# CHAPTER 12

# PRINCIPAL COMPONENT ANALYSIS

## 12.1 INTRODUCTION

In principal component analysis, we seek to maximize the variance of a linear combination of the variables. For example, we might want to rank students on the basis of their scores on achievement tests in English, mathematics, reading, and so on. An average score would provide a single scale on which to compare the students, but with unequal weights we can spread the students out further on the scale and obtain a better ranking.

Essentially, principal component analysis is a one-sample technique applied to data with no groupings among the observations as in Chapters 8 and 9 and no partitioning of the variables into subsets $\mathbf{y}$ and $\mathbf{x}$ as in Chapters 10 and 11. All the linear combinations that we have considered previously were related to other variables or to the data structure. In regression, we have linear combinations of the independent variables that best predict the dependent variable(s); in canonical correlation, we have linear combinations of a subset of variables that maximally correlate with linear combinations of another subset of variables; and discriminant analysis involves linear combinations that maximally separate groups of observations. Principal components, on the other hand, are concerned only with the core structure of a single

*Methods of Multivariate Analysis, Third Edition.* By Alvin C. Rencher and William F. Christensen 

sample of observations on $p$ variables. None of the variables is designated as dependent, and no grouping of observations is assumed. [For a discussion of the use of principal components with data consisting of several samples or groups, see Rencher (1998, Section 9.9)].

The first principal component is the linear combination with maximal variance; we are essentially searching for a dimension along which the observations are maximally separated or spread out. The second principal component is the linear combination with maximal variance in a direction orthogonal to the first principal component, and so on. In general, the principal components define dimensions that are different from those defined by discriminant functions or canonical variates.

In some applications, the principal components are an end in themselves and may be amenable to interpretation. More often they are obtained for use as input to another analysis. For example, two situations in regression where principal components may be useful are (1) if the number of independent variables is large relative to the number of observations, a test may be ineffective or even impossible; and (2) if the independent variables are highly correlated, the estimates of regression coefficients may be unstable. In such cases, the independent variables can be reduced to a smaller number of principal components that will yield a better test or more stable estimates of the regression coefficients. For details of this application, see Rencher (1998, Section 9.8).

As another illustration, suppose that in a MANOVA application $p$ is close to $\nu_E$, so that a test has low power, or that $p > \nu_E$, in which case we have so many dependent variables that a test cannot be made. In such cases, we can replace the dependent variables with a smaller set of principal components and then carry out the test.

In these illustrations, principal components are used to reduce the number of dimensions. A useful dimension reduction device is to evaluate the first two principal components for each observation vector and construct a scatterplot to check for multivariate normality, outliers, and so on.

Finally, we note that in the term *principal components* we use the adjective *principal*, describing what kind of components—main, primary, fundamental, major, and so on. We do not use the noun *principle* as a modifier for *components*.

## 12.2 GEOMETRIC AND ALGEBRAIC BASES OF PRINCIPAL COMPONENTS

### 12.2.1 Geometric Approach

As noted in Section 12.1, principal components analysis deals with a single sample of $n$ observation vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ that form a swarm of points in a $p$-dimensional space. Principal component analysis can be applied to any distribution of $\mathbf{y}$, but it will be easier to visualize geometrically if the swarm of points is ellipsoidal.

If the variables $y_1, y_2, \ldots, y_p$ in $\mathbf{y}$ are correlated, the ellipsoidal swarm of points is not oriented parallel to any of the axes represented by $y_1, y_2, \ldots, y_p$. We wish to find the natural axes of the swarm of points (the axes of the ellipsoid) with origin at

$\bar{\mathbf{y}}$, the mean vector of $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$. This is done by translating the origin to $\bar{\mathbf{y}}$ and then rotating the axes. After rotation so that the axes become the natural axes of the ellipsoid, the new variables (principal components) will be uncorrelated.

We could indicate the translation of the origin to $\bar{\mathbf{y}}$ by writing $\mathbf{y}_i - \bar{\mathbf{y}}$ but will not usually do so for economy of notation. We will write $\mathbf{y}_i - \bar{\mathbf{y}}$ when there is an explicit need; otherwise we assume that $\mathbf{y}_i$ has been centered.

The axes can be rotated by multiplying each $\mathbf{y}_i$ by an orthogonal matrix $\mathbf{A}$ [see (2.101), where the orthogonal matrix was denoted by $\mathbf{C}$]:

$$\mathbf{z}_i = \mathbf{A}\mathbf{y}_i. \tag{12.1}$$

Since $\mathbf{A}$ is orthogonal, $\mathbf{A}'\mathbf{A} = \mathbf{I}$, and the distance to the origin is unchanged:

$$\mathbf{z}_i'\mathbf{z}_i = (\mathbf{A}\mathbf{y}_i)'(\mathbf{A}\mathbf{y}_i) = \mathbf{y}_i'\mathbf{A}'\mathbf{A}\mathbf{y}_i = \mathbf{y}_i'\mathbf{y}_i$$

[see (2.103)]. Thus an orthogonal matrix transforms $\mathbf{y}_i$ to a point $\mathbf{z}_i$ that is the same distance from the origin, and the axes are effectively rotated.

Finding the axes of the ellipsoid is equivalent to finding the orthogonal matrix $\mathbf{A}$ that rotates the axes to line up with the natural extensions of the swarm of points so that the new variables (principal components) $z_1, z_2, \ldots, z_p$ in $\mathbf{z} = \mathbf{A}\mathbf{y}$ are uncorrelated. Thus we want the sample covariance matrix of $\mathbf{z}$, $\mathbf{S}_z = \mathbf{A}\mathbf{S}\mathbf{A}'$ [see (3.64)], to be diagonal,

$$\mathbf{S}_z = \mathbf{A}\mathbf{S}\mathbf{A}' = \begin{pmatrix} s_{z_1}^2 & 0 & \cdots & 0 \\ 0 & s_{z_2}^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & s_{z_p}^2 \end{pmatrix}, \tag{12.2}$$

where $\mathbf{S}$ is the sample covariance matrix of $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$. By (2.111), $\mathbf{C}'\mathbf{S}\mathbf{C} = \mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$, where the $\lambda_i$'s are eigenvalues of $\mathbf{S}$ and $\mathbf{C}$ is an orthogonal matrix whose columns are normalized eigenvectors of $\mathbf{S}$. Thus the orthogonal matrix $\mathbf{A}$ that diagonalizes $\mathbf{S}$ is the transpose of the matrix $\mathbf{C}$:

$$\mathbf{A} = \mathbf{C}' = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_p' \end{pmatrix}, \tag{12.3}$$

where $\mathbf{a}_i$ is the $i$th normalized ($\mathbf{a}_i'\mathbf{a}_i = 1$) eigenvector of $\mathbf{S}$. The *principal components* are the transformed variables $z_1 = \mathbf{a}_1'\mathbf{y}$, $z_2 = \mathbf{a}_2'\mathbf{y}, \ldots, z_p = \mathbf{a}_p'\mathbf{y}$ in $\mathbf{z} = \mathbf{A}\mathbf{y}$. For example, $z_1 = a_{11}y_1 + a_{12}y_2 + \cdots + a_{1p}y_p$.

By (2.111), the diagonal elements of $\mathbf{A}\mathbf{S}\mathbf{A}'$ on the right side of (12.2) are eigenvalues of $\mathbf{S}$. Hence the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ of $\mathbf{S}$ are the (sample) variances of the principal components $z_i = \mathbf{a}_i'\mathbf{y}$:

$$s_{z_i}^2 = \lambda_i. \tag{12.4}$$

Since the rotation lines up with the natural extensions of the swarm of points, $z_1 = \mathbf{a}_1'\mathbf{y}$ has the largest (sample) variance and $z_p = \mathbf{a}_p'\mathbf{y}$ has the smallest variance. This also follows from (12.4), because the variance of $z_1$ is $\lambda_1$, the largest eigenvalue, and the variance of $z_p$ is $\lambda_p$, the smallest eigenvalue. If some of the eigenvalues are small, we can neglect them and represent the points fairly well with fewer than $p$ dimensions. For example, if $p = 3$ and $\lambda_3$ is small, then the swarm of points is an "elliptical pancake" and a two-dimensional representation will adequately portray the configuration of points.

Because the eigenvalues are variances of the principal components, we can speak of "the proportion of variance explained" by the first $k$ components:

$$\begin{aligned}\text{Proportion of variance} &= \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \\ &= \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\sum_{j=1}^{p} s_{jj}},\end{aligned} \tag{12.5}$$

since $\sum_{i=1}^{p} \lambda_i = \text{tr}(\mathbf{S})$ by (2.107). Thus we try to represent the $p$-dimensional points $(y_{i1}, y_{i2}, \ldots, y_{ip})$ with a few principal components $(z_{i1}, z_{i2}, \ldots, z_{ik})$ that account for a large proportion of the total variance. If a few variables have relatively large variances, they will figure disproportionately in $\sum_j s_{jj}$ and in the principal components. For example, if $s_{22}$ is strikingly larger than the other variances, then in $z_1 = a_{11}y_1 + a_{12}y_2 + \cdots + a_{1p}y_p$, the coefficient $a_{12}$ will be large and all other $a_{1j}$ will be small.

