Chapter 8

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 $X_1, X_2, \dots, X_p$. 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 $\mathbf{\Sigma}$ (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 $\mathbf{\Sigma}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Consider the linear combinations

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

Then, using (2-45), we obtain

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

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

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 $\operatorname{Var}(Y_1) = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1$. It is clear that $\operatorname{Var}(Y_1) = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1$ 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

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

At the ith step,

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

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

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

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

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

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

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

Thus,

$$\sum_{i=1}^{p} \operatorname{Var}(X_i) = \operatorname{tr}(\mathbf{\Sigma}) = \operatorname{tr}(\mathbf{\Lambda}) = \sum_{i=1}^{p} \operatorname{Var}(Y_i) \qquad \blacksquare$$

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

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

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 $\mathbf{\Sigma}$, 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 $\mathbf{\Sigma}$.

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

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

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

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

It may be verified that the eigenvalue–eigenvector pairs are

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

Therefore, the principal components become

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

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,

$$\begin{aligned} \operatorname{Var}(Y_1) &= \operatorname{Var}(.383X_1 - .924X_2) \\ &= (.383)^2 \operatorname{Var}(X_1) + (-.924)^2 \operatorname{Var}(X_2) \\ &\quad + 2(.383)(-.924) \operatorname{Cov}(X_1, X_2) \\ &= .147(1) + .854(5) - .708(-2) \\ &= 5.83 = \lambda_1 \\ \operatorname{Cov}(Y_1, Y_2) &= \operatorname{Cov}(.383X_1 - .924X_2, X_3) \\ &= .383 \operatorname{Cov}(X_1, X_3) - .924 \operatorname{Cov}(X_2, X_3) \\ &= .383(0) - .924(0) = 0 \end{aligned}$$

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 $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We know from (4-7) that the density of $\mathbf{X}$ is constant on the $\boldsymbol{\mu}$ centered ellipsoids

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

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 $\boldsymbol{\Sigma}$. 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 $\mathbf{A} = \boldsymbol{\Sigma}^{-1}$, we can write

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

[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.

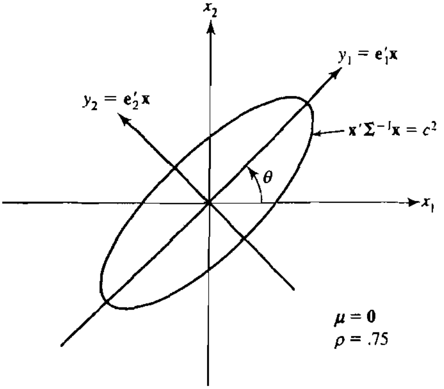

x_2
$y_1 = \mathbf{e}_1'\mathbf{x}$
$y_2 = \mathbf{e}_2'\mathbf{x}$
$\mathbf{x}'\Sigma^{-1}\mathbf{x} = c^2$
θ
x_1
$\mu = 0$
$\rho = .75$

Figure 8.1 The constant density ellipse $\mathbf{x}'\Sigma^{-1}\mathbf{x} = c^2$ 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

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

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

$$\begin{aligned} \lambda_1 &= 100.16, \qquad & \mathbf{e}_1' &= [.040, .999] \\ \lambda_2 &= .84, \qquad & \mathbf{e}_2' &= [.999, -.040] \end{aligned}$$

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

$$\begin{aligned} \lambda_1 &= 1 + \rho = 1.4, \qquad & \mathbf{e}_1' &= [.707, .707] \\ \lambda_2 &= 1 - \rho = .6, \qquad & \mathbf{e}_2' &= [.707, -.707] \end{aligned}$$

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 $\boldsymbol{\Sigma}$ 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

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

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

$$\begin{aligned}
\mathbf{e}_2' &= \left[\frac{1}{\sqrt{1 \times 2}}, \frac{-1}{\sqrt{1 \times 2}}, 0, \ldots, 0\right] \\
\mathbf{e}_3' &= \left[\frac{1}{\sqrt{2 \times 3}}, \frac{1}{\sqrt{2 \times 3}}, \frac{-2}{\sqrt{2 \times 3}}, 0, \ldots, 0\right] \\
&\vdots \\
\mathbf{e}_i' &= \left[\frac{1}{\sqrt{(i-1)i}}, \ldots, \frac{1}{\sqrt{(i-1)i}}, \frac{-(i-1)}{\sqrt{(i-1)i}}, 0, \ldots, 0\right] \\
&\vdots \\
\mathbf{e}_p' &= \left[\frac{1}{\sqrt{(p-1)p}}, \ldots, \frac{1}{\sqrt{(p-1)p}}, \frac{-(p-1)}{\sqrt{(p-1)p}}\right]
\end{aligned}$$

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 $\boldsymbol{\Sigma}$. 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,

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

In addition, (8-20)

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

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:

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

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 $\hat{\mathbf{\Sigma}} = \mathbf{S}_n$, the maximum likelihood estimate of the covariance matrix $\mathbf{\Sigma}$, if the $\mathbf{X}_j$ are normally distributed. (See Result 4.11.) In this case, provided that the eigenvalues of $\mathbf{\Sigma}$ 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 $\hat{\mathbf{\Sigma}}$ because the assumption of normality is not required in this section. Also, $\hat{\mathbf{\Sigma}}$ 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 $\hat{\mathbf{\Sigma}}$ 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 $\hat{\mathbf{\Sigma}}$ give the same sample correlation matrix $\mathbf{R}$, so if the variables are standardized, the choice of $\mathbf{S}$ or $\hat{\mathbf{\Sigma}}$ 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.

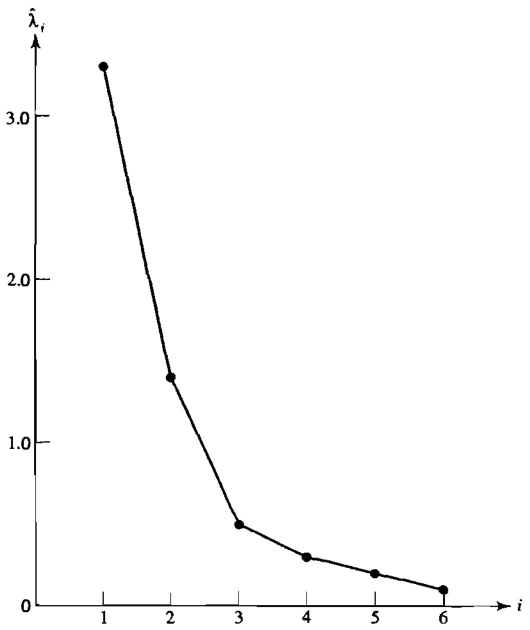


Figure 8.2 A scree plot.

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

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

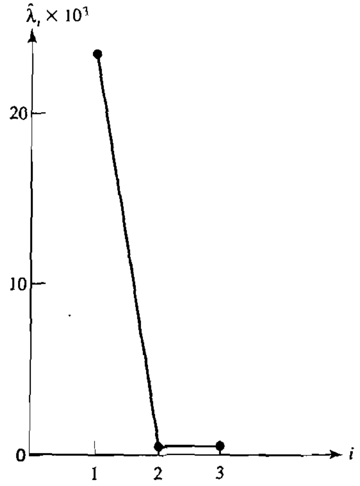


Figure 8.3 A scree plot for the turtle data.

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;                               }  PROGRAM COMMANDS
x1 = log(length); x2 =log(width); x3 =log(height);       }
proc princomp cov data = turtle out = result;            }
var x1 x2 x3;                                            }
```

```
                         Principal Components Analysis

24 Observations                                                          OUTPUT
 3 Variables

                              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 (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 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 $\lambda_1 = \lambda_2$, the axes of the ellipse (circle) of constant distance are not uniquely determined and can lie in any two perpendicular directions, including the

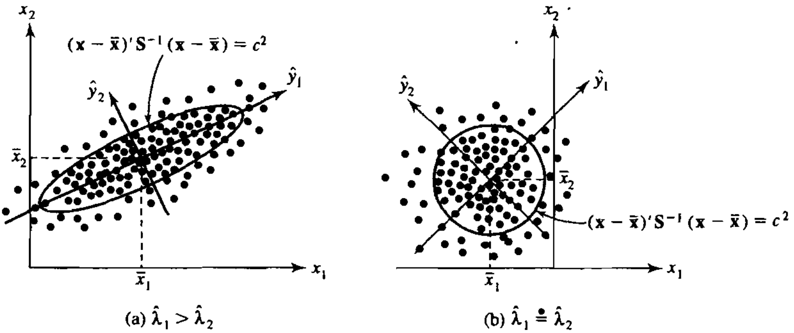


Figure 8.4 Sample principal components and ellipses of constant distance.

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

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

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

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

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

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

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

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

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

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

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

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

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,

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

In addition, (8-29)

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

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

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

and

$$\mathbf{R} = \begin{bmatrix} 1.000 & .632 & .511 & .115 & .155 \\ .632 & 1.000 & .574 & .322 & .213 \\ .511 & .574 & 1.000 & .183 & .146 \\ .115 & .322 & .183 & 1.000 & .683 \\ .155 & .213 & .146 & .683 & 1.000 \end{bmatrix}$$

We note that **R** is the covariance matrix of the standardized observations

$$z_1 = \frac{x_1 - \bar{x}_1}{\sqrt{s_{11}}}, z_2 = \frac{x_2 - \bar{x}_2}{\sqrt{s_{22}}}, \ldots, z_5 = \frac{x_5 - \bar{x}_5}{\sqrt{s_{55}}}$$

The eigenvalues and corresponding normalized eigenvectors of **R**, determined by a computer, are

$$\begin{aligned}
\hat{\lambda}_1 &= 2.437, & \hat{\mathbf{e}}_1' &= [\ .469,\ .532,\ .465,\ .387,\ .361] \\
\hat{\lambda}_2 &= 1.407, & \hat{\mathbf{e}}_2' &= [-.368, -.236, -.315,\ .585,\ .606] \\
\hat{\lambda}_3 &= .501, & \hat{\mathbf{e}}_3' &= [-.604, -.136,\ .772,\ .093, -.109] \\
\hat{\lambda}_4 &= .400, & \hat{\mathbf{e}}_4' &= [\ .363, -.629,\ .289, -.381,\ .493] \\
\hat{\lambda}_5 &= .255, & \hat{\mathbf{e}}_5' &= [\ .384, -.496,\ .071,\ .595, -.498]
\end{aligned}$$

Using the standardized variables, we obtain the first two sample principal components:

$$\begin{aligned}
\hat{y}_1 &= \hat{\mathbf{e}}_1'\mathbf{z} = .469z_1 + .532z_2 + .465z_3 + .387z_4 + .361z_5 \\
\hat{y}_2 &= \hat{\mathbf{e}}_2'\mathbf{z} = -\ .368z_1 - .236z_2 - .315z_3 + .585z_4 + .606z_5
\end{aligned}$$

These components, which account for

$$\left(\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{p}\right)100\% = \left(\frac{2.437 + 1.407}{5}\right)100\% = 77\%$$

of the total (standardized) sample variance, have interesting interpretations. The first component is a roughly equally weighted sum, or "index," of the five stocks. This component might be called a *general stock-market component*, or, simply, a *market component.*

The second component represents a contrast between the banking stocks (JP Morgan, Citibank, Wells Fargo) and the oil stocks (Royal Dutch Shell, ExxonMobil). It might be called an *industry component*. Thus, we see that most of the variation in these stock returns is due to market activity and uncorrelated industry activity. This interpretation of stock price behavior also has been suggested by King [12].

The remaining components are not easy to interpret and, collectively, represent variation that is probably specific to each stock. In any event, they do not explain much of the total sample variance. ■

Example 8.6 (Components from a correlation matrix with a special structure) Geneticists are often concerned with the inheritance of characteristics that can be measured several times during an animal's lifetime. Body weight (in grams) for $n = 150$ female mice were obtained immediately after the birth of their first four litters.[4] The sample mean vector and sample correlation matrix were, respectively,

$$\bar{\mathbf{x}}' = [39.88, 45.08, 48.11, 49.95]$$

and

$$\mathbf{R} = \begin{bmatrix} 1.000 & .7501 & .6329 & .6363 \\ .7501 & 1.000 & .6925 & .7386 \\ .6329 & .6925 & 1.000 & .6625 \\ .6363 & .7386 & .6625 & 1.000 \end{bmatrix}$$

The eigenvalues of this matrix are

$$\hat{\lambda}_1 = 3.085, \quad \hat{\lambda}_2 = .382, \quad \hat{\lambda}_3 = .342, \quad \text{and} \quad \hat{\lambda}_4 = .217$$

We note that the first eigenvalue is nearly equal to $1 + (p - 1)\bar{r} = 1 + (4 - 1)(.6854) = 3.056$, where $\bar{r}$ is the arithmetic average of the off-diagonal elements of $\mathbf{R}$. The remaining eigenvalues are small and about equal, although $\hat{\lambda}_4$ is somewhat smaller than $\hat{\lambda}_2$ and $\hat{\lambda}_3$. Thus, there is some evidence that the corresponding population correlation matrix $\boldsymbol{\rho}$ may be of the "equal-correlation" form of (8-15). This notion is explored further in Example 8.9.

