Population Principal Components

Algebraically, principal components are particular linear combinations of the p random variables $X_1, X_2, \ldots, X_p$. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with $X_1, X_2, \ldots, X_p$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and more parsinonious 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.

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

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

Then, using what we already know about covariances of linear combinations, we obtain

$$\begin{array}{rcll} \mathrm{Var}(Y_i) & = & \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i & i = 1, 2, \ldots, p \\ \mathrm{Cov}(Y_i, Y_k) & = & \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_k & i, k = 1, 2, \ldots, p \end{array}$$

The principal components are those *uncorrelated* linear combinations $Y_1, Y_2, \ldots, Y_p$ whose variances as defined above are as large as possible.

The first principal component is the linear combination with maximum variance. That is, it maximizes $\text{Var}(Y_1) = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1$. It is clear that $\text{Var}(Y_1) = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1$ 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

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

At the ith step,

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

Result. Let $\mathbf{\Sigma}$ be the covariance matrix associated with the random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$. Let $\mathbf{\Sigma}$ 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

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

With these choices,

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

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

Proof. It has been proved that for the maximization of quadratic forms of the unit sphere, with $\mathbf{B} = \mathbf{\Sigma}$, the following holds

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

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

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

Similarly,

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

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 $k = 1, 2, \ldots, p-1$,

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

But $\mathbf{e}_{k+1}'(\mathbf{\Sigma}\mathbf{e}_{k+1}) = \lambda_{k+1}\mathbf{e}_{k+1}'\mathbf{e}_{k+1} = \lambda_{k+1}$ 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 $\mathbf{\Sigma}$ 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 $\mathbf{\Sigma}\mathbf{e}_k = \lambda_k\mathbf{e}_k$, premultiplication by $\mathbf{e}_i'$ gives

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

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

Thus, the principal components are uncorrelated and have variances equal to the eigenvalues of $\mathbf{\Sigma}$.

Result. Let $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have covariance matrix $\mathbf{\Sigma}$, with eigenvalue-eigenvector pairs $(\lambda_i, \mathbf{e}_i)$, $i = 1, 2, \ldots, p$, 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 the definition of the *trace* of a matrix, $\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \text{tr}(\mathbf{\Sigma})$. Using spectral decomposition, we can write $\mathbf{\Sigma} = \mathbf{P\Lambda P}'$ where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues and $\mathbf{P} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p]$ so that $\mathbf{PP}' = \mathbf{P}'\mathbf{P} = \mathbf{I}$. Applying the equivalence $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$, we have

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

Thus,

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

We therefore have that

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

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

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

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 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

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

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 $\mathrm{Cov}(X_k, Y_i) = \mathrm{Cov}(\mathbf{a}_k'\mathbf{X}, \mathbf{e}_i'\mathbf{X}) = \mathbf{a}_k'\mathbf{\Sigma}\mathbf{e}_i$. Since $\mathbf{\Sigma}\mathbf{e}_i = \lambda_i\mathbf{e}_i$, it follows that $\mathrm{Cov}(X_k, Y_i) = \mathbf{a}_k'\lambda_i\mathbf{e}_i = \lambda_i e_{ik}$. Then $\mathrm{Var}(Y_i) = \lambda_i$ and $\mathrm{Var}(X_k) = \sigma_{kk}$ yield

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

□

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

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

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

It may be verified that the eigenvalue-eigenvector pairs are

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

Therefore, the principal components become

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

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

We have

$$\begin{array}{rcl} \text{Var}(Y_1) & = & \text{Var}(.383X_1 - .924X_2) \\ & = & (.383)^2\text{Var}(X_1) + (-.924)^2\text{Var}(X_2) \\ & & +2(.383)(-.924)\text{Cov}(X_1, X_2) \\ & = & .147(1) + .854(5) - .708(-2) \\ & = & 5.83 = \lambda_1 \\ \text{Cov}(Y_1, Y_2) & = & \text{Cov}(.383X_1 - .924X_2, X3) \\ & = & .383\text{Cov}(X_1, X_3) - .924\text{Cov}(X_2, X_3) \\ & = & .383(0) - .924(0) = 0 \end{array}$$

It is also readily apparent that

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

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, we obtain

$$\begin{aligned} \rho_{Y_1,X_1} &= \frac{e_{11}\sqrt{\lambda_1}}{\sqrt{\sigma_{11}}} = \frac{.383\sqrt{5.83}}{\sqrt{1}} = .925 \\ \rho_{Y_1,X_2} &= \frac{e_{12}\sqrt{\lambda_1}}{\sqrt{\sigma_{22}}} = \frac{-.924\sqrt{5.83}}{\sqrt{5}} = -.998 \end{aligned}$$

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 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,

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

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 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 $\pm c\sqrt{\lambda_i}\mathbf{e}_i, \quad i = 1, 2, \ldots, p$, 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; no generality is lost by doing this.