When a ratio analogous to (12.5) is used for discriminant functions and canonical variates [see (8.13) and (11.9)], it is frequently referred to as *percent of variance*. However, in the case of discriminant functions and canonical variates, the eigenvalues are not variances, as they are in principal components.

If the variables are highly correlated, the essential dimensionality is much smaller than $p$. In this case, the first few eigenvalues will be large, and (12.5) will be close to 1 for a small value of $k$. On the other hand, if the correlations among the variables are all small, the dimensionality is close to $p$ and the eigenvalues will be nearly equal. In this case, no useful reduction in dimension is achieved, because the principal components essentially duplicate the variables.

Any two principal components $z_i = \mathbf{a}_i'\mathbf{y}$ and $z_j = \mathbf{a}_j'\mathbf{y}$ are orthogonal for $i \neq j$, that is, $\mathbf{a}_i'\mathbf{a}_j = 0$, because $\mathbf{a}_i$ and $\mathbf{a}_j$ are eigenvectors of the symmetric matrix $\mathbf{S}$ (see Section 2.11.6). Principal components also have the secondary property of being uncorrelated in the sample [see (12.2) and (3.63)]; that is, the covariance of $z_i$ and $z_j$ is zero:

$$s_{z_i z_j} = \mathbf{a}_i'\mathbf{S}\mathbf{a}_j = 0 \quad \text{for} \quad i \neq j. \tag{12.6}$$

Discriminant functions and canonical variates, on the other hand, have the weaker property of being uncorrelated but not the stronger property of orthogonality. Thus when we plot the first two discriminant functions or canonical variates on perpendicular coordinate axes, there is some distortion of their true relationship because the actual angle between their axes is not 90°.

If we change the scale on one or more of the $y$'s, the shape of the swarm of points will change, and we will need different components to represent the new points. Hence the principal components are not scale invariant. We therefore need to be concerned with the units in which the variables are measured. If possible, all variables should be expressed in the same units. If the variables have widely disparate variances, we could standardize them before extracting eigenvalues and eigenvectors. This is equivalent to finding principal components of the correlation matrix $\mathbf{R}$ and is treated in Section 12.5.

If one variable has a much greater variance than the other variables, the swarm of points will be elongated and will be nearly parallel to the axis corresponding to the variable with large variance. The first principal component will largely represent that variable, and the other principal components will have negligibly small variances. Such principal components (based on $\mathbf{S}$) do not involve the other $p-1$ variables, and we may prefer to analyze the correlation matrix $\mathbf{R}$.

### ■ EXAMPLE 12.2.1

To illustrate principal components as a rotation when $p = 2$, we use two variables from the sons data of Table 3.8: $y_1$ is head length and $y_2$ is head width for the first son. The mean vector and covariance matrix are

$$\overline{\mathbf{y}} = \begin{pmatrix} 185.7 \\ 151.1 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 95.29 & 52.87 \\ 52.87 & 54.36 \end{pmatrix}.$$

The eigenvalues and eigenvectors of $\mathbf{S}$ are

$$\lambda_1 = 131.52, \quad \lambda_2 = 18.14,$$
$$\mathbf{a}_1' = (a_{11}, a_{12}) = (.825, .565), \qquad \mathbf{a}_2' = (a_{21}, a_{22}) = (-.565, .825).$$

The symmetric pattern in the eigenvectors is due to their orthogonality: $\mathbf{a}_1'\mathbf{a}_2 = a_{11}a_{21} + a_{12}a_{22} = 0$.

The observations are plotted in Figure 12.1, along with the (translated and) rotated axes. The major axis is the line passing through $\overline{\mathbf{y}}' = (185.7, 151.1)$ in the direction determined by $\mathbf{a}_1' = (.825, .565)$; the slope is $a_{12}/a_{11} = .565/.825$. Alternatively, the equation of the major axis can be obtained by setting $z_2 = 0$:

$$\begin{aligned} z_2 = 0 &= a_{21}(y_1 - \overline{y}_1) + a_{22}(y_2 - \overline{y}_2) \\ &= -.565(y_1 - 185.7) + .825(y_2 - 151.1). \end{aligned}$$

Note that the line formed by the major axis can be considered to be a regression line. It is fit to the points so that the perpendicular distance of the points to the line is minimized, rather than the usual vertical distance (see Section 12.3).

□

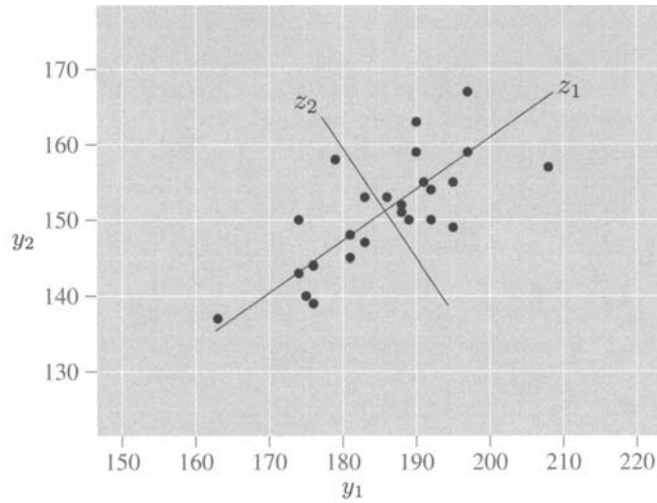


**Figure 12.1** Principal component transformation for the sons data.

### 12.2.2 Algebraic Approach

An algebraic approach to principal components can be briefly described as follows. As noted in Section 12.1, we seek a linear combination with maximal variance. By (3.55), the sample variance of $z = \mathbf{a}'\mathbf{y}$ is $\mathbf{a}'\mathbf{S}\mathbf{a}$. Since $\mathbf{a}'\mathbf{S}\mathbf{a}$ has no maximum if $\mathbf{a}$ is unrestricted, we seek the maximum of

$$\lambda = \frac{\mathbf{a}'\mathbf{S}\mathbf{a}}{\mathbf{a}'\mathbf{a}}. \tag{12.7}$$

By an argument similar to that used in (8.8)–(8.12), the maximum value of $\lambda$ is given by the largest eigenvalue in the expression

$$(\mathbf{S} - \lambda\mathbf{I})\mathbf{a} = \mathbf{0} \tag{12.8}$$

(see Problem 12.1). The eigenvector $\mathbf{a}_1$ corresponding to the largest eigenvalue $\lambda_1$ is the coefficient vector in $z_1 = \mathbf{a}_1'\mathbf{y}$, the linear combination with maximum variance.

Unlike discriminant analysis or canonical correlation, there is no inverse involved before obtaining eigenvectors for principal components. Therefore, $\mathbf{S}$ can be singular, in which case some of the eigenvalues are zero and can be ignored. A singular $\mathbf{S}$ would arise, for example, when $n < p$, that is, when the sample size is less than the number of variables.

This tolerance of principal component analysis for a singular $\mathbf{S}$ is important in certain research situations. For example, suppose that one has a one-way MANOVA with 10 observations in each of three groups and that $p = 50$, so that there are 50 variables in each of these 30 observation vectors. A MANOVA test involving $\mathbf{E}^{-1}\mathbf{H}$ cannot be carried out directly in this case because $\mathbf{E}$ is singular, but we could reduce

the 50 variables to a small number of principal components and then do a MANOVA test on the components. The principal components would be based on $\mathbf{S}$ obtained from the 30 observations (ignoring groups). For entry into the MANOVA program, we would evaluate the principal components for each observation vector. If we are retaining $k$ components, we calculate

$$\begin{aligned} z_{1i} &= \mathbf{a}_1'\mathbf{y}_i \\ z_{2i} &= \mathbf{a}_2'\mathbf{y}_i \\ &\vdots \\ z_{ki} &= \mathbf{a}_k'\mathbf{y}_i \end{aligned} \tag{12.9}$$

for $i = 1, 2, \ldots, 30$. These are sometimes referred to as *component scores*. In vector form, (12.9) can be rewritten as

$$\mathbf{z}_i = \mathbf{A}_k\mathbf{y}_i, \tag{12.10}$$

where

$$\mathbf{z}_i = \begin{pmatrix} z_{1i} \\ z_{2i} \\ \vdots \\ z_{ki} \end{pmatrix} \quad \text{and} \quad \mathbf{A}_k = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_k' \end{pmatrix}.$$

We then use $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{30}$ as input to the MANOVA program.

Note that in this case with $p > n$, the $k$ components would not likely be stable; that is, they would be different in a new sample. However, this is of no concern here because we are using the components only to extract information from the sample at hand in order to compare the three groups.

