x}$ where $\mathbf{a}_2' = [a_{21}, a_{22}, \ldots, a_{2q}]$ and $\mathbf{x}' = [x_1, x_2, \ldots, x_q]$. which has the greatest variance subject to the following two conditions:

$$\mathbf{a}_2'\mathbf{a}_2 = 1,$$
$$\mathbf{a}_2'\mathbf{a}_1 = 0.$$

  The second condition ensures that y_1 and y_2 are uncorrelated.
- Similarly, the jth principal component is that linear combination $y_j = \mathbf{a}_j'\mathbf{x}$ which has the greatest variance subject to the conditions

$$\mathbf{a}_j'\mathbf{a}_j = 1,$$
$$\mathbf{a}_j'\mathbf{a}_i = 0 \quad (i < j).$$

- To find the coefficients defining the first principal component we need to choose the elements of the vector $\mathbf{a}_1$ so as to maximize the variance of y_1 subject to the constraint $\mathbf{a}_1'\mathbf{a}_1 = 1$.
- To maximize a function of several variables subject to one or more constraints, the method of *Lagrange multipliers* is used. This leads to the solution that $\mathbf{a}_1$ is the eigenvector of the sample covariance matrix, $\mathbf{S}$, corresponding to its largest eigenvalue. Full details are given in Morrison (1990), and an example with $q = 2$ appears in Subsection 3.2.4.
- The other components are derived in similar fashion, with $\mathbf{a}_j$ being the eigenvector of $\mathbf{S}$ associated with its jth largest eigenvalue.
- If the eigenvalues of $\mathbf{S}$ are $\lambda_1, \lambda_2, \ldots, \lambda_q$, then since $\mathbf{a}_i'\mathbf{a}_i = 1$, the variance of the ith principal component is given by λ_i.
- The total variance of the q principal components will equal the total variance of the original variables so that

$$\sum_{i=1}^{q} \lambda_i = s_1^2 + s_2^2 + \cdots + s_q^2$$

where s_i^2 is the sample variance of x_i. We can write this more concisely as

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

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

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

- The first m principal components, where $m < q$ account for a proportion $P^{(m)}$ of the total variation in the original data, where

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

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

The account of principal components given in Display 3.1 is in terms of the eigenvalues and eigenvectors of the covariance matrix, **S**. In practice, however, it is far more usual to extract the components from the correlation matrix, **R**. The reasons are not difficult to identify. If we imagine a set of multivariate data where the variables $x_1, x_2, \ldots, x_q$ are of completely different types, for example, length, temperature, blood pressure, anxiety rating, etc., then the structure of the principal components derived from the covariance matrix will depend on the essentially arbitrary choice of choice of units of measurement; for example, changing lengths from centimeters to inches will alter the derived components.

Additionally if there are large differences between the variances of the original variables, those whose variances are largest will tend to dominate the early components; an example illustrating this problem is given in Jolliffe (2002). Extracting the components as the eigenvectors of **R**, which is equivalent to calculating the principal components from the original variables after each has been standardized to have unit variance, overcomes these problems. It should be noted, however, that there is rarely any simple correspondence between the components derived from **S** and those derived from **R**. And choosing to work with **R** rather than with **S** involves a definite but possibly arbitrary decision to make variables "equally important."

The correlations or covariances between the original variables and the derived components are often useful in interpreting a principal components analysis. They can be obtained as shown in Display 3.2.

Display 3.2
Correlations and Covariances of Variables and Components

- The covariance of variable i with component j is given by

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

- The correlation of variable x_i with component y_j is therefore

$$r_{x_i,y_j} = \frac{\lambda_j a_{ji}}{\sqrt{\text{Var}(x_i)\,\text{Var}(y_j)}} = \frac{\lambda_j a_{ji}}{s_i\sqrt{\lambda_j}} = \frac{a_{ji}\sqrt{\lambda_j}}{s_i}.$$

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

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

since in this case the standard deviation, s_i, is unity.

3.2.1 *Rescaling Principal Components*

It is often useful to rescale principal components so that the coefficients that define them are analogous in some respects to the factor loadings in exploratory factor analysis (see Chapter 4). Again the necessary algebra is relatively simple and is outlined in Display 3.3.

Display 3.3
Rescaling Principal Components

- Let the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_q$, which define the principal components, be used to form a $q \times q$ matrix, $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_q]$.
- Arrange the eigenvalues $\lambda_1, \ldots, \lambda_q$ along the main diagonal of a diagonal matrix, Λ.
- Then it can be shown that the covariance matrix of the observed variables $x_1, x_2, \ldots, x_q$ is given by

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

(We are assuming here that $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_q$ have been derived from $\mathbf{S}$ rather than from $\mathbf{R}$.)

- Rescaling the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_q$ so that the sum of squares of their elements is equal to the corresponding eigenvalue, i.e., calculating $\mathbf{a}_i^* = \lambda_i^{1/2}\mathbf{a}_i$, allows $\mathbf{S}$ to may be written more simply as

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

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

- In the case where components arise from a correlation matrix this rescaling leads to coefficients that are the correlations between the components and the original variables (see Display 3.2). The rescaled coefficients are analogous to factor loadings as we shall see in the next chapter. It is often these rescaled coefficients that are presented as the results of a principal components analysis.
- If the matrix $\mathbf{A}^*$ is formed from say the first m components rather than from all q, then $\mathbf{A}^*(\mathbf{A}^*)'$ gives the predicted value of $\mathbf{S}$ based on these m components.

3.2.2 *Choosing the Number of Components*

As described earlier, principal components analysis is seen to be a technique for transforming a set of observed variables into a new set of variables that are uncorrelated with one another. The variation in the original q variables is only *completely* accounted for by *all* q principal components. The usefulness of these transformed variables, however, stems from their property of accounting for the variance in decreasing proportions. The first component, for example, accounts for the maximum amount of variation possible for any linear combination of the original variables. But how useful is this artificial variation constructed from the observed variables? To answer this question we would first need to know the proportion of the total variance of the original variables for which it accounted. If, for example, 80% of the variation in a multivariate data set involving six variables could be accounted for by a simple weighted average of the variable values, then almost all the variation can be expressed along a single continuum rather than in six-dimensional space.