We have seen before that

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

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

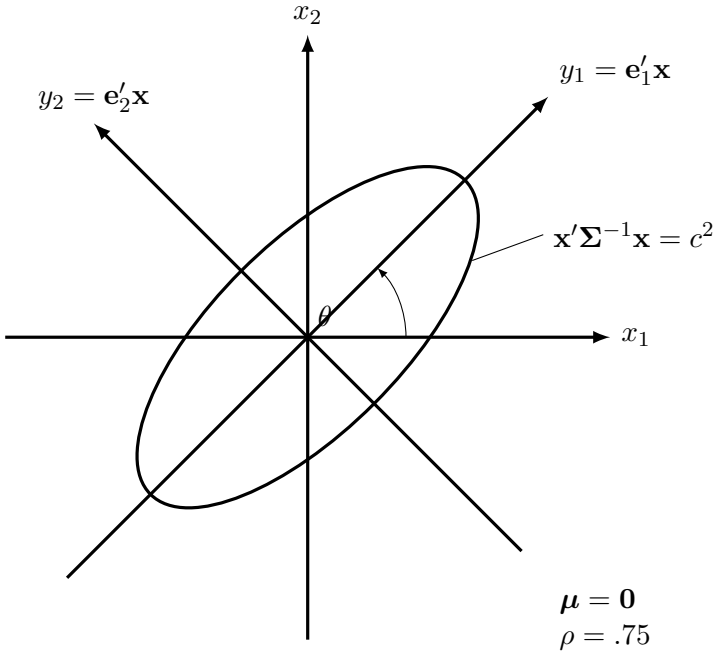
$$c^2 = \frac{1}{\lambda_1} y_1^2 + \frac{1}{\lambda_2} y_2^2 + \cdots + \frac{1}{\lambda_p} y_p^2$$

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 direction 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 $\mathbf{e}_i = [e_{i1}, e_{12}, \ldots e_{i_p}]$ 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 $\boldsymbol{\mu} = 0$ and $\rho = .75$ are shown in the figure on the following page. 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.



The constant density ellipse $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} = c^2$ and the principal components y_1, y_2 for 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

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

In matrix notation,

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

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

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

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 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,

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

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. This last Result follows from the sequence of the three consecutive Results preceding it, 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 therefore have that the total (standardized variables) population variance is simply p, the sum of the diagonal elements of the matrix $\boldsymbol{\rho}$. Using what we have previously learned about the proportion of total variance due to a principal component, replacing $\mathbf{X}$ with $\mathbf{Z}$, we find that the proportion of total variance explained by the kth principal component of $\mathbf{Z}$ is

$$\left(\begin{array}{l} \text{Proportion of (standardized)} \\ \text{population variance due to} \\ k\text{th principal component} \end{array} \right) = \frac{\lambda_k}{p}, \qquad k = 1, 2, \ldots, p$$

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

Example: Principal components obtained from covariance and correlation matrices are different. Consider the covariance matrix

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

and the derived correlation matrix

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

The eigenvalue-eigenvector pairs from $\mathbf{\Sigma}$ are

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

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

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

The respective principal components become

$$\mathbf{\Sigma}\text{:} \quad \begin{array}{rcrcl} Y_1 & = & .040X_1 & + & .999X_2 \\ Y_2 & = & .999X_1 & - & .040X_2 \end{array}$$

and

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

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

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

of 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 the previous Result, 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 $\lambda_1/p = 1.4/2 = .7$ 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 $\mathbf{\Sigma}$. □

The preceding example demonstrates that the principal components derived from $\mathbf{\Sigma}$ are different from those derived from $\boldsymbol{\rho}$. Furthermore, one se 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 the annual sales in the \$10,000 to \$350,000 range and X_2 in 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 wieghting 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.

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 $\mathbf{\Sigma}$ is the diagonal matrix

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & 0 & \cdots & 0 \\ 0 & \sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{pp} \end{bmatrix}$$

Setting $\mathbf{e}_1' = [0, \ldots, 0, 1, 0, \ldots, 0]$, with 1 in the ith position, we observe that

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

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 diagonal covariance matrix, nothing is gained by extracting the principal components. From another point of view, if $\mathbf{X}$ is distributed as $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, 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. Furthermore, standardization does not substantially alter the situation for diagonal $\mathbf{\Sigma}$, in which case $\boldsymbol{\rho} = \mathbf{I}$. The principal components determined from $\boldsymbol{\rho}$ are also the original variables $Z_1, \ldots, Z_p$. Moreover, the eigenvalues all equal 1, meaning that 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

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

The resulting correlation matrix

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

is also the covariance matrix of the standardized variables. In this situation, the variables $X_1, X_2, \ldots, X_p$ are equally correlated.

It is not difficult to show that the p eigenvalues of the correlation matrix above can be divided into two groups. When ρ is positive, the largest is

$$\lambda_i = 1 + (p-1)\rho$$

with associated eigenvector

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

The remaining $p-1$ eigenvalues are

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

and one choice for their eigenvectors is

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

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 standardized variables. It might be regarded as an "index" with equal weights. This principal component explains a proportion

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

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 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 $\boldsymbol{\rho}$ (the equal-correlation matrix from above), then the ellipsoids of constant density are called "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).

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 $\overline{\mathbf{x}}$, the sample covariance matrix $\mathbf{S}$, and the sample correlation matrix $\mathbf{R}$.

Our objective here 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

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

have sample mean $\mathbf{a}_1'\overline{\mathbf{x}}$ and sample variance $\mathbf{a}_1'\mathbf{S}\mathbf{a}_\mathbf{1}$. Also, the pairs of values $(\mathbf{a}_1'\mathbf{x}_j, \mathbf{a}_2'\mathbf{x}_j)$, for two linear combinations, have sample covariance $\mathbf{a}_1'\mathbf{S}\mathbf{a}_2$.

