Chapter

PRINCIPAL COMPONENTS

8.1 Introduction

A principal component analysis is concerned with explaining the variance–covariance structure of a set of variables through a few *linear* combinations of these variables. Its general objectives are (1) data reduction and (2) interpretation.

Although p components are required to reproduce the total system variability, often much of this variability can be accounted for by a small number k of the principal components. If so, there is (almost) as much information in the k components as there is in the original p variables. The k principal components can then replace the initial p variables, and the original data set, consisting of n measurements on p variables, is reduced to a data set consisting of n measurements on k principal components.

An analysis of principal components often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily result. A good example of this is provided by the stock market data discussed in Example 8.5.

Analyses of principal components are more of a means to an end rather than an end in themselves, because they frequently serve as intermediate steps in much larger investigations. For example, principal components may be inputs to a multiple regression (see Chapter 7) or cluster analysis (see Chapter 12). Moreover, (scaled) principal components are one "factoring" of the covariance matrix for the factor analysis model considered in Chapter 9.

8.2 Population Principal Components

Algebraically, principal components are particular linear combinations of the p random variables $X_1, X_2, \ldots, X_p$. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system

with $X_1, X_2, \ldots, X_p$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and more parsimonious description of the covariance structure.

As we shall see, principal components depend solely on the covariance matrix $\boldsymbol{\Sigma}$ (or the correlation matrix $\boldsymbol{\rho}$) of $X_1, X_2, \ldots, X_p$. Their development does not require a multivariate normal assumption. On the other hand, principal components derived for multivariate normal populations have useful interpretations in terms of the constant density ellipsoids. Further, inferences can be made from the sample components when the population is multivariate normal. (See Section 8.5.)

Let the random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix $\boldsymbol{\Sigma}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Consider the linear combinations

$$\begin{aligned} Y_1 &= \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ Y_2 &= \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \qquad\qquad\qquad\qquad \vdots \\ Y_p &= \mathbf{a}_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{aligned} \tag{8-1}$$

Then, using (2-45), we obtain

$$\text{Var}(Y_i) = \mathbf{a}_i'\boldsymbol{\Sigma}\mathbf{a}_i \qquad i = 1, 2, \ldots, p \tag{8-2}$$

$$\text{Cov}(Y_i, Y_k) = \mathbf{a}_i'\boldsymbol{\Sigma}\mathbf{a}_k \qquad i, k = 1, 2, \ldots, p \tag{8-3}$$

The principal components are those *uncorrelated* linear combinations $Y_1, Y_2, \ldots, Y_p$ whose variances in (8-2) are as large as possible.

The first principal component is the linear combination with maximum variance. That is, it maximizes $\text{Var}(Y_1) = \mathbf{a}_1'\boldsymbol{\Sigma}\mathbf{a}_1$. It is clear that $\text{Var}(Y_1) = \mathbf{a}_1'\boldsymbol{\Sigma}\mathbf{a}_1$ can be increased by multiplying any $\mathbf{a}_1$ by some constant. To eliminate this indeterminacy, it is convenient to restrict attention to coefficient vectors of unit length. We therefore define

$$\begin{aligned} \text{First principal component} &= \text{linear combination } \mathbf{a}_1'\mathbf{X} \text{ that maximizes} \\ &\quad \text{Var}(\mathbf{a}_1'\mathbf{X}) \text{ subject to } \mathbf{a}_1'\mathbf{a}_1 = 1 \\ \text{Second principal component} &= \text{linear combination } \mathbf{a}_2'\mathbf{X} \text{ that maximizes} \\ &\quad \text{Var}(\mathbf{a}_2'\mathbf{X}) \text{ subject to } \mathbf{a}_2'\mathbf{a}_2 = 1 \text{ and} \\ &\quad \text{Cov}(\mathbf{a}_1'\mathbf{X}, \mathbf{a}_2'\mathbf{X}) = 0 \end{aligned}$$

At the ith step,

$$\begin{aligned} i\text{th principal component} &= \text{linear combination } \mathbf{a}_i'\mathbf{X} \text{ that maximizes} \\ &\quad \text{Var}(\mathbf{a}_i'\mathbf{X}) \text{ subject to } \mathbf{a}_i'\mathbf{a}_i = 1 \text{ and} \\ &\quad \text{Cov}(\mathbf{a}_i'\mathbf{X}, \mathbf{a}_k'\mathbf{X}) = 0 \quad \text{for} \quad k < i \end{aligned}$$

Result 8.1. Let Σ be the covariance matrix associated with the random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$. Let Σ have the eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then the *ith principal component* is given by

$$Y_i = \mathbf{e}_i'\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, \qquad i = 1, 2, \ldots, p \tag{8-4}$$

With these choices,

$$\begin{aligned} \text{Var}(Y_i) &= \mathbf{e}_i'\Sigma\mathbf{e}_i = \lambda_i \qquad i = 1, 2, \ldots, p \\ \text{Cov}(Y_i, Y_k) &= \mathbf{e}_i'\Sigma\mathbf{e}_k = 0 \qquad i \neq k \end{aligned} \tag{8-5}$$

If some λ_i are equal, the choices of the corresponding coefficient vectors, $\mathbf{e}_i$, and hence Y_i, are not unique.

Proof. We know from (2-51), with $\mathbf{B} = \Sigma$, that

$$\max_{\mathbf{a} \neq \mathbf{0}} \frac{\mathbf{a}'\Sigma\mathbf{a}}{\mathbf{a}'\mathbf{a}} = \lambda_1 \qquad \text{(attained when } \mathbf{a} = \mathbf{e}_1\text{)}$$

But $\mathbf{e}_1'\mathbf{e}_1 = 1$ since the eigenvectors are normalized. Thus,

$$\max_{\mathbf{a} \neq \mathbf{0}} \frac{\mathbf{a}'\Sigma\mathbf{a}}{\mathbf{a}'\mathbf{a}} = \lambda_1 = \frac{\mathbf{e}_1'\Sigma\mathbf{e}_1}{\mathbf{e}_1'\mathbf{e}_1} = \mathbf{e}_1'\Sigma\mathbf{e}_1 = \text{Var}(Y_1)$$

Similarly, using (2-52), we get

$$\max_{\mathbf{a} \perp \mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k} \frac{\mathbf{a}'\Sigma\mathbf{a}}{\mathbf{a}'\mathbf{a}} = \lambda_{k+1} \quad k = 1, 2, \ldots, p-1$$

For the choice $\mathbf{a} = \mathbf{e}_{k+1}$, with $\mathbf{e}_{k+1}'\mathbf{e}_i = 0$, for $i = 1, 2, \ldots, k$ and $k = 1, 2, \ldots, p-1$,

$$\mathbf{e}_{k+1}'\Sigma\mathbf{e}_{k+1}/\mathbf{e}_{k+1}'\mathbf{e}_{k+1} = \mathbf{e}_{k+1}'\Sigma\mathbf{e}_{k+1} = \text{Var}(Y_{k+1})$$

But $\mathbf{e}_{k+1}'(\Sigma\mathbf{e}_{k+1}) = \lambda_{k+1}\mathbf{e}_{k+1}'\mathbf{e}_{k+1} = \lambda_{k+1}$ so $\text{Var}(Y_{k+1}) = \lambda_{k+1}$. It remains to show that $\mathbf{e}_i$ perpendicular to $\mathbf{e}_k$ (that is, $\mathbf{e}_i'\mathbf{e}_k = 0, i \neq k$) gives $\text{Cov}(Y_i, Y_k) = 0$. Now, the eigenvectors of Σ are orthogonal if all the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ are distinct. If the eigenvalues are not all distinct, the eigenvectors corresponding to common eigenvalues may be chosen to be orthogonal. Therefore, for any two eigenvectors $\mathbf{e}_i$ and $\mathbf{e}_k$, $\mathbf{e}_i'\mathbf{e}_k = 0, i \neq k$. Since $\Sigma\mathbf{e}_k = \lambda_k\mathbf{e}_k$, premultiplication by $\mathbf{e}_i'$ gives

$$\text{Cov}(Y_i, Y_k) = \mathbf{e}_i'\Sigma\mathbf{e}_k = \mathbf{e}_i'\lambda_k\mathbf{e}_k = \lambda_k\mathbf{e}_i'\mathbf{e}_k = 0$$

for any $i \neq k$, and the proof is complete. ■

From Result 8.1, the principal components are uncorrelated and have variances equal to the eigenvalues of Σ.