The first principal component

$$\hat{y}_1 = \hat{\mathbf{e}}_1'\mathbf{z} = .49z_1 + .52z_2 + .49z_3 + .50z_4$$

accounts for $100(\hat{\lambda}_1/p)\% = 100(3.058/4)\% = 76\%$ of the total variance. Although the average postbirth weights increase over time, the *variation* in weights is fairly well explained by the first principal component with (nearly) equal coefficients. ■

Comment. An unusually small value for the *last* eigenvalue from either the sample covariance or correlation matrix can indicate an unnoticed linear dependency in the data set. If this occurs, one (or more) of the variables is redundant and should be deleted. Consider a situation where x_1, x_2, and x_3 are subtest scores and the total score x_4 is the sum $x_1 + x_2 + x_3$. Then, although the linear combination $\mathbf{e}'\mathbf{x} = [1, 1, 1, -1]\mathbf{x} = x_1 + x_2 + x_3 - x_4$ is always zero, rounding error in the computation of eigenvalues may lead to a small nonzero value. If the linear expression relating x_4 to (x_1, x_2, x_3) was initially overlooked, the smallest eigenvalue–eigenvector pair should provide a clue to its existence. (See the discussion in Section 3.4, pages 131–133.)

Thus, although "large" eigenvalues and the corresponding eigenvectors are important in a principal component analysis, eigenvalues very close to zero should not be routinely ignored. The eigenvectors associated with these latter eigenvalues may point out linear dependencies in the data set that can cause interpretive and computational problems in a subsequent analysis.

[4] Data courtesy of J. J. Rutledge.

8.4 Graphing the Principal Components

Plots of the principal components can reveal suspect observations, as well as provide checks on the assumption of normality. Since the principal components are linear combinations of the original variables, it is not unreasonable to expect them to be nearly normal. It is often necessary to verify that the first few principal components are approximately normally distributed when they are to be used as the input data for additional analyses.

The last principal components can help pinpoint suspect observations. Each observation can be expressed as a linear combination

$$\begin{aligned}\mathbf{x}_j &= (\mathbf{x}_j'\hat{\mathbf{e}}_1)\hat{\mathbf{e}}_1 + (\mathbf{x}_j'\hat{\mathbf{e}}_2)\hat{\mathbf{e}}_2 + \cdots + (\mathbf{x}_j'\hat{\mathbf{e}}_p)\hat{\mathbf{e}}_p \\ &= \hat{y}_{j1}\hat{\mathbf{e}}_1 + \hat{y}_{j2}\hat{\mathbf{e}}_2 + \cdots + \hat{y}_{jp}\hat{\mathbf{e}}_p\end{aligned}$$

of the complete set of eigenvectors $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_p$ of $\mathbf{S}$. Thus, the magnitudes of the last principal components determine how well the first few fit the observations. That is, $\hat{y}_{j1}\hat{\mathbf{e}}_1 + \hat{y}_{j2}\hat{\mathbf{e}}_2 + \cdots + \hat{y}_{j,q-1}\hat{\mathbf{e}}_{q-1}$ differs from $\mathbf{x}_j$ by $\hat{y}_{jq}\hat{\mathbf{e}}_q + \cdots + \hat{y}_{jp}\hat{\mathbf{e}}_p$, the square of whose length is $\hat{y}_{jq}^2 + \cdots + \hat{y}_{jp}^2$. Suspect observations will often be such that at least one of the coordinates $\hat{y}_{jq}, \ldots, \hat{y}_{jp}$ contributing to this squared length will be large. (See Supplement 8A for more general approximation results.)

The following statements summarize these ideas.

1. To help check the normal assumption, construct scatter diagrams for pairs of the first few principal components. Also, make *Q–Q* plots from the sample values generated by *each* principal component.
2. Construct scatter diagrams and *Q–Q* plots for the last few principal components. These help identify suspect observations.

Example 8.7 (Plotting the principal components for the turtle data) We illustrate the plotting of principal components for the data on male turtles discussed in Example 8.4. The three sample principal components are

$$\begin{aligned}\hat{y}_1 &= .683(x_1 - 4.725) + .510(x_2 - 4.478) + .523(x_3 - 3.703) \\ \hat{y}_2 &= -.159(x_1 - 4.725) - .594(x_2 - 4.478) + .788(x_3 - 3.703) \\ \hat{y}_3 &= -.713(x_1 - 4.725) + .622(x_2 - 4.478) + .324(x_3 - 3.703)\end{aligned}$$

where $x_1 = \ln(\text{length})$, $x_2 = \ln(\text{width})$, and $x_3 = \ln(\text{height})$, respectively.

Figure 8.5 shows the *Q–Q* plot for $\hat{y}_2$ and Figure 8.6 shows the scatter plot of $(\hat{y}_1, \hat{y}_2)$. The observation for the first turtle is circled and lies in the lower right corner of the scatter plot and in the upper right corner of the *Q–Q* plot; it may be suspect. This point should have been checked for recording errors, or the turtle should have been examined for structural anomalies. Apart from the first turtle, the scatter plot appears to be reasonably elliptical. The plots for the other sets of principal components do not indicate any substantial departures from normality. ■

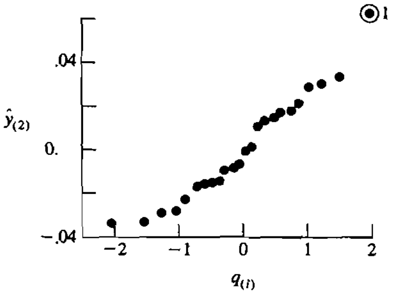


Figure 8.5 A Q–Q plot for the second principal component $\hat{y}_2$ from the data on male turtles.

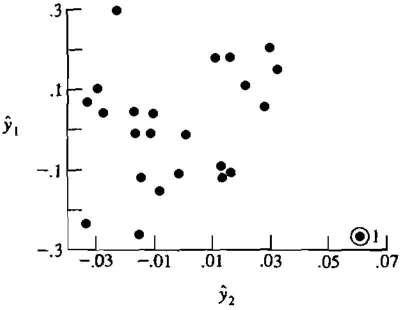


Figure 8.6 Scatter plot of the principal components $\hat{y}_1$ and $\hat{y}_2$ of the data on male turtles.

The diagnostics involving principal components apply equally well to the checking of assumptions for a multivariate multiple regression model. In fact, having fit any model by any method of estimation, it is prudent to consider the

$$\text{Residual vector} = (\text{observation vector}) - \begin{pmatrix}\text{vector of predicted}\\ \text{(estimated) values}\end{pmatrix}$$

or

$$\underset{(p\times1)}{\hat{\boldsymbol{\varepsilon}}_j} = \underset{(p\times1)}{\mathbf{y}_j} - \underset{(p\times1)}{\hat{\boldsymbol{\beta}}'\mathbf{z}_j} \qquad j = 1,2,\ldots,n \tag{8-31}$$

for the multivariate linear model. Principal components, derived from the covariance matrix of the residuals,

$$\frac{1}{n-p}\sum_{j=1}^{n}(\hat{\boldsymbol{\varepsilon}}_j - \bar{\hat{\boldsymbol{\varepsilon}}}_j)(\hat{\boldsymbol{\varepsilon}}_j - \bar{\hat{\boldsymbol{\varepsilon}}}_j)' \tag{8-32}$$

can be scrutinized in the same manner as those determined from a random sample. You should be aware that there *are* linear dependencies among the residuals from a linear regression analysis, so the last eigenvalues will be zero, within rounding error.

8.5 Large Sample Inferences

We have seen that the eigenvalues and eigenvectors of the covariance (correlation) matrix are the essence of a principal component analysis. The eigenvectors determine the directions of maximum variability, and the eigenvalues specify the variances. When the first few eigenvalues are much larger than the rest, most of the total variance can be "explained" in fewer than p dimensions.

In practice, decisions regarding the quality of the principal component approximation must be made on the basis of the eigenvalue–eigenvector pairs $(\hat{\lambda}_i, \hat{\mathbf{e}}_i)$ extracted from $\mathbf{S}$ or $\mathbf{R}$. Because of sampling variation, these eigenvalues and eigenvectors will differ from their underlying population counterparts. The sampling distributions of $\hat{\lambda}_i$ and $\hat{\mathbf{e}}_i$ are difficult to derive and beyond the scope of this book. If you are interested, you can find some of these derivations for multivariate normal populations in [1], [2], and [5]. We shall simply summarize the pertinent large sample results.

Large Sample Properties of $\hat{\lambda}_i$ and $\hat{\mathbf{e}}_i$

Currently available results concerning large sample confidence intervals for $\hat{\lambda}_i$ and $\hat{\mathbf{e}}_i$ assume that the observations $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are a random sample from a normal population. It must also be assumed that the (unknown) eigenvalues of $\mathbf{\Sigma}$ are distinct and positive, so that $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$. The one exception is the case where the number of equal eigenvalues is known. Usually the conclusions for distinct eigenvalues are applied, unless there is a strong reason to believe that $\mathbf{\Sigma}$ has a special structure that yields equal eigenvalues. Even when the normal assumption is violated, the confidence intervals obtained in this manner still provide some indication of the uncertainty in $\hat{\lambda}_i$ and $\hat{\mathbf{e}}_i$.

Anderson [2] and Girshick [5] have established the following large sample distribution theory for the eigenvalues $\hat{\boldsymbol{\lambda}}' = [\hat{\lambda}_1, \ldots, \hat{\lambda}_p]$ and eigenvectors $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_p$ of $\mathbf{S}$:

1. Let Λ be the diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_p$ of $\mathbf{\Sigma}$, then $\sqrt{n}\,(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})$ is approximately $N_p(\mathbf{0}, 2\Lambda^2)$.
2. Let
$$\mathbf{E}_i = \lambda_i \sum_{\substack{k=1 \\ k \neq i}}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \mathbf{e}_k \mathbf{e}_k'$$
then $\sqrt{n}\,(\hat{\mathbf{e}}_i - \mathbf{e}_i)$ is approximately $N_p(\mathbf{0}, \mathbf{E}_i)$.
3. Each $\hat{\lambda}_i$ is distributed independently of the elements of the associated $\hat{\mathbf{e}}_i$.

Result 1 implies that, for n large, the $\hat{\lambda}_i$ are independently distributed. Moreover, $\hat{\lambda}_i$ has an approximate $N(\lambda_i, 2\lambda_i^2/n)$ distribution. Using this normal distribution, we obtain $P[|\hat{\lambda}_i - \lambda_i| \le z(\alpha/2)\lambda_i\sqrt{2/n}] = 1 - \alpha$. A large sample $100(1 - \alpha)\%$ confidence interval for λ_i is thus provided by

$$\frac{\hat{\lambda}_i}{(1 + z(\alpha/2)\sqrt{2/n})} \le \lambda_i \le \frac{\hat{\lambda}_i}{(1 - z(\alpha/2)\sqrt{2/n})} \tag{8-33}$$

where $z(\alpha/2)$ is the upper $100(\alpha/2)$th percentile of a standard normal distribution. Bonferroni-type simultaneous $100(1 - \alpha)\%$ intervals for m λ_i's are obtained by replacing $z(\alpha/2)$ with $z(\alpha/2m)$. (See Section 5.4.)

Result 2 implies that the $\hat{\mathbf{e}}_i$'s are normally distributed about the corresponding $\mathbf{e}_i$'s for large samples. The elements of each $\hat{\mathbf{e}}_i$ are correlated, and the correlation depends to a large extent on the separation of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ (which is unknown) and the sample size n. Approximate standard errors for the coefficients $\hat{e}_{ik}$ are given by the square roots of the diagonal elements of $(1/n)\hat{\mathbf{E}}_i$ where $\hat{\mathbf{E}}_i$ is derived from $\mathbf{E}_i$ by substituting $\hat{\lambda}_i$'s for the λ_i's and $\hat{\mathbf{e}}_i$'s for the $\mathbf{e}_i$'s.

Example 8.8 (Constructing a confidence interval for λ_1) We shall obtain a 95% confidence interval for λ_1, the variance of the first population principal component, using the stock price data listed in Table 8.4 in the Exercises.

