Ch 8. Principal Components

- Principal component analysis tries to explain the variance-covariance structure of a set of variables through **a few linear** combinations of these variables.

  - Instead of *p* components to reproduce the total variability, much of this variability can often be accounted for by a smaller number *k* of the principal components.

- Objectives of a principal component analysis

  (1) Data reduction: the *k* principal components can 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.

  (2) Interpretation: an analysis of principal components often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily result.

- Principal component analysis frequently serves as an intermediate step in other investigations such as multiple regression and cluster analysis.

8.2. Population Principal Components

- Principal components are 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 direction with maximum variability and provide a simpler and more parsimonious description of the covariance structure.
- Principal components depend solely on the covariance matrix Σ (or the correlation matrix ρ) of $X_1, X_2, \ldots, X_p$. (They do not require a multivariate normal assumption.)
  - Principal components from multivariate normal populations have useful interpretations in terms of the constant density ellipsoids and allow inferences based on the multivariate normal distribution.

8.2. Population Principal Components

- The random vector $X' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix Σ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

- Consider the linear combinations

$$
\begin{aligned}
Y_1 &= a_1'X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 &= a_2'X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
&\vdots \\
Y_p &= a_p'X = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p
\end{aligned}
$$

Then, $Var(Y_i) = a_i'\Sigma a_i \qquad i = 1,2,\ldots,p$

$Cov(Y_i, Y_k) = a_i'\Sigma a_k \qquad i, k = 1,2,\ldots,p$

- The principal components are **uncorrelated** linear combinations $Y_1, Y_2, \ldots, Y_p$ whose variance are as large as possible.

(1) The first principal component = linear combination $a_1'X$ that maximizes $Var(a_1'X)$ subject to $a_1'a_1 = 1$.

(2) The second principal component = linear combination $a_2'X$ that maximizes $Var(a_2'X)$ subject to $a_2'a_2 = 1$ and $Cov(a_1'X, a_2'X) = 0$.

$\vdots$

(i) The ith principal component = linear combination $a_i'X$ that maximizes $Var(a_i'X)$ subject to $a_i'a_i = 1$ and $Cov(a_i'X, a_k'X) = 0$ for $k < i$.

- Note that since $Var(Y_i) = a_i'\Sigma a_i$ can be increased by multiplying any a_i by some constant, it is convenient to restrict $a_i'a_i = 1$ to eliminate indeterminacy.

8.2. Population Principal Components

- Result 8.1 Let Σ be the covariance matrix associated with the random vector $X' = [X_1, X_2, \ldots, X_p]$. Let Σ have the eigenvalue-eigenvector pairs $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.
  Then the ***i*th principal component** is given by
  $$Y_i = e_i'X = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, \quad i = 1,2,\ldots, p.$$
  With these choices,
  $$Var(Y_i) = e_i'\Sigma e_i = \lambda_i \qquad i = 1,2,\ldots, p$$
  $$Cov(Y_i, Y_k) = e_i'\Sigma e_k = 0 \qquad i \neq k$$
  If some λ_i are equal, the choices of the corresponding coefficient vectors, e_i, and hence Y_i, are not unique.

Proof.

To determine the coefficients that maximize $Var(Y_1)$, introduce the constraint by means of the Lagrange multiplier and differentiate with respect to a_1:

$$\frac{\partial}{\partial a_1}\left[a_1'\Sigma a_1 - \lambda\left(a_1'a_1 - 1\right)\right] = 0.$$

Since $\frac{\partial}{\partial a_1}\left[a_1'\Sigma a_1 - \lambda\left(a_1'a_1 - 1\right)\right] = 2\Sigma a_1 - 2\lambda a_1 = 2\left(\Sigma - \lambda I\right)a_1,$

the coefficients must satisfy the p simultaneous linear equations

$$\left(\Sigma - \lambda I\right)a_1 = 0.$$

If the solution to these equations is to be other than the null vector, the value of λ must be chosen as

$$\left|\Sigma - \lambda I\right| = 0,$$

and thus λ is a eigenvalue of the covariance matrix Σ and a_1 is its associated eigenvector .

8.2. Population Principal Components

- <u>Result 8.1</u>

Proof. (continued)

To determine which of the p eigenvalues should be used, premultiply the equations by a_1':

$$a_1'(\Sigma - \lambda I)a_1 = a_1'\Sigma a_1 - \lambda a_1' a_1 = a_1'\Sigma a_1 - \lambda = 0.$$

It follows that $\lambda = a_1'\Sigma a_1 = Var(Y_1)$. The coefficient vector was chosen to maximize this variance, and λ must be the greatest eigenvalue of Σ.