### ■ EXAMPLE 12.2.2

Consider the football data of Table 8.3. In Example 8.8, we saw that high school football players (group 1) differed from the other two groups, college football players and college-age non-football players. Therefore, to obtain a homogeneous group of observations, we delete group 1 and use groups 2 and 3 combined. The covariance matrix is as follows:

$$\mathbf{S} = \begin{pmatrix} .370 & .602 & .149 & .044 & .107 & .209 \\ .602 & 2.629 & .801 & .666 & .103 & .377 \\ .149 & .801 & .458 & .011 & -.013 & .120 \\ .044 & .666 & .011 & 1.474 & .252 & -.054 \\ .107 & .103 & -.013 & .252 & .488 & -.036 \\ .209 & .377 & .120 & -.054 & -.036 & .324 \end{pmatrix}.$$

The total variance is

$$\sum_{j=1}^{6} s_{jj} = \sum_{i=1}^{6} \lambda_i = 5.743.$$

The eigenvalues of S are as follows:

| Eigenvalue | Proportion of Variance | Cumulative Proportion |
|---|---|---|
| 3.323 | .579 | .579 |
| 1.374 | .239 | .818 |
| .476 | .083 | .901 |
| .325 | .057 | .957 |
| .157 | .027 | .985 |
| .088 | .015 | 1.000 |

The first two principal components account for 81.8% of the total variance. The corresponding eigenvectors are as follows:

| | $\mathbf{a}_1$ | $\mathbf{a}_2$ |
|---|---|---|
| WDIM | .207 | −.142 |
| CIRCUM | .873 | −.219 |
| FBEYE | .261 | −.231 |
| EYEHD | .326 | .891 |
| EARHD | .066 | .222 |
| JAW | .128 | −.187 |

Thus the first two principal components are

$$z_1 = \mathbf{a}_1'\mathbf{y} = .207y_1 + .873y_2 + .261y_3 + .326y_4 + .066y_5 + .128y_6,$$
$$z_2 = \mathbf{a}_2'\mathbf{y} = -.142y_1 - .219y_2 - .231y_3 + .891y_4 + .222y_5 - .187y_6.$$

Notice that the large coefficient in $z_1$ and the large coefficient in $z_2$, .873 and .891, respectively, correspond to the two largest variances on the diagonal of S. The two variables with large variances, $y_2$ and $y_4$, have a notable influence on the first two principal components. However, $z_1$ and $z_2$ are still meaningful linear functions. If the six variances were closer in size, the six variables would enter more evenly into the first two principal components. On the other hand, if the variances of $y_2$ and $y_4$ were substantially larger, $z_1$ and $z_2$ would be essentially equal to $y_2$ and $y_4$, respectively.

Note that $y_2$ and $y_3$ did not contribute at all when this data set was used to separate groups in Examples 8.5, 8.9, 9.3.1, and 9.6(a). However, these two variables are very useful here in the first two dimensions showing the spread of individual observations. □

## 12.3 PRINCIPAL COMPONENTS AND PERPENDICULAR REGRESSION

It was noted in Section 12.2.1 that principal components constitute a rotation of axes. Another geometric property of the line formed by the first principal component is

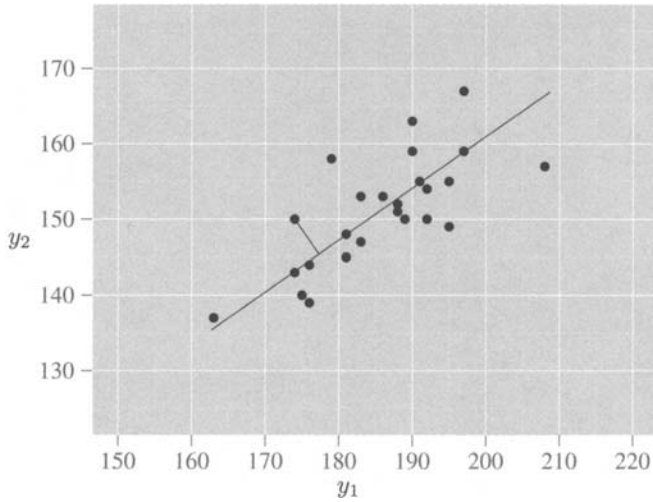


**Figure 12.2** The first principal component as a perpendicular regression line.

that it minimizes the total sum of squared perpendicular distances from the points to the line. This is easily demonstrated in the bivariate case. The first principal component line is plotted in Figure 12.2 for the first two variables of the sons data, as in Example 12.2.1. The perpendicular distance from each point to the line is simply $z_2$, the second coordinate in the transformed coordinates $(z_1, z_2)$. Hence the sum of squares of perpendicular distances is

$$\sum_{i=1}^{n} z_{2i}^2 = \sum_{i=1}^{n} [\mathbf{a}_2'(\mathbf{y}_i - \bar{\mathbf{y}})]^2, \tag{12.11}$$

where $\mathbf{a}_2$ is the second eigenvector of $\mathbf{S}$ and we use $\mathbf{y}_i - \bar{\mathbf{y}}$ because the axes have been translated to the new origin $\bar{\mathbf{y}}$. Since $\mathbf{a}_2'(\mathbf{y}_i - \bar{\mathbf{y}}) = (\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{a}_2$, we can write (12.11) in the form

$$\begin{aligned}
\sum_{i=1}^{n} z_{2i}^2 &= \sum_{i=1}^{n} \mathbf{a}_2'(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{a}_2 \\
&= \mathbf{a}_2' \left[ \sum_i (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})' \right] \mathbf{a}_2 \qquad \text{[by (2.44)]} \\
&= (n-1)\mathbf{a}_2'\mathbf{S}\mathbf{a}_2 = (n-1)\lambda_2 \qquad \text{[by (3.27)]},
\end{aligned} \tag{12.12}$$

which is a minimum [see remarks following (12.4)].

For the two variables $y_1$ and $y_2$, as plotted in Figure 12.2, the ordinary regression line of $y_2$ on $y_1$ minimizes the sum of squares of vertical distances from the points to the line. Similarly, the regression of $y_1$ on $y_2$ minimizes the sum of squares of

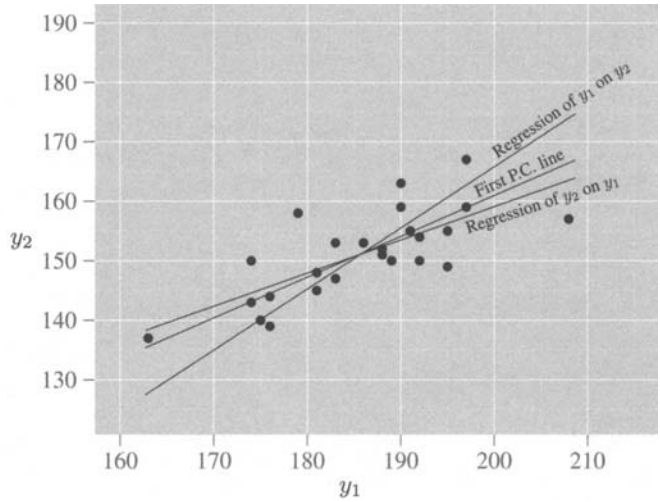


**Figure 12.3** Regression lines compared with first principal component (P.C.) line.

horizontal distances from the points to the line. The first principal component line represents a "perpendicular" regression line that lies between the other two. The three lines are compared in Figure 12.3 for the partial sons data. The equation of the first principal component line is easily obtained by setting $z_2 = 0$:

$$\begin{aligned} z_2 = \mathbf{a}_2'(\mathbf{y} - \overline{\mathbf{y}}) &= 0, \\ a_{21}(y_1 - \overline{y}_1) + a_{22}(y_2 - \overline{y}_2) &= 0, \\ -.565(y_1 - \overline{y}_1) + .825(y_2 - \overline{y}_2) &= 0. \end{aligned}$$

## 12.4 PLOTTING OF PRINCIPAL COMPONENTS

The plots in Figures 12.1 and 12.2 were illustrations of principal components as a rotation of axes when $p = 2$. When $p > 2$, we can plot the first two components as a dimension reduction device. We simply evaluate the first two components $(z_1, z_2)$ for each observation vector and plot these $n$ points. The plot is equivalent to a projection of the $p$-dimensional data swarm onto the plane that shows the greatest spread of the points.

The plot of the first two components may reveal some important features of the data set. In Example 12.4(a), we show a principal component plot that exhibits a pattern typical of a sample from a multivariate normal distribution. One of the objectives of plotting is to check for departures from normality, such as outliers or nonlinearity. In Examples 12.4(b) and 12.4(c), we illustrate principal component plots showing a nonnormal pattern characterized by the presence of outliers. Jackson

**Table 12.1** Principal Components for the Ramus Bone Data of Table 3.7

| Eigenvalues | | First Two Eigenvectors | | |
|---|---|---|---|---|
| Number | Value | Variable | $\mathbf{a}_1$ | $\mathbf{a}_2$ |
| 1 | 25.05 | AGE 8 | .474 | .592 |
| 2 | 1.74 | AGE 8.5 | .492 | .406 |
| 3 | .22 | AGE 9 | .515 | −.304 |
| 4 | .11 | AGE 9.5 | .517 | −.627 |

(1980) provided a test for adequacy of representation of observation vectors in terms of principal components.

Gnanadesikan (1997, p. 308) pointed out that, in general, the first few principal components are sensitive to outliers that inflate variances or distort covariances, and the last few are sensitive to outliers that introduce artificial dimensions or mask singularities. We could examine the bivariate plots of at least the first two and the last two principal components in a search for outliers that may exert undue influence.

Devlin et al. (1981) recommended the extraction of principal components from robust estimates of $\mathbf{S}$ or $\mathbf{R}$ that reduce the influence of outliers. Campbell (1980) and Ruymgaart (1981) discussed direct robust estimation of principal components. Critchley (1985) developed methods for detection of influential observations in principal component analysis.

Another feature of the data that a plot of the first two components may reveal is a tendency of the points to cluster. The plot may reveal groupings of points; this is illustrated in Example 12.4(d).