The principal components analysis would have provided a highly parsimonious summary (reducing the dimensionality of the data from six to one) that might be useful in later analysis.

So the question we need to ask is how many components are needed to provide an adequate summary of a given data set? A number of informal and more formal techniques are available. Here we shall concentrate on the former; examples of the use of formal inferential methods are given in Jolliffe (2002) and Rencher (1995).

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

- Retain just enough components to explain some specified, large percentage of the total variation of the original variables. Values between 70% and 90% are usually suggested, although smaller values might be appropriate as q or n, the sample size, increases.
- Exclude those principal components whose eigenvalues are less than the average, $\sum_{i=1}^{q} \lambda_i / q$. Since $\sum_{i}^{q} \lambda_i = \text{trace}(\mathbf{S})$ the average eigenvalue is also the average variance of the original variables. This method then retains those components that account for more variance than the average for the variables.
- When the components are extracted from the correlation matrix, $\text{trace}(\mathbf{R}) = q$, and the average is therefore one; components with eigenvalues less than one are therefore excluded. This rule was originally suggested by Kaiser (1958), but Jolliffe (1972), on the basis of a number of simulation studies, proposed that a more appropriate procedure would be to exclude components extracted from a correlation matrix whose associated eigenvalues are less than 0.7.
- Cattell (1965) suggests examination of the plot of the λ_i against i, the so-called *scree diagram*. The number of components selected is the value of i corresponding to an "elbow" in the curve, this point being considered to be where "large" eigenvalues cease and "small" eigenvalues begin. A modification described by Jolliffe (1986) is the *log-eigenvalue diagram* consisting of a plot of $\log(\lambda_i)$ against i.

3.2.3 *Calculating Principal Component Scores*

If we decide that we need say m principal components to adequately represent our data (using one or other of the methods described in the previous subsection), then we will generally wish to calculate the scores on each of these components for each individual in our sample. If, for example, we have derived the components from the covariance matrix, $\mathbf{S}$, then the m principal component scores for individual i with original $q \times 1$ vector of variable values $\mathbf{x}_i$, are obtained as

$$
\begin{aligned}
y_{i1} &= \mathbf{a}_1' \mathbf{x}_i \\
y_{i2} &= \mathbf{a}_2' \mathbf{x}_i \\
&\vdots \\
y_{im} &= \mathbf{a}_m' \mathbf{x}_i
\end{aligned}
$$

If the components are derived from the correlation matrix, then $\mathbf{x}_i$ would contain individual i's standardized scores for each variable.

The principal component scores calculated as above have variances equal to λ_j for $j = 1, \ldots, m$. Many investigators might prefer to have scores with means zero and variances equal to unity. Such scores can be found as follows:

$$\mathbf{z} = \mathbf{\Lambda}_m^{-1}\mathbf{A}_m'\mathbf{x}$$

where $\mathbf{\Lambda}_m$ is an $m \times m$ diagonal matrix with $\lambda_1, \lambda_2, \ldots, \lambda_m$ on the main diagonal, $\mathbf{A}_m = [\mathbf{a}_1, \ldots, \mathbf{a}_m]$, and $\mathbf{x}$ is the $q \times 1$ vector of standardized scores.

We should note here that the first m principal component scores are the same whether we retain all possible q components or just the first m. As we shall see in the next chapter, this is *not* the case with the calculation of factor scores.

3.2.4 *Principal Components of Bivariate Data with Correlation Coefficient r*

Before we move on to look at some practical examples of the application of principal components analysis it will be helpful to look in a little more detail at the mathematics of the method in one very simple case. This we do in Display 3.4 for bivariate data where the variables have correlation coefficient r.

Display 3.4
Principal Components of Bivariate Data with Correlation *r*

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

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

- In order to find the principal components of the data r we need to find the eigenvalues and eigenvectors of $\mathbf{R}$.
- The eigenvalues are roots of the equation

$$|\mathbf{R} - \lambda\mathbf{I}| = 0.$$

- This leads to a quadratic equation in λ,

$$(1 - \lambda)^2 - r^2 = 0,$$

  giving eigenvalues $\lambda_1 = 1 + r$, $\lambda_2 = 1 - r$. Note that the sum of the eigenvalues is 2, equal to trace ($\mathbf{R}$).
- The eigenvector corresponding to λ_1 is obtained by solving the equation

$$\mathbf{R}\mathbf{a}_1 = \lambda_1\mathbf{a}_1$$

- This leads to the equations

$$a_{11} + ra_{12} = (1 + r)a_{11}, \qquad ra_{11} + a_{12} = (1 + r)a_{12}.$$

- The two equations are identical and both reduce to $a_{11} = a_{12}$.
- If we now introduce the normalization constraint, $\mathbf{a}_1'\mathbf{a}_1 = 1$ we find that

$$a_{11} = a_{12} = \frac{1}{\sqrt{2}}.$$

- Similarly, we find the second eigenvector to be given by $a_{21} = 1/\sqrt{2}$ and $a_{22} = -1/\sqrt{2}$.
- The two principal components are then given by

$$y_1 = \frac{1}{\sqrt{2}}(x_1 + x_2), \qquad y_2 = \frac{1}{\sqrt{2}}(x_1 - x_2).$$

- Notice that if $r < 0$ the order of the eigenvalues and hence of the principal components is reversed; if $r = 0$ the eigenvalues are both equal to 1 and any two solutions at right angles could be chosen to represent the two components.
- Two further points:
  1. There is an arbitrary sign in the choice of the elements of $\mathbf{a}_i$; it is customary to choose a_{i1} to be positive.
  2. The components do not depend on r, although the proportion of variance explained by each does change with r. As r tends to 1 the proportion of variance accounted for by y_1, namely $(1 + r)/2$, also tends to one.
- When $r = 1$, the points all line on a straight line and the variation in the data is unidimensional.