The coefficients of the ith principal component are found by introducing i constraints by the Lagrange multipliers λ and γs and differentiating with respect to a_i:

$$\frac{\partial}{\partial a_i}\left[a_i'\Sigma a_i - \lambda(a_i'a_i - 1) - \sum_{j<i}\gamma_j a_i'a_j\right] = 0, \text{ where } j < i.$$

Since $\frac{\partial}{\partial a_i}\left[a_i'\Sigma a_i - \lambda(a_i'a_i - 1) - \sum\gamma_j a_i'a_j\right] = 2\Sigma a_i - 2\lambda a_i - 2\sum\gamma_j a_j,$

by setting the equation to zero and premultiplying both terms by a_i', it follows that

$$a_i'(\Sigma - \lambda I)a_i - \sum\gamma_j a_i'a_j = 0.$$

By the constraint, $a_i'a_j = a_j'a_i = 0$ and hence γs are all 0.

It follows that

$$(\Sigma - \lambda I)a_i - \sum\gamma_j a_i = (\Sigma - \lambda I)a_i = 0,$$

meaning that the coefficient of the ith component are the elements of the eigenvector corresponding to the ith greatest eigenvalue.

The eigenvectors of Σ are orthogonal if all the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_p$ are distinct.
Then, $Cov(Y_i, Y_k) = e_i'\Sigma e_k = e_i'\lambda_k e_k = \lambda_k e_i'e_k = 0$ since $e_i'e_k = 0$ for $i \neq k$.

8.2. Population Principal Components

- Result 8.2 Let $X' = [X_1, X_2,\ldots, X_p]$ have covariance matrix Σ, with eigenvalue-eigenvector pairs $(\lambda_1, e_1), (\lambda_2, e_2),\ldots, (\lambda_p, e_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Let $Y_1 = e_1'X$, $Y_2 = e_2'X$,…, $Y_p = e_p'X$ be the principal components. Then

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

Proof.

Note that $\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = tr(\Sigma)$.

With $A = \Sigma$, we can write $\Sigma = P\Lambda P'$ where Λ is the diagonal matrix of eigenvalues and $P = [e_1, e_2,\ldots, e_p]$ so that $PP' = P'P = I$.

Since $tr(\Sigma) = tr(P\Lambda P') = tr(\Lambda P'P) = tr(\Lambda) = \lambda_1 + \lambda_2 + \cdots + \lambda_p$,

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

8.2. Population Principal Components

- Result 8.2 indicates that

$$\text{Total population variance} = \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp}$$
$$= \lambda_1 + \lambda_2 + \cdots + \lambda_p.$$

Therefore, the proportion of total variance due to (explained by) the *k*th principal component is

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

- If most 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.

8.2. Population Principal Components

- The magnitude of e_{ik} from the coefficient vector $e_i' = [e_{i1}, e_{i2}, \ldots, e_{ip}]$ 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 $Y_1 = e_1'X$, $Y_2 = e_2'X, \ldots, Y_p = e_p'X$ are the principal components obtained from the covariance matrix Σ, then

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

are the correlation coefficients between the components Y_i and the variables X_k. Here $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$ are eigenvalue-eigenvector pairs for Σ.

Proof.

Set $a_k' = [0, \ldots, 0, 1, 0, \ldots, 0]$ so that $X_k = a_k'X$ and $Cov(X_k, Y_i) = Cov(a_k'X, e_i'X) = a_k'\Sigma e_i$. Since $\Sigma e_i = \lambda_i e_i$, $Cov(X_k, Y_i) = a_k'\lambda_i e_i = \lambda_i e_{ik}$.

Then $Var(Y_i) = \lambda_i$ and $Var(X_k) = \sigma_{kk}$ yield

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

8.2. Population Principal Components

- Suppose X is distributed as $N_p(\mu, \Sigma)$. The density of X is constant on the μ centered ellipsoids

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

which have axes $\pm c\sqrt{\lambda_i}e_i$, i =1,2,…, p, where the (λ_i, e_i) are the eigenvalue-eigenvector pairs of Σ.