### ■ EXAMPLE 12.4(a)

For the modified football data in Example 12.2.2, the first two principal components were given as follows:

$$z_1 = \mathbf{a}_1'\mathbf{y} = .207y_1 + .873y_2 + .261y_3 + .326y_4 + .066y_5 + .128y_6,$$
$$z_2 = \mathbf{a}_2'\mathbf{y} = -.142y_1 - .219y_2 - .231y_3 + .891y_4 + .222y_5 - .187y_6.$$

These are evaluated for each observation vector and plotted in Figure 12.4. (For convenience in scaling, $\mathbf{y} - \overline{\mathbf{y}}$ was used in the computations.) The pattern is typical of that from a multivariate normal distribution. Note that the variance along the $z_1$ axis is greater than the variance in the $z_2$ direction, as expected. □

### ■ EXAMPLE 12.4(b)

In Figures 4.10 and 4.11, the $Q-Q$ plot and bivariate scatterplots for the ramus bone data of Table 3.7 exhibit a nonnormal pattern. A principal component analysis using the covariance matrix is given in Table 12.1, and the first two

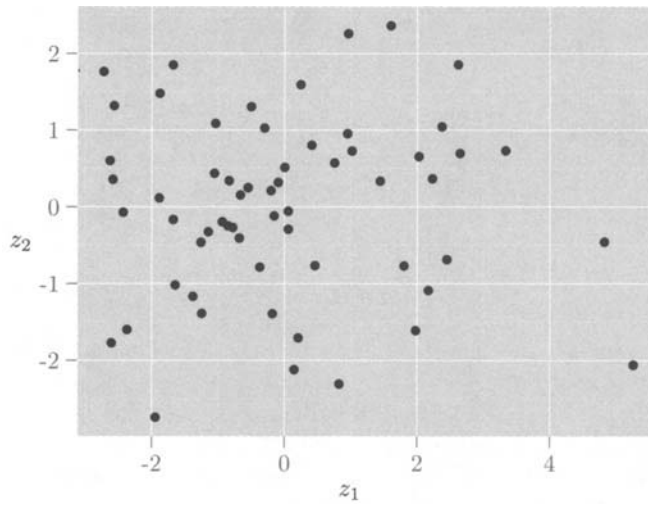


**Figure 12.4** Plot of first two components for the modified football data.

principal components are plotted in Figure 12.5. The presence of three outliers that cause a nonnormal pattern is evident. These outliers do not appear when the four variables are examined individually. □

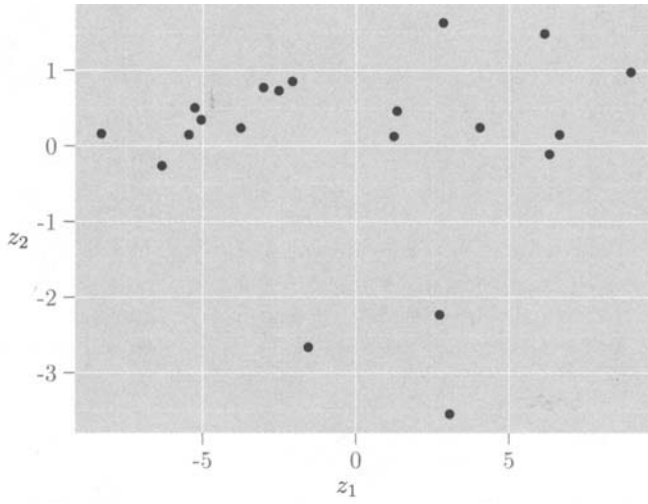


**Figure 12.5** First two principal components for the ramus bone data in Table 3.7.

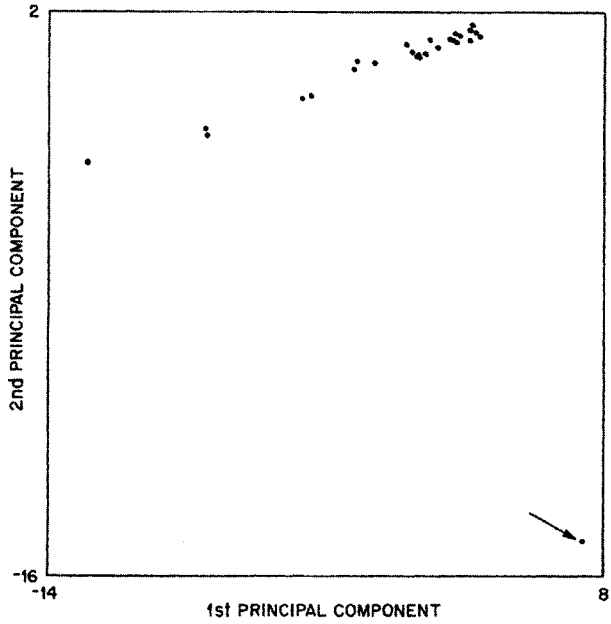


**Figure 12.6** First two principal components for economics data.

## ■ EXAMPLE 12.4(c)

A rather extreme example of the effect of an outlier is given by Devlin et al. (1981). The data set involved $p = 14$ economic variables for $n = 29$ chemical companies. The first two principal components are plotted in Figure 12.6. The sample correlation $r_{z_1 z_2}$ is indeed zero for all 29 points, as it must be [see (12.6)], but if the apparent outlier is excluded from the computation, then $r_{z_1 z_2} = .99$ for the remaining 28 points. If the outlier were deleted from the data set, the axes of the principal components would pass through the natural extensions of the data swarm. □

## ■ EXAMPLE 12.4(d)

Jeffers (1967) applied principal component analysis to a sample of 40 alate adelges (winged aphids) on which the following 19 variables had been measured:

| | |
|---|---|
| LENGTH | body length |
| WIDTH | body width |
| FORWING | forewing length |
| HINWING | hindwing length |
| SPIRAC | number of spiracles |
| ANTSEG 1 | length of antennal segment I |
| ANTSEG 2 | length of antennal segment II |

| | |
|---|---|
| ANTSEG 3 | length of antennal segment III |
| ANTSEG 4 | length of antennal segment IV |
| ANTSEG 5 | length of antennal segment V |
| ANTSPIN | number of antennal spines |
| TARSUS 3 | leg length, tarsus III |
| TIBIA 3 | leg length, tibia III |
| FEMUR 3 | leg length, femur III |
| ROSTRUM | rostrum |
| OVIPOS | ovipositor |
| OVSPIN | number of ovipositor spines |
| FOLD | anal fold |
| HOOKS | number of hindwing hooks |

**Table 12.2** Correlation Matrix for Winged Aphid Variables (Lower Triangle)

| | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_5$ | $y_6$ | $y_7$ | $y_8$ | $y_9$ | $y_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $y_1$ | 1.000 | | | | | | | | | |
| $y_2$ | .934 | 1.000 | | | | | | | | |
| $y_3$ | .927 | .941 | 1.000 | | | | | | | |
| $y_4$ | .909 | .944 | .933 | 1.000 | | | | | | |
| $y_5$ | .524 | .487 | .543 | .499 | 1.000 | | | | | |
| $y_6$ | .799 | .821 | .856 | .833 | .703 | 1.000 | | | | |
| $y_7$ | .854 | .865 | .886 | .889 | .719 | .923 | 1.000 | | | |
| $y_8$ | .789 | .834 | .846 | .885 | .253 | .699 | .751 | 1.000 | | |
| $y_9$ | .835 | .863 | .862 | .850 | .462 | .752 | .793 | .745 | 1.000 | |
| $y_{10}$ | .845 | .878 | .863 | .881 | .567 | .836 | .913 | .787 | .805 | 1.000 |
| $y_{11}$ | −.458 | −.496 | −.522 | −.488 | −.174 | −.317 | −.383 | −.497 | −.356 | −.371 |
| $y_{12}$ | .917 | .942 | .940 | .945 | .516 | .846 | .907 | .861 | .848 | .902 |
| $y_{13}$ | .939 | .961 | .956 | .952 | .494 | .849 | .914 | .876 | .877 | .901 |
| $y_{14}$ | .953 | .954 | .946 | .949 | .452 | .823 | .886 | .878 | .883 | .891 |
| $y_{15}$ | .895 | .899 | .882 | .908 | .551 | .831 | .891 | .794 | .818 | .848 |
| $y_{16}$ | .691 | .652 | .694 | .623 | .815 | .812 | .855 | .410 | .620 | .712 |
| $y_{17}$ | .327 | .305 | .356 | .272 | .746 | .553 | .567 | .067 | .300 | .384 |
| $y_{18}$ | −.676 | −.712 | −.667 | −.736 | −.233 | −.504 | −.502 | −.758 | −.666 | −.629 |
| $y_{19}$ | .702 | .729 | .746 | .777 | .285 | .499 | .592 | .793 | .671 | .668 |

| | $y_{11}$ | $y_{12}$ | $y_{13}$ | $y_{14}$ | $y_{15}$ | $y_{16}$ | $y_{17}$ | $y_{18}$ | $y_{19}$ |
|---|---|---|---|---|---|---|---|---|---|
| $y_{11}$ | 1.000 | | | | | | | | |
| $y_{12}$ | −.465 | 1.000 | | | | | | | |
| $y_{13}$ | −.447 | .981 | 1.000 | | | | | | |
| $y_{14}$ | −.439 | .971 | .991 | 1.000 | | | | | |
| $y_{15}$ | −.405 | .908 | .920 | .921 | 1.000 | | | | |
| $y_{16}$ | −.198 | .725 | .714 | .676 | .720 | 1.000 | | | |
| $y_{17}$ | −.032 | .396 | .360 | .298 | .378 | .781 | 1.000 | | |
| $y_{18}$ | .492 | −.657 | −.655 | −.678 | −.633 | −.186 | .169 | 1.000 | |
| $y_{19}$ | −.425 | .696 | .724 | .731 | .694 | .287 | .026 | −.775 | 1.000 |