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 the pairs $(\mathbf{a}_i'\mathbf{x}_j, \mathbf{a}_k'\mathbf{x}_j), k < i$

The first principal component maximizes

$$\frac{\mathbf{a}_1'\mathbf{S}\mathbf{a}_1}{\mathbf{a}_1'\mathbf{a}_1}$$

Using familiar logic, 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 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, using techniques like those we have used in previous proofs, we obtain the following results concerning sample principal components:

If $\mathbf{S} = \{\mathbf{s}_{ik}\}$ 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

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

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

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

In addition,

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

and

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

We shall denote the sample principal components by $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_p$, irrespective of whether they are obtained from $\mathbf{S}$ or $\mathbf{R}$.[1]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

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

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

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

generated by substituting each observation $\mathbf{x}_j$ for the arbitrary $\mathbf{x}$, then

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

That is, the sample mean of each principal component is zero. The sample variances are still given by the $\hat{\lambda}_i's$.

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

Example: 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 Table 8.5 in the textbook) for 61 tracks produced the following summary statistics:

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

and

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

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

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. □

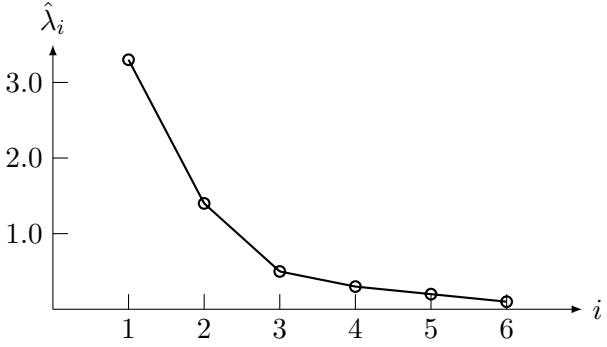
The Number of Principal Components

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.

A useful visual aid to determining an appropriate number of principal components is a *scree plot.*[2] 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. The figure below shows a scree plot for a situation with six principal components.

An elbow occurs in this plot 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.

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



A scree plot

Example: Summarizing sample variability with one sample principal component In a study of size and shape relationships for painted turtles, Jolicoeur and Mosimann measured carapace length, width and height.

length x_1	width x_2	height x_3
93	74	37
94	78	35
96	80	35
101	84	39
102	85	38
103	81	37
104	83	39
106	83	39
107	82	38
112	89	40
113	88	40
114	86	40

... continued length	width	height
116	90	43
117	90	41
117	91	41
119	93	41
120	89	40
120	93	44
121	95	42
125	93	45
127	96	45
128	95	45
131	95	46
135	106	47

Their data suggest an analysis in terms of logarithms. (Jolicoeur generally suggests a logarithmic transformation in studies of size-and-shape relationships.)

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 yields the following summary:

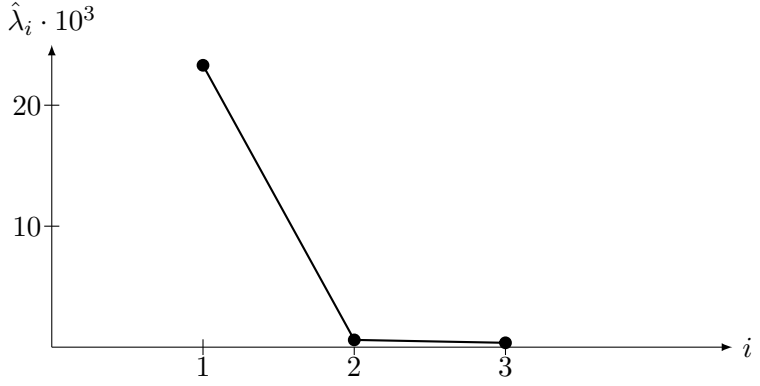
Variable	$\hat{\mathbf{e}}_1(r_{\hat{y}_1,x_k})$	$\hat{\mathbf{e}}_2$	$\hat{\mathbf{e}}_3$
ln(length)	.683 (.99)	−.159	−.713
ln(width)	.510 (.97)	−.594	.622
ln(height)	.523 (.97)	.788	.324
Variance ($\hat{\lambda}_i$):	$23.30 \cdot 10^{-3}$	$.60 \cdot 10^{-3}$	$.36 \cdot 10^{-3}$
Cumulative percentage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

A scree plot is shown in the figure on the next slide. The very distinct elbow in this plot occurs at $i = 2$. There is clearly one dominant principal component.

The first principal component, which explains 6% of the total variance, has an interesting subject-matter interpretation. Since

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

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



A scree plot for the turtle data

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 realization 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 $p \times \mathbf{1}$ vectors $\mathbf{x}$ satisfying

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

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 earlier, but is not required for the development of the properties of the sample principal components summarized in the equations describing *sample variance* and *sample covariance*.

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 be expressed in the new coordinates, which coincide with the axes of the contour expressed by the equation above, which defines a hyperellipsoid that is centered at $\hat{\mathbf{x}}$ and whose axes are given by the eigenvectors of $\mathbf{S}^{-1}$ or, equivalently, of $\mathbf{S}$. 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 $\hat{\mathbf{e}}_i$. 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 the figure for $p = 2$. Figure 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 shows a constant distance ellipse, centered at $\bar{\mathbf{x}}$, with $\hat{\lambda}_1 \doteq \hat{\lambda}_2$. If $\hat{\lambda}_1 = \hat{\lambda}_2$, the axes of the ellipse (circle) of constant distance are not uniquely determined and can lie in any two perpendicular directions, including the 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 homogenous 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.