- A point lying on the ith axis of the ellipsoid will have coordinates proportional to $e_i' = [e_{i1}, e_{i2},\ldots, e_{ip}]$ in the coordinate system that has origin μ and axes that are parallel to the original axes $x_1, x_2,\ldots, x_p$.

- When $\mu = 0$, with $A = \Sigma^{-1}$,

$$c^2 = x'\Sigma^{-1}x = \frac{1}{\lambda_1}(e_1'x)^2 + \frac{1}{\lambda_2}(e_2'x)^2 + \cdots + \frac{1}{\lambda_p}(e_p'x)^2,$$

where $e_1'x$, $e_2'x$,…, $e_p'x$ are recognized as the principal components of x. Setting $y_1 = e_1'x, y_2 = e_2'x,\ldots, y_p = e_p'x$,

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

- This equation defines an ellipsoid (since $\lambda_1, \lambda_2, \cdots , \lambda_p$ are positive) in a coordinate system with axes $y_1, y_2,\ldots, y_p$ lying in the direction of $e_1, e_2,\ldots, e_p$, respectively.

- If λ_1 is the largest eigenvalue, then the major axis lies in the direction e_1. The remaining minor axes lie in the direction defined by $e_2,\ldots, e_p$.

8.2. Population Principal Components

- The principal components $y_1 = e_1'x$, $y_2 = e_2'x, \ldots, y_p = e_p'x$ lie in the directions of axes of a constant density ellipsoid.

  - Any point on the ith ellipsoid axis has x coordinates proportional to $e_i' = [e_{i1}, e_{i2}, \ldots, e_{ip}]$, and necessarily, principal component coordinates of the form $[0,\ldots,0, y_i, 0,\ldots,0]$.

- When $\mu \neq 0$, the mean-centered principal component $y_i = e_i'(x - \mu)$ has mean 0 and lies in the direction e_i.
- For a constant density ellipse and the principal components for a bivariate normal random vector with $\mu = 0$ and $\rho = .75$, see Figure 8.1.

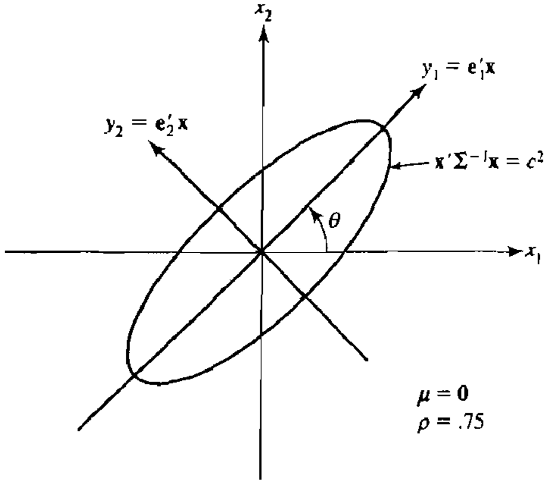

x_2
$y_1 = \mathbf{e}_1'\mathbf{x}$
$y_2 = \mathbf{e}_2'\mathbf{x}$
$\mathbf{x}'\Sigma^{-1}\mathbf{x} = c^2$
θ
x_1
$\mu = 0$
$\rho = .75$

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

$$Z_1 = \frac{(X_1 - \mu_1)}{\sqrt{\sigma_{11}}}$$

$$Z_2 = \frac{(X_2 - \mu_2)}{\sqrt{\sigma_{22}}}$$

$$\vdots$$

$$Z_p = \frac{(X_p - \mu_p)}{\sqrt{\sigma_{pp}}}$$

- In matrix notation,

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

where the diagonal standard deviation matrix $V^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$.

- $E(Z) = 0$ and $Cov(Z) = (V^{1/2})^{-1}\Sigma(V^{1/2})^{-1} = \rho$.

- The principal components of Z may be obtained from the eigenvectors of the correlation matrix ρ of X. All previous results apply.

- The (λ_i, e_i) derived from Σ are, in general, not the same as the ones derived from ρ.

Principal Components Obtained from Standardized Variables

- Result 8.4 The *i*th principal component of the standardized variables $Z' = [Z_1, Z_2, \dots, Z_p]$ with $Cov(Z) = \rho$ is given by

$$Y_i = e_i'Z = e_i'(V^{1/2})^{-1}(X - \mu), \quad i = 1,2,\dots,p.$$

Moreover,

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

and

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

In this case, $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$ are the eigenvalue-eigenvector pairs for ρ, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

- The total (standardized variables) population variance is simply p, the sum of the diagonal elements of the matrix ρ, so

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

where the λ_k's are the eigenvalues of ρ.