Result 8.2. Let $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have covariance matrix Σ, with eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Let $Y_1 = \mathbf{e}_1'\mathbf{X}, Y_2 = \mathbf{e}_2'\mathbf{X}, \ldots, Y_p = \mathbf{e}_p'\mathbf{X}$ be the principal components. Then

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} \text{Var}(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \text{Var}(Y_i)$$

Proof. From Definition 2A.28, $\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \text{tr}(\Sigma)$. From (2-20) with $\mathbf{A} = \Sigma$, we can write $\Sigma = \mathbf{P}\Lambda\mathbf{P}'$ where Λ is the diagonal matrix of eigenvalues and $\mathbf{P} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p]$ so that $\mathbf{P}\mathbf{P}' = \mathbf{P}'\mathbf{P} = \mathbf{I}$. Using Result 2A.12(c), we have

$$\text{tr}(\Sigma) = \text{tr}(\mathbf{P}\Lambda\mathbf{P}') = \text{tr}(\Lambda\mathbf{P}'\mathbf{P}) = \text{tr}(\Lambda) = \lambda_1 + \lambda_2 + \cdots + \lambda_p$$

Thus,

$$\sum_{i=1}^{p} \text{Var}(X_i) = \text{tr}(\Sigma) = \text{tr}(\Lambda) = \sum_{i=1}^{p} \text{Var}(Y_i)$$ ■

Result 8.2 says that

$$\begin{aligned} \text{Total population variance} &= \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_p \end{aligned} \tag{8-6}$$

and consequently, the proportion of total variance due to (explained by) the kth principal component is

$$\left(\begin{array}{c}\text{Proportion of total}\\ \text{population variance}\\ \text{due to } k\text{th principal}\\ \text{component}\end{array}\right) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \qquad k = 1, 2, \ldots, p \tag{8-7}$$

If most (for instance, 80 to 90%) of the total population variance, for large p, can be attributed to the first one, two, or three components, then these components can "replace" the original p variables without much loss of information.

Each component of the coefficient vector $\mathbf{e}_i' = [e_{i1}, \ldots, e_{ik}, \ldots, e_{ip}]$ also merits inspection. The magnitude of e_{ik} measures the importance of the kth variable to the ith principal component, irrespective of the other variables. In particular, e_{ik} is proportional to the correlation coefficient between Y_i and X_k.

Result 8.3. If $Y_1 = \mathbf{e}_1'\mathbf{X}, Y_2 = \mathbf{e}_2'\mathbf{X}, \ldots, Y_p = \mathbf{e}_p'\mathbf{X}$ are the principal components obtained from the covariance matrix Σ, then

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \qquad i, k = 1, 2, \ldots, p \tag{8-8}$$

are the correlation coefficients between the components Y_i and the variables X_k. Here $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ are the eigenvalue–eigenvector pairs for Σ.

Proof. Set $\mathbf{a}_k' = [0, \ldots, 0, 1, 0, \ldots, 0]$ so that $X_k = \mathbf{a}_k'\mathbf{X}$ and $\text{Cov}(X_k, Y_i) = \text{Cov}(\mathbf{a}_k'\mathbf{X}, \mathbf{e}_i'\mathbf{X}) = \mathbf{a}_k'\Sigma\mathbf{e}_i$, according to (2-45). Since $\Sigma\mathbf{e}_i = \lambda_i\mathbf{e}_i$, $\text{Cov}(X_k, Y_i) = \mathbf{a}_k'\lambda_i\mathbf{e}_i = \lambda_i e_{ik}$. Then $\text{Var}(Y_i) = \lambda_i$ [see (8-5)] and $\text{Var}(X_k) = \sigma_{kk}$ yield

$$\rho_{Y_i, X_k} = \frac{\text{Cov}(Y_i, X_k)}{\sqrt{\text{Var}(Y_i)}\sqrt{\text{Var}(X_k)}} = \frac{\lambda_i e_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_{kk}}} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad i, k = 1, 2, \ldots, p$$ ■

Although the correlations of the variables with the principal components often help to interpret the components, they measure only the univariate contribution of an individual X to a component Y. That is, they do not indicate the importance of an X to a component Y in the presence of the other X's. For this reason, some

statisticians (see, for example, Rencher [16]) recommend that only the coefficients e_{ik}, and not the correlations, be used to interpret the components. Although the coefficients and the correlations can lead to different rankings as measures of the importance of the variables to a given component, it is our experience that these rankings are often not *appreciably* different. In practice, variables with relatively large coefficients (in absolute value) tend to have relatively large correlations, so the two measures of importance, the first multivariate and the second univariate, frequently give similar results. We recommend that both the coefficients and the correlations be examined to help interpret the principal components.

The following hypothetical example illustrates the contents of Results 8.1, 8.2, and 8.3.

Example 8.1 (Calculating the population principal components) Suppose the random variables X_1, X_2 and X_3 have the covariance matrix

$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

It may be verified that the eigenvalue–eigenvector pairs are

$$\lambda_1 = 5.83, \quad \mathbf{e}_1' = [.383, -.924, 0]$$
$$\lambda_2 = 2.00, \quad \mathbf{e}_2' = [0, 0, 1]$$
$$\lambda_3 = 0.17, \quad \mathbf{e}_3' = [.924, .383, 0]$$

Therefore, the principal components become

$$Y_1 = \mathbf{e}_1'\mathbf{X} = .383X_1 - .924X_2$$
$$Y_2 = \mathbf{e}_2'\mathbf{X} = X_3$$
$$Y_3 = \mathbf{e}_3'\mathbf{X} = .924X_1 + .383X_2$$

The variable X_3 is one of the principal components, because it is uncorrelated with the other two variables.

Equation (8-5) can be demonstrated from first principles. For example,

$$\begin{aligned} \text{Var}(Y_1) &= \text{Var}(.383X_1 - .924X_2) \\ &= (.383)^2 \text{Var}(X_1) + (-.924)^2 \text{Var}(X_2) \\ &\quad + 2(.383)(-.924) \text{Cov}(X_1, X_2) \\ &= .147(1) + .854(5) - .708(-2) \\ &= 5.83 = \lambda_1 \end{aligned}$$

$$\begin{aligned} \text{Cov}(Y_1, Y_2) &= \text{Cov}(.383X_1 - .924X_2, X_3) \\ &= .383 \text{Cov}(X_1, X_3) - .924 \text{Cov}(X_2, X_3) \\ &= .383(0) - .924(0) = 0 \end{aligned}$$

It is also readily apparent that

$$\sigma_{11} + \sigma_{22} + \sigma_{33} = 1 + 5 + 2 = \lambda_1 + \lambda_2 + \lambda_3 = 5.83 + 2.00 + .17$$

validating Equation (8-6) for this example. The proportion of total variance accounted for by the first principal component is $\lambda_1/(\lambda_1 + \lambda_2 + \lambda_3) = 5.83/8 = .73$. Further, the first two components account for a proportion $(5.83 + 2)/8 = .98$ of the population variance. In this case, the components Y_1 and Y_2 could replace the original three variables with little loss of information.

Next, using (8-8), we obtain

$$\rho_{Y_1,X_1} = \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{.383\sqrt{5.83}}{\sqrt{1}} = .925$$

$$\rho_{Y_1,X_2} = \frac{e_{12}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = \frac{-.924\sqrt{5.83}}{\sqrt{5}} = -.998$$

Notice here that the variable X_2, with coefficient $-.924$, receives the greatest weight in the component Y_1. It also has the largest correlation (in absolute value) with Y_1. The correlation of X_1, with Y_1, .925, is almost as large as that for X_2, indicating that the variables are about equally important to the first principal component. The relative sizes of the coefficients of X_1 and X_2 suggest, however, that X_2 contributes more to the determination of Y_1 than does X_1. Since, in this case, both coefficients are reasonably large and they have opposite signs, we would argue that both variables aid in the interpretation of Y_1.