Standardizing the Sample Principal Components

Sample principal components are, in general, not invariant with respect to changes in scale. 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

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

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

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

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

yields the sample mean vector

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

and sample covariance matrix

$$
\begin{aligned}
\mathbf{S}_z &= \frac{1}{n-1}\left(\mathbf{Z}-\frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{Z}\right)'\left(\mathbf{Z}-\frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{Z}\right) \\
&= \frac{1}{n-1}(\mathbf{Z}-\mathbf{1}\bar{\mathbf{z}}')'(\mathbf{Z}-\mathbf{1}\bar{\mathbf{z}}') \\
&= \frac{1}{n-1}\mathbf{Z}'\mathbf{Z} \\
&= \frac{1}{n-1}\begin{bmatrix}
\frac{(n-1)s_{11}}{s_{11}} & \frac{(n-1)s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \cdots & \frac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} \\
\frac{(n-1)s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \frac{(n-1)s_{22}}{s_{22}} & \cdots & \frac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} \\
\vdots & \vdots & \ddots & vdots \\
\frac{(n-1)s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}} & \frac{(n-1)s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}} & \cdots & \frac{(n-1)s_{pp}}{s_{pp}}
\end{bmatrix} = \mathbf{R}
\end{aligned}
$$

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

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,

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

In addition,

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

and

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

Using which, we see that the proportion of the total sample variance explained by the ith sample principal component is

$$\left(\begin{array}{l} \text{Proportion of (standardized)} \\ \text{sample variance due to } i\text{th} \\ \text{sample principal component} \end{array} \right) = \frac{\hat{\lambda}_i}{p} \qquad i = 1, 2, \ldots, p$$

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: Sample principal components from standardized data The weekly rates of return for five stocks (J P Morgan, Citibank, Wells Fargo, Royal Dutch Shell and Exxon Mobil) 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 as follows:

Stock-price Data (Weekly Rate of Return)

Week	J P Morgan	Citibank	Wells Fargo	Royal Dutch Shell	Exxon Mobil
1	0.013 03	−0.007 84	−0.003 19	−0.044 77	0.005 22
2	0.008 49	0.016 69	−0.006 21	0.011 96	0.013 49
3	−0.017 92	−0.008 64	0.010 04	0	−0.006 14
4	0.021 56	−0.003 49	0.017 44	−0.028 59	−0.006 95
5	0.010 82	0.003 72	−0.010 13	0.029 19	0.040 98
⋮	⋮	⋮	⋮	⋮	⋮
99	0.003 37	−0.015 31	−0.023 82	−0.001 67	−0.017 23
100	0.003 36	0.002 90	−0.003 05	−0.001 22	−0.009 70
101	0.017 01	0.009 51	0.018 20	−0.016 18	−0.007 56
102	0.010 39	−0.002 66	0.004 43	−0.002 48	−0.016 45
103	−0.012 79	−0.014 37	−0.018 74	−0.004 98	−0.016 37

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 J P Morgan, Citibank, Wells Fargo, Royal Dutch Shell and Exxon Mobil, 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 $\mathbf{R}$ 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 $\mathbf{R}$, determined by a computer, are:

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

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 (J P Morgan, Citibank, Wells Fargo) and the oil stocks (Royal Dutch Shell, Exxon Mobil). 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 behaviour has also been suggested by King.

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: 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.[3] The sample mean vector and sample correlation matrix were, respectively,

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

and

$$\mathbf{R} = \begin{bmatrix} 1.0000 & .7501 & .6329 & .6363 \\ .7501 & 1.0000 & .6925 & .7386 \\ .6329 & .6925 & 1.0000 & .6625 \\ .6363 & .7386 & .6625 & 1.0000 \end{bmatrix}$$

The eigenvalues of this matrix are:

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

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 *Brho* may be of the "equal-correlation" form—where all the off-diagonal matrix elements are equal. This notion will be explored further in one of the following examples.

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. □

[3] Data courtesy of J. J. Rutledge

Comment. An unusually small value for the *last* eigenvalue from either the sample covariance or correlation matrix can indicate an unnoticed linear dependancy 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 subset scores and the total score, x_4, is the sum $x_1 + x_2 + x_3$. Then, although the linear combination $\mathbf{ex} = [1, 1, 1, -1]\mathbf{x} = x_1 + x_2 + x_3 - x_4$ is always zero, rounding error in the computation 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.

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.

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, $\hat{y}_{j1}\hat{\mathbf{e}}_1 + \hat{y}_{j2}\hat{\mathbf{e}}_2 + \cdots + \hat{y}_{jq-1}\hat{\mathbf{e}}_{q-1}$ 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.

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: Plotting the principal components for the turtle data We illustrate the plotting of principal components for the data on male turtles discussed in one of the previous examples. 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.