Principal Components Obtained from Standardized Variables

- Example 8.2 Principal components obtained from covariance and correlation matrices are different.

  - The principal components derived from Σ are different from those derived from ρ.

  - One set of principal components is not a simple function of the other. (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 measurements are not commensurate.

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.

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

setting $e_i' = [\, e_{i1}, e_{i2}, \ldots, e_{ip} \,] = [\, 0, \ldots 0, 1, 0, \ldots, 0]$, with 1 in the ith position,

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

- (σ_{ii}, e_i) is the ith eigenvalue-eigenvector pair.

- Since the linear combination $e_i'X = X_i$, the set of principal components is just the original set of uncorrelated random variables.

- If X is distributed as $N_p(\mu, \Sigma)$, the contours of constant density are ellipsoids whose axes already lie in the directions of maximum variations and there is no need to rotate the coordinate system.

8.3. Summarizing Sample Variation by Principal Components

- The data $x_1, x_2, \ldots, x_n$ represent n independent samples from p-dimensional population with mean vector μ and covariance matrix Σ.

  - $\bar{x}$: sample mean vector

  S : sample covariance matrix

  R : sample correlation matrix

  - The uncorrelated combinations with the largest variances are called the **sample principal components**.

- The sample principal components are defined as those linear combinations which have maximum sample variance. For $j = 1,2,\ldots, n$,

  (1) first sample principal component = linear combination $a_1'x$ that maximizes the sample variance of $a_1'x$ subject to $a_1'a_1 = 1$;

  (2) second sample principal component = linear combination $a_2'x$ that maximizes the sample variance of $a_2'x$ subject to $a_2'a_2 = 1$ and zero sample covariance for the pairs $(a_1'x, a_2'x)$;

  $\vdots$

  (i) the ith principal component = linear combination $a_i'x$ that maximizes the sample variance of $a_i'x$ subject to $a_i'a_i = 1$ and zero sample covariance for all pairs $(a_i'x, a_k'x)$, $k < i$.

8.3. Summarizing Sample Variation by Principal Components

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

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

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

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

$$\text{Sample covariance}(\hat{y}_i, \hat{y}_k) = 0, \qquad i \ne k.$$

In addition,

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

and

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

8.3. Summarizing Sample Variation by Principal Components

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

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

for any observation vector x.

- The *values* of the ith component

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

is generated by substituting each observation x_j for the arbitrary x, then

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

- The sample variances are still given by the $\hat{\lambda}_i$'s.

The Number of Principal Components

- How many components should be retained?

  - There is no definitive answer.

  - Should consider 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.

  - A component associated with an eigenvalue near zero may indicate an unsuspected linear dependency in the data.

- A usual aid to determining an appropriate number of principal component is a **scree plot**.

  - 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.

  - The number of components is taken to be the point at which the remaining eigenvalues are relatively small and all about the same size (look for an elbow).

  - See Figure 8.2 (p. 445).

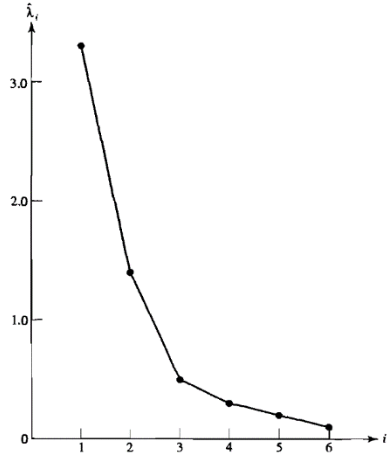


Figure 8.2 A scree plot.

Interpretation of the Sample Principal Components

- Assume $X \sim N_p(\mu, \Sigma)$. Then the sample principal components, $\hat{y}_i = \hat{e}_i'(x - \bar{x})$ are realization of population principal component $y_i = e_i'(x - \mu)$, which have an $N_p(0, \Lambda)$.

  - Note that the diagonal matrix Λ has entries $\lambda_1, \lambda_2, \cdots, \lambda_p$ and (λ_i, e_i) are the eigenvalue-eigenvector pairs of Σ.