3.3 An Example of Principal Components Analysis: Air Pollution in U.S. Cities

To illustrate a number of aspects of principal components analysis we shall apply the technique to the data shown in Table 3.1, which is again concerned with air pollution in the United States. For 41 cities in the United States the following seven variables were recorded:

SO2:	Sulphur dioxide content of air in micrograms per cubic meter
Temp:	Average annual temperature in °F
Manuf:	Number of manufacturing enterprises employing 20 or more workers

Table 3.1 Air Pollution in U.S. Cities. From *Biometry*, 2/E, Robert R. Sokal and F. James Rohlf. Copyright © 1969, 1981 by W.H. Freeman and Company. Used with permission.

City	*SO2*	Temp	Manuf	Pop	Wind	Precip	Days
Phoenix	10	70.3	213	582	6.0	7.05	36
Little Rock	13	61.0	91	132	8.2	48.52	100
San Francisco	12	56.7	453	716	8.7	20.66	67
Denver	17	51.9	454	515	9.0	12.95	86
Hartford	56	49.1	412	158	9.0	43.37	127
Wilmington	36	54.0	80	80	9.0	40.25	114
Washington	29	57.3	434	757	9.3	38.89	111
Jacksonville	14	68.4	136	529	8.8	54.47	116
Miami	10	75.5	207	335	9.0	59.80	128
Atlanta	24	61.5	368	497	9.1	48.34	115
Chicago	110	50.6	3344	3369	10.4	34.44	122
Indianapolis	28	52.3	361	746	9.7	38.74	121
Des Moines	17	49.0	104	201	11.2	30.85	103
Wichita	8	56.6	125	277	12.7	30.58	82
Louisville	30	55.6	291	593	8.3	43.11	123
New Orleans	9	68.3	204	361	8.4	56.77	113
Baltimore	47	55.0	625	905	9.6	41.31	111
Detroit	35	49.9	1064	1513	10.1	30.96	129
Minneapolis	29	43.5	699	744	10.6	25.94	137
Kansas	14	54.5	381	507	10.0	37.00	99
St Louis	56	55.9	775	622	9.5	35.89	105
Omaha	14	51.5	181	347	10.9	30.18	98
Albuquerque	11	56.8	46	244	8.9	7.77	58
Albany	46	47.6	44	116	8.8	33.36	135
Buffalo	11	47.1	391	463	12.4	36.11	166
Cincinnati	23	54.0	462	453	7.1	39.04	132
Cleveland	65	49.7	1007	751	10.9	34.99	155
Columbus	26	51.5	266	540	8.6	37.01	134
Philadelphia	69	54.6	1692	1950	9.6	39.93	115
Pittsburgh	61	50.4	347	520	9.4	36.22	147
Providence	94	50.0	343	179	10.6	42.75	125
Memphis	10	61.6	337	624	9.2	49.10	105
Nashville	18	59.4	275	448	7.9	46.00	119
Dallas	9	66.2	641	844	10.9	35.94	78
Houston	10	68.9	721	1233	10.8	48.19	103

(*Continued*)

Table 3.1 (*Continued*)

City	*SO2*	Temp	Manuf	Pop	Wind	Precip	Days
Salt Lake City	28	51.0	137	176	8.7	15.17	89
Norfolk	31	59.3	96	308	10.6	44.68	116
Richmond	26	57.8	197	299	7.6	42.59	115
Seattle	29	51.1	379	531	9.4	38.79	164
Charleston	31	55.2	35	71	6.5	40.75	148
Milwaukee	16	45.7	569	717	11.8	29.07	123

Data assumed to be available as data frame `usair.dat` with variable names as specified in the table.

Pop: Population size (1970 census) in thousands
Wind: Average annual wind speed in miles per hour
Precip: Average annual precipitation in inches
Days: Average number of days with precipitation per year

The data were originally collected to investigate the determinants of pollution presumably by regressing *SO2* on the other six variables. Here, however, we shall examine how principal components analysis can be used to explore various aspects of the data, before looking at how such an analysis can also be used to address the determinants of pollution question.

To begin we shall ignore the *SO2* variable and concentrate on the others, two of which relate to human ecology (*Pop*, *Manuf*) and four to climate (*Temp*, *Wind*, *Precip*, *Days*). A case can be made to use negative temperature values in subsequent analyses, since then all six variables are such that high values represent a less attractive environment. This is, of course, a personal view, but as we shall see later, the simple transformation of *Temp* does aid interpretation.

Prior to undertaking a principal components analysis (or any other analysis) on a set of multivariate data, it is usually imperative to graph the data in some way so as to gain an insight into its overall structure and/or any "peculiarities" that may have an impact on the analysis. Here it is useful to construct a scatterplot matrix of the six variables, with histograms for each variable on the main diagonal. How to do this using the S-PLUS GUI (assuming the dataframe `usair.dat` has already been attached) has already been described in Chapter 2 (see Section 2.5). The diagram that results is shown in Figure 3.1.

A clear message from Figure 3.1 is that there is at least one city, and probably more than one, that should be considered an outlier. On the *Manuf* variable, for example, Chicago with a value of 3344 has about twice as many manufacturing enterprises employing 20 or more workers than has the city with the second highest number (Philadelphia). We shall return to this potential problem later in the chapter, but for the moment we shall carry on with a principal components analysis of the data for *all* 41 cities.