**Table 12.3** Eigenvalues of the Correlation Matrix of the Winged Aphid Data

| Component | Eigenvalue | Percent of Variance | Cumulative Percent |
|---|---|---|---|
| 1 | 13.861 | 73.0 | 73.0 |
| 2 | 2.370 | 12.5 | 85.4 |
| 3 | .748 | 3.9 | 89.4 |
| 4 | .502 | 2.6 | 92.0 |
| 5 | .278 | 1.4 | 93.5 |
| 6 | .266 | 1.4 | 94.9 |
| 7 | .193 | 1.0 | 95.9 |
| 8 | .157 | .8 | 96.7 |
| 9 | .140 | .7 | 97.4 |
| 10 | .123 | .6 | 98.1 |
| 11 | .092 | .4 | 98.6 |
| 12 | .074 | .4 | 99.0 |
| 13 | .060 | .3 | 99.3 |
| 14 | .042 | .2 | 99.5 |
| 15 | .036 | .2 | 99.7 |
| 16 | .024 | .1 | 99.8 |
| 17 | .020 | .1 | 99.9 |
| 18 | .011 | .1 | 100.0 |
| 19 | .003 | .0 | 100.0 |
| | 19.000 | | |

An objective in the study was to determine the number of distinct taxa present in the habitat where the sample was taken. Since adelges are difficult to identify by the usual taxonomic methods, principal component analysis was used to search for groupings among the 40 individuals in the sample.

The correlation matrix is given in Table 12.2, and the eigenvalues and first four eigenvectors are in Tables 12.3 and 12.4, respectively. The eigenvectors are scaled so that the largest value in each is 1. The first principal component is largely an index of size. The second component is associated with SPIRAC, OVIPOS, OVSPIN, and FOLD.

The first two components were computed for each of the 40 individuals and plotted in Figure 12.7. Since the first two components account for 85% of the total variance, the plot represents the data with very little distortion. There are four major groups, apparently corresponding to species. The groupings form an interesting S-shape. □

## 12.5 PRINCIPAL COMPONENTS FROM THE CORRELATION MATRIX

Generally, extracting components from **S** rather than **R** remains closer to the spirit and intent of principal component analysis, especially if the components are to be

**Table 12.4** Eigenvectors for the First Four Components of the Winged Aphid Data

| | Eigenvectors | | | |
|---|---|---|---|---|
| Variable | 1 | 2 | 3 | 4 |
| LENGTH | .96 | −.06 | .03 | −.12 |
| WIDTH | .98 | −.12 | .01 | −.16 |
| FORWING | .99 | −.06 | −.06 | −.11 |
| HINWING | .98 | −.16 | .03 | −.00 |
| SPIRAC | .61 | .74 | −.20 | 1.00 |
| ANTSEG 1 | .91 | .33 | .04 | .02 |
| ANTSEG 2 | .96 | .30 | .00 | −.04 |
| ANTSEG 3 | .88 | −.43 | .06 | −.18 |
| ANTSEG 4 | .90 | −.08 | .18 | −.01 |
| ANTSEG 5 | .94 | .05 | .11 | .03 |
| ANTSPIN | −.49 | .37 | 1.00 | .27 |
| TARSUS 3 | .99 | −.02 | .03 | −.29 |
| TIBIA 3 | 1.00 | −.05 | .09 | −.31 |
| FEMUR 3 | .99 | −.12 | .12 | −.31 |
| ROSTRUM | .96 | .02 | .08 | −.06 |
| OVIPOS | .76 | .73 | −.03 | −.09 |
| OVSPIN | .41 | 1.00 | −.16 | −.06 |
| FOLD | −.71 | .64 | .04 | −.80 |
| HOOKS | .76 | −.52 | .06 | .72 |

used in further computations. However, in some cases, the principal components will be more interpretable if $\mathbf{R}$ is used. For example, if the variances differ widely or if the measurement units are not commensurate, the components of $\mathbf{S}$ will be dominated by the variables with large variances. The other variables will contribute very little. For a more balanced representation in such cases, components of $\mathbf{R}$ may be used (see, for example, Problem 12.9).

As with any change of scale, when the variables are standardized in transforming from $\mathbf{S}$ to $\mathbf{R}$, the shape of the swarm of points will change. Note, however, that after transforming to $\mathbf{R}$, any further changes of scale on the variables would not affect the components because changes of scale do not change $\mathbf{R}$. Thus the principal components from $\mathbf{R}$ are scale invariant.

To illustrate how the eigenvalues and eigenvectors change when converting from $\mathbf{S}$ to $\mathbf{R}$, we use a simple bivariate example in which one variance is substantially larger than the other. Suppose that $\mathbf{S}$ and the corresponding $\mathbf{R}$ have the values

$$\mathbf{S} = \begin{pmatrix} 1 & 4 \\ 4 & 25 \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} 1 & .8 \\ .8 & 1 \end{pmatrix}.$$

The eigenvalues and eigenvectors from $\mathbf{S}$ are

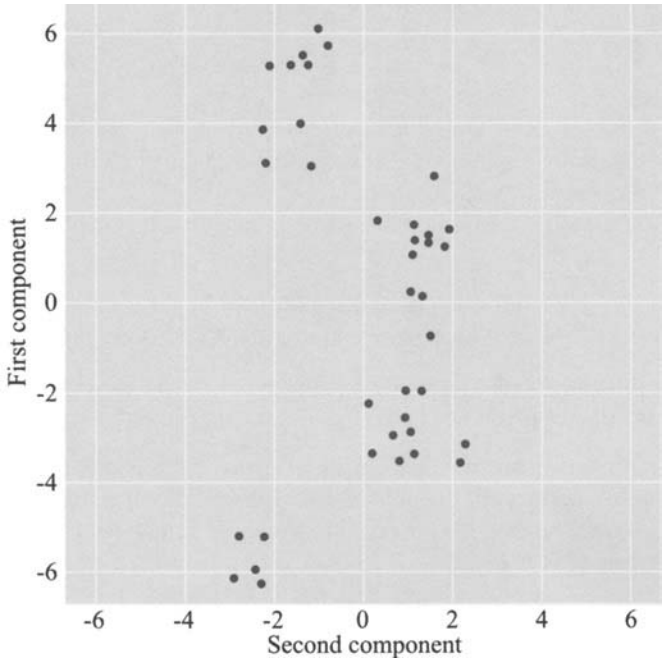


**Figure 12.7** Plotted values of the first two components for individual insects.

$$\lambda_1 = 25.65, \quad \mathbf{a}_1' = (.160, .987),$$
$$\lambda_2 = .35, \quad \mathbf{a}_2' = (.987, -.160).$$

The patterns we see in $\lambda_1, \lambda_2, \mathbf{a}_1$, and $\mathbf{a}_2$ are quite predictable. The symmetry in $\mathbf{a}_1$ and $\mathbf{a}_2$ is due to their orthogonality, $\mathbf{a}_1'\mathbf{a}_2 = 0$. The large variance of $y_2$ in $\mathbf{S}$ is reflected in the first principal component $z_1 = .160y_1 + .987y_2$, where $y_2$ is weighted heavily. Thus the first principal component $z_1$ essentially duplicates $y_2$ and does not show the mutual effect of $y_1$ and $y_2$. As expected, $z_1$ accounts for virtually all of the total variance:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{25.65}{26} = .9865.$$

The eigenvalues and eigenvectors of $\mathbf{R}$ are

$$\lambda_1 = 1.8, \quad \mathbf{a}_1' = (.707, .707),$$
$$\lambda_2 = .2, \quad \mathbf{a}_2' = (.707, -.707).$$

The first principal component of $\mathbf{R}$,

$$z_1 = .707\frac{y_1 - \overline{y}_1}{1} + .707\frac{y_2 - \overline{y}_2}{5},$$

accounts for a high proportion of variance,

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{1.8}{2} = .9,$$

because the variables are fairly highly correlated ($r = .8$). But the standardized variables $(y_1 - \overline{y}_1)/1$ and $(y_2 - \overline{y}_2)/5$ are equally weighted in $z_1$, due to the equality of the diagonal elements ("variances") of $\mathbf{R}$.

We now list some general comparisons of principal components from $\mathbf{R}$ with those from $\mathbf{S}$:

1. The percent of variance in (12.5) accounted for by the components of $\mathbf{R}$ will differ from the percent for $\mathbf{S}$, as illustrated above.

2. The coefficients of the principal components from $\mathbf{R}$ differ from those obtained from $\mathbf{S}$, as illustrated above.