The following figures show the Q–Q plots for $\hat{y}_2$ and $(\hat{y}_1, \hat{y}_2)$. The observation for the first turtle is additionally marked with a square 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 or principal components do not indicate any substantial departures from normality. □

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

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

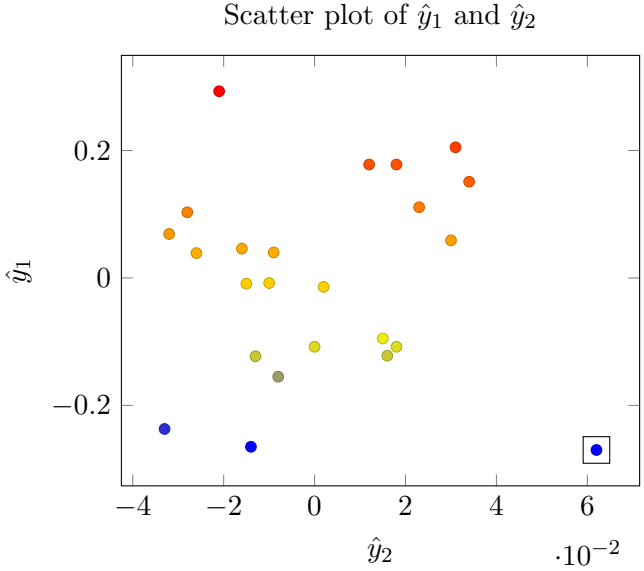
or

$$\underset{(p\times 1)}{\hat{\mathbf{e}}_j} = \underset{(p\times 1)}{\mathbf{y}_j} - \underset{(p\times 1)}{\hat{\boldsymbol{\beta}}'\mathbf{z}_j} \qquad j = 1, 2, \ldots, n$$

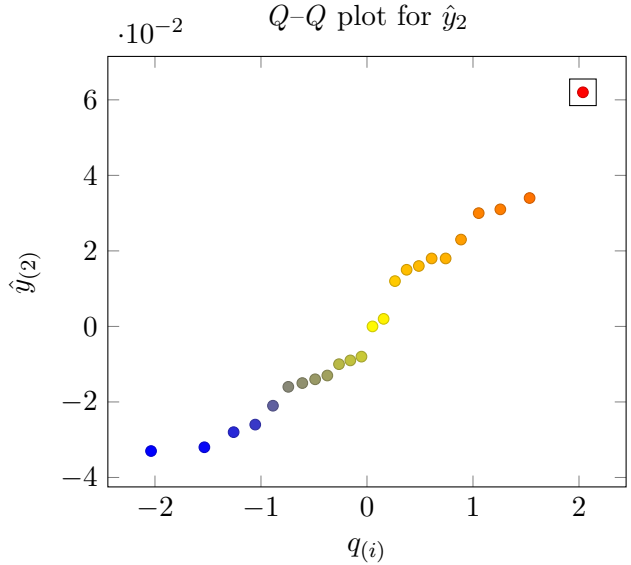
for the multivariate linear model. Principal components, derived from the covariance matrix of hte residuals,

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

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, withinn rounding error.



A scatter plot of the principal components $\hat{y}_1$ and $\hat{y}_2$ of the data on male turtles.



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

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 bases 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. We shall 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 $\hat{\mathbf{e}}_i$.

Andersson and Girshick 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 $\boldsymbol{\Lambda}$ be the diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_p$ of $\boldsymbol{\Sigma}$, then $\sqrt{n}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda})$ is approximately $N_p(\mathbf{0}, 2\boldsymbol{\Lambda}^2)$.

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 $N_p(\mathbf{0}, \mathbf{E})$.

3. Each $\hat{\lambda}_i$ is distributed independently of the associated $\hat{\mathbf{e}}_i$.

Our first result 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 $P[\,|\hat{\lambda}_i - \lambda_i| \leq z(\alpha/2)\lambda_i\sqrt{2/n}] = 1 - \alpha$. A large sample $100(1-\alpha)\%$ confidence interval for λ_i is thus provided by
$$\frac{\hat{\lambda}_i}{(1 + z(\alpha/2)\sqrt{2/n})} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{(1 - z(\alpha/2)\sqrt{2/n})}$$
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 $m\lambda_i$'s are obtained by replacing $z(\alpha/2)$ with $z(\alpha/m)$.

Our second result 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: 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 we explored in one of our previous examples:

Stock-price Data (Weekly Rate of Return)

Week	J P Morgan	Citibank	Wells Fargo	Royal Dutch Shell	Exxon Mobil
1	0.013 03	−0.007 84	−0.003 19	−0.044 77	0.005 22
2	0.008 49	0.016 69	−0.006 21	0.011 96	0.013 49
3	−0.017 92	−0.008 64	0.010 04	0	−0.006 14
4	0.021 56	−0.003 49	0.017 44	−0.028 59	−0.006 95
5	0.010 82	0.003 72	−0.010 13	0.029 19	0.040 98
⋮	⋮	⋮	⋮	⋮	⋮
99	0.003 37	−0.015 31	−0.023 82	−0.001 67	−0.017 23
100	0.003 36	0.002 90	−0.003 05	−0.001 22	−0.009 70
101	0.017 01	0.009 51	0.018 20	−0.016 18	−0.007 56
102	0.010 39	−0.002 66	0.004 43	−0.002 48	−0.016 45
103	−0.012 79	−0.014 37	−0.018 74	−0.004 98	−0.016 37

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 the previous equation, with $i = 1$ to construct a 95% confidence interval for λ_1. The values of $\hat{\lambda}_1$ and $z(.025)$ are as follows:

$$\hat{\lambda}_1 = .0014 \qquad\qquad z(.025) = 1.96$$

Therefore, with 95% confidence,

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

$$.0011 \leq \lambda_1 \leq .0019$$

□

Whenever an eigenvalue is large, such as 100 or even 1 000, the intervals generated by the equation we have just used 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 $\mathbf{\Sigma}$ are not distinct and the previous results do *not* apply. To test for this structure, let

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

and

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

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

Lawley's procedure requires the quantities

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

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

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

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

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: Testing for equicorrelation structure Let us again use the same data on female mice – 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 from the last equation.