For the data in Table 3.1 it seems necessary to extract the principal components from the correlation rather than the covariance matrix, since the six variables to be

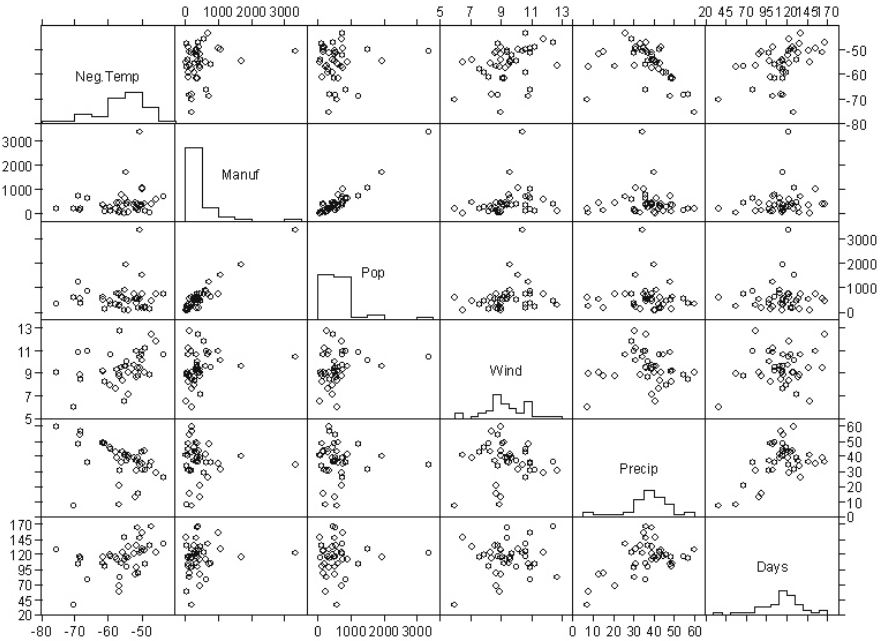

Figure 3.1 Scatterplot matrix of six variables in the air pollution data.

used are on very different scales. The correlation matrix and the principal components of the data can be obtained in R and S-PLUS® using the following command line code;

```
cor(usair.dat[,-1])
usair.pc<-princomp(usair.dat[,-1],cor=TRUE)
summary(usair.pc,loadings=TRUE)
```

The resulting output is shown in Table 3.2. (This output results from using S-PLUS; with R the signs of the coefficients of the first principal component are reversed.) One thing to note about the correlations is the very high value for *Manuf* and *Pop*, a finding returned to in Exercise 3.8. From Table 3.2 we see that the first three components all have variances (eigenvalues) greater than one and together account for almost 85% of the variance of the original variables. Scores on these three components might be used to summarize the data in further analyses with little loss of information. We shall illustrate this possibility later.

Most users of principal components analysis search for an interpretation of the derived coefficients that allow them to be "labelled" in some sense. This requires examining the coefficients defining each component (in Table 3.2 these are scaled so that their sums of squares equal unity—"blanks" indicate near-zero values), we see that the first component might be regarded as some index of "quality of life" with high values indicating a relatively poor environment (in the author's terms at least). The second component is largely concerned with a city's rainfall, having

Table 3.2 S-PLUS Results from the Principal Components Analysis of the Air Pollution Data

	Neg temp	Manuf	Pop	Wind	Precip	Days
Neg temp	1.000	0.190	0.063	0.350	−0.386	0.430
Manuf	0.190	1.000	0.955	0.238	−0.032	0.132
Pop	0.0627	0.955	1.000	0.213	−0.026	0.042
Wind	0.350	0.238	0.213	1.000	−0.013	0.164
Precip	−0.386	−0.032	−0.026	−0.013	1.000	0.496
Days	0.430	0.132	0.042	0.164	0.496	1.000

Importance of components:

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6
Standard deviation	1.482	1.225	1.181	0.872	0.338	0.186
Proportion of variance	0.366	0.250	0.232	0.127	0.019	0.006
Cumulative proportion	0.366	0.616	0.848	0.975	0.994	1.000

Loadings:

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6
Neg temp	0.330	0.128	0.672	−0.306	0.558	0.136
Manuf	0.612	−0.168	−0.273	−0.137	0.102	−0.703
Pop	0.578	−0.222	−0.350	—	—	0.695
Wind	0.354	0.131	0.297	0.869	−0.113	—
Precip	—	0.623	−0.505	0.171	0.568	—
Days	0.238	0.708	—	−0.311	−0.580	—

high coefficients for *Precip* and *Days*, and might be labeled as the "wet weather" component. Component three is essentially a contrast between *Precip* and *Neg temp*, and will separate cities having high temperatures and high rainfall from those that are colder but drier. A suitable label might be simply "climate type."

Attempting to label components in this way is not without its critics; the following quotation from Marriott (1974) should act as a salutary warning about the dangers of overinterpretation.

> It must be emphasized that no mathematical method is, or could be, designed to give physically meaningful results. If a mathematical expression of this sort has an obvious physical meaning, it must be attributed to a lucky change, or to the fact that the data have a strongly marked structure that shows up in analysis. Even in the latter case, quite small sampling fluctuations can upset the interpretation; for example, the first two principal components may appear in reverse order, or may become confused altogether. Reification then requires considerable skill and experience if it is to give a true picture of the physical meaning of the data.

Even if we do not care to label the three components they can still be used as the basis of various graphical displays of the cities. In fact, this is often the most useful aspect of a principal components analysis because regarding the principal components analysis as a means to providing an informative view of multivariate data has the advantage of making it less urgent or tempting to try to interpret and label the components. The first few component scores provide a low-dimensional "map" of the observations in which the Euclidean distances between the points

representing the individuals best approximate in some sense the Euclidean distances between the individuals based on the original variables. We shall return to this point in Chapter 5.

So we will begin by looking at the scatterplot of the first two principal components created using the following R and S-PLUS commands;

```
#choose square plotting area and make limits on both the x
#and y axes the same
#
par(pty="s")
plot(usair.pc$scores[,1],usair.pc$scores[,2],
ylim=range(usair.pc$scores[,1]),
xlab="PC1",ylab="PC2",type="n",lwd=2)
#
#now add abbreviated city names
#
text(usair.pc$scores[,1],usair.pc$scores[,2],
labels=abbreviate(row.names(usair.dat)),cex=0.7,lwd=2)
```

The resulting diagram is given in Figure 3.2. Similar diagrams for components 1 and 3 and 2 and 3 are given in Figures 3.3 and 3.4. (These diagrams are from the S-PLUS results.) The plots again demonstrate clearly that Chicago is an outlier and suggest that Phoenix and Philadelphia may also be suspects in this respect. Phoenix appears to offer the best quality of life (on the limited basis on the six variables recorded), and Buffalo is a city to avoid if you prefer a drier environment. We leave further interpretation to readers.