3. If we express the components from $\mathbf{R}$ in terms of the original variables, they still will not agree with the components from $\mathbf{S}$. By transforming the standardized variables back to the original variables in the above illustration, the components of $\mathbf{R}$ become

$$\begin{aligned} z_1 &= .707\frac{y_1 - \overline{y}_1}{1} + .707\frac{y_2 - \overline{y}_2}{5} \\ &= .707y_1 + .141y_2 + \text{const}, \\ z_2 &= .707\frac{y_1 - \overline{y}_1}{1} - .707\frac{y_2 - \overline{y}_2}{5} \\ &= .707y_1 - .141y_2 + \text{const}. \end{aligned}$$

   As expected, these are very different from the components extracted directly from $\mathbf{S}$. This problem arises, of course, because of the lack of scale invariance of the components of $\mathbf{S}$.

4. The principal components from $\mathbf{R}$ are scale invariant, because $\mathbf{R}$ itself is scale invariant.

5. The components from a given matrix $\mathbf{R}$ are not unique to that $\mathbf{R}$. For example, in the bivariate case, the eigenvalues of

$$\mathbf{R} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

   are given by

$$\lambda_1 = 1 + r, \quad \lambda_2 = 1 - r, \tag{12.13}$$

   and the eigenvectors are $\mathbf{a}_1' = (.707, .707)$ and $\mathbf{a}_2' = (.707, -.707)$, which give principal components

$$\begin{aligned} z_1 &= .707\frac{y_1 - \overline{y}_1}{s_1} + .707\frac{y_2 - \overline{y}_2}{s_2}, \\ z_2 &= .707\frac{y_1 - \overline{y}_1}{s_1} - .707\frac{y_2 - \overline{y}_2}{s_2}. \end{aligned} \tag{12.14}$$

The components in (12.14) do not depend on $r$. For example, they serve equally well for $r = .01$ and for $r = .99$. For $r = .01$, the proportion of variance explained by $z_1$ is $\lambda_1/(\lambda_1 + \lambda_2) = (1 + .01)/(1 + .01 + 1 - .01) = 1.01/2 = .505$. For $r = .99$, the ratio is $1.99/2 = .995$. Thus the statement that the first component from a correlation matrix accounts for, say, 90% of the variance is not very meaningful. In general, for $p > 2$, the components from $\mathbf{R}$ depend only on the ratios (relative values) of the correlations, not on their actual values, and components of a given $\mathbf{R}$ matrix will serve for other $\mathbf{R}$ matrices [see Rencher (1998, Section 9.4)].

## 12.6 DECIDING HOW MANY COMPONENTS TO RETAIN

In every application, a decision must be made on how many principal components should be retained in order to effectively summarize the data. The following guidelines have been proposed:

1. Retain sufficient components to account for a specified percentage of the total variance, say 80%.

2. Retain the components whose eigenvalues are greater than the average of the eigenvalues, $\sum_{i=1}^{p} \lambda_i/p$. For a correlation matrix, this average is 1.

3. Use the *scree graph*, a plot of $\lambda_i$ versus $i$, and look for a natural break between the "large" eigenvalues and the "small" eigenvalues.

4. Test the significance of the "larger" components, that is, the components corresponding to the larger eigenvalues.

We now discuss the above four criteria for choosing the components to keep. Note, however, that the smallest components may carry valuable information that should not be routinely ignored (see Section 12.7).

In method 1, the challenge lies in selecting an appropriate threshold percentage. If we aim too high, we run the risk of including components that are either *sample specific* or *variable specific*. By *sample specific* we mean that a component may not generalize to the population or to other samples. A *variable specific* component is dominated by a single variable and does not represent a composite summary of several variables.

Method 2 is widely used and is the default in many software packages. By (2.107), $\sum_i \lambda_i = \text{tr}(\mathbf{S})$, and the average eigenvalue is also the average variance of the individual variables. Thus method 2 retains those components that account for more variance than the average variance of the variables. In cases where the data can be successfully summarized in a relatively small number of dimensions, there is often a wide gap between the two eigenvalues that fall on both sides of the average. In Example 12.2.2, the average eigenvalue (of $\mathbf{S}$) for the football data is .957, which is

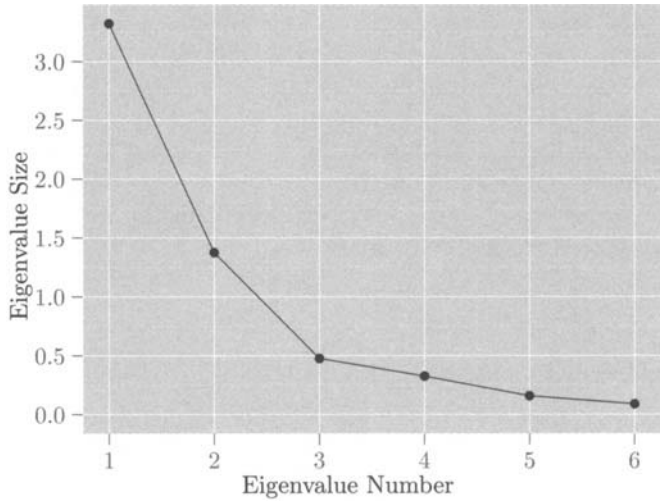


**Figure 12.8** Scree graph for eigenvalues of modified football data.

amply bracketed by $\lambda_2 = 1.37$ and $\lambda_3 = .48$. In the winged aphid data in Example 12.4(d), the second and third eigenvalues (of $\mathbf{R}$) are 2.370 and .748, leaving a comfortable margin on both sides of 1. In some cases, one may wish to move the cutoff point slightly to accommodate a visible gap in eigenvalues.

The scree graph in method 3 is named for its similarity in appearance to a cliff with rocky debris at its bottom. The scree graph for the modified football data of Example 12.2.2 exhibits an ideal pattern, as shown in Figure 12.8. The first two eigenvalues form a steep curve, followed by a bend and then a straight-line trend with shallow slope. The recommendation is to retain those eigenvalues in the steep curve *before* the first one on the straight line. Thus in Figure 12.8, two components would be retained. In practice, the turning point between the steep curve and the straight line may not be as distinct as this or there may be more than one discernible bend. In such cases, this approach is not as conclusive. The scree graph for the winged aphid data in Example 12.4(d) is plotted in Figure 12.9. The plot would suggest that two components be retained (possibly four).

The remainder of this section is devoted to method 4, tests of significance. The tests assume multivariate normality, which is not required for estimation of principal components.

It may be useful to make a preliminary test of complete independence of the variables, as in Section 7.4.3: $H_0$: $\boldsymbol{\Sigma} = \text{diag}\,(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp})$, or equivalently, $H_0$: $\mathbf{P}_\rho = \mathbf{I}$. The test statistic is given in (7.37) and (7.38). If the results indicate that the variables are independent, there is no point in extracting principal components, since (except for sampling fluctuation) the variables themselves already form the principal components.

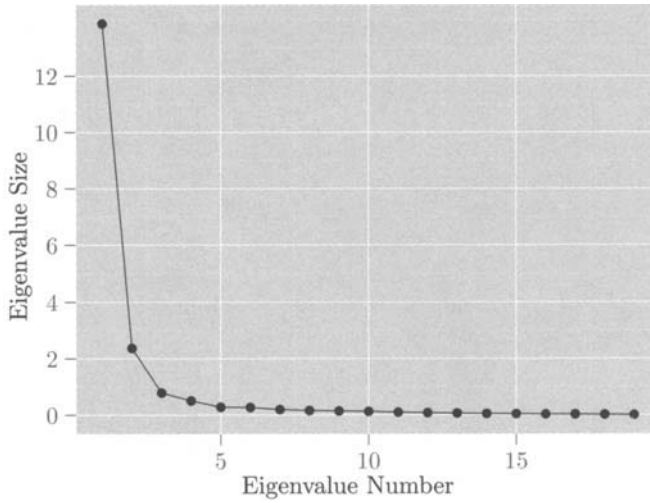


**Figure 12.9** Scree graph for eigenvalues of winged aphid data.

To test the significance of the "larger" components, we test the hypothesis that the last $k$ population eigenvalues are small and equal, $H_{0k}$: $\gamma_{p-k+1} = \gamma_{p-k+2} = \cdots = \gamma_p$, where $\gamma_1, \gamma_2, \ldots, \gamma_p$ denote the population eigenvalues, namely, the eigenvalues of $\boldsymbol{\Sigma}$. The implication is that the first sample components capture all the essential dimensions, while the last components reflect noise. If $H_0$ is true, the last $k$ sample eigenvalues will tend to have the pattern shown by the straight line with small slope in the ideal scree graph, such as in Figure 12.8 or 12.9.

To test $H_{0k}$: $\gamma_{p-k+1} = \cdots = \gamma_p$ using a likelihood ratio approach, we calculate the average of the last $k$ eigenvalues of $\mathbf{S}$,

$$\bar{\lambda} = \sum_{i=p-k+1}^{p} \frac{\lambda_i}{k}$$

and use the test statistic

$$u = \left(n - \frac{2p+11}{6}\right)\left(k \ln \bar{\lambda} - \sum_{i=p-k+1}^{p} \ln \lambda_i\right), \tag{12.15}$$

which has an approximate $\chi^2$-distribution. We reject $H_0$ if $u \geq \chi^2_{\alpha,\nu}$, where $\nu = \frac{1}{2}(k-1)(k+2)$.