Finally,

$$\rho_{Y_2,X_1} = \rho_{Y_2,X_2} = 0 \quad \text{and} \quad \rho_{Y_2,X_3} = \frac{\sqrt{\lambda_2}}{\sqrt{\sigma_{33}}} = \frac{\sqrt{2}}{\sqrt{2}} = 1 \quad \text{(as it should)}$$

The remaining correlations can be neglected, since the third component is unimportant. ■

It is informative to consider principal components derived from multivariate normal random variables. Suppose $\mathbf{X}$ is distributed as $N_p(\boldsymbol{\mu}, \Sigma)$. We know from (4-7) that the density of $\mathbf{X}$ is constant on the $\boldsymbol{\mu}$ centered ellipsoids

$$(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2$$

which have axes $\pm c\sqrt{\lambda_i}\,\mathbf{e}_i$, $i = 1, 2, \ldots, p$, where the $(\lambda_i, \mathbf{e}_i)$ are the eigenvalue–eigenvector pairs of Σ. A point lying on the ith axis of the ellipsoid will have coordinates proportional to $\mathbf{e}_i' = [e_{i1}, e_{i2}, \ldots, e_{ip}]$ in the coordinate system that has origin $\boldsymbol{\mu}$ and axes that are parallel to the original axes $x_1, x_2, \ldots, x_p$. It will be convenient to set $\boldsymbol{\mu} = \mathbf{0}$ in the argument that follows.[1]

From our discussion in Section 2.3 with $\mathbf{A} = \Sigma^{-1}$, we can write

$$c^2 = \mathbf{x}'\Sigma^{-1}\mathbf{x} = \frac{1}{\lambda_1}(\mathbf{e}_1'\mathbf{x})^2 + \frac{1}{\lambda_2}(\mathbf{e}_2'\mathbf{x})^2 + \cdots + \frac{1}{\lambda_p}(\mathbf{e}_p'\mathbf{x})^2$$

[1]This can be done without loss of generality because the normal random vector $\mathbf{X}$ can always be translated to the normal random vector $\mathbf{W} = \mathbf{X} - \boldsymbol{\mu}$ and $E(\mathbf{W}) = \mathbf{0}$. However, $\text{Cov}(\mathbf{X}) = \text{Cov}(\mathbf{W})$.

where $\mathbf{e}_1'\mathbf{x}, \mathbf{e}_2'\mathbf{x}, \ldots, \mathbf{e}_p'\mathbf{x}$ are recognized as the principal components of $\mathbf{x}$. Setting $y_1 = \mathbf{e}_1'\mathbf{x}, y_2 = \mathbf{e}_2'\mathbf{x}, \ldots, y_p = \mathbf{e}_p'\mathbf{x}$, we have

$$c^2 = \frac{1}{\lambda_1} y_1^2 + \frac{1}{\lambda_2} y_2^2 + \cdots + \frac{1}{\lambda_p} y_p^2$$

and this equation defines an ellipsoid (since $\lambda_1, \lambda_2, \ldots, \lambda_p$ are positive) in a coordinate system with axes $y_1, y_2, \ldots, y_p$ lying in the directions of $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$, respectively. If λ_1 is the largest eigenvalue, then the major axis lies in the direction $\mathbf{e}_1$. The remaining minor axes lie in the directions defined by $\mathbf{e}_2, \ldots, \mathbf{e}_p$.

To summarize, the principal components $y_1 = \mathbf{e}_1'\mathbf{x}, y_2 = \mathbf{e}_2'\mathbf{x}, \ldots, y_p = \mathbf{e}_p'\mathbf{x}$ lie in the directions of the axes of a constant density ellipsoid. Therefore, any point on the ith ellipsoid axis has $\mathbf{x}$ coordinates proportional to $\mathbf{e}_i' = [e_{i1}, e_{i2}, \ldots, e_{ip}]$ and, necessarily, principal component coordinates of the form $[0, \ldots, 0, y_i, 0, \ldots, 0]$.

When $\boldsymbol{\mu} \neq \mathbf{0}$, it is the mean-centered principal component $y_i = \mathbf{e}_i'(\mathbf{x} - \boldsymbol{\mu})$ that has mean 0 and lies in the direction $\mathbf{e}_i$.

A constant density ellipse and the principal components for a bivariate normal random vector with $\boldsymbol{\mu} = \mathbf{0}$ and $\rho = .75$ are shown in Figure 8.1. We see that the principal components are obtained by rotating the original coordinate axes through an angle θ until they coincide with the axes of the constant density ellipse. This result holds for $p > 2$ dimensions as well.

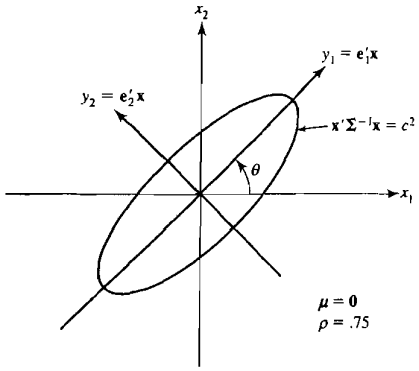

Figure 8.1 The constant density ellipse $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} = c^2$ and the principal components y_1, y_2 for a bivariate normal random vector $\mathbf{X}$ having mean $\mathbf{0}$.

Principal Components Obtained from Standardized Variables

Principal components may also be obtained for the standardized variables

$$\begin{aligned} Z_1 &= \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}} \\ Z_2 &= \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}} \\ &\vdots \qquad \vdots \\ Z_p &= \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}} \end{aligned} \tag{8-9}$$

In matrix notation,

$$\mathbf{Z} = (\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \boldsymbol{\mu}) \tag{8-10}$$

where the diagonal standard deviation matrix $\mathbf{V}^{1/2}$ is defined in (2-35). Clearly, $E(\mathbf{Z}) = \mathbf{0}$ and

$$\text{Cov}(\mathbf{Z}) = (\mathbf{V}^{1/2})^{-1}\boldsymbol{\Sigma}(\mathbf{V}^{1/2})^{-1} = \boldsymbol{\rho}$$

by (2-37). The principal components of $\mathbf{Z}$ may be obtained from the eigenvectors of the *correlation* matrix $\boldsymbol{\rho}$ of $\mathbf{X}$. All our previous results apply, with some simplifications, since the variance of each Z_i is unity. We shall continue to use the notation Y_i to refer to the ith principal component and $(\lambda_i, \mathbf{e}_i)$ for the eigenvalue–eigenvector pair from either $\boldsymbol{\rho}$ or $\boldsymbol{\Sigma}$. *However, the* $(\lambda_i, \mathbf{e}_i)$ *derived from* $\boldsymbol{\Sigma}$ *are, in general, not the same as the ones derived from* $\boldsymbol{\rho}$.