We can also construct a three-dimensional plot of the cities using these three component scores. The initial step is to construct a new data frame containing the first three principal component scores and the city names using

```
usair1.dat <- data.frame(cities=row.names(usair.dat),
    usair.dat, usair.pc$scores[,1:3])
attach(usair1.dat)
```

We shall now use the S-PLUS GUI to construct a drop-line three-dimensional plot of the data. Details of how to construct such a plot were given in Chapter 2, but it may be helpful to go through them again here;

- Click **Graph** on the tool bar;
- Select **3D**;
- In **Insert Graph** dialogue, choose **3D Scatter with drop line (x, y)**, and click **OK**;
- In the **3D Line/Scatter Plot [1] dialogue** select **Data Set** `usair.dat`;
- Select *Comp 1* for **x Column**, *Comp 2* for **y Column**, *Comp 3* for **z Column** and *Cities* for **w Column**;
- Check **Symbol** tab;
- Check **Use Text as Symbol** button;
- Specify text to use as **w Column**;
- Change **Font** to bold and **Height** to 0.15, click **OK**

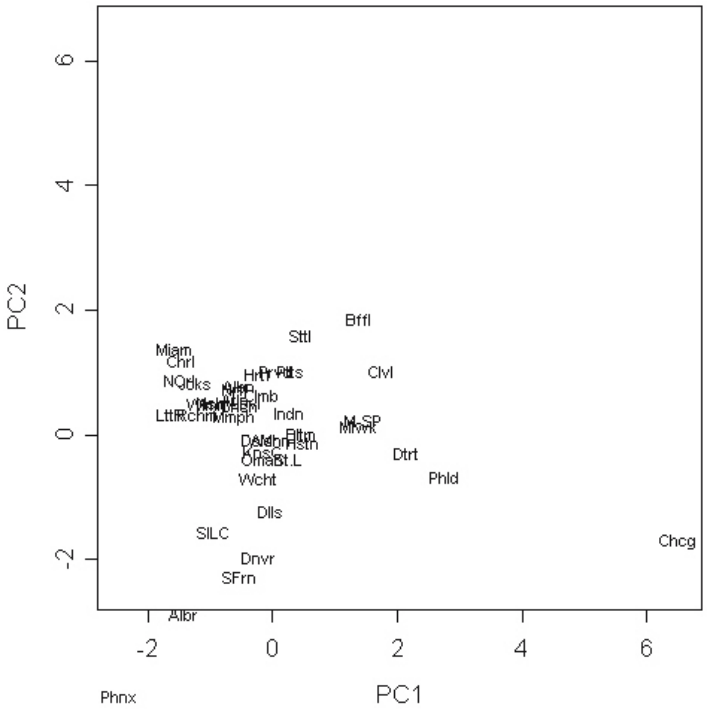

Figure 3.2 Scatterplot of the air pollution data in the space of the first two principal components.

The resulting diagram is shown in Figure 3.5. Again the problem with Chicago is very clear.

We will now use the three component scores for each city to investigate perhaps the prime question for these data, namely what characteristics of a city are predictive of its level of sulfur dioxide pollution? It may first be helpful to have a record of the component scores found from

```
usair.pc$scores[,1:3]
```

The scores are shown in Table 3.3. Before undertaking a formal regression analysis of the data we might look at *SO2* plotted against each of the three principal component scores. We can construct these plots in both R and S-PLUS as follows:

```
par(mfrow=c(1,3))
plot(usair.pc$scores[,1],SO2,xlab="PC1")
plot(usair.pc$scores[,2],SO2,xlab="PC2")
plot(usair.pc$scores[,3],SO2,xlab="PC3")
```

The plots are shown in Figure 3.6.

Interpretation of the plots is somewhat hampered by the presence of the outliers such as Chicago, but it does appear that pollution *is* related to the first principal component score but not, perhaps, to the other two. We can examine this more

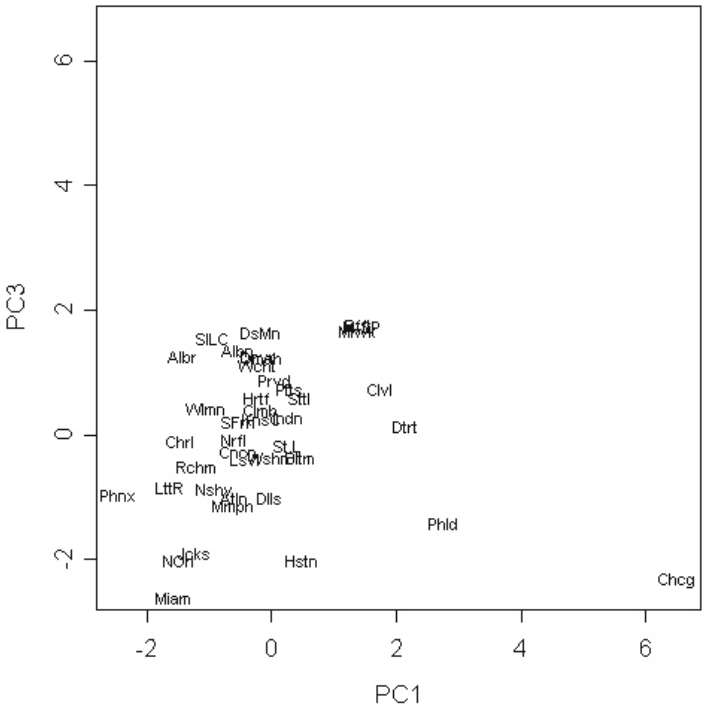

Figure 3.3 Scatterplot of the air pollution data in the space of the first and third principal components.

formally by regressing sulphur dioxide concentration on the first three principal components scores. The necessary R and S-PLUS command is;

```
summary(lm(SO2 ~ usair.pc$scores[, 1] + usair.pc$scores[, 2] +
usair.pc$scores[, 3]))
```

The resulting output is shown in Table 3.4. Clearly pollution is predicted only by the first principal component score. As "quality of life"—as measured by the human ecology and climate variable—gets worse (i.e., first PC score increases), pollution also tends to increase. (Note that because we are using principal component scores as explanatory variables in this regression the correlations of coefficients are all zero.)