Assume that the stock rates of return represent independent drawings from an $N_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ population, where $\boldsymbol{\Sigma}$ is positive definite with distinct eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_5 > 0$. Since $n = 103$ is large, we can use (8-33) with $i = 1$ to construct a 95% confidence interval for λ_1. From Exercise 8.10, $\hat{\lambda}_1 = .0014$ and in addition, $z(.025) = 1.96$. Therefore, with 95% confidence,

$$\frac{.0014}{\left(1 + 1.96\sqrt{\frac{2}{103}}\right)} \leq \lambda_1 \leq \frac{.0014}{\left(1 - 1.96\sqrt{\frac{2}{103}}\right)} \quad \text{or} \quad .0011 \leq \lambda_1 \leq .0019 \qquad \blacksquare$$

Whenever an eigenvalue is large, such as 100 or even 1000, the intervals generated by (8-33) can be quite wide, for reasonable confidence levels, even though n is fairly large. In general, the confidence interval gets wider at the same rate that $\hat{\lambda}_i$ gets larger. Consequently, some care must be exercised in dropping or retaining principal components based on an examination of the $\hat{\lambda}_i$'s.

Testing for the Equal Correlation Structure

The special correlation structure $\text{Cov}(X_i, X_k) = \sqrt{\sigma_{ii}\sigma_{kk}}\,\rho$, or $\text{Corr}(X_i, X_k) = \rho$, all $i \neq k$, is one important structure in which the eigenvalues of $\boldsymbol{\Sigma}$ are not distinct and the previous results do not apply.

To test for this structure, let

$$H_0\colon \boldsymbol{\rho} = \underset{(p\times p)}{\boldsymbol{\rho}_0} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

and

$$H_1\colon \boldsymbol{\rho} \neq \boldsymbol{\rho}_0$$

A test of H_0 versus H_1 may be based on a likelihood ratio statistic, but Lawley [14] has demonstrated that an equivalent test procedure can be constructed from the off-diagonal elements of $\mathbf{R}$.

Lawley's procedure requires the quantities

$$\bar{r}_k = \frac{1}{p-1} \sum_{\substack{i=1 \\ i \neq k}}^{p} r_{ik} \quad k = 1, 2, \ldots, p; \quad \bar{r} = \frac{2}{p(p-1)} \sum\sum_{i<k} r_{ik}$$

$$\hat{\gamma} = \frac{(p-1)^2[1-(1-\bar{r})^2]}{p-(p-2)(1-\bar{r})^2} \tag{8-34}$$

It is evident that $\bar{r}_k$ is the average of the off-diagonal elements in the kth column (or row) of $\mathbf{R}$ and $\bar{r}$ is the overall average of the off-diagonal elements.

The large sample approximate α-level test is to reject H_0 in favor of H_1 if

$$T = \frac{(n-1)}{(1-\bar{r})^2}\left[\sum\sum_{i<k}(r_{ik}-\bar{r})^2 - \hat{\gamma}\sum_{k=1}^{p}(\bar{r}_k-\bar{r})^2\right] > \chi^2_{(p+1)(p-2)/2}(\alpha) \tag{8-35}$$

where $\chi^2_{(p+1)(p-2)/2}(\alpha)$ is the upper (100α)th percentile of a chi-square distribution with $(p+1)(p-2)/2$ d.f.

Example 8.9 (Testing for equicorrelation structure) From Example 8.6, the sample correlation matrix constructed from the $n = 150$ post-birth weights of female mice is

$$\mathbf{R} = \begin{bmatrix} 1.0 & .7501 & .6329 & .6363 \\ .7501 & 1.0 & .6925 & .7386 \\ .6329 & .6925 & 1.0 & .6625 \\ .6363 & .7386 & .6625 & 1.0 \end{bmatrix}$$

We shall use this correlation matrix to illustrate the large sample test in (8-35).

Here $p = 4$, and we set

$$H_0: \boldsymbol{\rho} = \boldsymbol{\rho}_0 = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

$$H_1: \boldsymbol{\rho} \neq \boldsymbol{\rho}_0$$

Using (8-34) and (8-35), we obtain

$$\bar{r}_1 = \frac{1}{3}(.7501 + .6329 + .6363) = .6731, \quad \bar{r}_2 = .7271,$$

$$\bar{r}_3 = .6626, \quad \bar{r}_4 = .6791$$

$$\bar{r} = \frac{2}{4(3)}(.7501 + .6329 + .6363 + .6925 + .7386 + .6625) = .6855$$

$$\sum\sum_{i<k}(r_{ik}-\bar{r})^2 = (.7501 - .6855)^2 + (.6329 - .6855)^2 + \cdots + (.6625 - .6855)^2 = .01277$$

$$\sum_{k=1}^{4} (\bar{r}_k - \bar{r})^2 = (.6731 - .6855)^2 + \cdots + (.6791 - .6855)^2 = .00245$$

$$\hat{\gamma} = \frac{(4-1)^2[1-(1-.6855)^2]}{4-(4-2)(1-.6855)^2} = 2.1329$$

and

$$T = \frac{(150-1)}{(1-.6855)^2}[.01277 - (2.1329)(.00245)] = 11.4$$

Since $(p+1)(p-2)/2 = 5(2)/2 = 5$, the 5% critical value for the test in (8-35) is $\chi_5^2(.05) = 11.07$. The value of our test statistic is approximately equal to the large sample 5% critical point, so the evidence against H_0 (equal correlations) is strong, but not overwhelming.

As we saw in Example 8.6, the smallest eigenvalues $\hat{\lambda}_2$, $\hat{\lambda}_3$, and $\hat{\lambda}_4$ are slightly different, with $\hat{\lambda}_4$ being somewhat smaller than the other two. Consequently, with the large sample size in this problem, small differences from the equal correlation structure show up as statistically significant. ■

Assuming a multivariate normal population, a large sample test that all variables are independent (all the off-diagonal elements of $\mathbf{\Sigma}$ are zero) is contained in Exercise 8.9.

8.6 Monitoring Quality with Principal Components

In Section 5.6, we introduced multivariate control charts, including the quality ellipse and the T^2 chart. Today, with electronic and other automated methods of data collection, it is not uncommon for data to be collected on 10 or 20 process variables. Major chemical and drug companies report measuring over 100 process variables, including temperature, pressure, concentration, and weight, at various positions along the production process. Even with 10 variables to monitor, there are 45 pairs for which to create quality ellipses. Clearly, another approach is required to both visually display important quantities and still have the sensitivity to detect special causes of variation.

Checking a Given Set of Measurements for Stability

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample from a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. We consider the first two sample principal components, $\hat{y}_{j1} = \hat{\mathbf{e}}_1'(\mathbf{x}_j - \bar{\mathbf{x}})$ and $\hat{y}_{j2} = \hat{\mathbf{e}}_2'(\mathbf{x}_j - \bar{\mathbf{x}})$. Additional principal components could be considered, but two are easier to inspect visually and, of any two components, the first two explain the largest cumulative proportion of the total sample variance.

If a process is stable over time, so that the measured characteristics are influenced only by variations in common causes, then the values of the first two principal components should be stable. Conversely, if the principal components remain stable over time, the common effects that influence the process are likely to remain constant. To monitor quality using principal components, we consider a two-part procedure. The first part of the procedure is to construct an ellipse format chart for the pairs of values $(\hat{y}_{j1}, \hat{y}_{j2})$ for $j = 1, 2, \ldots, n$.

By (8-20), the sample variance of the first principal component $\hat{y}_1$ is given by the largest eigenvalue $\hat{\lambda}_1$, and the sample variance of the second principal component $\hat{y}_2$ is the second-largest eigenvalue $\hat{\lambda}_2$. The two sample components are uncorrelated, so the quality ellipse for n large (see Section 5.6) reduces to the collection of pairs of possible values $(\hat{y}_1, \hat{y}_2)$ such that

$$\frac{\hat{y}_1^2}{\hat{\lambda}_1} + \frac{\hat{y}_2^2}{\hat{\lambda}_2} \le \chi_2^2(\alpha) \tag{8-36}$$

Example 8.10 (An ellipse format chart based on the first two principal components) Refer to the police department overtime data given in Table 5.8. Table 8.1 contains the five normalized eigenvectors and eigenvalues of the sample covariance matrix **S**.

The first two sample components explain 82% of the total variance. The sample values for all five components are displayed in Table 8.2.

Table 8.1 Eigenvectors and Eigenvalues of Sample Covariance Matrix for Police Department Data

Variable	$\hat{\mathbf{e}}_1$	$\hat{\mathbf{e}}_2$	$\hat{\mathbf{e}}_3$	$\hat{\mathbf{e}}_4$	$\hat{\mathbf{e}}_5$
Appearances overtime (x_1)	.046	−.048	.629	−.643	.432
Extraordinary event (x_2)	.039	.985	−.077	−.151	−.007
Holdover hours (x_3)	−.658	.107	.582	.250	−.392
COA hours (x_4)	.734	.069	.503	.397	−.213
Meeting hours (x_5)	−.155	.107	.081	.586	.784
$\hat{\lambda}_i$	2,770,226	1,429,206	628,129	221,138	99,824

Table 8.2 Values of the Principal Components for the Police Department Data

Period	$\hat{y}_{j1}$	$\hat{y}_{j2}$	$\hat{y}_{j3}$	$\hat{y}_{j4}$	$\hat{y}_{j5}$
1	2044.9	588.2	425.8	−189.1	−209.8
2	−2143.7	−686.2	883.6	−565.9	−441.5
3	−177.8	−464.6	707.5	736.3	38.2
4	−2186.2	450.5	−184.0	443.7	−325.3
5	−878.6	−545.7	115.7	296.4	437.5
6	563.2	−1045.4	281.2	620.5	142.7
7	403.1	66.8	340.6	−135.5	521.2
8	−1988.9	−801.8	−1437.3	−148.8	61.6
9	132.8	563.7	125.3	68.2	611.5
10	−2787.3	−213.4	7.8	169.4	−202.3
11	283.4	3936.9	−0.9	276.2	−159.6
12	761.6	256.0	−2153.6	−418.8	28.2
13	−498.3	244.7	966.5	−1142.3	182.6
14	2366.2	−1193.7	−165.5	270.6	−344.9
15	1917.8	−782.0	−82.9	−196.8	−89.9
16	2187.7	−373.8	170.1	−84.1	−250.2

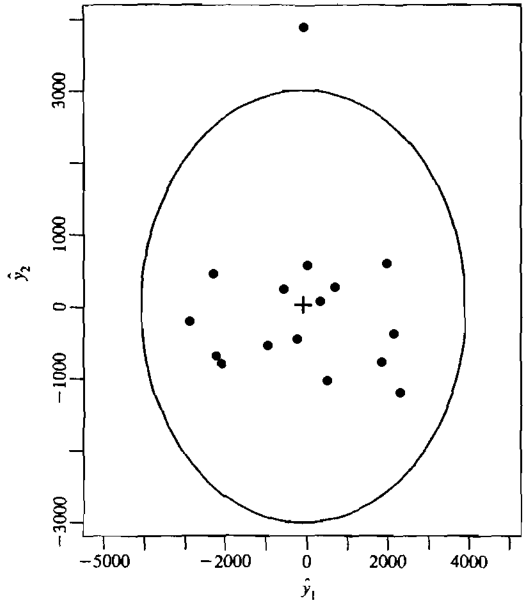


Figure 8.7 The 95% control ellipse based on the first two principal components of overtime hours.

Let us construct a 95% ellipse format chart using the first two sample principal components and plot the 16 pairs of component values in Table 8.2.

Although $n = 16$ is not large, we use $\chi_2^2(.05) = 5.99$, and the ellipse becomes

$$\frac{\hat{y}_1^2}{\hat{\lambda}_1} + \frac{\hat{y}_2^2}{\hat{\lambda}_2} \leq 5.99$$

This ellipse centered at (0, 0), is shown in Figure 8.7, along with the data.

One point is out of control, because the second principal component for this point has a large value. Scanning Table 8.2, we see that this is the value 3936.9 for period 11. According to the entries of $\hat{\mathbf{e}}_2$ in Table 8.1, the second principal component is essentially extraordinary event overtime hours. The principal component approach has led us to the same conclusion we came to in Example 5.9. ■

In the event that special causes are likely to produce shocks to the system, the second part of our two-part procedure—that is, a second chart—is required. This chart is created from the information in the principal components not involved in the ellipse format chart.