Result 8.4. The ith principal component of the standardized variables $\mathbf{Z}' = [Z_1, Z_2, \ldots, Z_p]$ with $\text{Cov}(\mathbf{Z}) = \boldsymbol{\rho}$, is given by

$$Y_i = \mathbf{e}_i'\mathbf{Z} = \mathbf{e}_i'(\mathbf{V}^{1/2})^{-1}(\mathbf{X} - \boldsymbol{\mu}), \qquad i = 1, 2, \ldots, p$$

Moreover,

$$\sum_{i=1}^{p} \text{Var}(Y_i) = \sum_{i=1}^{p} \text{Var}(Z_i) = p \tag{8-11}$$

and

$$\rho_{Y_i, Z_k} = e_{ik}\sqrt{\lambda_i} \qquad i, k = 1, 2, \ldots, p$$

In this case, $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ are the eigenvalue–eigenvector pairs for $\boldsymbol{\rho}$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Proof. Result 8.4 follows from Results 8.1, 8.2, and 8.3, with $Z_1, Z_2, \ldots, Z_p$ in place of $X_1, X_2, \ldots, X_p$ and $\boldsymbol{\rho}$ in place of $\boldsymbol{\Sigma}$. ■

We see from (8-11) that the total (standardized variables) population variance is simply p, the sum of the diagonal elements of the matrix $\boldsymbol{\rho}$. Using (8-7) with $\mathbf{Z}$ in place of $\mathbf{X}$, we find that the proportion of total variance explained by the kth principal component of $\mathbf{Z}$ is

$$\begin{pmatrix} \text{Proportion of (standardized)} \\ \text{population variance due} \\ \text{to } k\text{th principal component} \end{pmatrix} = \frac{\lambda_k}{p}, \qquad k = 1, 2, \ldots, p \tag{8-12}$$

where the λ_k's are the eigenvalues of $\boldsymbol{\rho}$.

Example 8.2 (Principal components obtained from covariance and correlation matrices are different) Consider the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$$

and the derived correlation matrix

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & .4 \\ .4 & 1 \end{bmatrix}$$

The eigenvalue–eigenvector pairs from $\boldsymbol{\Sigma}$ are

$$\lambda_1 = 100.16, \quad \mathbf{e}_1' = [.040, .999]$$
$$\lambda_2 = .84, \quad \mathbf{e}_2' = [.999, -.040]$$

Similarly, the eigenvalue–eigenvector pairs from $\boldsymbol{\rho}$ are

$$\lambda_1 = 1 + \rho = 1.4, \quad \mathbf{e}_1' = [.707, .707]$$
$$\lambda_2 = 1 - \rho = .6, \quad \mathbf{e}_2' = [.707, -.707]$$

The respective principal components become

$$\boldsymbol{\Sigma}: \quad \begin{aligned} Y_1 &= .040X_1 + .999X_2 \\ Y_2 &= .999X_1 - .040X_2 \end{aligned}$$

and

$$\boldsymbol{\rho}: \quad \begin{aligned} Y_1 = .707Z_1 + .707Z_2 &= .707\left(\frac{X_1 - \mu_1}{1}\right) + .707\left(\frac{X_2 - \mu_2}{10}\right) \\ &= .707(X_1 - \mu_1) + .0707(X_2 - \mu_2) \\ Y_2 = .707Z_1 - .707Z_2 &= .707\left(\frac{X_1 - \mu_1}{1}\right) - .707\left(\frac{X_2 - \mu_2}{10}\right) \\ &= .707(X_1 - \mu_1) - .0707(X_2 - \mu_2) \end{aligned}$$

Because of its large variance, X_2 completely dominates the first principal component determined from $\boldsymbol{\Sigma}$. Moreover, this first principal component explains a proportion

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{100.16}{101} = .992$$

of the total population variance.

When the variables X_1 and X_2 are standardized, however, the resulting variables contribute equally to the principal components determined from $\boldsymbol{\rho}$. Using Result 8.4, we obtain

$$\rho_{Y_1, Z_1} = e_{11}\sqrt{\lambda_1} = .707\sqrt{1.4} = .837$$

and

$$\rho_{Y_1, Z_2} = e_{21}\sqrt{\lambda_1} = .707\sqrt{1.4} = .837$$

In this case, the first principal component explains a proportion

$$\frac{\lambda_1}{p} = \frac{1.4}{2} = .7$$

of the total (standardized) population variance.

Most strikingly, we see that the relative importance of the variables to, for instance, the first principal component is greatly affected by the standardization.

When the first principal component obtained from $\boldsymbol{\rho}$ is expressed in terms of X_1 and X_2, the relative magnitudes of the weights .707 and .0707 are in direct opposition to those of the weights .040 and .999 attached to these variables in the principal component obtained from $\boldsymbol{\Sigma}$. ■

The preceding example demonstrates that the principal components derived from $\boldsymbol{\Sigma}$ are different from those derived from $\boldsymbol{\rho}$. Furthermore, one set of principal components is not a simple function of the other. This suggests that the standardization is not inconsequential.

Variables should probably be standardized if they are measured on scales with widely differing ranges or if the units of measurement are not commensurate. For example, if X_1 represents annual sales in the \$10,000 to \$350,000 range and X_2 is the ratio (net annual income)/(total assets) that falls in the .01 to .60 range, then the total variation will be due almost exclusively to dollar sales. In this case, we would expect a single (important) principal component with a heavy weighting of X_1. Alternatively, if both variables are standardized, their subsequent magnitudes will be of the same order, and X_2 (or Z_2) will play a larger role in the construction of the principal components. This behavior was observed in Example 8.2.

Principal Components for Covariance Matrices with Special Structures

There are certain patterned covariance and correlation matrices whose principal components can be expressed in simple forms. Suppose $\boldsymbol{\Sigma}$ is the diagonal matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{bmatrix} \tag{8-13}$$

Setting $\mathbf{e}_i' = [0, \ldots, 0, 1, 0, \ldots, 0]$, with 1 in the ith position, we observe that

$$\begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1\sigma_{ii} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{or} \quad \boldsymbol{\Sigma}\mathbf{e}_i = \sigma_{ii}\mathbf{e}_i$$

and we conclude that $(\sigma_{ii}, \mathbf{e}_i)$ is the ith eigenvalue–eigenvector pair. Since the linear combination $\mathbf{e}_i'\mathbf{X} = X_i$, the set of principal components is just the original set of uncorrelated random variables.

For a covariance matrix with the pattern of (8-13), nothing is gained by extracting the principal components. From another point of view, if $\mathbf{X}$ is distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the contours of constant density are ellipsoids whose axes already lie in the directions of maximum variation. Consequently, there is no need to rotate the coordinate system.

Standardization does not substantially alter the situation for the Σ in (8-13). In that case, $\boldsymbol{\rho} = \mathbf{I}$, the $p \times p$ identity matrix. Clearly, $\boldsymbol{\rho}\mathbf{e}_i = 1\mathbf{e}_i$, so the eigenvalue 1 has multiplicity p and $\mathbf{e}_i' = [0, \ldots, 0, 1, 0, \ldots, 0]$, $i = 1, 2, \ldots, p$, are convenient choices for the eigenvectors. Consequently, the principal components determined from $\boldsymbol{\rho}$ are also the original variables $Z_1, \ldots, Z_p$. Moreover, in this case of equal eigenvalues, the multivariate normal ellipsoids of constant density are spheroids.

Another patterned covariance matrix, which often describes the correspondence among certain biological variables such as the sizes of living things, has the general form

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \cdots & \sigma^2 \end{bmatrix} \tag{8-14}$$

The resulting correlation matrix

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \tag{8-15}$$

is also the covariance matrix of the standardized variables. The matrix in (8-15) implies that the variables $X_1, X_2, \ldots, X_p$ are equally correlated.