- From the sample values x_j, can approximate μ by $\bar{x}$ and Σ by S. If S is positive definite, the contour consisting of all $p\times 1$ vectors x satisfying

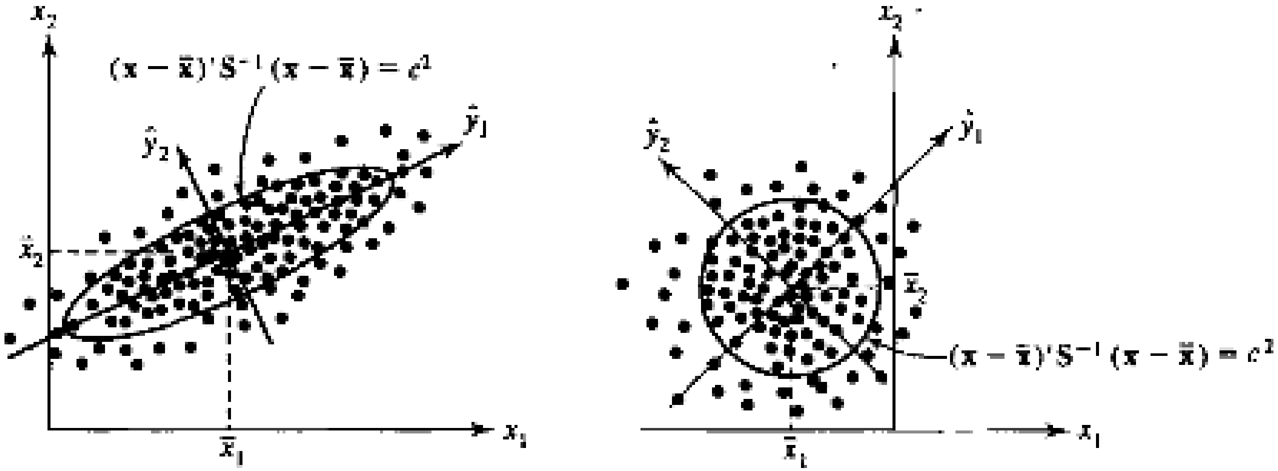
$$(x - \bar{x})' S^{-1} (x - \bar{x}) = c^2$$

  estimates the constant density contour $(x - \mu)' \Sigma^{-1} (x - \mu)$ of the underlying normal density.

  - 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.
  - This contour defines a hyperellipsoid that is centered at $\bar{x}$ and whose axes are given by the eigenvectors of S^{-1} or equivalently, of 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 S.

Interpretation of the Sample Principal Components

- Since $\hat{e}_i$ has length 1, the absolute value of the *i*th principal component, $|\hat{y}_i| = |\,\hat{e}_i'(x-\bar{x})\,|$, gives the length of the projection of the vector $(x-\bar{x})$ on the unit vector $\hat{e}_i$.
- The sample principal components $\hat{y}_i = \hat{e}_i'(x-\bar{x}), i = 1,2,\ldots, p$, lie along the axes of the hyperellipsoid, and their absolute values are the lengths of the projection of $(x-\bar{x})$ in the directions of the axes $\hat{e}_i$.
- The sample principal components can be viewed as the result of translating the origin of the original coordinate system to $\bar{x}$ and then rotating the coordinate axes until they pass through the scatter in the directions of maximum variance (See p. 449, Figure 8.4).
- When the eigenvalues of *S* are nearly equal, the sample variation is homogeneous in all directions and it is 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 $\hat{e}_i$ 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.


x_2
$(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}})=c^2$
$\hat{y}_1$
$\hat{y}_2$
$\bar{x}_2$
$\bar{x}_1$
x_1
x_2
$\hat{y}_2$
$\hat{y}_1$
$\bar{x}_2$
$(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}})=c^2$
$\bar{x}_1$
x_1

(a) $\hat{\lambda}_1 > \hat{\lambda}_2$

(b) $\hat{\lambda}_1 \doteq \hat{\lambda}_2$

Figure 8.4 Sample principal components and ellipses of constant distance.

Standardizing the Sample Principal Components

- Sample principal components are, in general, not invariant with respect to changes in scale.
- If Z_1, Z_2,…, Z_p are standardized observations with covariance R, the ith sample principal component is

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

where $(\hat{\lambda}_i, \hat{e}_i)$ is the ith eigenvalue-eigenvector pair of R with

$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$. Also,

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

$$\text{Sample covariance}(\hat{y}_i, \hat{y}_k) = 0, \qquad i \neq k$$

In addition,

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

and

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

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

Standardizing the Sample Principal Components

- A rule of thumb

  - Retain 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 theoretical support and should not be applied blindly.