Consider the deviation vector $\mathbf{X} - \boldsymbol{\mu}$, and assume that $\mathbf{X}$ is distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Even without the normal assumption, $\mathbf{X}_j - \boldsymbol{\mu}$ can be expressed as the sum of its projections on the eigenvectors of $\boldsymbol{\Sigma}$

$$\begin{aligned}\mathbf{X} - \boldsymbol{\mu} = (\mathbf{X} - \boldsymbol{\mu})'\mathbf{e}_1\mathbf{e}_1 + (\mathbf{X} - \boldsymbol{\mu})'\mathbf{e}_2\mathbf{e}_2 \\ + (\mathbf{X} - \boldsymbol{\mu})'\mathbf{e}_3\mathbf{e}_3 + \cdots + (\mathbf{X} - \boldsymbol{\mu})'\mathbf{e}_p\mathbf{e}_p\end{aligned}$$

or

$$\mathbf{X} - \boldsymbol{\mu} = Y_1\mathbf{e}_1 + Y_2\mathbf{e}_2 + Y_3\mathbf{e}_3 + \cdots + Y_p\mathbf{e}_p \tag{8-37}$$

where $Y_i = (\mathbf{X} - \boldsymbol{\mu})'\mathbf{e}_i$ is the population ith principal component centered to have mean 0. The approximation to $\mathbf{X} - \boldsymbol{\mu}$ by the first two principal components has the form $Y_1\mathbf{e}_1 + Y_2\mathbf{e}_2$. This leaves an unexplained component of

$$\mathbf{X} - \boldsymbol{\mu} - Y_1\mathbf{e}_1 - Y_2\mathbf{e}_2$$

Let $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p]$ be the orthogonal matrix whose columns are the eigenvectors of $\boldsymbol{\Sigma}$. The orthogonal transformation of the unexplained part,

$$\mathbf{E}'(\mathbf{X} - \boldsymbol{\mu} - Y_1\mathbf{e}_1 - Y_2\mathbf{e}_2) = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_p \end{bmatrix} - \begin{bmatrix} Y_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ Y_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ Y_3 \\ \vdots \\ Y_p \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \mathbf{Y}_{(2)} \end{bmatrix}$$

so the last $p - 2$ principal components are obtained as an orthogonal transformation of the approximation errors. Rather than base the T^2 chart on the approximation errors, we can, equivalently, base it on these last principal components. Recall that

$$\text{Var}(Y_i) = \lambda_i \quad \text{for} \quad i = 1, 2, \ldots, p$$

and $\text{Cov}(Y_i, Y_k) = 0$ for $i \neq k$. Consequently, the statistic $\mathbf{Y}'_{(2)}\boldsymbol{\Sigma}^{-1}_{\mathbf{Y}_{(2)},\mathbf{Y}_{(2)}}\mathbf{Y}_{(2)}$, based on the last $p - 2$ population principal components, becomes

$$\frac{Y_3^2}{\lambda_3} + \frac{Y_4^2}{\lambda_4} + \cdots + \frac{Y_p^2}{\lambda_p} \tag{8-38}$$

This is just the sum of the squares of $p - 2$ independent standard normal variables, $\lambda_k^{-1/2}Y_k$, and so has a chi-square distribution with $p - 2$ degrees of freedom.

In terms of the sample data, the principal components and eigenvalues must be estimated. Because the coefficients of the linear combinations $\hat{\mathbf{e}}_i$ are also estimates, the principal components do not have a normal distribution even when the population is normal. However, it is customary to create a T^2-chart based on the statistic

$$T_j^2 = \frac{\hat{y}_{j3}^2}{\hat{\lambda}_3} + \frac{\hat{y}_{j4}^2}{\hat{\lambda}_4} + \cdots + \frac{\hat{y}_{jp}^2}{\hat{\lambda}_p}$$

which involves the estimated eigenvalues and vectors. Further, it is usual to appeal to the large sample approximation described by (8-38) and set the upper control limit of the T^2-chart as $\text{UCL} = c^2 = \chi^2_{p-2}(\alpha)$.

This T^2-statistic is based on high-dimensional data. For example, when $p = 20$ variables are measured, it uses the information in the 18-dimensional space perpendicular to the first two eigenvectors $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$. Still, this T^2 based on the unexplained variation in the original observations is reported as highly effective in picking up special causes of variation.

Example 8.11 (A T^2-chart for the unexplained [orthogonal] overtime hours) Consider the quality control analysis of the police department overtime hours in Example 8.10. The first part of the quality monitoring procedure, the quality ellipse based on the first two principal components, was shown in Figure 8.7. To illustrate the second step of the two-step monitoring procedure, we create the chart for the other principal components.

Since $p = 5$, this chart is based on $5 - 2 = 3$ dimensions, and the upper control limit is $\chi_3^2(.05) = 7.81$. Using the eigenvalues and the values of the principal components, given in Example 8.10, we plot the time sequence of values

$$T_j^2 = \frac{\hat{y}_{j3}^2}{\hat{\lambda}_3} + \frac{\hat{y}_{j4}^2}{\hat{\lambda}_4} + \frac{\hat{y}_{j5}^2}{\hat{\lambda}_5}$$

where the first value is $T^2 = .891$ and so on. The T^2-chart is shown in Figure 8.8.

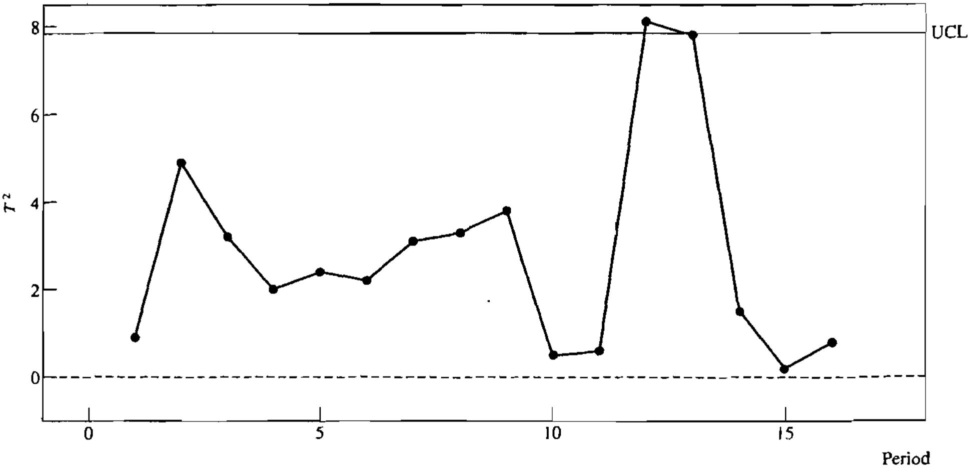


Figure 8.8 A T^2-chart based on the last three principal components of overtime hours.

Since points 12 and 13 exceed or are near the upper control limit, something has happened during these periods. We note that they are just beyond the period in which the extraordinary event overtime hours peaked.

From Table 8.2, $\hat{y}_{3j}$ is large in period 12, and from Table 8.1, the large coefficients in $\mathbf{e}_3$ belong to legal appearances, holdover, and COA hours. Was there some adjusting of these other categories following the period extraordinary hours peaked? ■

Controlling Future Values

Previously, we considered checking whether a given series of multivariate observations was stable by considering separately the first two principal components and then the last $p - 2$. Because the chi-square distribution was used to approximate the UCL of the T^2-chart and the critical distance for the ellipse format chart, no further modifications are necessary for monitoring future values.

Example 8.12 (Control ellipse for future principal components) In Example 8.10, we determined that case 11 was out of control. We drop this point and recalculate the eigenvalues and eigenvectors based on the covariance of the remaining 15 observations. The results are shown in Table 8.3.

Table 8.3 Eigenvectors and Eigenvalues from the 15 Stable Observations

	$\hat{e}_1$	$\hat{e}_2$	$\hat{e}_3$	$\hat{e}_4$	$\hat{e}_5$
Appearances overtime (x_1)	.049	.629	.304	.479	.530
Extraordinary event (x_2)	.007	−.078	.939	−.260	−.212
Holdover hours (x_3)	−.662	.582	−.089	−.158	−.437
COA hours (x_4)	.731	.503	−.123	−.336	−.291
Meeting hours (x_5)	−.159	.081	−.058	−.752	.632
$\hat{\lambda}_i$	2,964,749.9	672,995.1	396,596.5	194,401.0	92,760.3

The principal components have changed. The component consisting primarily of extraordinary event overtime is now the third principal component and is not included in the chart of the first two. Because our initial sample size is only 16, dropping a single case can make a substantial difference. Usually, at least 50 or more observations are needed, from stable operation of the process, in order to set future limits.

Figure 8.9 gives the 99% prediction (8-36) ellipse for future pairs of values for the new first two principal components of overtime. The 15 stable pairs of principal components are also shown. ■

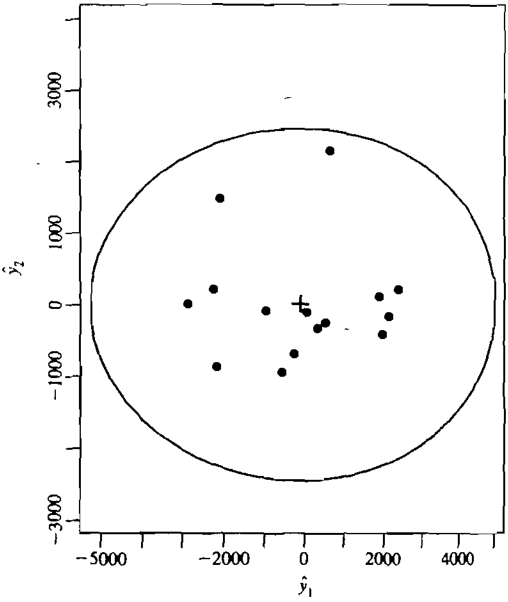


Figure 8.9 A 99% ellipse format chart for the first two principal components of future values of overtime.

In some applications of multivariate control in the chemical and pharmaceutical industries, more than 100 variables are monitored simultaneously. These include numerous process variables as well as quality variables. Typically, the space orthogonal to the first few principal components has a dimension greater than 100 and some of the eigenvalues are very small. An alternative approach (see [13]) to constructing a control chart, that avoids the difficulty caused by dividing a small squared principal component by a very small eigenvalue, has been successfully applied. To implement this approach, we proceed as follows.

For each stable observation, take the sum of squares of its unexplained component

$$d_{Uj}^2 = (\mathbf{x}_j - \bar{\mathbf{x}} - \hat{y}_{j1}\hat{\mathbf{e}}_1 - \hat{y}_{j2}\mathbf{e}_2)'(\mathbf{x}_j - \bar{\mathbf{x}} - \hat{y}_{j1}\hat{\mathbf{e}}_1 - \hat{y}_{j2}\hat{\mathbf{e}}_2)$$

Note that, by inserting $\hat{\mathbf{E}}\hat{\mathbf{E}}' = \mathbf{I}$, we also have

$$d_{Uj}^2 = (\mathbf{x}_j - \bar{\mathbf{x}} - \hat{y}_{j1}\hat{\mathbf{e}}_1 - \hat{y}_{j2}\hat{\mathbf{e}}_2)'\hat{\mathbf{E}}\hat{\mathbf{E}}'(\mathbf{x}_j - \bar{\mathbf{x}} - \hat{y}_{j1}\hat{\mathbf{e}}_1 - \hat{y}_{j2}\hat{\mathbf{e}}_2) = \sum_{k=3}^{p} \hat{y}_{jk}^2$$

which is just the sum of squares of the neglected principal components.

Using either form, the d_{Uj}^2 are plotted versus j to create a control chart. The lower limit of the chart is 0 and the upper limit is set by approximating the distribution of d_{Uj}^2 as the distribution of a constant c times a chi-square random variable with ν degrees of freedom.

For the chi-square approximation, the constant c and degrees of freedom ν are chosen to match the sample mean and variance of the d_{Uj}^2, $j = 1, 2, \ldots, n$. In particular, we set

$$\overline{d_U^2} = \frac{1}{n}\sum_{j=1}^{n} d_{Uj}^2 = c\,\nu$$

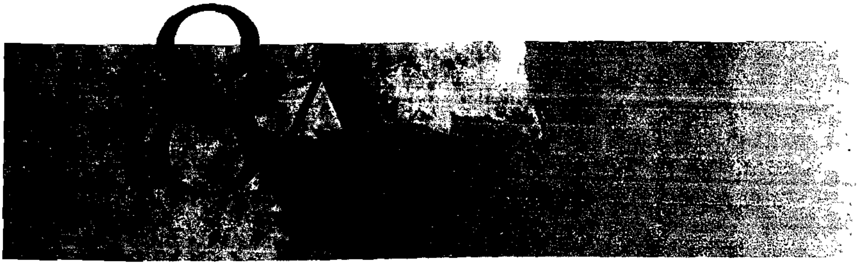
$$s_{d^2}^2 = \frac{1}{n-1}\sum_{j=1}^{n} (d_{Uj}^2 - \overline{d_U^2})^2 = 2c^2\nu$$

and determine

$$c = \frac{s_{d^2}^2}{2\overline{d_U^2}} \quad \text{and} \quad \nu = 2\frac{(\overline{d_U^2})^2}{s_{d^2}^2}$$

The upper control limit is then $c\chi_\nu^2(\alpha)$, where $\alpha = .05$ or $.01$.