It is not difficult to show (see Exercise 8.5) that the p eigenvalues of the correlation matrix (8-15) can be divided into two groups. When ρ is positive, the largest is

$$\lambda_1 = 1 + (p - 1)\rho \tag{8-16}$$

with associated eigenvector

$$\mathbf{e}_1' = \left[\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}, \ldots, \frac{1}{\sqrt{p}}\right] \tag{8-17}$$

The remaining $p - 1$ eigenvalues are

$$\lambda_2 = \lambda_3 = \cdots = \lambda_p = 1 - \rho$$

and one choice for their eigenvectors is

$$\mathbf{e}_2' = \left[\frac{1}{\sqrt{1 \times 2}}, \frac{-1}{\sqrt{1 \times 2}}, 0, \ldots, 0\right]$$

$$\mathbf{e}_3' = \left[\frac{1}{\sqrt{2 \times 3}}, \frac{1}{\sqrt{2 \times 3}}, \frac{-2}{\sqrt{2 \times 3}}, 0, \ldots, 0\right]$$

$$\vdots \qquad\qquad \vdots$$

$$\mathbf{e}_i' = \left[\frac{1}{\sqrt{(i-1)i}}, \ldots, \frac{1}{\sqrt{(i-1)i}}, \frac{-(i-1)}{\sqrt{(i-1)i}}, 0, \ldots, 0\right]$$

$$\vdots \qquad\qquad \vdots$$

$$\mathbf{e}_p' = \left[\frac{1}{\sqrt{(p-1)p}}, \ldots, \frac{1}{\sqrt{(p-1)p}}, \frac{-(p-1)}{\sqrt{(p-1)p}}\right]$$

The first principal component

$$Y_1 = \mathbf{e}_1'\mathbf{Z} = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} Z_i$$

is proportional to the sum of the p standarized variables. It might be regarded as an "index" with equal weights. This principal component explains a proportion

$$\frac{\lambda_1}{p} = \frac{1 + (p - 1)\rho}{p} = \rho + \frac{1 - \rho}{p} \tag{8-18}$$

of the total population variation. We see that $\lambda_1/p \doteq \rho$ for ρ close to 1 or p large. For example, if $\rho = .80$ and $p = 5$, the first component explains 84% of the total variance. When ρ is near 1, the last $p - 1$ components collectively contribute very little to the total variance and can often be neglected. In this special case, retaining only the first principal component $Y_1 = (1/\sqrt{p})[1, 1, \ldots, 1]\mathbf{X}$, a measure of total size, still explains the same proportion (8-18) of total variance.

If the standardized variables $Z_1, Z_2, \ldots, Z_p$ have a multivariate normal distribution with a covariance matrix given by (8-15), then the ellipsoids of constant density are "cigar shaped," with the major axis proportional to the first principal component $Y_1 = (1/\sqrt{p})[1, 1, \ldots, 1]\mathbf{Z}$. This principal component is the projection of $\mathbf{Z}$ on the equiangular line $\mathbf{1}' = [1, 1, \ldots, 1]$. The minor axes (and remaining principal components) occur in spherically symmetric directions perpendicular to the major axis (and first principal component).

8.3 Summarizing Sample Variation by Principal Components

We now have the framework necessary to study the problem of summarizing the variation in n measurements on p variables with a few judiciously chosen linear combinations.

Suppose the data $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ represent n independent drawings from some p-dimensional population with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. These data yield the sample mean vector $\bar{\mathbf{x}}$, the sample covariance matrix $\mathbf{S}$, and the sample correlation matrix $\mathbf{R}$.

Our objective in this section will be to construct uncorrelated linear combinations of the measured characteristics that account for much of the variation in the sample. The uncorrelated combinations with the largest variances will be called the *sample principal components*.

Recall that the n values of any linear combination

$$\mathbf{a}_1'\mathbf{x} = a_{11}x_{j1} + a_{12}x_{j2} + \cdots + a_{1p}x_{jp}, \qquad j = 1, 2, \ldots, n$$

have sample mean $\mathbf{a}_1'\bar{\mathbf{x}}$ and sample variance $\mathbf{a}_1'\mathbf{S}\mathbf{a}_1$. Also, the pairs of values $(\mathbf{a}_1'\mathbf{x}_j, \mathbf{a}_2'\mathbf{x}_j)$, for two linear combinations, have sample covariance $\mathbf{a}_1'\mathbf{S}\mathbf{a}_2$ [see (3-36)].

The sample principal components are defined as those linear combinations which have maximum sample variance. As with the population quantities, we restrict the coefficient vectors $\mathbf{a}_i$ to satisfy $\mathbf{a}_i'\mathbf{a}_i = 1$. Specifically,

First *sample* principal component = linear combination $\mathbf{a}_1'\mathbf{x}_j$ that maximizes the sample variance of $\mathbf{a}_1'\mathbf{x}_j$ subject to $\mathbf{a}_1'\mathbf{a}_1 = 1$

Second *sample* principal component = linear combination $\mathbf{a}_2'\mathbf{x}_j$ that maximizes the sample variance of $\mathbf{a}_2'\mathbf{x}_j$ subject to $\mathbf{a}_2'\mathbf{a}_2 = 1$ and zero sample covariance for the pairs $(\mathbf{a}_1'\mathbf{x}_j, \mathbf{a}_2'\mathbf{x}_j)$

At the ith step, we have

ith *sample* principal component = linear combination $\mathbf{a}_i'\mathbf{x}_j$ that maximizes the sample variance of $\mathbf{a}_i'\mathbf{x}_j$ subject to $\mathbf{a}_i'\mathbf{a}_i = 1$ and zero sample covariance for all pairs $(\mathbf{a}_i'\mathbf{x}_j, \mathbf{a}_k'\mathbf{x}_j)$, $k < i$

The first principal component maximizes $\mathbf{a}_1'\mathbf{S}\mathbf{a}_1$ or, equivalently,

$$\frac{\mathbf{a}_1'\mathbf{S}\mathbf{a}_1}{\mathbf{a}_1'\mathbf{a}_1} \tag{8-19}$$

By (2-51), the maximum is the largest eigenvalue $\hat{\lambda}_1$ attained for the choice $\mathbf{a}_1 =$ eigenvector $\hat{\mathbf{e}}_1$ of $\mathbf{S}$. Successive choices of $\mathbf{a}_i$ maximize (8-19) subject to $0 = \mathbf{a}_i'\mathbf{S}\hat{\mathbf{e}}_k = \mathbf{a}_i'\hat{\lambda}_k\hat{\mathbf{e}}_k$, or $\mathbf{a}_i$ perpendicular to $\hat{\mathbf{e}}_k$. Thus, as in the proofs of Results 8.1–8.3, we obtain the following results concerning sample principal components:

If $\mathbf{S} = \{s_{ik}\}$ is the $p \times p$ sample covariance matrix with eigenvalue-eigenvector pairs $(\hat{\lambda}_1, \hat{\mathbf{e}}_1), (\hat{\lambda}_2, \hat{\mathbf{e}}_2), \ldots, (\hat{\lambda}_p, \hat{\mathbf{e}}_p)$, the ith sample principal component is given by

$$\hat{y}_i = \hat{\mathbf{e}}_i'\mathbf{x} = \hat{e}_{i1}x_1 + \hat{e}_{i2}x_2 + \cdots + \hat{e}_{ip}x_p, \qquad i = 1, 2, \ldots, p$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$ and $\mathbf{x}$ is any observation on the variables $X_1, X_2, \ldots, X_p$. Also,

$$\begin{aligned} \text{Sample variance}(\hat{y}_k) &= \hat{\lambda}_k, \quad k = 1, 2, \ldots, p \\ \text{Sample covariance}(\hat{y}_i, \hat{y}_k) &= 0, \quad i \neq k \end{aligned} \tag{8-20}$$

In addition,

$$\text{Total sample variance} = \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p$$

and

$$r_{\hat{y}_i, x_k} = \frac{\hat{e}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}, \quad i, k = 1, 2, \ldots, p$$

We shall denote the sample principal components by $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_p$, irrespective of whether they are obtained from $\mathbf{S}$ or $\mathbf{R}$.[2] The components constructed from $\mathbf{S}$ and $\mathbf{R}$ are *not* the same, in general, but it will be clear from the context which matrix is being used, and the single notation $\hat{y}_i$ is convenient. It is also convenient to label the component coefficient vectors $\hat{\mathbf{e}}_i$ and the component variances $\hat{\lambda}_i$ for both situations.