To carry out this procedure, we could begin by testing $H_{02}$: $\gamma_{p-1} = \gamma_p$. If this is accepted, we could then test $H_{03}$: $\gamma_{p-2} = \gamma_{p-1} = \gamma_p$ and continue testing in this fashion until $H_{0k}$ is rejected for some value of $k$.

In practice, when the variables are fairly highly correlated and the data can be successfully represented by a small number of principal components, the first three

methods will typically agree on the number of components to retain, and the test in method 4 will often indicate a larger number of components.

## ■ EXAMPLE 12.6

We apply the above four criteria to the modified football data of Example 12.2.2.

For method 1, we simply examine the eigenvalues and their proportion of variance explained, as obtained in Example 12.2.2:

| Eigenvalue | Proportion of Variance | Cumulative Proportion |
|---|---|---|
| 3.323 | .579 | .579 |
| 1.374 | .239 | .818 |
| .476 | .083 | .901 |
| .325 | .057 | .957 |
| .157 | .027 | .985 |
| .088 | .015 | 1.000 |

To account for 82% of the variance, we would keep two components. This percent of variance is high enough for most descriptive purposes. For certain other applications, such as input to another analysis, we might wish to retain three components, which would account for 90% of the variance.

To apply method 2, we find the average eigenvalue to be

$$\overline{\lambda} = \sum_{i=1}^{6} \frac{\lambda_i}{6} = \frac{5.742824}{6} = .957.$$

Since only $\lambda_1$ and $\lambda_2$ exceed .957, we would retain two components.

For method 3, the scree graph in Figure 12.8 indicates conclusively that two components should be retained.

To implement method 4, we carry out the significance tests in (12.15). The values of the test statistic $u$ for $k = 2, 3, \ldots, 6$ are as follows:

| Eigenvalue | $k$ | $u$ | df | $\chi^2_{.05}$ |
|---|---|---|---|---|
| 3.32341 | 6 | 245.57 | 20 | 31.41 |
| 1.37431 | 5 | 123.93 | 14 | 23.68 |
| .47607 | 4 | 44.10 | 9 | 16.92 |
| .32468 | 3 | 23.84 | 5 | 11.07 |
| .15650 | 2 | 4.62 | 2 | 5.99 |
| .08785 | 1 | | | |

The tests indicate that only the last two (population) eigenvalues are equal and we should retain the first four. This differs from the results of the other three criteria, which are in close agreement that two components should be retained.

□

## 12.7 INFORMATION IN THE LAST FEW PRINCIPAL COMPONENTS

Up to this point, we have focused on using the first few principal components to summarize and simplify the data. However, the last few components may carry useful information in some applications.

Since the eigenvalues serve as variances of the principal components, the last few principal components have smaller variances. If the variance of a component is zero or close to zero, the component represents a linear relationship among the variables that is essentially constant; that is, the relationship holds for all $\mathbf{y}_i$'s in the sample. Thus if the last eigenvalue is near zero, it signifies the presence of a collinearity that may provide new information for the researcher. Suppose, for example, that there are five variables and $y_5 = \sum_{j=1}^{4} y_j/4$. Then $\mathbf{S}$ is singular, and barring round-off error, $\lambda_5$ will be zero. Thus $s_{z_5}^2 = 0$, and $z_5$ is constant. As noted early in Section 12.2, the $\mathbf{y}_i$'s are centered, because the origin of the principal components is translated to $\overline{\mathbf{y}}$. Hence the constant value of $z_5$ is its mean, which is zero:

$$z_5 = \mathbf{a}_5'\mathbf{y} = a_{51}y_1 + a_{52}y_2 + \cdots + a_{55}y_5 = 0.$$

Since this must reflect the dependency of $y_5$ on $y_1, y_2, y_3$, and $y_4$, the eigenvector $\mathbf{a}_5'$ will be proportional to $(1, 1, 1, 1, -4)$.

## 12.8 INTERPRETATION OF PRINCIPAL COMPONENTS

In Section 12.5, we noted that principal components obtained from $\mathbf{R}$ are not compatible with those obtained from $\mathbf{S}$. Because of this lack of scale invariance of principal components from $\mathbf{S}$, the coefficients cannot be converted to standardized form, as can be done with coefficients in discriminant functions in Chapter 8 and canonical variates in Chapter 11. Hence interpretation of principal components is not as clear-cut as with previous linear functions that we have discussed. We must choose between components of $\mathbf{S}$ or $\mathbf{R}$, knowing they will have a different interpretation. If the variables have widely disparate variances, we can use $\mathbf{R}$ instead of $\mathbf{S}$ to improve interpretation.

For certain patterns of elements in $\mathbf{S}$ or $\mathbf{R}$, the form of the principal components can be predicted. This aid to interpretation is discussed in Section 12.8.1. As with discriminant functions and canonical variates, some writers have advocated rotation and the use of correlations between the variables and the principal components. We argue against the use of these two approaches to interpretation in Sections 12.8.2 and 12.8.3.

### 12.8.1 Special Patterns in S or R

In the covariance or correlation matrix, we may recognize a distinguishing pattern from which the structure of the principal components can be deduced. For example, we noted in Section 12.2 that if one variable has a much larger variance than

the other variables, this variable will dominate the first component, which will account for most of the variance. Another case in which a component will duplicate a variable occurs when the variable is uncorrelated with the other variables. We now demonstrate this by showing that if all $p$ variables are uncorrelated, the variables themselves are the principal components. If the variables were uncorrelated (orthogonal), $\mathbf{S}$ would have the form

$$\mathbf{S} = \begin{pmatrix} s_{11} & 0 & \cdots & 0 \\ 0 & s_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & s_{pp} \end{pmatrix}, \tag{12.16}$$

and the characteristic equation would be

$$0 = |\mathbf{S} - \lambda\mathbf{I}| = \prod_{i=1}^{p}(s_{ii} - \lambda) \quad \text{[by (2.83)]},$$

which has solutions

$$\lambda_i = s_{ii}, \qquad i = 1, 2, \ldots, p. \tag{12.17}$$

The corresponding normalized eigenvectors have a 1 in the $i$th position and 0's elsewhere:

$$\mathbf{a}_i' = (0, \ldots, 0, 1, 0, \ldots, 0). \tag{12.18}$$

Thus the $i$th component is

$$z_i = \mathbf{a}_i'\mathbf{y} = y_i.$$

In practice, the sample correlations (of continuous random variables) will not be zero, but if the correlations are all small, the principal components will largely duplicate the variables.

By the Perron–Frobenius theroem in Section 2.11.4, if all correlations or covariances are positive, all elements of the first eigenvector $\mathbf{a}_1$ are positive. Since the remaining eigenvectors $\mathbf{a}_2, \mathbf{a}_3, \ldots, \mathbf{a}_p$ are orthogonal to $\mathbf{a}_1$, they must have both positive and negative elements. When all elements of $\mathbf{a}_1$ are positive, the first component is a weighted average of the variables and is sometimes referred to as a measure of *size*. Likewise, the positive and negative coefficients in subsequent components may be regarded as defining *shape*. This pattern is often seen when the variables are various measurements of an organism.

■ **EXAMPLE 12.8.1**

In the modified football data of Example 12.2.2, there are a few negative covariances in $\mathbf{S}$, but they are small, and all elements of the first eigenvector remain positive. The second eigenvector therefore has positive and negative elements:

First Two Eigenvectors

| | $\mathbf{a}_1$ | $\mathbf{a}_2$ |
|---|---|---|
| WDIM | .207 | −.142 |
| CIRCUM | .873 | −.219 |
| FBEYE | .261 | −.231 |
| EYEHD | .326 | .891 |
| EARHD | .066 | .222 |
| JAW | .128 | −.187 |

With all positive coefficients, the first component $z_1$ is an overall measure of head size ($z_1$ increases if all six variables increase). The second component $z_2$ is a shape component that contrasts the vertical measurements EYEHD and EARHD with the three lateral measurements and CIRCUM ($z_2$ increases if EYEHD and EARHD increase and the other four variables decrease). □

### 12.8.2 Rotation

The principal components are initially obtained by rotating axes in order to line up with the natural extensions of the system, whereupon the new variables become uncorrelated and reflect the directions of maximum variance. If the resulting components do not have a satisfactory interpretation, they can be further rotated, seeking dimensions in which many of the coefficients of the linear combinations are near zero to simplify interpretation.

However, the new rotated components are correlated, and they do not successively account for maximum variance. They are therefore no longer principal components in the usual sense, and their routine use is questionable. For improved interpretation, one may wish to try factor analysis (Chapter 13), in which rotation does not destroy any properties. (In factor analysis, the rotation does not involve the variables $y_1, y_2, \ldots, y_p$, but another space, that of the factor loadings.)

### 12.8.3 Correlations Between Variables and Principal Components

The use of correlations between variables and principal components is widely recommended as an aid to interpretation. It was noted in Sections 8.7.3 and 11.5.2 that analogous correlations for discriminant functions and canonical variates are not useful in a multivariate context because they provide only univariate information about how each variable operates by itself, ignoring the other variables. Rencher (1992b) obtained a similar result for principal components.