Supplement 8A



The Geometry of the Sample Principal Component Approximation

In this supplement, we shall present interpretations for approximations to the data based on the first r sample principal components. The interpretations of both the p-dimensional scatter plot and the n-dimensional representation rely on the algebraic result that follows. We consider approximations of the form $\underset{(n\times p)}{\mathbf{A}} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n]'$ to the mean corrected data matrix

$$[\mathbf{x}_1 - \bar{\mathbf{x}}, \mathbf{x}_2 - \bar{\mathbf{x}}, \dots, \mathbf{x}_n - \bar{\mathbf{x}}]'$$

The error of approximation is quantified as the sum of the np squared errors

$$\sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}} - \mathbf{a}_j)'(\mathbf{x}_j - \bar{\mathbf{x}} - \mathbf{a}_j) = \sum_{j=1}^{n}\sum_{i=1}^{p} (x_{ji} - \bar{x}_i - a_{ji})^2 \qquad \text{(8A-1)}$$

Result 8A.1 Let $\underset{(n\times p)}{\mathbf{A}}$ be any matrix with rank(A) $\leq r < \min(p, n)$. Let $\hat{\mathbf{E}}_r = [\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \dots, \hat{\mathbf{e}}_r]$, where $\hat{\mathbf{e}}_i$ is the ith eigenvector of $\mathbf{S}$. The error of approximation sum of squares in (8A-1) is minimized by the choice

$$\hat{\mathbf{A}} = \begin{bmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}})' \\ (\mathbf{x}_2 - \bar{\mathbf{x}})' \\ \vdots \\ (\mathbf{x}_n - \bar{\mathbf{x}})' \end{bmatrix} \hat{\mathbf{E}}_r \hat{\mathbf{E}}_r' = [\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \dots, \hat{\mathbf{y}}_r]\hat{\mathbf{E}}_r'$$

so the jth column of its transpose $\hat{\mathbf{A}}'$ is

$$\hat{\mathbf{a}}_j = \hat{y}_{j1}\hat{\mathbf{e}}_1 + \hat{y}_{j2}\hat{\mathbf{e}}_2 + \cdots + \hat{y}_{jr}\hat{\mathbf{e}}_r$$

where

$$[\hat{y}_{j1}, \hat{y}_{j2}, \ldots, \hat{y}_{jr}]' = [\hat{\mathbf{e}}_1'(\mathbf{x}_j - \bar{\mathbf{x}}), \hat{\mathbf{e}}_2'(\mathbf{x}_j - \bar{\mathbf{x}}), \ldots, \hat{\mathbf{e}}_r'(\mathbf{x}_j - \bar{\mathbf{x}})]'$$

are the values of the first r sample principal components for the jth unit. Moreover,

$$\sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}} - \hat{\mathbf{a}}_j)'(\mathbf{x}_j - \bar{\mathbf{x}} - \hat{\mathbf{a}}_j) = (n-1)(\hat{\lambda}_{r+1} + \cdots + \hat{\lambda}_p)$$

where $\hat{\lambda}_{r+1} \geq \cdots \geq \hat{\lambda}_p$ are the smallest eigenvalues of $\mathbf{S}$.

Proof. Consider first any $\mathbf{A}$ whose transpose $\mathbf{A}'$ has columns $\mathbf{a}_j$ that are a linear combination of a *fixed* set of r perpendicular vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$, so that $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r]$ satisfies $\mathbf{U}'\mathbf{U} = \mathbf{I}$. For fixed $\mathbf{U}$, $\mathbf{x}_j - \bar{\mathbf{x}}$ is best approximated by its projection on the space spanned by $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$ (see Result 2A.3), or

$$(\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{u}_1\mathbf{u}_1 + (\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{u}_2\mathbf{u}_2 + \cdots + (\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{u}_r\mathbf{u}_r$$

$$= [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r]\begin{bmatrix} \mathbf{u}_1'(\mathbf{x}_j - \bar{\mathbf{x}}) \\ \mathbf{u}_2'(\mathbf{x}_j - \bar{\mathbf{x}}) \\ \vdots \\ \mathbf{u}_r'(\mathbf{x}_j - \bar{\mathbf{x}}) \end{bmatrix} = \mathbf{U}\mathbf{U}'(\mathbf{x}_j - \bar{\mathbf{x}}) \qquad \text{(8A-2)}$$

This follows because, for an arbitrary vector $\mathbf{b}_j$,

$$\begin{aligned} \mathbf{x}_j - \bar{\mathbf{x}} - \mathbf{U}\mathbf{b}_j &= \mathbf{x}_j - \bar{\mathbf{x}} - \mathbf{U}\mathbf{U}'(\mathbf{x}_j - \bar{\mathbf{x}}) + \mathbf{U}\mathbf{U}'(\mathbf{x}_j - \bar{\mathbf{x}}) - \mathbf{U}\mathbf{b}_j \\ &= (\mathbf{I} - \mathbf{U}\mathbf{U}')(\mathbf{x}_j - \bar{\mathbf{x}}) + \mathbf{U}(\mathbf{U}'(\mathbf{x}_j - \bar{\mathbf{x}}) - \mathbf{b}_j) \end{aligned}$$

so the error sum of squares is

$$\begin{aligned} (\mathbf{x}_j - \bar{\mathbf{x}} - \mathbf{U}\mathbf{b}_j)'(\mathbf{x}_j - \bar{\mathbf{x}} - \mathbf{U}\mathbf{b}_j) &= (\mathbf{x}_j - \bar{\mathbf{x}})'(\mathbf{I} - \mathbf{U}\mathbf{U}')(\mathbf{x}_j - \bar{\mathbf{x}}) + 0 \\ &\quad + (\mathbf{U}'(\mathbf{x}_j - \bar{\mathbf{x}}) - \mathbf{b}_j)'(\mathbf{U}'(\mathbf{x}_j - \bar{\mathbf{x}}) - \mathbf{b}_j) \end{aligned}$$

where the cross product vanishes because $(\mathbf{I} - \mathbf{U}\mathbf{U}')\mathbf{U} = \mathbf{U} - \mathbf{U}\mathbf{U}'\mathbf{U} = \mathbf{U} - \mathbf{U} = \mathbf{0}$. The last term is positive unless $\mathbf{b}_j$ is chosen so that $\mathbf{b}_j = \mathbf{U}'(\mathbf{x}_j - \bar{\mathbf{x}})$ and $\mathbf{U}\mathbf{b}_j = \mathbf{U}\mathbf{U}'(\mathbf{x}_j - \bar{\mathbf{x}})$ is the projection of $\mathbf{x}_j - \bar{\mathbf{x}}$ on the plane.

Further, with the choice $\mathbf{a}_j = \mathbf{U}\mathbf{b}_j = \mathbf{U}\mathbf{U}'(\mathbf{x}_j - \bar{\mathbf{x}})$, (8A-1) becomes

$$\begin{aligned} &\sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}} - \mathbf{U}\mathbf{U}'(\mathbf{x}_j - \bar{\mathbf{x}}))'(\mathbf{x}_j - \bar{\mathbf{x}} - \mathbf{U}\mathbf{U}'(\mathbf{x}_j - \bar{\mathbf{x}})) \\ &\qquad = \sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}})'(\mathbf{I} - \mathbf{U}\mathbf{U}')(\mathbf{x}_j - \bar{\mathbf{x}}) \\ &\qquad = \sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}}) - \sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{U}\mathbf{U}'(\mathbf{x}_j - \bar{\mathbf{x}}) \qquad \text{(8A-3)} \end{aligned}$$

We are now in a position to minimize the error over choices of $\mathbf{U}$ by maximizing the last term in (8A-3). By the properties of trace (see Result 2A.12),

$$\begin{aligned} \sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{U}\mathbf{U}'(\mathbf{x}_j - \bar{\mathbf{x}}) &= \sum_{j=1}^{n} \operatorname{tr}[(\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{U}\mathbf{U}'(\mathbf{x}_j - \bar{\mathbf{x}})] \\ &= \sum_{j=1}^{n} \operatorname{tr}[\mathbf{U}\mathbf{U}'(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'] \\ &= (n-1)\operatorname{tr}[\mathbf{U}\mathbf{U}'\mathbf{S}] = (n-1)\operatorname{tr}[\mathbf{U}'\mathbf{S}\mathbf{U}] \qquad \text{(8A-4)} \end{aligned}$$

That is, the best choice for $\mathbf{U}$ maximizes the sum of the diagonal elements of $\mathbf{U}'\mathbf{S}\mathbf{U}$. From (8-19), selecting $\mathbf{u}_1$ to maximize $\mathbf{u}_1'\mathbf{S}\mathbf{u}_1$, the first diagonal element of $\mathbf{U}'\mathbf{S}\mathbf{U}$, gives $\mathbf{u}_1 = \hat{\mathbf{e}}_1$. For $\mathbf{u}_2$ perpendicular to $\hat{\mathbf{e}}_1$, $\mathbf{u}_2'\mathbf{S}\mathbf{u}_2$ is maximized by $\hat{\mathbf{e}}_2$. [See (2-52).] Continuing, we find that $\hat{\mathbf{U}} = [\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_r] = \hat{\mathbf{E}}_r$ and $\hat{\mathbf{A}}' = \hat{\mathbf{E}}_r\hat{\mathbf{E}}_r'[\mathbf{x}_1 - \bar{\mathbf{x}}, \mathbf{x}_2 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_n - \bar{\mathbf{x}}]$, as asserted.

With this choice the ith diagonal element of $\hat{\mathbf{U}}'\mathbf{S}\hat{\mathbf{U}}$ is $\hat{\mathbf{e}}_i'\mathbf{S}\hat{\mathbf{e}}_i = \hat{\mathbf{e}}_i'(\hat{\lambda}_i\hat{\mathbf{e}}_i) = \hat{\lambda}_i$ so $\operatorname{tr}[\hat{\mathbf{U}}'\mathbf{S}\hat{\mathbf{U}}] = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_r$. Also, $\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})'(\mathbf{x}_j - \bar{\mathbf{x}}) = \operatorname{tr}\left[\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'\right]$ $= (n-1)\operatorname{tr}(\mathbf{S}) = (n-1)(\hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p)$. Let $\mathbf{U} = \hat{\mathbf{U}}$ in (8A-3), and the error bound follows. ■

The p-Dimensional Geometrical Interpretation

The geometrical interpretations involve the determination of best approximating planes to the p-dimensional scatter plot. The plane through the origin, determined by $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$, consists of all points $\mathbf{x}$ with

$$\mathbf{x} = b_1\mathbf{u}_1 + b_2\mathbf{u}_2 + \cdots + b_r\mathbf{u}_r = \mathbf{U}\mathbf{b}, \qquad \text{for some } \mathbf{b}$$

This plane, translated to pass through $\mathbf{a}$, becomes $\mathbf{a} + \mathbf{U}\mathbf{b}$ for some $\mathbf{b}$.

We want to select the r-dimensional plane $\mathbf{a} + \mathbf{U}\mathbf{b}$ that *minimizes the sum of squared distances* $\sum_{j=1}^{n} d_j^2$ *between the observations* $\mathbf{x}_j$ *and the plane.* If $\mathbf{x}_j$ is approximated by $\mathbf{a} + \mathbf{U}\mathbf{b}_j$ with $\sum_{j=1}^{n}\mathbf{b}_j = \mathbf{0}$,[5] then

$$\begin{aligned}
&\sum_{j=1}^{n}(\mathbf{x}_j - \mathbf{a} - \mathbf{U}\mathbf{b}_j)'(\mathbf{x}_j - \mathbf{a} - \mathbf{U}\mathbf{b}_j)\\
&\quad = \sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}} - \mathbf{U}\mathbf{b}_j + \bar{\mathbf{x}} - \mathbf{a})'(\mathbf{x}_j - \bar{\mathbf{x}} - \mathbf{U}\mathbf{b}_j + \bar{\mathbf{x}} - \mathbf{a})\\
&\quad = \sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}} - \mathbf{U}\mathbf{b}_j)'(\mathbf{x}_j - \bar{\mathbf{x}} - \mathbf{U}\mathbf{b}_j) + n(\bar{\mathbf{x}} - \mathbf{a})'(\bar{\mathbf{x}} - \mathbf{a})\\
&\quad \geq \sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}} - \hat{\mathbf{E}}_r\hat{\mathbf{E}}_r'(\mathbf{x}_j - \bar{\mathbf{x}}))'(\mathbf{x}_j - \bar{\mathbf{x}} - \hat{\mathbf{E}}_r\hat{\mathbf{E}}_r'(\mathbf{x}_j - \bar{\mathbf{x}}))
\end{aligned}$$

by Result 8A.1, since $[\mathbf{U}\mathbf{b}_1, \ldots, \mathbf{U}\mathbf{b}_n] = \mathbf{A}'$ has rank $(\mathbf{A}) \leq r$. The lower bound is reached by taking $\mathbf{a} = \bar{\mathbf{x}}$, so the plane passes through the sample mean. This plane is determined by $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_r$. The coefficients of $\hat{\mathbf{e}}_k$ are $\hat{\mathbf{e}}_k'(\mathbf{x}_j - \bar{\mathbf{x}}) = \hat{y}_{jk}$, the kth sample principal component evaluated at the jth observation.