The observations $\mathbf{x}_j$ are often "centered" by subtracting $\bar{\mathbf{x}}$. This has no effect on the sample covariance matrix $\mathbf{S}$ and gives the ith principal component

$$\hat{y}_i = \hat{\mathbf{e}}_i'(\mathbf{x} - \bar{\mathbf{x}}), \qquad i = 1, 2, \ldots, p \tag{8-21}$$

for any observation vector $\mathbf{x}$. If we consider the *values* of the ith component

$$\hat{y}_{ji} = \hat{\mathbf{e}}_i'(\mathbf{x}_j - \bar{\mathbf{x}}), \qquad j = 1, 2, \ldots, n \tag{8-22}$$

generated by substituting each observation $\mathbf{x}_j$ for the arbitrary $\mathbf{x}$ in (8-21), then

$$\bar{\hat{y}}_i = \frac{1}{n}\sum_{j=1}^{n} \hat{\mathbf{e}}_i'(\mathbf{x}_j - \bar{\mathbf{x}}) = \frac{1}{n}\hat{\mathbf{e}}_i'\left(\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})\right) = \frac{1}{n}\hat{\mathbf{e}}_i'\mathbf{0} = 0 \tag{8-23}$$

That is, the sample mean of each principal component is zero. The sample variances are still given by the $\hat{\lambda}_i$'s, as in (8-20).

Example 8.3 (Summarizing sample variability with two sample principal components) A census provided information, by tract, on five socioeconomic variables for the Madison, Wisconsin, area. The data from 61 tracts are listed in Table 8.5 in the exercises at the end of this chapter. These data produced the following summary statistics:

$$\bar{\mathbf{x}}' = \underset{\substack{\text{total}\\ \text{population}\\ \text{(thousands)}}}{[4.47,} \quad \underset{\substack{\text{professional}\\ \text{degree}\\ \text{(percent)}}}{3.96,} \quad \underset{\substack{\text{employed}\\ \text{age over 16}\\ \text{(percent)}}}{71.42,} \quad \underset{\substack{\text{government}\\ \text{employment}\\ \text{(percent)}}}{26.91,} \quad \underset{\substack{\text{median}\\ \text{home value}\\ \text{(\$100,000)}}}{1.64]}$$

and

$$\mathbf{S} = \begin{bmatrix} 3.397 & -1.102 & 4.306 & -2.078 & 0.027 \\ -1.102 & 9.673 & -1.513 & 10.953 & 1.203 \\ 4.306 & -1.513 & 55.626 & -28.937 & -0.044 \\ -2.078 & 10.953 & -28.937 & 89.067 & 0.957 \\ 0.027 & 1.203 & -0.044 & 0.957 & 0.319 \end{bmatrix}$$

Can the sample variation be summarized by one or two principal components?

[2]Sample principal components also can be obtained from $\hat{\boldsymbol{\Sigma}} = \mathbf{S}_n$, the maximum likelihood estimate of the covariance matrix $\boldsymbol{\Sigma}$, if the $\mathbf{X}_j$ are normally distributed. (See Result 4.11.) In this case, provided that the eigenvalues of $\boldsymbol{\Sigma}$ are distinct, the sample principal components can be viewed as the maximum likelihood estimates of the corresponding population counterparts. (See [1].) We shall not consider $\hat{\boldsymbol{\Sigma}}$ because the assumption of normality is not required in this section. Also, $\hat{\boldsymbol{\Sigma}}$ has eigenvalues $[(n-1)/n]\hat{\lambda}_i$, and corresponding eigenvectors $\hat{\mathbf{e}}_i$, where $(\hat{\lambda}_i, \hat{\mathbf{e}}_i)$ are the eigenvalue–eigenvector pairs for $\mathbf{S}$. Thus, both $\mathbf{S}$ and $\hat{\boldsymbol{\Sigma}}$ give the same sample principal components $\hat{\mathbf{e}}_i'\mathbf{x}$ [see (8-20)] and the same proportion of explained variance $\hat{\lambda}_i/(\hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p)$. Finally, both $\mathbf{S}$ and $\hat{\boldsymbol{\Sigma}}$ give the same sample correlation matrix $\mathbf{R}$, so if the variables are standardized, the choice of $\mathbf{S}$ or $\hat{\boldsymbol{\Sigma}}$ is irrelevant.

We find the following:

Coefficients for the Principal Components
(Correlation Coefficients in Parentheses)

Variable	$\hat{\mathbf{e}}_1\,(r_{\hat{y}_1,x_k})$	$\hat{\mathbf{e}}_2\,(r_{\hat{y}_2,x_k})$	$\hat{\mathbf{e}}_3$	$\hat{\mathbf{e}}_4$	$\hat{\mathbf{e}}_5$
Total population	−0.039(−.22)	0.071(.24)	0.188	0.977	−0.058
Profession	0.105(.35)	0.130(.26)	−0.961	0.171	−0.139
Employment (%)	−0.492(−.68)	0.864(.73)	0.046	−0.091	0.005
Government employment (%)	0.863(.95)	0.480(.32)	0.153	−0.030	0.007
Medium home value	0.009(.16)	0.015(.17)	−0.125	0.082	0.989
Variance ($\hat{\lambda}_i$):	107.02	39.67	8.37	2.87	0.15
Cumulative percentage of total variance	67.7	92.8	98.1	99.9	1.000

The first principal component explains 67.7% of the total sample variance. The first two principal components, collectively, explain 92.8% of the total sample variance. Consequently, sample variation is summarized very well by two principal components and a reduction in the data from 61 observations on 5 observations to 61 observations on 2 principal components is reasonable.

Given the foregoing component coefficients, the first principal component appears to be essentially a weighted difference between the percent employed by government and the percent total employment. The second principal component appears to be a weighted sum of the two. ■

As we said in our discussion of the population components, the component coefficients $\hat{e}_{ik}$ and the correlations $r_{\hat{y}_i,x_k}$ should both be examined to interpret the principal components. The correlations allow for differences in the variances of the original variables, but only measure the importance of an individual X without regard to the other X's making up the component. We notice in Example 8.3, however, that the correlation coefficients displayed in the table confirm the interpretation provided by the component coefficients.

The Number of Principal Components

There is always the question of how many components to retain. There is no definitive answer to this question. Things to consider include the amount of total sample variance explained, the relative sizes of the eigenvalues (the variances of the sample components), and the subject-matter interpretations of the components. In addition, as we discuss later, a component associated with an eigenvalue near zero and, hence, deemed unimportant, may indicate an unsuspected linear dependency in the data.

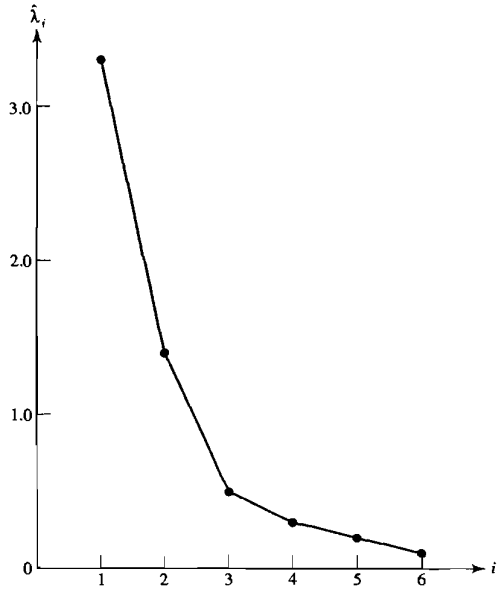

Figure 8.2 A scree plot.

A useful visual aid to determining an appropriate number of principal components is a *scree plot*.[3] With the eigenvalues ordered from largest to smallest, a scree plot is a plot of $\hat{\lambda}_i$ versus i—the magnitude of an eigenvalue versus its number. To determine the appropriate number of components, we look for an elbow (bend) in the scree plot. The number of components is taken to be the point at which the remaining eigenvalues are relatively small and all about the same size. Figure 8.2 shows a scree plot for a situation with six principal components.

An elbow occurs in the plot in Figure 8.2 at about $i = 3$. That is, the eigenvalues after $\hat{\lambda}_2$ are all relatively small and about the same size. In this case, it appears, without any other evidence, that two (or perhaps three) sample principal components effectively summarize the total sample variance.

Example 8.4 (Summarizing sample variability with one sample principal component) In a study of size and shape relationships for painted turtles, Jolicoeur and Mosimann [11] measured carapace length, width, and height. Their data, reproduced in Exercise 6.18, Table 6.9, suggest an analysis in terms of logarithms. (Jolicoeur [10] generally suggests a logarithmic transformation in studies of size-and-shape relationships.) Perform a principal component analysis.