Now we need to consider what to do about the obvious outliers in the data such as Chicago. The simplest approach would be to remove the relevant cities and then repeat the analyses above. The problem with such an approach is deciding when to stop removing cities, and we shall leave that as an exercise for the reader (see Exercise 3.7). Here we shall use a different approach that involves what is known as the *minimum volume ellipsoid*, a robust estimator of the correlation matrix of the data proposed by Rousseeuw (1985) and described in less technical terms in Rousseeuw and van Zomeren (1990). The essential feature of the estimator is

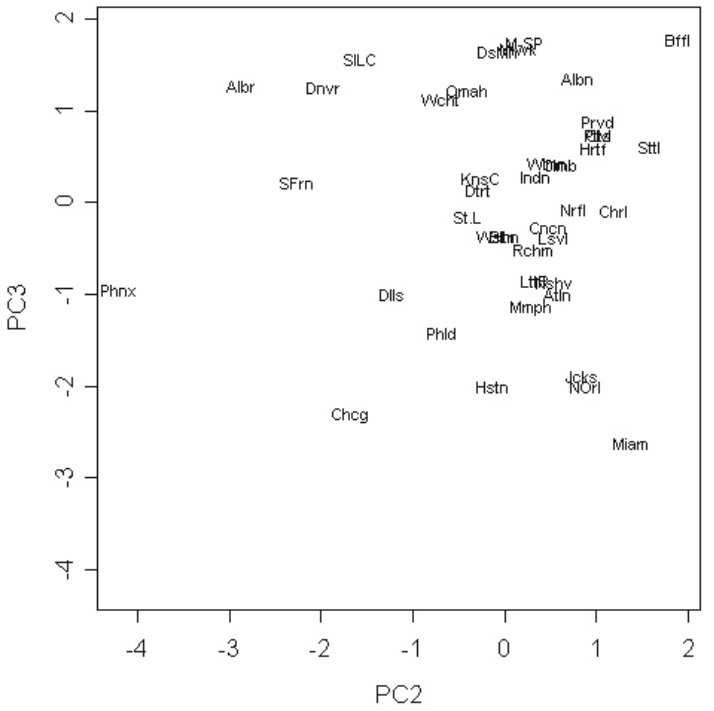

Figure 3.4 Scatterplot of the air pollution data in the space of the second and third principal components.

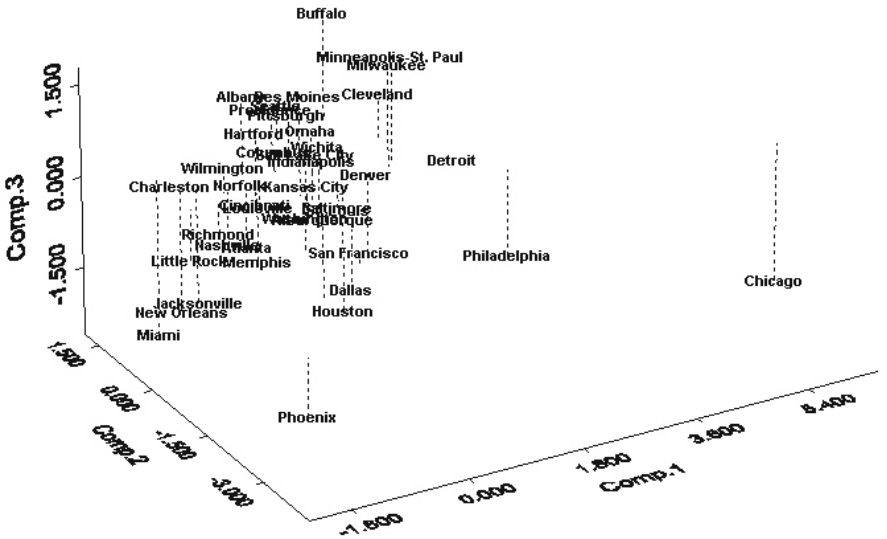

Figure 3.5 Drop line plot of air pollution data in the space of the first three principal components.

Table 3.3 First Three Principal Components Scores for Each City in the Air Pollution Data Set

City	Comp 1	Comp 2	Comp 3
Phoenix	−2.440	−4.191	−0.942
Little Rock	−1.612	0.342	−0.840
San Francisco	−0.502	−2.255	0.227
Denver	−0.207	−1.963	1.266
Hartford	−0.219	0.976	0.595
Wilmington	−0.996	0.501	0.433
Washington	−0.023	−0.055	−0.354
Jacksonville	−1.228	0.849	−1.876
Miami	−1.533	1.405	−2.607
Atlanta	−0.599	0.587	−0.995
Chicago	6.514	−1.668	−2.286
Indianapolis	0.308	0.360	0.285
Des Moines	−0.132	−0.061	1.650
Wichita	−0.197	−0.676	1.131
Louisville	−0.424	0.541	−0.374
New Orleans	−1.454	0.901	−1.992
Baltimore	0.509	0.029	−0.364
Detroit	2.167	−0.271	0.147
Minneapolis	1.500	0.247	1.751
Kansas	−0.131	−0.252	0.275
St Louis	0.286	−0.384	−0.156
Omaha	−0.134	−0.385	1.236
Albuquerque	−1.417	−2.866	1.275
Albany	−0.539	0.792	1.363
Buffalo	1.391	1.880	1.776
Cincinnati	−0.508	0.486	−0.266
Cleveland	1.766	1.039	0.747
Columbus	−0.119	0.640	0.423
Philadelphia	2.797	−0.658	−1.415
Pittsburgh	0.322	1.027	0.748
Providence	0.070	10.34	0.888
Memphis	−0.578	0.325	−1.115
Nashville	−0.910	0.543	−0.859
Dallas	−0.007	−1.212	−0.998
Houston	0.508	−0.113	−1.994