The approximating plane interpretation of sample principal components is illustrated in Figure 8.10.

An alternative interpretation can be given. The investigator places a plane through $\bar{\mathbf{x}}$ and moves it about to obtain the *largest spread* among the shadows of the

[5] If $\sum_{j=1}^{n}\mathbf{b}_j = n\bar{\mathbf{b}} \neq \mathbf{0}$, use $\mathbf{a} + \mathbf{U}\mathbf{b}_j = (\mathbf{a} + \mathbf{U}\bar{\mathbf{b}}) + \mathbf{U}(\mathbf{b}_j - \bar{\mathbf{b}}) = \mathbf{a}^* + \mathbf{U}\mathbf{b}_j^*$.

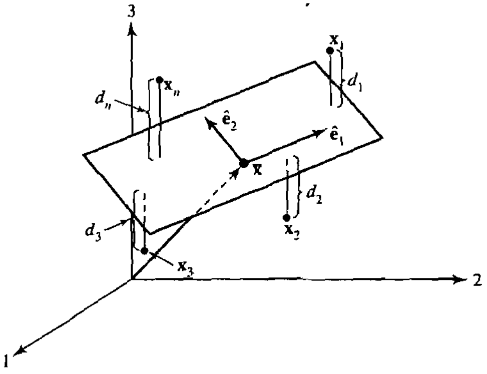


Figure 8.10 The $r = 2$-dimensional plane that approximates the scatter plot by minimizing $\sum_{j=1}^{n} d_j^2$.

observations. From (8A-2), the projection of the deviation $\mathbf{x}_j - \bar{\mathbf{x}}$ on the plane $\mathbf{Ub}$ is $\mathbf{v}_j = \mathbf{UU}'(\mathbf{x}_j - \bar{\mathbf{x}})$. Now, $\bar{\mathbf{v}} = \mathbf{0}$ and the *sum of the squared lengths of the projection deviations*

$$\sum_{j=1}^{n} \mathbf{v}_j'\mathbf{v}_j = \sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{UU}'(\mathbf{x}_j - \bar{\mathbf{x}}) = (n-1)\operatorname{tr}[\mathbf{U}'\mathbf{SU}]$$

is maximized by $\mathbf{U} = \hat{\mathbf{E}}$. Also, since $\bar{\mathbf{v}} = \mathbf{0}$,

$$(n-1)\mathbf{S}_{\mathbf{v}} = \sum_{j=1}^{n} (\mathbf{v}_j - \bar{\mathbf{v}})(\mathbf{v}_j - \bar{\mathbf{v}})' = \sum_{j=1}^{n} \mathbf{v}_j\mathbf{v}_j'$$

and this plane also maximizes the total variance

$$\operatorname{tr}(\mathbf{S}_{\mathbf{v}}) = \frac{1}{(n-1)}\operatorname{tr}\left[\sum_{j=1}^{n} \mathbf{v}_j\mathbf{v}_j'\right] = \frac{1}{(n-1)}\operatorname{tr}\left[\sum_{j=1}^{n} \mathbf{v}_j'\mathbf{v}_j\right]$$

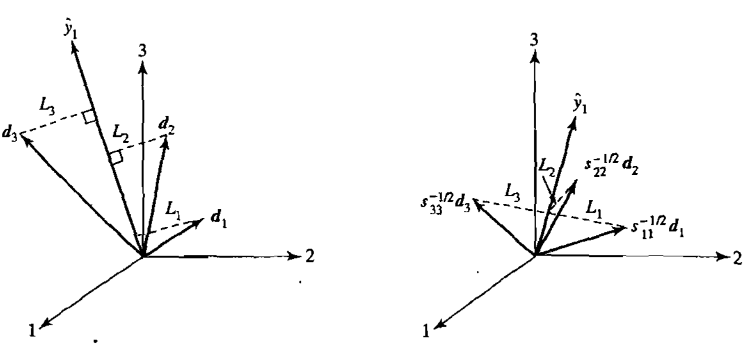
The n-Dimensional Geometrical Interpretation

Let us now consider, by columns, the approximation of the mean-centered data matrix by $\mathbf{A}$. For $r = 1$, the ith column $[x_{1i} - \bar{x}_i, x_{2i} - \bar{x}_i, \ldots, x_{ni} - \bar{x}_i]'$ is approximated by a multiple $c_i\mathbf{b}'$ of a fixed vector $\mathbf{b}' = [b_1, b_2, \ldots, b_n]$. The square of the length of the error of approximation is

$$L_i^2 = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i - c_ib_j)^2$$

Considering $\underset{(n\times p)}{\mathbf{A}}$ to be of rank one, we conclude from Result 8A.1 that

$$\hat{\mathbf{A}} = \begin{bmatrix} \hat{\mathbf{e}}_1\hat{\mathbf{e}}_1'(\mathbf{x}_1 - \bar{\mathbf{x}}) \\ \hat{\mathbf{e}}_1\hat{\mathbf{e}}_1'(\mathbf{x}_2 - \bar{\mathbf{x}}) \\ \vdots \\ \hat{\mathbf{e}}_1\hat{\mathbf{e}}_1'(\mathbf{x}_n - \bar{\mathbf{x}}) \end{bmatrix} = \begin{bmatrix} \hat{y}_{11} \\ \hat{y}_{21} \\ \vdots \\ \hat{y}_{1n} \end{bmatrix} \hat{\mathbf{e}}_1'$$



(a) Principal component of S

(b) Principal component of **R**

Figure 8.11 The first sample principal component, $\hat{y}_1$, minimizes the sum of the squares of the distances, L_i^2, from the deviation vectors, $d_i' = [x_{1i} - \bar{x}_i, x_{2i} - \bar{x}_i, \ldots, x_{ni} - \bar{x}_i]$, to a line.

minimizes the sum of squared lengths $\sum_{i=1}^{p} L_i^2$. That is, the best direction is determined by the vector of values of the first principal component. This is illustrated in Figure 8.11(a). Note that the longer deviation vectors (the larger s_{ii}'s) have the most influence on the minimization of $\sum_{i=1}^{p} L_i^2$.

If the variables are first standardized, the resulting vector $[(x_{1i} - \bar{x}_i)/\sqrt{s_{ii}}, (x_{2i} - \bar{x}_i)/\sqrt{s_{ii}}, \ldots, (x_{ni} - \bar{x}_i)/\sqrt{s_{ii}}]$ has length $n - 1$ for all variables, and each vector exerts equal influence on the choice of direction. [See Figure 8.11(b).]

In either case, the vector **b** is moved around in n-space to minimize the sum of the squares of the distances $\sum_{i=1}^{p} L_i^2$. In the former case L_i^2 is the squared distance between $[x_{1i} - \bar{x}_i, x_{2i} - \bar{x}_i, \ldots, x_{ni} - \bar{x}_i]'$ and its projection on the line determined by **b**. The second principal component minimizes the same quantity among all vectors perpendicular to the first choice.

Exercises

8.1. Determine the population principal components Y_1 and Y_2 for the covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

Also, calculate the proportion of the total population variance explained by the first principal component.

8.2. Convert the covariance matrix in Exercise 8.1 to a correlation matrix $\boldsymbol{\rho}$.

(a) Determine the principal components Y_1 and Y_2 from $\boldsymbol{\rho}$ and compute the proportion of total population variance explained by Y_1.

(b) Compare the components calculated in Part a with those obtained in Exercise 8.1. Are they the same? Should they be?

(c) Compute the correlations ρ_{Y_1,Z_1}, ρ_{Y_1,Z_2}, and ρ_{Y_2,Z_1}.

8.3. Let

$$\mathbf{\Sigma} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

Determine the principal components Y_1, Y_2, and Y_3. What can you say about the eigenvectors (and principal components) associated with eigenvalues that are not distinct?

8.4. Find the principal components and the proportion of the total population variance explained by each when the covariance matrix is

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 & \sigma^2\rho & 0 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ 0 & \sigma^2\rho & \sigma^2 \end{bmatrix}, \quad -\frac{1}{\sqrt{2}} < \rho < \frac{1}{\sqrt{2}}$$

8.5. (a) Find the eigenvalues of the correlation matrix

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

Are your results consistent with (8-16) and (8-17)?

(b) Verify the eigenvalue–eigenvector pairs for the $p \times p$ matrix $\boldsymbol{\rho}$ given in (8-15).

8.6. Data on x_1 = sales and x_2 = profits for the 10 largest companies in the world were listed in Exercise 1.4 of Chapter 1.
From Example 4.12

$$\bar{\mathbf{x}} = \begin{bmatrix} 155.60 \\ 14.70 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 7476.45 & 303.62 \\ 303.62 & 26.19 \end{bmatrix}$$

(a) Determine the sample principal components and their variances for these data. (You may need the quadratic formula to solve for the eigenvalues of $\mathbf{S}$.)

(b) Find the proportion of the total sample variance explained by $\hat{y}_1$.

(c) Sketch the constant density ellipse $(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) = 1.4$, and indicate the principal components $\hat{y}_1$ and $\hat{y}_2$ on your graph.

(d) Compute the correlation coefficients $r_{\hat{y}_1,x_k}$, $k = 1, 2$. What interpretation, if any, can you give to the first principal component?

8.7. Convert the covariance matrix $\mathbf{S}$ in Exercise 8.6 to a sample correlation matrix $\mathbf{R}$.

(a) Find the sample principal components $\hat{y}_1$, $\hat{y}_2$ and their variances.

(b) Compute the proportion of the total sample variance explained by $\hat{y}_1$.

(c) Compute the correlation coefficients $r_{\hat{y}_1,z_k}$, $k = 1, 2$. Interpret $\hat{y}_1$.

(d) Compare the components obtained in Part a with those obtained in Exercise 8.6(a). Given the original data displayed in Exercise 1.4, do you feel that it is better to determine principal components from the sample covariance matrix or sample correlation matrix? Explain.

8.8. Use the results in Example 8.5.

(a) Compute the correlations r_{y_i, z_k} for $i = 1, 2$ and $k = 1, 2, \ldots, 5$. Do these correlations reinforce the interpretations given to the first two components? Explain.

(b) Test the hypothesis

$$H_0\colon \boldsymbol{\rho} = \boldsymbol{\rho}_0 = \begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix}$$

versus

$$H_1\colon \boldsymbol{\rho} \neq \boldsymbol{\rho}_0$$

at the 5% level of significance. List any assumptions required in carrying out this test.

8.9. *(A test that all variables are independent.)*

(a) Consider that the normal theory likelihood ratio test of H_0: $\boldsymbol{\Sigma}$ is the diagonal matrix

$$\begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{bmatrix}, \quad \sigma_{ii} > 0$$

Show that the test is as follows: Reject H_0 if

$$\Lambda = \frac{|\mathbf{S}|^{n/2}}{\prod_{i=1}^{p} s_{ii}^{n/2}} = |\mathbf{R}|^{n/2} < c$$

For a large sample size, $-2 \ln \Lambda$ is approximately $\chi^2_{p(p-1)/2}$. Bartlett [3] suggests that the test statistic $-2[1 - (2p + 11)/6n] \ln \Lambda$ be used in place of $-2 \ln \Lambda$. This results in an improved chi-square approximation. The large sample α critical point is $\chi^2_{p(p-1)/2}(\alpha)$. Note that testing $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ is the same as testing $\boldsymbol{\rho} = \mathbf{I}$.

(b) Show that the likelihood ratio test of H_0: $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$ rejects H_0 if

$$\Lambda = \frac{|\mathbf{S}|^{n/2}}{(\mathrm{tr}(\mathbf{S})/p)^{np/2}} = \left[\frac{\prod_{i=1}^{p} \hat{\lambda}_i}{\left(\frac{1}{p} \sum_{i=1}^{p} \hat{\lambda}_i \right)^p} \right]^{n/2} = \left[\frac{\text{geometric mean } \hat{\lambda}_i}{\text{arithmetic mean } \hat{\lambda}_i} \right]^{np/2} < c$$

For a large sample size, Bartlett [3] suggests that

$$-2[1 - (2p^2 + p + 2)/6pn] \ln \Lambda$$

is approximately $\chi^2_{(p+2)(p-1)/2}$. Thus, the large sample α critical point is $\chi^2_{(p+2)(p-1)/2}(\alpha)$. This test is called a *sphericity test*, because the constant density contours are spheres when $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$.