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 two equations we have just introduced, we obtain

$$\begin{aligned} \bar{r}_1 &= \frac{1}{3}(.7501 + .6329 + .6363) = .6731, \\ \bar{r}_2 &= .7271, \\ \bar{r}_3 &= .6626, \\ \bar{r}_4 &= .6791 \\ \bar{r} &= \frac{2}{4(3)}(.7501 + .6329 + .6363 + .6925 + .7386 + .6625) = .6855 \end{aligned}$$

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

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

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

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 last test is $\chi^2_5(.05) = 11.07$. 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 before, 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. □

Monitoring Quality with Principal Components

Multivariate control charts, including quality ellipse and the T^2 chart, have already been introduced. 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, 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 $\boldsymbol{\Sigma}$. 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-art 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 $j = 1, 2, \ldots, n$.

We know that 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 reduces to the collection of pairs of possible values $(\hat{y}_1, \hat{y}_2)$ such that

$$\frac{\hat{y}_1^2}{\hat{\lambda}_1} + \frac{\hat{y}_2^2}{\hat{\lambda}_2} \leq \chi_2^2(\alpha)$$

Example: An ellipse format chart based on the first two principal components Here we have a table with police department overtime, from which we render the five normalized eigenvectors and eigenvalues of the sample covariance matrix **S**:

Five Types of Overtime Hours for the Madison, Wisconsin, Police Department				
x_1	x_2	x_3	x_4	x_5
Legal Appearances Hours	Extraordinary Event Hours	Holdover Hours	COA[4] Hours	Meeting Hours
3387	2200	1181	14,861	236
3109	875	3532	11,367	310
2670	957	2502	13,329	1182
3125	1758	4510	12,328	1208
3469	868	3032	12,847	1385
3120	398	2130	13,979	1053
3671	1603	1982	13,528	1046
4531	523	4675	12,699	1100
3678	2034	2354	13,534	1349
3238	1136	4606	11,609	1150
3135	5326	3044	14,189	1216
5217	1658	3340	15,052	660
3728	1945	2111	12,236	299
3506	344	1291	15,482	206
3824	807	1365	14,900	239
3516	1223	1175	15,078	161

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)	0.046	−0.048	0.629	−0.643	0.432
Extraordinary event (x_2)	0.039	0.985	−0.077	−0.151	−0.007
Holdover hours (x_3)	−0.658	0.107	0.582	0.250	−0.392
COA hours (x_4)	0.734	0.069	0.503	0.397	−0.213
Meeting hours (x_5)	−0.155	0.107	0.081	0.586	0.784
$\hat{\lambda}_i$	2,770,226	1,429,206	628,129	221,138	99,824

The first two sample components explain 82% of the total variance.

The sample values for all five components are as follows:

[4] Compensatory overtime allowed.

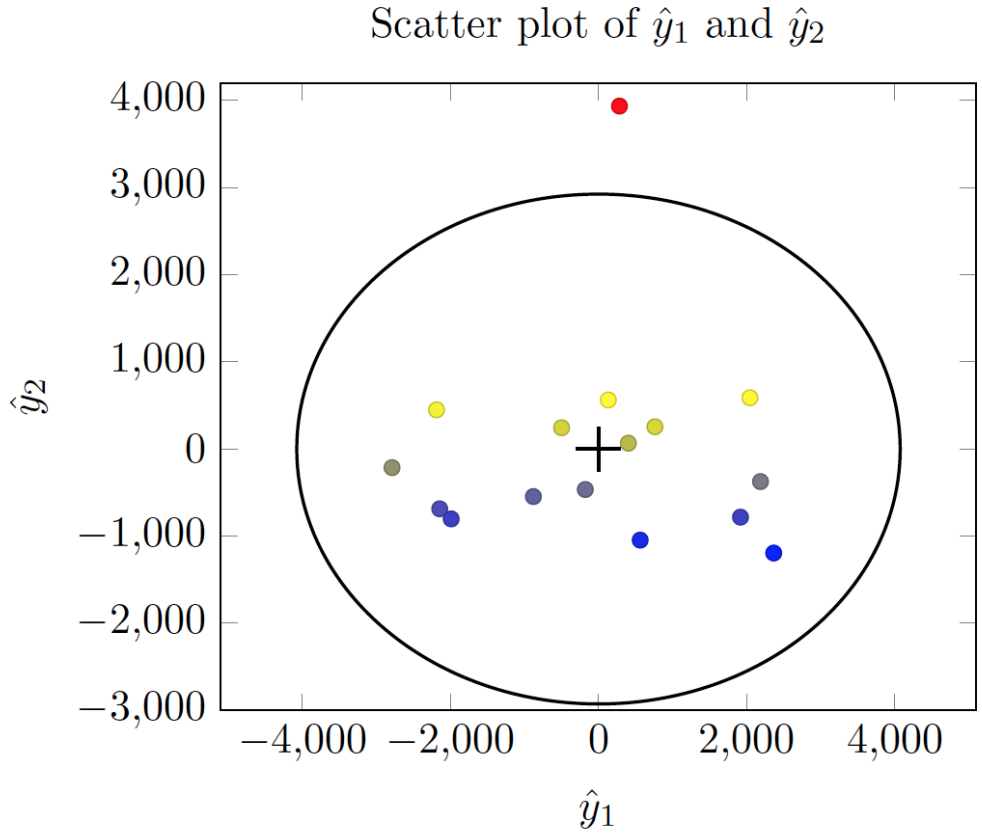
Values of the Principal Components for the Police Department Data					
Period	$\hat{y}_{j1}$	$\hat{y}_{j2}$	$\hat{y}_{j3}$	$\hat{y}_{j4}$	$\hat{y}_{j5}$
1	2044.9	588.2	425.8	−189.1	−209.8
2	−2143.7	−686.2	883.6	−565.9	−441.5
3	−177.8	−464.6	707.5	736.3	38.2
4	−2186.2	450.5	−184.0	443.7	−325.3
5	−878.6	−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