- 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.

8.4. Graphing the Principal Components

- Plots of the principal components can reveal suspect observations, as well as provide checks on the assumptions of normality.

  - Since the principal components are linear combinations of the original variables, it is not unreasonable to expect them to be nearly normal.

  - When the first few principal components are used as the input data for additional analyses, it is necessary to verify that they are approximately normally distributed.

- The last principal components can help pinpoint suspect observations.

  - Each observation can be expressed as a linear combination

$$x_j = (x_j'\hat{e}_1)\hat{e}_1 + (x_j'\hat{e}_2)\hat{e}_2 + \cdots + (x_j'\hat{e}_p)\hat{e}_p = \hat{y}_{j1}\hat{e}_1 + \hat{y}_{j2}\hat{e}_2 + \cdots + \hat{y}_{jp}\hat{e}_p$$

  of the complete set of eigenvectors $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_p$ of S.

  - The magnitudes of the last principal components determine how well the first few fit the observations. That is, $\hat{y}_{j1}\hat{e}_1 + \hat{y}_{j2}\hat{e}_2 + \cdots + \hat{y}_{j,q-1}\hat{e}_{q-1}$ differs from x_j by $\hat{y}_{jq}\hat{e}_q + \cdots + \hat{y}_{jp}\hat{e}_p$, 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 contributing to this squared length will be large.

8.4. Graphing the Principal Components

- 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.

$\hat{e}_i$

8.5. Large Sample Inferences

- Decisions regarding the quality of the principal component approximation must be made on the basis of the eigenvalue-eigenvector pairs $\left(\hat{\lambda}_i, \hat{e}_i\right)$ extracted from S or R.

- Because of sampling variation, these eigenvalues and eigenvectors will differ from their underlying population counterparts.

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

- Assume that the observations $X_1, X_2, \ldots, X_n$ are a random sample from a normal population.

- Assume that the (unknown) eigenvalues of Σ are distinct and positive, so that $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$.

1. Let Λ be the diagonal matrix of eigenvalue $\lambda_1, \lambda_2, \cdots, \lambda_p$ of Σ, then $\sqrt{n}\left(\hat{\lambda} - \lambda\right)$ is approximately $N_p(0, 2\Lambda^2)$.

2. Let $E_i = \lambda_i \sum_{\substack{k=1 \\ k \neq i}}^{p} \frac{\lambda_k}{\left(\lambda_k - \lambda_i\right)^2} e_k e_k'$.

   then $\sqrt{n}\left(\hat{e}_i - e_i\right)$ is approximately $N_p(0, E_i)$.

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

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

- Result 1 implies that, for large *n*, the $\hat{\lambda}_i$ are independently distributed.
  - $\hat{\lambda}_i$ has an approximate $N(\lambda_i, 2\lambda_i^2/n)$ distribution.
  - A large sample $100(1 - \alpha)\%$ confidence interval for λ_i is provided by

$$\frac{\hat{\lambda}_i}{\left(1+z\left(\frac{\alpha}{2}\right)\sqrt{\frac{2}{n}}\right)} \le \lambda_i \le \frac{\hat{\lambda}_i}{\left(1-z\left(\frac{\alpha}{2}\right)\sqrt{\frac{2}{n}}\right)}.$$

  - Bonferroni-type simultaneous $100(1 - \alpha)\%$ intervals for m λ_i's are obtained by replacing $z(\alpha/2)$ with $z(\alpha/2m)$.
- Result 2 implies that the $\hat{e}_i$'s are normally distributed about the corresponding e_i's for large samples.
  - The elements of each $\hat{e}_i$ are correlated, and the correlation depends to a large extent on the separation of the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_p$ (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 $(1/n)\hat{E}_i$ where $\hat{E}_i$ is derived from E_i by substituting $\hat{\lambda}_i$'s for the λ_i's and $\hat{e}_i$'s for the e_i's.

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

- Whenever an eigenvalue is large, such as 100 or even 1000, a large sample 100(1 - α)% confidence interval for λ_i can be quite wide, for reasonable confidence levels, even if *n* is fairly large.
  - In general, the confidence interval gets wider at the same rate that $\hat{\lambda}_i$ gets large.
  - Some care must be given in dropping or retaining principal components based on an examination of the $\hat{\lambda}_i$'s.