[3] Scree is the rock debris at the bottom of a cliff.

The natural logarithms of the dimensions of 24 male turtles have sample mean vector $\bar{\mathbf{x}}' = [4.725, 4.478, 3.703]$ and covariance matrix

$$\mathbf{S} = 10^{-3}\begin{bmatrix} 11.072 & 8.019 & 8.160 \\ 8.019 & 6.417 & 6.005 \\ 8.160 & 6.005 & 6.773 \end{bmatrix}$$

A principal component analysis (see Panel 8.1 on page 447 for the output from the SAS statistical software package) yields the following summary:

Coefficients for the Principal Components
(Correlation Coefficients in Parentheses)

Variable	$\hat{\mathbf{e}}_1 (r_{\hat{y}_1, x_k})$	$\hat{\mathbf{e}}_2$	$\hat{\mathbf{e}}_3$
ln (length)	.683 (.99)	−.159	−.713
ln (width)	.510 (.97)	−.594	.622
ln (height)	.523 (.97)	.788	.324
Variance ($\hat{\lambda}_i$):	23.30×10^{-3}	$.60 \times 10^{-3}$	$.36 \times 10^{-3}$
Cumulative percentage of total variance	96.1	98.5	100

A scree plot is shown in Figure 8.3. The very distinct elbow in this plot occurs at $i = 2$. There is clearly one dominant principal component.

The first principal component, which explains 96% of the total variance, has an interesting subject-matter interpretation. Since

$$\hat{y}_1 = .683 \ln(\text{length}) + .510 \ln(\text{width}) + .523 \ln(\text{height})$$
$$= \ln[(\text{length})^{.683}(\text{width})^{.510}(\text{height})^{.523}]$$

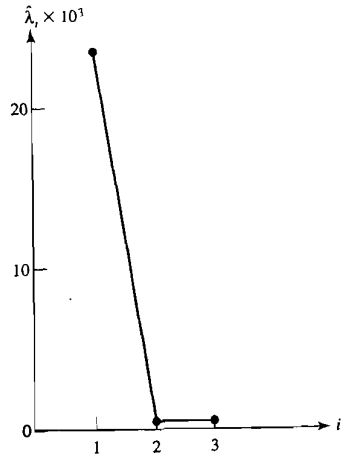

Figure 8.3 A scree plot for the turtle data.

PANEL 8.1 SAS ANALYSIS FOR EXAMPLE 8.4 USING PROC PRINCOMP.

```
title 'Principal Component Analysis';
data turtle;
infile 'E8-4.dat';
input length width height;
x1 = log(length); x2 =log(width); x3 =log(height);
proc princomp cov data = turtle out = result;
var x1 x2 x3;
```

PROGRAM COMMANDS

Principal Components Analysis

24 Observations
3 Variables

OUTPUT

Simple Statistics

	X1	X2	X3
Mean	4.725443647	4.477573765	3.703185794
StD	0.105223590	0.080104466	0.082296771

Covariance Matrix

	X1	X2	X3
X1	0.0110720040	0.0080191419	0.0081596480
X2	0.0080191419	0.0064167255	0.0060052707
X3	0.0081596480	0.0060052707	0.0067727585

Total Variance = 0.024261488

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	0.023303	0.022705	0.960508	0.96051
PRIN2	0.000598	0.000238	0.024661	0.98517
PRIN3	0.000360		0.014832	1.00000

Eigenvectors

	PRIN1	PRIN2	PRIN3
X1	0.683102	−.159479	−.712697
X2	0.510220	−.594012	0.621953
X3	0.522539	0.788490	0.324401

the first principal component may be viewed as the ln (volume) of a box with adjusted dimensions. For instance, the adjusted height is $(\text{height})^{.523}$, which accounts, in some sense, for the rounded shape of the carapace. ■

Interpretation of the Sample Principal Components

The sample principal components have several interpretations. First, suppose the underlying distribution of $\mathbf{X}$ is nearly $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the sample principal components, $\hat{y}_i = \hat{\mathbf{e}}_i'(\mathbf{x} - \bar{\mathbf{x}})$ are realizations of population principal components $Y_i = \mathbf{e}_i'(\mathbf{X} - \boldsymbol{\mu})$, which have an $N_p(\mathbf{0}, \boldsymbol{\Lambda})$ distribution. The diagonal matrix $\boldsymbol{\Lambda}$ has entries $\lambda_1, \lambda_2, \ldots, \lambda_p$ and $(\lambda_i, \mathbf{e}_i)$ are the eigenvalue–eigenvector pairs of $\boldsymbol{\Sigma}$.

Also, from the sample values $\mathbf{x}_j$, we can approximate $\boldsymbol{\mu}$ by $\bar{\mathbf{x}}$ and $\boldsymbol{\Sigma}$ by $\mathbf{S}$. If $\mathbf{S}$ is positive definite, the contour consisting of all $p \times 1$ vectors $\mathbf{x}$ satisfying

$$(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) = c^2 \tag{8-24}$$

estimates the constant density contour $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2$ of the underlying normal density. The approximate contours can be drawn on the scatter plot to indicate the normal distribution that generated the data. The normality assumption is useful for the inference procedures discussed in Section 8.5, but it is not required for the development of the properties of the sample principal components summarized in (8-20).

Even when the normal assumption is suspect and the scatter plot may depart somewhat from an elliptical pattern, we can still extract the eigenvalues from $\mathbf{S}$ and obtain the sample principal components. Geometrically, the data may be plotted as n points in p-space. The data can then be expressed in the new coordinates, which coincide with the axes of the contour of (8-24). Now, (8-24) defines a hyperellipsoid that is centered at $\bar{\mathbf{x}}$ and whose axes are given by the eigenvectors of $\mathbf{S}^{-1}$ or, equivalently, of $\mathbf{S}$. (See Section 2.3 and Result 4.1, with $\mathbf{S}$ in place of $\boldsymbol{\Sigma}$.) The lengths of these hyperellipsoid axes are proportional to $\sqrt{\hat{\lambda}_i}$, $i = 1, 2, \ldots, p$, where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$ are the eigenvalues of $\mathbf{S}$.

Because $\hat{\mathbf{e}}_i$ has length 1, the absolute value of the ith principal component, $|\hat{y}_i| = |\hat{\mathbf{e}}_i'(\mathbf{x} - \bar{\mathbf{x}})|$, gives the length of the projection of the vector $(\mathbf{x} - \bar{\mathbf{x}})$ on the unit vector $\hat{\mathbf{e}}_i$. [See (2-8) and (2-9).] Thus, the sample principal components $\hat{y}_i = \hat{\mathbf{e}}_i'(\mathbf{x} - \bar{\mathbf{x}})$, $i = 1, 2, \ldots, p$, lie along the axes of the hyperellipsoid, and their absolute values are the lengths of the projections of $\mathbf{x} - \bar{\mathbf{x}}$ in the directions of the axes $\hat{\mathbf{e}}_i$. Consequently, the sample principal components can be viewed as the result of translating the origin of the original coordinate system to $\bar{\mathbf{x}}$ and then rotating the coordinate axes until they pass through the scatter in the directions of maximum variance.