Hint:

(a) $\max_{\boldsymbol{\mu},\boldsymbol{\Sigma}} L(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is given by (5-10), and $\max L(\boldsymbol{\mu},\boldsymbol{\Sigma}_0)$ is the product of the univariate likelihoods, $\max_{\mu_i \sigma_{ii}} (2\pi)^{-n/2}\sigma_{ii}^{-n/2}\exp\left[-\sum_{j=1}^{n}(x_{ji}-\mu_i)^2/2\sigma_{ii}\right]$. Hence $\hat{\mu}_i = n^{-1}\sum_{j=1}^{n} x_{ji}$ and $\hat{\sigma}_{ii} = (1/n)\sum_{j=1}^{n}(x_{ji}-\bar{x}_i)^2$. The divisor n cancels in Λ, so $\mathbf{S}$ may be used.

(b) Verify $\hat{\sigma}^2 = \left[\sum_{j=1}^{n}(x_{j1}-\bar{x}_1)^2 + \cdots + \sum_{j=1}^{n}(x_{jp}-\bar{x}_p)^2\right]\Big/ np$ under H_0. Again, the divisors n cancel in the statistic, so $\mathbf{S}$ may be used. Use Result 5.2 to calculate the chi-square degrees of freedom.

The following exercises require the use of a computer.

8.10. The weekly rates of return for five stocks listed on the New York Stock Exchange are given in Table 8.4. (See the stock-price data on the following website: www.prenhall.com/statistics.)

(a) Construct the sample covariance matrix $\mathbf{S}$, and find the sample principal components in (8-20). (Note that the sample mean vector $\bar{\mathbf{x}}$ is displayed in Example 8.5.)

(b) Determine the proportion of the total sample variance explained by the first three principal components. Interpret these components.

(c) Construct Bonferroni simultaneous 90% confidence intervals for the variances λ_1, λ_2, and λ_3 of the first three population components Y_1, Y_2, and Y_3.

(d) Given the results in Parts a–c, do you feel that the stock rates-of-return data can be summarized in fewer than five dimensions? Explain.

Table 8.4 Stock-Price Data (Weekly Rate Of Return)

Week	J P Morgan	Citibank	Wells Fargo	Royal Dutch Shell	Exxon Mobil
1	0.01303	−0.00784	−0.00319	−0.04477	0.00522
2	0.00849	0.01669	−0.00621	0.01196	0.01349
3	−0.01792	−0.00864	0.01004	0	−0.00614
4	0.02156	−0.00349	0.01744	−0.02859	−0.00695
5	0.01082	0.00372	−0.01013	0.02919	0.04098
6	0.01017	−0.01220	−0.00838	0.01371	0.00299
7	0.01113	0.02800	0.00807	0.03054	0.00323
8	0.04848	−0.00515	0.01825	0.00633	0.00768
9	−0.03449	−0.01380	−0.00805	−0.02990	−0.01081
10	−0.00466	0.02099	−0.00608	−0.02039	−0.01267
⋮	⋮	⋮	⋮	⋮	⋮
94	0.03732	0.03593	0.02528	0.05819	0.01697
95	0.02380	0.00311	−0.00688	0.01225	0.02817
96	0.02568	0.05253	0.04070	−0.03166	−0.01885
97	−0.00606	0.00863	0.00584	0.04456	0.03059
98	0.02174	0.02296	0.02920	0.00844	0.03193
99	0.00337	−0.01531	−0.02382	−0.00167	−0.01723
100	0.00336	0.00290	−0.00305	−0.00122	−0.00970
101	0.01701	0.00951	0.01820	−0.01618	−0.00756
102	0.01039	−0.00266	0.00443	−0.00248	−0.01645
103	−0.01279	−0.01437	−0.01874	−0.00498	−0.01637

8.11. Consider the census-tract data listed in Table 8.5. Suppose the observations on X_5 = median value home were recorded in ten thousands, rather than hundred thousands, of dollars; that is, multiply all the numbers listed in the sixth column of the table by 10.

(a) Construct the sample covariance matrix **S** for the census-tract data when X_5 = median value home is recorded in ten thousands of dollars. (Note that this covariance matrix can be obtained from the covariance matrix given in Example 8.3 by multiplying the off-diagonal elements in the fifth column and row by 10 and the diagonal element s_{55} by 100. Why?)

(b) Obtain the eigenvalue–eigenvector pairs and the first two sample principal components for the covariance matrix in Part a.

(c) Compute the proportion of total variance explained by the first two principal components obtained in Part b. Calculate the correlation coefficients, $r_{\hat{y}_i, x_k}$, and interpret these components if possible. Compare your results with the results in Example 8.3. What can you say about the effects of this change in scale on the principal components?

8.12. Consider the air-pollution data listed in Table 1.5. Your job is to summarize these data in fewer than $p = 7$ dimensions if possible. Conduct a principal component analysis of the data using both the covariance matrix **S** and the correlation matrix **R**. What have you learned? Does it make any difference which matrix is chosen for analysis? Can the data be summarized in three or fewer dimensions? Can you interpret the principal components?

Table 8.5 Census-tract Data

Tract	Total population (thousands)	Professional degree (percent)	Employed age over 16 (percent)	Government employment (percent)	Median home value ($100,000)
1	2.67	5.71	69.02	30.3	1.48
2	2.25	4.37	72.98	43.3	1.44
3	3.12	10.27	64.94	32.0	2.11
4	5.14	7.44	71.29	24.5	1.85
5	5.54	9.25	74.94	31.0	2.23
6	5.04	4.84	53.61	48.2	1.60
7	3.14	4.82	67.00	37.6	1.52
8	2.43	2.40	67.20	36.8	1.40
9	5.38	4.30	83.03	19.7	2.07
10	7.34	2.73	72.60	24.5	1.42
	⋮	⋮	⋮	⋮	⋮
52	7.25	1.16	78.52	23.6	1.50
53	5.44	2.93	73.59	22.3	1.65
54	5.83	4.47	77.33	26.2	2.16
55	3.74	2.26	79.70	20.2	1.58
56	9.21	2.36	74.58	21.8	1.72
57	2.14	6.30	86.54	17.4	2.80
58	6.62	4.79	78.84	20.0	2.33
59	4.24	5.82	71.39	27.1	1.69
60	4.72	4.71	78.01	20.6	1.55
61	6.48	4.93	74.23	20.9	1.98

Note: Observations from adjacent census tracts are likely to be correlated. That is, these 61 observations may not constitute a random sample. Complete data set available at www.prenhall.com/statistics.

8.13. In the radiotherapy data listed in Table 1.7 (see also the radiotherapy data on the website www.prenhall.com/statistics), the $n = 98$ observations on $p = 6$ variables represent patients' reactions to radiotherapy.

(a) Obtain the covariance and correlation matrices **S** and **R** for these data.

(b) Pick one of the matrices **S** or **R** (justify your choice), and determine the eigenvalues and eigenvectors. Prepare a table showing, in decreasing order of size, the percent that each eigenvalue contributes to the total sample variance.

(c) Given the results in Part b, decide on the number of important sample principal components. Is it possible to summarize the radiotherapy data with a single reaction-index component? Explain.

(d) Prepare a table of the correlation coefficients between each principal component you decide to retain and the original variables. If possible, interpret the components.

8.14. Perform a principal component analysis using the sample covariance matrix of the sweat data given in Example 5.2. Construct a Q–Q plot for each of the important principal components. Are there any suspect observations? Explain.

8.15. The four sample standard deviations for the postbirth weights discussed in Example 8.6 are

$$\sqrt{s_{11}} = 32.9909, \quad \sqrt{s_{22}} = 33.5918, \quad \sqrt{s_{33}} = 36.5534, \quad \text{and} \quad \sqrt{s_{44}} = 37.3517$$

Use these and the correlations given in Example 8.6 to construct the sample covariance matrix **S**. Perform a principal component analysis using **S**.

8.16. Over a period of five years in the 1990s, yearly samples of fishermen on 28 lakes in Wisconsin were asked to report the time they spent fishing and how many of each type of game fish they caught. Their responses were then converted to a catch rate per hour for

$$\begin{array}{lll} x_1 = \text{Bluegill} & x_2 = \text{Black crappie} & x_3 = \text{Smallmouth bass} \\ x_4 = \text{Largemouth bass} & x_5 = \text{Walleye} & x_6 = \text{Northern pike} \end{array}$$

The estimated correlation matrix (courtesy of Jodi Barnet)

$$\mathbf{R} = \begin{bmatrix} 1 & .4919 & .2636 & .4653 & -.2277 & .0652 \\ .4919 & 1 & .3127 & .3506 & -.1917 & .2045 \\ .2635 & .3127 & 1 & .4108 & .0647 & .2493 \\ .4653 & .3506 & .4108 & 1 & -.2249 & .2293 \\ -.2277 & -.1917 & .0647 & -.2249 & 1 & -.2144 \\ .0652 & .2045 & .2493 & .2293 & -.2144 & 1 \end{bmatrix}$$

is based on a sample of about 120. (There were a few missing values.)

Fish caught by the same fisherman live alongside of each other, so the data should provide some evidence on how the fish group. The first four fish belong to the centrarchids, the most plentiful family. The walleye is the most popular fish to eat.

(a) Comment on the pattern of correlation within the centrarchid family x_1 through x_4. Does the walleye appear to group with the other fish?

(b) Perform a principal component analysis using only x_1 through x_4. Interpret your results.

(c) Perform a principal component analysis using all six variables. Interpret your results.

8.17. Using the data on bone mineral content in Table 1.8, perform a principal component analysis of **S**.

8.18. The data on national track records for women are listed in Table 1.9.

(a) Obtain the sample correlation matrix **R** for these data, and determine its eigenvalues and eigenvectors.

(b) Determine the first two principal components for the standardized variables. Prepare a table showing the correlations of the standardized variables with the components, and the cumulative percentage of the total (standardized) sample variance explained by the two components.

(c) Interpret the two principal components obtained in Part b. (Note that the first component is essentially a normalized unit vector and might measure the athletic excellence of a given nation. The second component might measure the relative strength of a nation at the various running distances.)

(d) Rank the nations based on their score on the first principal component. Does this ranking correspond with your inituitive notion of athletic excellence for the various countries?

8.19. Refer to Exercise 8.18. Convert the national track records for women in Table 1.9 to speeds measured in meters per second. Notice that the records for 800 m, 1500 m, 3000 m, and the marathon are given in minutes. The marathon is 26.2 miles, or 42,195 meters, long. Perform a principal components analysis using the covariance matrix **S** of the speed data. Compare the results with the results in Exercise 8.18. Do your interpretations of the components differ? If the nations are ranked on the basis of their score on the first principal component, does the subsequent ranking differ from that in Exercise 8.18? Which analysis do you prefer? Why?

8.20. The data on national track records for men are listed in Table 8.6. (See also the data on national track records for men on the website www.prenhall.com/statistics) Repeat the principal component analysis outlined in Exercise 8.18 for the men. Are the results consistent with those obtained from the women's data?

8.21. Refer to Exercise 8.20. Convert the national track records for men in Table 8.6 to speeds measured in meters per second. Notice that the records for 800 m, 1500 m, 5000 m, 10,000 m and the marathon are given in minutes. The marathon is 26.2 miles, or 42,195 meters, long. Perform a principal component analysis using the covariance matrix **S** of the speed data. Compare the results with the results in Exercise 8.20. Which analysis do you prefer? Why?

8.22. Consider the data on bulls in Table 1.10. Utilizing the seven variables YrHgt, FtFrBody, PrctFFB, Frame, BkFat, SaleHt, and SaleWt, perform a principal component analysis using the covariance matrix **S** and the correlation matrix **R**. Your analysis should include the following:

(a) Determine the appropriate number of components to effectively summarize the sample variability. Construct a scree plot to aid your determination.

(b) Interpret the sample principal components.

(c) Do you think it is possible to develop a "body size" or "body configuration" index from the data on the seven variables above? Explain.

(d) Using the values for the first two principal components, plot the data in a two-dimensional space with $\hat{y}_1$ along the vertical axis and $\hat{y}_2$ along the horizontal axis. Can you distinguish groups representing the three breeds of cattle? Are there any outliers?

(e) Construct a Q–Q plot using the first principal component. Interpret the plot.