(*Continued*)

Table 3.3 (*Continued*)

City	Comp 1	Comp 2	Comp 3
Salt Lake City	−0.912	−1.547	1.565
Norfolk	−0.589	0.752	−0.061
Richmond	−1.172	0.335	−0.509
Seattle	0.482	1.597	0.609
Charleston	−1.430	1.211	−0.079
Milwaukee	1.391	0.158	1.691

selecting a covariance matrix (**C**) and mean vector (**M**) such that the determinant of **C** is minimized subject to the number of observations for which

$$(\mathbf{x}_i - \mathbf{M})' \mathbf{C}^{-1} (\mathbf{x}_i - \mathbf{M}) \leq a^2$$

is greater than or equal to h where h is the integer part of $(n + q + 1)/2$. The number a^2 is a fixed constant, usually chosen as $\chi^2_{q,0.50}$, when we expect the

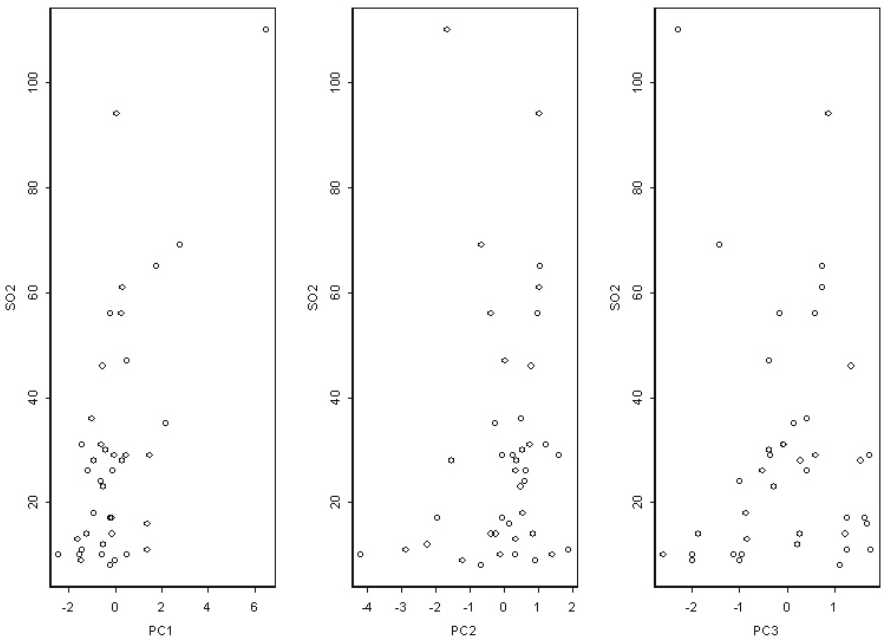

Figure 3.6 Plots of sulphur dioxide concentration against first three principal component scores.

Table 3.4 Results of Regressing Sulphur Dioxide Concentration on First Three Principal Component Scores

Residuals:

Min	1Q	Median	3Q	Max
−36.42	−10.98	−3.184	12.09	61.27

Coefficients:

	Value	Std error	t value	Pr (?$\|t\|$)
(Intercept)	30.0488	2.9072	10.3360	0.0000
usair.pc$scores[, 1]	9.9420	1.9617	5.0679	0.0000
usair.pc$scores[, 2]	2.2396	2.3738	0.9435	0.3516
usair.pc$scores[, 3]	−0.3750	2.4617	−0.1523	0.8798

Residual standard error: 18.62 on 37 degrees of freedom
Multiple R-squared: 0.4182
F-statistic: 8.866 on 3 and 37 degrees of freedom, the p-value is 0.0001473

Correlation of coefficients:

	(Intercept)	usair.pc$scores[, 1]	usair.pc$scores[, 2]
usair.pc$scores[, 1]	0		
usair.pc$scores[, 2]	0	0	
usair.pc$scores[, 3]	0	0	0

majority of the data to come from a normal distribution. The estimator has a high breakdown point, but is computationally expensive; see Rousseeuw and van Zomren (1990) for further details.

The necessary R and S-PLUS function to apply this estimator is `cov.mve` (in R the *lqs* library needs to be loaded to make the function available). The following code applies the function and then uses principal components on the robustly estimated correlation matrix:

```
#in R load lqs library
library(lqs)
usair.mve<-cov.mve(usair.dat[,-1],cor=TRUE)
usair.mve$cor
usair.pc1<-princomp(usair.dat[,-1],covlist=usair.mve,cor=TRUE)
summary(usair.pc1,loadings=T)
```

The resulting correlation matrix and principal components are shown in Table 3.5. (Different estimates will result each time this code is used.)