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

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 the figure along with the data.



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

One point is out of control, because the second principal component for this point has a large value. Scanning the table holding the values of the principal components, we see that this is the value 3936.9 for period 11. According to the $\hat{\mathbf{e}}_2$ entries in the table of eigenvectors and eigenvalues above, the second principal component is essentially "extraordinary event overtime hours." The principal component approach has led us to a conclusion that is congruent with the meaning of the data in question—there was a special cause for that overtime: during that period, the United States bombed a foreign capital and students at Madison were protesting. A majority of the extraordinary overtime was used in that four-week period. Although, by its very definition, extraordinary overtime occurs only when special events occur and is therefore unpredictable, it still has a certain predictable stability. □

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

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

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

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

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

$$\frac{Y_3^2}{\lambda_3} + \frac{Y_4^2}{\lambda_4} + \cdots + \frac{Y_p^2}{\lambda_p}$$

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 $\frac{Y_3^2}{\lambda_3} + \frac{Y_4^2}{\lambda_4} + \cdots + \frac{Y_p^2}{\lambda_p}$, 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: A T^2-chart for the unexplained [orthogonal] overtime hours Consider the quality control analysis of the police department overtime hours in the previous example. The first part of the quality monitoring procedure, the quality ellipse based on the first two principal components has already been shown. To illustrate the second step of the two-step monitoring procedure, we create the chart for the other principal components.

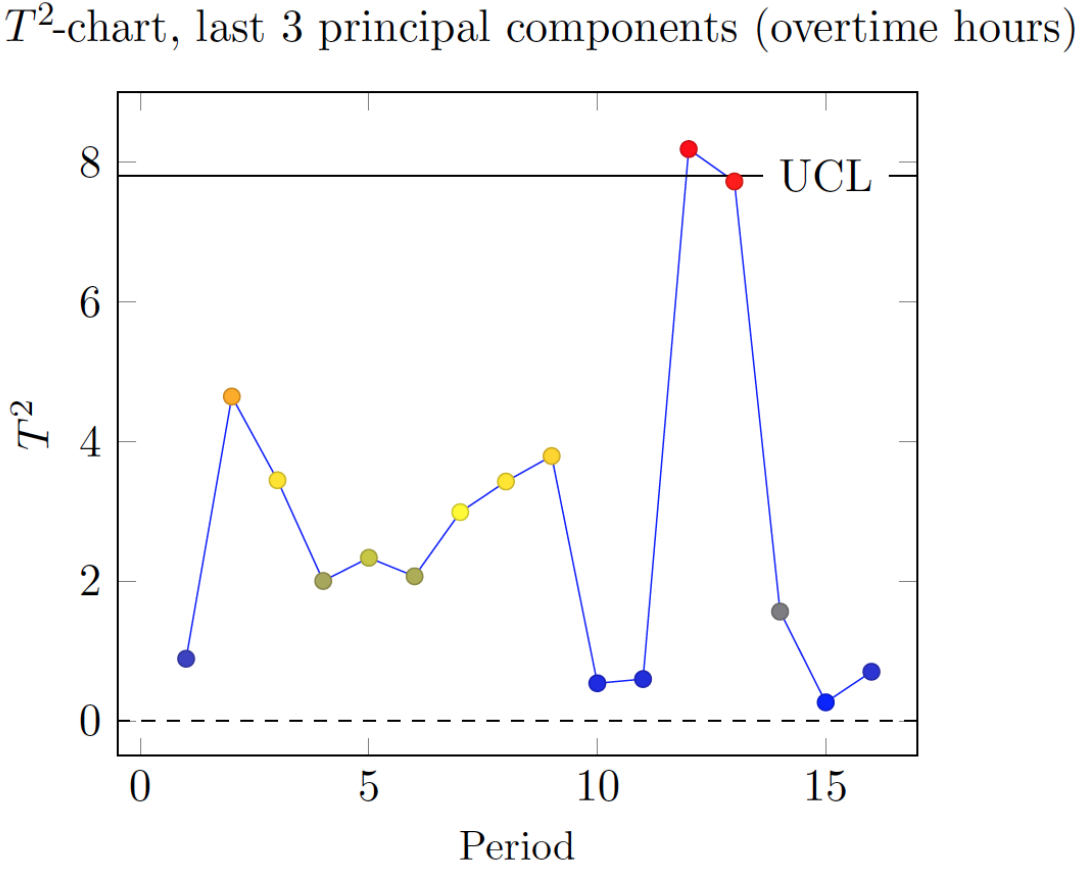
Since $p = 5$, this chart is based on $5 - 2 = 3$ dimensions, and the upper control limit is $\chi^2_3(.05) = 7.81$. Using the eigenvalues and the values of the principal components given in the previous example, we plot the time sequence of values