Table 8.6 National Track Records for Men

Country	100 m (s)	200 m (s)	400 m (s)	800 m (min)	1500 m (min)	5000 m (min)	10,000 m (min)	Marathon (min)
Argentina	10.23	20.37	46.18	1.77	3.68	13.33	27.65	129.57
Australia	9.93	20.06	44.38	1.74	3.53	12.93	27.53	127.51
Austria	10.15	20.45	45.80	1.77	3.58	13.26	27.72	132.22
Belgium	10.14	20.19	45.02	1.73	3.57	12.83	26.87	127.20
Bermuda	10.27	20.30	45.26	1.79	3.70	14.64	30.49	146.37
Brazil	10.00	19.89	44.29	1.70	3.57	13.48	28.13	126.05
Canada	9.84	20.17	44.72	1.75	3.53	13.23	27.60	130.09
Chile	10.10	20.15	45.92	1.76	3.65	13.39	28.09	132.19
China	10.17	20.42	45.25	1.77	3.61	13.42	28.17	129.18
Columbia	10.29	20.85	45.84	1.80	3.72	13.49	27.88	131.17
Cook Islands	10.97	22.46	51.40	1.94	4.24	16.70	35.38	171.26
Costa Rica	10.32	20.96	46.42	1.87	3.84	13.75	28.81	133.23
Czech Republic	10.24	20.61	45.77	1.75	3.58	13.42	27.80	131.57
Denmark	10.29	20.52	45.89	1.69	3.52	13.42	27.91	129.43
DominicanRepublic	10.16	20.65	44.90	1.81	3.73	14.31	30.43	146.00
Finland	10.21	20.47	45.49	1.74	3.61	13.27	27.52	131.15
France	10.02	20.16	44.64	1.72	3.48	12.98	27.38	126.36
Germany	10.06	20.23	44.33	1.73	3.53	12.91	27.36	128.47
Great Britain	9.87	19.94	44.36	1.70	3.49	13.01	27.30	127.13
Greece	10.11	19.85	45.57	1.75	3.61	13.48	28.12	132.04
Guatemala	10.32	21.09	48.44	1.82	3.74	13.98	29.34	132.53
Hungary	10.08	20.11	45.43	1.76	3.59	13.45	28.03	132.10
India	10.33	20.73	45.48	1.76	3.63	13.50	28.81	132.00
Indonesia	10.20	20.93	46.37	1.83	3.77	14.21	29.65	139.18
Ireland	10.35	20.54	45.58	1.75	3.56	13.07	27.78	129.15
Israel	10.20	20.89	46.59	1.80	3.70	13.66	28.72	134.21
Italy	10.01	19.72	45.26	1.73	3.35	13.09	27.28	127.29
Japan	10.00	20.03	44.78	1.77	3.62	13.22	27.58	126.16
Kenya	10.28	20.43	44.18	1.70	3.44	12.66	26.46	124.55
Korea, South	10.34	20.41	45.37	1.74	3.64	13.84	28.51	127.20
Korea, North	10.60	21.23	46.95	1.82	3.77	13.90	28.45	129.26
Luxembourg	10.41	20.77	47.90	1.76	3.67	13.64	28.77	134.03
Malaysia	10.30	20.92	46.41	1.79	3.76	14.11	29.50	149.27
Mauritius	10.13	20.06	44.69	1.80	3.83	14.15	29.84	143.07
Mexico	10.21	20.40	44.31	1.78	3.63	13.13	27.14	127.19
Myanmar(Burma)	10.64	21.52	48.63	1.80	3.80	14.19	29.62	139.57
Netherlands	10.19	20.19	45.68	1.73	3.55	13.22	27.44	128.31
New Zealand	10.11	20.42	46.09	1.74	3.54	13.21	27.70	128.59
Norway	10.08	20.17	46.11	1.71	3.62	13.11	27.54	130.17
Papua New Guinea	10.40	21.18	46.77	1.80	4.00	14.72	31.36	148.13
Philippines	10.57	21.43	45.57	1.80	3.82	13.97	29.04	138.44
Poland	10.00	19.98	44.62	1.72	3.59	13.29	27.89	129.23
Portugal	9.86	20.12	46.11	1.75	3.50	13.05	27.21	126.36
Romania	10.21	20.75	45.77	1.76	3.57	13.25	27.67	132.30
Russia	10.11	20.23	44.60	1.71	3.54	13.20	27.90	129.16
Samoa	10.78	21.86	49.98	1.94	4.01	16.28	34.71	161.50
Singapore	10.37	21.14	47.60	1.84	3.86	14.96	31.32	144.22
Spain	10.17	20.59	44.96	1.73	3.48	13.04	27.24	127.23
Sweden	10.18	20.43	45.54	1.76	3.61	13.29	27.93	130.38
Switzerland	10.16	20.41	44.99	1.71	3.53	13.13	27.90	129.56
Taiwan	10.36	20.81	46.72	1.79	3.77	13.91	29.20	134.35
Thailand	10.23	20.69	46.05	1.81	3.77	14.25	29.67	139.33
Turkey	10.38	21.04	46.63	1.78	3.59	13.45	28.33	130.25
U.S.A.	9.78	19.32	43.18	1.71	3.46	12.97	27.23	125.38

Source: *IAAF/ATFS Track and Field Statistics Handbook* for the Helsinki 2005 Olympics. Courtesy of Ottavio Castellini.

8.23. A naturalist for the Alaska Fish and Game Department studies grizzly bears with the goal of maintaining a healthy population. Measurements on $n = 61$ bears provided the following summary statistics:

Variable	Weight (kg)	Body length (cm)	Neck (cm)	Girth (cm)	Head length (cm)	Head width (cm)
Sample mean $\bar{x}$	95.52	164.38	55.69	93.39	17.98	31.13

Covariance matrix

$$\mathbf{S} = \begin{bmatrix} 3266.46 & 1343.97 & 731.54 & 1175.50 & 162.68 & 238.37 \\ 1343.97 & 721.91 & 324.25 & 537.35 & 80.17 & 117.73 \\ 731.54 & 324.25 & 179.28 & 281.17 & 39.15 & 56.80 \\ 1175.50 & 537.35 & 281.17 & 474.98 & 63.73 & 94.85 \\ 162.68 & 80.17 & 39.15 & 63.73 & 9.95 & 13.88 \\ 238.37 & 117.73 & 56.80 & 94.85 & 13.88 & 21.26 \end{bmatrix}$$

(a) Perform a principal component analysis using the covariance matrix. Can the data be effectively summarized in fewer than six dimensions?

(b) Perform a principal component analysis using the correlation matrix.

(c) Comment on the similarities and differences between the two analyses.

8.24. Refer to Example 8.10 and the data in Table 5.8, page 240. Add the variable $x_6 =$ regular overtime hours whose values are (read across)

6187	7336	6988	6964	8425	6778	5922	7307
7679	8259	10954	9353	6291	4969	4825	6019

and redo Example 8.10.

8.25. Refer to the police overtime hours data in Example 8.10. Construct an alternate control chart, based on the sum of squares d^2_{Uj}, to monitor the unexplained variation in the original observations summarized by the additional principal components.

8.26. Consider the psychological profile data in Table 4.6. Using the five variables, Indep, Supp, Benev, Conform and Leader, performs a principal component analysis using the covariance matrix **S** and the correlation matrix **R**. Your analysis should include the following:

(a) Determine the appropriate number of components to effectively summarize the variability. Construct a scree plot to aid in your determination.

(b) Interpret the sample principal components.

(c) Using the values for the first two principal components, plot the data in a two-dimensional space with $\hat{y}_1$ along the vertical axis and $\hat{y}_2$ along the horizontal axis. Can you distinguish groups representing the two socioeconomic levels and/or the two genders? Are there any outliers?

(d) Construct a 95% confidence interval for λ_1, the variance of the first population principal component from the covariance matrix.

8.27. The pulp and paper properties data is given in Table 7.7. Using the four paper variables, BL (breaking length), EM (elastic modulus), SF (Stress at failure) and BS (burst strength), perform a principal component analysis using the covariance matrix **S** and the correlation matrix **R**. Your analysis should include the following:

(a) Determine the appropriate number of components to effectively summarize the variability. Construct a scree plot to aid in your determination.

(b) Interpret the sample principal components.

(c) Do you think it it is possible to develop a "paper strength" index that effectively contains the information in the four paper variables? Explain.

(d) Using the values for the first two principal components, plot the data in a two-dimensional space with $\hat{y}_1$ along the vertical axis and $\hat{y}_2$ along the horizontal axis. Identify any outliers in this data set.

8.28. Survey data were collected as part of a study to assess options for enhancing food security through the sustainable use of natural resources in the Sikasso region of Mali (West Africa). A total of $n = 76$ farmers were surveyed and observations on the nine variables

x_1 = Family (total number of individuals in household)

x_2 = DistRd (distance in kilometers to nearest passable road)

x_3 = Cotton (hectares of cotton planted in year 2000)

x_4 = Maize (hectares of maize planted in year 2000)

x_5 = Sorg (hectares of sorghum planted in year 2000)

x_6 = Millet (hectares of millet planted in year 2000)

x_7 = Bull (total number of bullocks or draft animals)

x_8 = Cattle (total); x_9 = Goats (total)

were recorded. The data are listed in Table 8.7 and on the website www.prenhall.com/statistics

(a) Construct two-dimensional scatterplots of Family versus DistRd, and DistRd versus Cattle. Remove any obvious outliers from the data set.

Table 8.7 Mali Family Farm Data

Family	DistRD	Cotton	Maize	Sorg	Millet	Bull	Cattle	Goats
12	80	1.5	1.00	3.0	.25	2	0	1
54	8	6.0	4.00	0	1.00	6	32	5
11	13	.5	1.00	0	0	0	0	0
21	13	2.0	2.50	1.0	0	1	0	5
61	30	3.0	5.00	0	0	4	21	0
20	70	0	2.00	3.0	0	2	0	3
29	35	1.5	2.00	0	0	0	0	0
29	35	2.0	3.00	2.0	0	0	0	0
57	9	5.0	5.00	0	0	4	5	2
23	33	2.0	2.00	1.0	0	2	1	7
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
20	0	1.5	1.00	3.0	0	1	6	0
27	41	1.1	.25	1.5	1.50	0	3	1
18	500	2.0	1.00	1.5	.50	1	0	0
30	19	2.0	2.00	4.0	1.00	2	0	5
77	18	8.0	4.00	6.0	4.00	6	8	6
21	500	5.0	1.00	3.0	4.00	1	0	5
13	100	.5	.50	0	1.00	0	0	4
24	100	2.0	3.00	0	.50	3	14	10
29	90	2.0	1.50	1.5	1.50	2	0	2
57	90	10.0	7.00	0	1.50	7	8	7

Source: Data courtesy of Jay Angerer.

(b) Perform a principal component analysis using the correlation matrix **R**. Determine the number of components to effectively summarize the variability. Use the proportion of variation explained and a scree plot to aid in your determination.

(c) Interpret the first five principal components. Can you identify, for example, a "farm size" component? A, perhaps, "goats and distance to road" component?

8.29. Refer to Exercise 5.28. Using the covariance matrix **S** for the first 30 cases of car body assembly data, obtain the sample principal components.

(a) Construct a 95% ellipse format chart using the first two principal components $\hat{y}_1$ and $\hat{y}_2$. Identify the car locations that appear to be out of control.

(b) Construct an alternative control chart, based on the sum of squares d_{Uj}^2, to monitor the variation in the original observations summarized by the remaining four principal components. Interpret this chart.

References

1. Anderson, T. W. *An Introduction to Multivariate Statistical Analysis* (3rd ed.). New York: John Wiley, 2003.
2. Anderson, T. W. "Asymptotic Theory for Principal Components Analysis." *Annals of Mathematical Statistics*, **34** (1963), 122–148.
3. Bartlett, M. S. "A Note on Multiplying Factors for Various Chi-Squared Approximations." *Journal of the Royal Statistical Society (B)*, **16** (1954), 296–298.
4. Dawkins, B. "Multivariate Analysis of National Track Records." *The American Statistician*, **43** (1989), 110–115.
5. Girschick, M. A. "On the Sampling Theory of Roots of Determinantal Equations." *Annals of Mathematical Statistics*, **10** (1939), 203–224.
6. Hotelling, H. "Analysis of a Complex of Statistical Variables into Principal Components." *Journal of Educational Psychology*, **24** (1933), 417–441, 498–520.
7. Hotelling, H. "The Most Predictable Criterion." *Journal of Educational Psychology*, **26** (1935), 139–142.
8. Hotelling, H. "Simplified Calculation of Principal Components." *Psychometrika*, **1** (1936), 27–35.
9. Hotelling, H. "Relations between Two Sets of Variates." *Biometrika*, **28** (1936), 321–377.
10. Jolicoeur, P. "The Multivariate Generalization of the Allometry Equation." *Biometrics*, **19** (1963), 497–499.
11. Jolicoeur, P., and J. E. Mosimann. "Size and Shape Variation in the Painted Turtle: A Principal Component Analysis." *Growth*, **24** (1960), 339–354.
12. King, B. "Market and Industry Factors in Stock Price Behavior." *Journal of Business*, **39** (1966), 139–190.
13. Kourti, T., and J. McGregor, "Multivariate SPC Methods for Process and Product Monitoring," *Journal of Quality Technology*, **28** (1996), 409–428.
14. Lawley, D. N. "On Testing a Set of Correlation Coefficients for Equality." *Annals of Mathematical Statistics*, **34** (1963), 149–151.
15. Rao, C. R. *Linear Statistical Inference and Its Applications* (2nd ed.). New York: Wiley-Interscience, 2002.
16. Rencher, A. C. "Interpretation of Canonical Discriminant Functions, Canonical Variates and Principal Components." *The American Statistician*, **46** (1992), 217–225.