We denote the correlation between the $i$th variable $y_i$ and the $j$th principal component $z_j$ by $r_{y_i z_j}$. Because of the orthogonality of the $z_j$'s, we have the simple relationship

$$r^2_{y_i z_1} + r^2_{y_i z_2} + \cdots + r^2_{y_i z_k} = R^2_{y_i|z_1,\ldots,z_k}, \tag{12.19}$$

where $k$ is the number of components retained and $R^2_{y_i|z_1,\ldots,z_k}$ is the squared multiple correlation of $y_i$ with the $z_j$'s. Thus $r^2_{y_i z_j}$ forms part of $R^2_{y_i|z_1,\ldots,z_k}$, which shows how $y_i$ relates to the $z$'s by itself, not what it contributes in the presence of the other

**Table 12.5** Eigenvectors Obtained from S, Correlations Between Variables and Principal Components, and $R^2$ for the First Two Principal Components

| Variable | Eigenvectors from S | | Correlations | | |
|---|---|---|---|---|---|
| | $\mathbf{a}_1$ | $\mathbf{a}_2$ | $r_{y_i z_1}$ | $r_{y_i z_2}$ | $R^2_{y_i \mid z_1, z_2}$ |
| 1 | .21 | −.14 | .62 | −.27 | .46 |
| 2 | .87 | −.22 | .98 | −.16 | .99 |
| 3 | .26 | −.23 | .70 | −.40 | .66 |
| 4 | .33 | .89 | .49 | .86 | .98 |
| 5 | .07 | .22 | .17 | .37 | .17 |
| 6 | .13 | −.19 | .41 | −.39 | .32 |

$y$'s. The correlations are therefore not informative about the joint contribution of the $y$'s in a principal component.

Note that the simple partitioning of $R^2$ into the sum of squares of correlations in (12.19) does not happen in practice when the independent variables ($x$'s) are correlated. However, here the $z$'s are principal components and are therefore orthogonal.

Since we do not recommend rotation or correlations for interpretation, we are left with the coefficients themselves, obtained from the eigenvectors of either **S** or **R**.

■ **EXAMPLE 12.8.3**

In Example 12.8.1, the eigenvectors of **S** from the modified football data gave a satisfactory interpretation of the first two principal components as head size and shape. We give these in Table 12.5, along with the correlations between each of the variables $y_1, y_2, \ldots, y_6$ and the first two principal components $z_1$ and $z_2$. For comparison we also give $R^2_{y_i \mid z_1, z_2}$ for each variable.

The correlations rank the variables somewhat differently in their contribution to the components, since they form part of the univariate information provided by $R^2$ for each variable by itself. For example, for the first component, the correlations rank the variables in the order 2, 3, 1, 4, 6, 5, whereas the coefficients (eigenvectors) from **S** rank them in the order 2, 4, 3, 1, 6, 5. □

## 12.9 SELECTION OF VARIABLES

We have previously discussed subset selection in connection with Wilks' $\Lambda$ (Section 6.11.2), discriminant analysis (Section 8.9), classification analysis (Section 9.6), and regression (Sections 10.2.7 and 10.8). In each case the criterion for selection of variables was the relationship of the variables to some external factor, such as dependent variable(s), separation of groups, or correct classification rates. In the context of principal components, we have no dependent variable, as in regression, and no groupings among the observations, as in discriminant analysis. With no external in-

fluence, we simply wish to find the subset that best captures the internal variation (and covariation) of the variables.

Jolliffe (1972, 1973) discussed eight selection methods and referred to the process as *discarding variables*. The eight methods were based on three basic approaches: multiple correlation, clustering of variables, and principal components. One of the correlation methods, for example, proceeds in a stepwise fashion, deleting at each step the variable that has the largest multiple correlation with the other variables. The clustering methods partition the variables into groups or clusters and select a variable from each cluster.

We describe Jolliffe's principal component methods in the context of selecting a subset of 10 variables out of 50 variables. One of his techniques associates a variable with each of the first 10 principal components and retains these 10 variables. Another approach is to associate a variable with each of the last 40 principal components and delete the 40 variables. To associate a variable with a principal component, we choose the variable corresponding to the largest coefficient (in absolute value) in the component, providing the variable has not previously been selected. We can use components extracted from either $\mathbf{S}$ or $\mathbf{R}$ . For example, in the two principal components for the football data in Example 12.2.2, we would choose variables 2 and 4, which clearly have the largest coefficients in the two components. Jolliffe's methods could also be applied iteratively, with the principal components being recomputed after a variable is retained or deleted.

Jolliffe (1972) compared the eight methods using both real and simulated data and found that the methods based on principal components performed well in comparison to the regression and cluster-based methods. But he concluded that no single method was uniformly best.

McCabe (1984) suggested several criteria for selection, most of which are based on the conditional covariance matrix of the variables not selected, given those selected. He denoted the selected variables as *principal variables*. Let $\mathbf{y}$ be partitioned as

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix},$$

where $\mathbf{y}_1$ contains the selected variables and $\mathbf{y}_2$ consists of the variables not selected. The corresponding covariance matrix is

$$\text{cov}(\mathbf{y}) = \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

By (4.8), the conditional covariance matrix is given by (assuming normality)

$$\text{cov}(\mathbf{y}_2|\mathbf{y}_1) = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12},$$

which is estimated by $\mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}$. To find a subset $\mathbf{y}_1$ of size $m$, two of McCabe's criteria are to choose the subset $\mathbf{y}_1$ that

1. Minimizes $|\mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}|$ and

2. Maximizes $\sum_{i=1}^{m^*} r_i^2$, where $r_i, i = 1, 2, \ldots, m^* = \min(m, p-m)$ are the canonical correlations between the $m$ selected variables in $\mathbf{y}_1$ and the $p-m$ deleted variables in $\mathbf{y}_2$.

Ideally, these criteria would be evaluated for all possible subsets so as to obtain the best subset of each size. McCabe suggested a regression approach for obtaining a percent of variance explained by a subset of variables to be compared with the percent of variance accounted for by the same number of principal components.

## PROBLEMS

**12.1** Show that the solutions to $\lambda = \mathbf{a}'\mathbf{S}\mathbf{a}/\mathbf{a}'\mathbf{a}$ in (12.7) are given by the eigenvalues and eigenvectors in (12.8), so that $\lambda$ in (12.7) is maximized by the largest eigenvalue of $\mathbf{S}$.

**12.2** Show that the eigenvalues of

$$\mathbf{R} = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

are $1 \pm r$, as in (12.13), and that the eigenvectors are as given in (12.14).

**12.3** **(a)** Give a justification based on the likelihood ratio for the test statistic $u$ in (12.15).

**(b)** Give a justification for the degrees of freedom $\nu = \frac{1}{2}(k-1)(k+2)$ for the test statistic in (12.15).

**12.4** Show that when $\mathbf{S}$ is diagonal as in (12.16), the eigenvectors have the form $\mathbf{a}_i' = (0, \ldots, 0, 1, 0, \ldots, 0)$, as given in (12.18).

**12.5** Show that $r_{y_i z_1}^2 + r_{y_i z_2}^2 + \cdots + r_{y_i z_k}^2 = R_{y_i | z_1, \ldots, z_k}^2$, as in (12.19).

**12.6** Carry out a principal component analysis of the diabetes data of Table 3.5. Use all five variables, including $y$'s and $x$'s. Use both $\mathbf{S}$ and $\mathbf{R}$. Which do you think is more appropriate here? Show the percent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either $\mathbf{S}$ or $\mathbf{R}$?

**12.7** Do a principal component analysis of the probe word data of Table 3.6. Use both $\mathbf{S}$ and $\mathbf{R}$. Which do you think is more appropriate here? Show the percent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either $\mathbf{S}$ or $\mathbf{R}$?

**12.8** Carry out a principal component analysis on all six variables of the glucose data of Table 3.9. Use both $\mathbf{S}$ and $\mathbf{R}$. Which do you think is more appropriate here? Show the percent of variance explained. Based on the average

eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either $\mathbf{S}$ or $\mathbf{R}$?

**12.9** Carry out a principal component analysis on the hematology data of Table 4.2. Use both $\mathbf{S}$ and $\mathbf{R}$. Which do you think is more appropriate here? Show the percent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either $\mathbf{S}$ or $\mathbf{R}$? Does the large variance of $y_3$ affect the pattern of the components of $\mathbf{S}$?

**12.10** Carry out a principal component analysis separately for males and females in the psychological data of Table 5.1. Compare the results for the two groups. Use $\mathbf{S}$.

**12.11** Carry out a principal component analysis separately for the two species in the beetle data of Table 5.5. Compare the results for the two groups. Use $\mathbf{S}$.

**12.12** Carry out a principal component analysis on the engineer data of Table 5.6 as follows:

**(a)** Use the pooled covariance matrix.

**(b)** Ignore groups and use a covariance matrix based on all 40 observations.

**(c)** Which of the approaches in (a) or (b) appears to be more successful?

**12.13** Repeat the previous problem for the dystrophy data of Table 5.7.

**12.14** Carry out a principal component analysis on all 10 variables of the Seishu data of Table 7.1. Use both $\mathbf{S}$ and $\mathbf{R}$. Which do you think is more appropriate here? Show the percent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either $\mathbf{S}$ or $\mathbf{R}$?

**12.15** Carry out a principal component analysis on the temperature data of Table 7.2. Use both $\mathbf{S}$ and $\mathbf{R}$. Which do you think is more appropriate here? Show the percent of variance explained. Based on the average eigenvalue or a scree plot, decide how many components to retain. Can you interpret the components of either $\mathbf{S}$ or $\mathbf{R}$?