Although the pattern of correlations in Table 3.5 is largely similar to that seen in Table 3.2, there are a number of individual correlation coefficients that differ considerably in the two correlation matrices; for example, those for *Precip* and *Neg temp* (−0.386 in Table 3.2 and −0.898 in Table 3.5), and *Wind* and *Precip* (−0.013 in Table 3.2 and −0.475 in Table 3.5). The effect on the principal components analysis of these differences is, however, considerable. The first component now has a considerable negative coefficient for *Precip* and the second component is considerably different from that in Table 3.2. Labelling the coefficients is not straightfoward (at least for the author) but again it might be of interest to regress sulphur dioxide

Table 3.5 Correlation Matrix and Principal Components from Using a Robust Estimator

	Neg temp	Manuf	Pop	Wind	Precip	Days
Neg temp	1.000	0.247	0.034	0.339	−0.898	0.393
Manuf	0.247	1.000	0.842	0.292	−0.310	0.213
Pop	0.034	0.842	1.000	0.243	−0.151	0.049
Wind	0.339	0.292	0.243	1.000	−0.475	−0.109
Precip	−0.898	−0.310	−0.151	−0.475	1.000	−0.138
Days	0.393	0.213	0.049	−0.109	−0.138	1.000

Importance of components:

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6
Standard deviation	1.620	1.240	1.066	0.719	0.360	0.234
Proportion of variance	0.437	0.256	0.189	0.086	0.216	0.009
Cumulative proportion	0.437	0.694	0.883	0.969	0.991	1.000

Loadings:

	Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6
Neg temp	0.485	−0.455	—	−0.226	—	0.706
Manuf	0.447	0.495	−0.151	—	0.723	—
Pop	0.351	0.627	—	−0.137	−0.670	0.113
Wind	0.370	—	0.561	0.739	—	—
Precip	−0.512	0.354	−0.164	0.347	0.119	0.671
Days	0.205	−0.168	−0.791	0.504	−0.122	−0.188

concentration on the first two or three principal component scores of this second analysis; see Exercise 3.8.

3.4 Summary

Principal components analysis is among the oldest of multivariate techniques having been introduced originally by Pearson (1901) and independently by Hotelling (1933). It remains, however, one of the most widely employed methods of multivariate analysis, useful both for providing a convenient method of displaying multivariate data in a lower-dimensional space and for possibly simplifying other analyses of the data. Modern competitors to principal components analysis that may offer more powerful analyses of complex multivariate data are *projection pursuit* (Jones and Sibson, 1987), and *independent components analysis* (Hyvarinen et al., 2001). The former is a technique for finding "interesting" directions in multidimensional data sets; a brief account of the method is given in Everitt and Dunn (2001). The later is a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements, or signals. An R implementation of both is described on the Internet at

http://CRAN.R-project.org/

Exercises

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

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

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

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

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

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

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

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

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

3.5 Use S-PLUS or R to find the principal components of the following correlation matrix calculated from measurements of seven physical characteristics in each of 3000 convicted criminals:

$$R = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array} \begin{pmatrix} 1.00 & & & & & & \\ 0.402 & 1.00 & & & & & \\ 0.396 & 0.618 & 1.00 & & & & \\ 0.301 & 0.150 & 0.321 & 1.00 & & & \\ 0.305 & 0.135 & 0.289 & 0.846 & 1.00 & & \\ 0.339 & 0.206 & 0.363 & 0.759 & 0.797 & 1.00 & \\ 0.340 & 0.183 & 0.345 & 0.661 & 0.800 & 0.736 & 1.00 \end{pmatrix}$$

Variables:

1. Head length
2. Head breadth
3. Face breadth
4. Left finger length
5. Left forearm length
6. Left foot length
7. Height

How would you interpret the derived components?

3.6 The data in Table 3.6 show the nutritional content of different foodstuffs (the quantity involved is always three ounces). Use S-PLUS or R to create a scatterplot matrix of the data labeling the foodstuffs appropriately in each panel. On the basis of this diagram undertake what you think is an appropriate principal components analysis and try to interpret your results.

3.7 As described in the text, the air pollution data in Table 3.1 suffers from containing one or perhaps more than one outlier. Investigate this potential problem in more detail and try to reach a conclusion as to how many cities' observations

Table 3.6 Contents of Foodstuffs. From *Clustering Algorithms*, Hartigan, J.A., 1975, John Wiley & Sons, Inc. Reprinted with kind permission of J.A. Hartigan.

	Energy	Protein	Fat	Calcium	Iron
BB Beef, braised	340	20	28	9	2.6
HR Hamburger	245	21	17	9	2.7
BR Beef roast	420	15	39	7	2.0
BS Beef, steak	375	19	32	9	2.5
BC Beef, canned	180	22	10	17	3.7
CB Chicken, broiled	115	20	3	8	1.4
CC Chicken, canned	170	25	7	12	1.5
BH Beef, heart	160	26	5	14	5.9
LL Lamb leg, roast	265	20	20	9	2.6
LS Lamb shoulder, roast	300	18	25	9	2.3
HS Smoked ham	340	20	28	9	2.5
PR Pork roast	340	19	29	9	2.5
PS Pork simmered	355	19	30	9	2.4
BT Beef tongue	205	18	14	7	2.5
VC Veal cutlet	185	23	9	9	2.7
FB Bluefish, baked	135	22	4	25	0.6
AR Clams, raw	70	11	1	82	6.0
AC Clams, canned	45	7	1	74	5.4
TC Crabmeat, canned	90	14	2	38	0.8
HF Haddock, fried	135	16	5	15	0.5
MB Mackerel, broiled	200	19	13	5	1.0
MC Mackerel, canned	155	16	9	157	1.8
PF Perch, fried	195	16	11	14	1.3
SC Salmon, canned	120	17	5	159	0.7
DC Sardines, canned	180	22	9	367	2.5
UC Tuna, canned	170	25	7	7	1.2
RC Shrimp, canned	110	23	1	98	2.6

might need to be dropped before applying principal components analysis. Then undertake the analysis on the reduced data set and compare the results from those given in the text derived from using a robust estimate of the correlation matrix.

3.8 Investigate the use of the principal component scores associated with the analysis using the robust estimator of the correlation matrix as explanatory variables in a regression with sulphur dioxide concentration as dependent variable. Compare the results both with those given in Table 3.4 and those obtained in Exercise 3.7.

3.9 Investigate the use of multiple regression on the air pollution data using the human ecology and climate variables to predict sulphur dioxide pollution, keeping in mind the possible problem of the large correlation between at least two of the predictors. Do the conclusions match up to those given in the text from using principal component scores as explanatory variables?