The geometrical interpretation of the sample principal components is illustrated in Figure 8.4 for $p = 2$. Figure 8.4(a) shows an ellipse of constant distance, centered at $\bar{\mathbf{x}}$, with $\hat{\lambda}_1 > \hat{\lambda}_2$. The sample principal components are well determined. They lie along the axes of the ellipse in the perpendicular directions of maximum sample variance. Figure 8.4(b) shows a constant distance ellipse, centered at $\bar{\mathbf{x}}$, with $\hat{\lambda}_1 \doteq \hat{\lambda}_2$. If $\hat{\lambda}_1 = \hat{\lambda}_2$, the axes of the ellipse (circle) of constant distance are not uniquely determined and can lie in any two perpendicular directions, including the

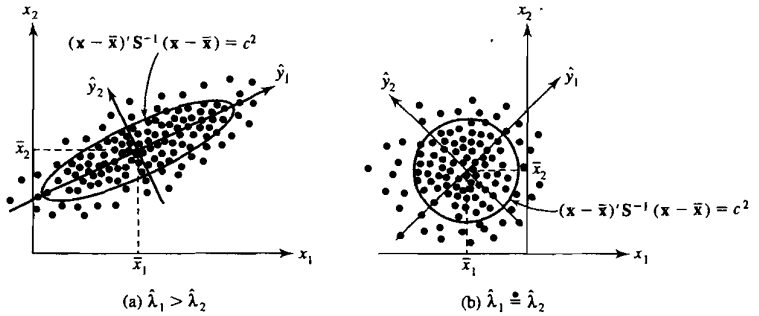

Figure 8.4 Sample principal components and ellipses of constant distance.

directions of the original coordinate axes. Similarly, the sample principal components can lie in any two perpendicular directions, including those of the original coordinate axes. When the contours of constant distance are nearly circular or, equivalently, when the eigenvalues of $\mathbf{S}$ are nearly equal, the sample variation is homogeneous in all directions. It is then not possible to represent the data well in fewer than p dimensions.

If the last few eigenvalues $\hat{\lambda}_i$ are sufficiently small such that the variation in the corresponding $\hat{\mathbf{e}}_i$ directions is negligible, the last few sample principal components can often be ignored, and the data can be adequately approximated by their representations in the space of the retained components. (See Section 8.4.)

Finally, Supplement 8A gives a further result concerning the role of the sample principal components when directly approximating the mean-centered data $\mathbf{x}_j - \bar{\mathbf{x}}$.

Standardizing the Sample Principal Components

Sample principal components are, in general, not invariant with respect to changes in scale. (See Exercises 8.6 and 8.7.) As we mentioned in the treatment of population components, variables measured on different scales or on a common scale with widely differing ranges are often standardized. For the sample, standardization is accomplished by constructing

$$\mathbf{z}_j = \mathbf{D}^{-1/2}(\mathbf{x}_j - \bar{\mathbf{x}}) = \begin{bmatrix} \dfrac{x_{j1} - \bar{x}_1}{\sqrt{s_{11}}} \\ \dfrac{x_{j2} - \bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \dfrac{x_{jp} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} \qquad j = 1, 2, \ldots, n \tag{8-25}$$

The $n \times p$ data matrix of standardized observations

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1' \\ \mathbf{z}_2' \\ \vdots \\ \mathbf{z}_n' \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{x_{11} - \bar{x}_1}{\sqrt{s_{11}}} & \dfrac{x_{12} - \bar{x}_2}{\sqrt{s_{22}}} & \cdots & \dfrac{x_{1p} - \bar{x}_p}{\sqrt{s_{pp}}} \\ \dfrac{x_{21} - \bar{x}_1}{\sqrt{s_{11}}} & \dfrac{x_{22} - \bar{x}_2}{\sqrt{s_{22}}} & \cdots & \dfrac{x_{2p} - \bar{x}_p}{\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{x_{n1} - \bar{x}_1}{\sqrt{s_{11}}} & \dfrac{x_{n2} - \bar{x}_2}{\sqrt{s_{22}}} & \cdots & \dfrac{x_{np} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} \tag{8-26}$$

yields the sample mean vector [see (3-24)]

$$\bar{\mathbf{z}} = \frac{1}{n}(\mathbf{1}'\mathbf{Z})' = \frac{1}{n}\mathbf{Z}'\mathbf{1} = \frac{1}{n}\begin{bmatrix} \sum_{j=1}^{n} \dfrac{x_{j1} - \bar{x}_1}{\sqrt{s_{11}}} \\ \sum_{j=1}^{n} \dfrac{x_{j2} - \bar{x}_2}{\sqrt{s_{22}}} \\ \vdots \\ \sum_{j=1}^{n} \dfrac{x_{jp} - \bar{x}_p}{\sqrt{s_{pp}}} \end{bmatrix} = \mathbf{0} \tag{8-27}$$

and sample covariance matrix [see (3-27)]

$$\mathbf{S}_z = \frac{1}{n-1}\left(\mathbf{Z} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{Z}\right)'\left(\mathbf{Z} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{Z}\right)$$

$$= \frac{1}{n-1}(\mathbf{Z} - \mathbf{1}\bar{\mathbf{z}}')'(\mathbf{Z} - \mathbf{1}\bar{\mathbf{z}}')$$

$$= \frac{1}{n-1}\mathbf{Z}'\mathbf{Z}$$

$$= \frac{1}{n-1}\begin{bmatrix} \dfrac{(n-1)s_{11}}{s_{11}} & \dfrac{(n-1)s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \cdots & \dfrac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} \\ \dfrac{(n-1)s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \dfrac{(n-1)s_{22}}{s_{22}} & \cdots & \dfrac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} & \dfrac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} & \cdots & \dfrac{(n-1)s_{pp}}{s_{pp}} \end{bmatrix} = \mathbf{R} \tag{8-28}$$

The sample principal components of the standardized observations are given by (8-20), with the matrix $\mathbf{R}$ in place of $\mathbf{S}$. Since the observations are already "centered" by construction, there is no need to write the components in the form of (8-21).

If $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ are standardized observations with covariance matrix $\mathbf{R}$, the ith sample principal component is

$$\hat{y}_i = \hat{\mathbf{e}}_i'\mathbf{z} = \hat{e}_{i1}z_1 + \hat{e}_{i2}z_2 + \cdots + \hat{e}_{ip}z_p, \qquad i = 1, 2, \ldots, p$$

where $(\hat{\lambda}_i, \hat{\mathbf{e}}_i)$ is the ith eigenvalue-eigenvector pair of $\mathbf{R}$ with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$. Also,

$$\text{Sample variance } (\hat{y}_i) = \hat{\lambda}_i \qquad i = 1, 2, \ldots, p$$
$$\text{Sample covariance } (\hat{y}_i, \hat{y}_k) = 0 \qquad i \neq k$$

In addition, (8-29)

$$\text{Total (standardized) sample variance} = \operatorname{tr}(\mathbf{R}) = p = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p$$

and

$$r_{\hat{y}_i, z_k} = \hat{e}_{ik}\sqrt{\hat{\lambda}_i}, \qquad i, k = 1, 2, \ldots, p$$

Using (8-29), we see that the proportion of the total sample variance explained by the ith sample principal component is

$$\begin{pmatrix} \text{Proportion of (standardized)} \\ \text{sample variance due to } i\text{th} \\ \text{sample principal component} \end{pmatrix} = \frac{\hat{\lambda}_i}{p} \qquad i = 1, 2, \ldots, p \tag{8-30}$$

A rule of thumb suggests retaining only those components whose variances $\hat{\lambda}_i$ are greater than unity or, equivalently, only those components which, individually, explain at least a proportion $1/p$ of the total variance. This rule does not have a great deal of theoretical support, however, and it should not be applied blindly. As we have mentioned, a scree plot is also useful for selecting the appropriate number of components.

Example 8.5 (Sample principal components from standardized data) The weekly rates of return for five stocks (JP Morgan, Citibank, Wells Fargo, Royal Dutch Shell, and ExxonMobil) listed on the New York Stock Exchange were determined for the period January 2004 through December 2005. The weekly rates of return are defined as (current week closing price—previous week closing price)/(previous week closing price), adjusted for stock splits and dividends. The data are listed in Table 8.4 in the Exercises. The observations in 103 successive weeks appear to be independently distributed, but the rates of return *across* stocks are correlated, because as one expects, stocks tend to move together in response to general economic conditions.

Let $x_1, x_2, \ldots, x_5$ denote observed weekly rates of return for JP Morgan, Citibank, Wells Fargo, Royal Dutch Shell, and ExxonMobil, respectively. Then

$$\bar{\mathbf{x}}' = [.0011, .0007, .0016, .0040, .0040]$$