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

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

Since points 12 and 13 exceed or are near to 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 the table holding the values of the principal components, $\hat{y}_{3j}$ is large in period 12, and from the table of eigenvectors and eigenvalues of sample covariance matrix, the large coeeficients in $\mathbf{e}_3$ belong to legal appearances, holdover and COA hours. Was there some adjusting of these other categories following the period in which extraordinary hours peaked? □



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

Controlling Future Values

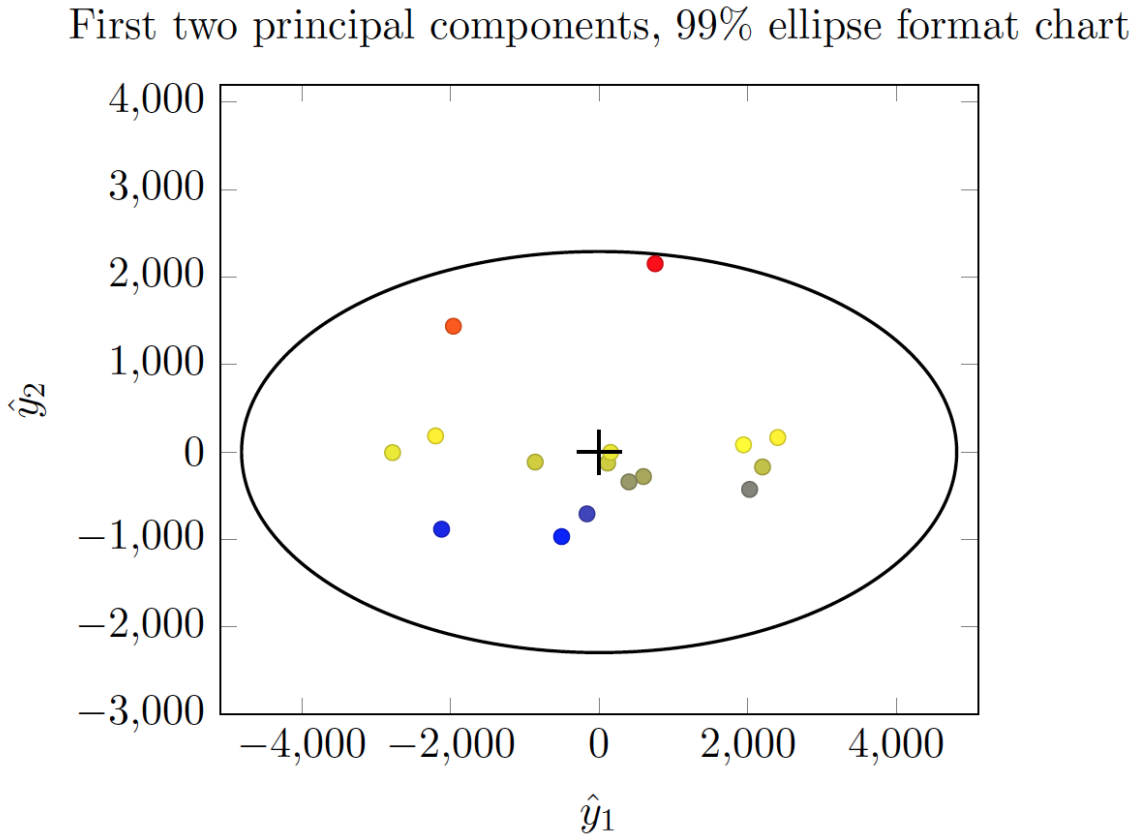
Previously, we considered checking whether a given series of multivariate observations was stable by considering separately the first two principal components and then the last $p-2$. Because the chi-square distribution was used to approximate the UCL of the T^2-chart and the critical distance for the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

Example: Control ellipse for future principal components In one of the previous examples, we termined 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 the table:

Eigenvectors and Eigenvalues from the *15 Stable* Observations					
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)	0.049	0.629	0.304	0.479	0.530
Extraordinary event (x_2)	0.007	−0.078	0.939	−0.260	−0.212
Holdover hours (x_3)	−0.662	0.582	−0.089	−0.158	−0.437
COA hours (x_4)	0.731	0.503	−0.123	−0.336	−0.291
Meeting hours (x_5)	−0.159	0.081	−0.058	−0.752	0.632
$\hat{\lambda}_i$	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 process, in order to set future limits.

The figure gives the 99% prediction ellipse for future pairs of values for the new first two principal components of overtime. The 5 stable pairs of principal components are also shown. □



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 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

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

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

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

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 $\alpha = 0.05$ or 0.01.