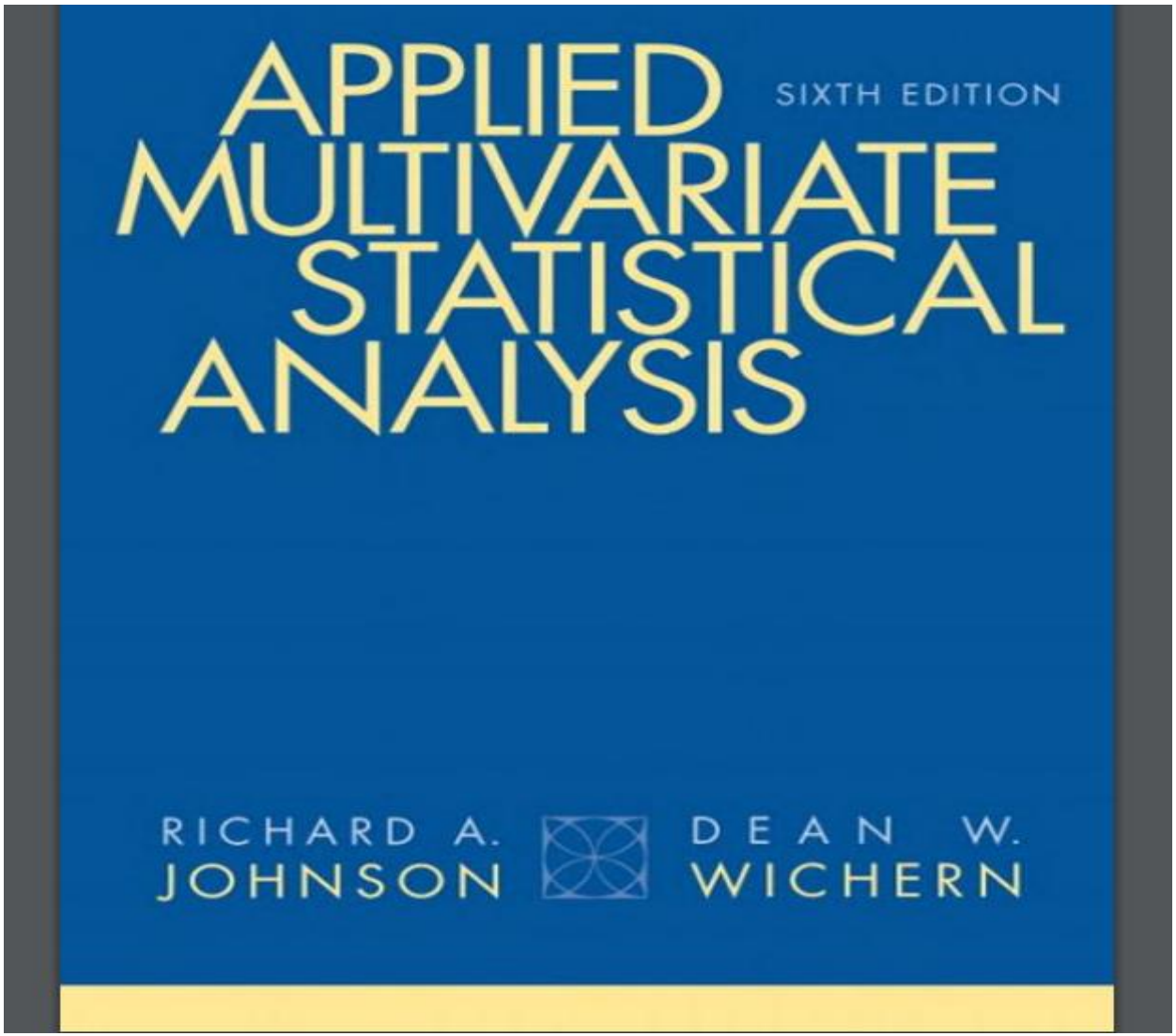
Advanced Multivariate Methods



Review of Factor Analysis

- Factor analysis assumes that the observed variables are linear combinations of some underlying (hypothetical or unobservable) factors.
- The unique factors are then (at least in exploratory factory analysis) assumed to be orthogonal to each other.
- Only common factors (which are assumed much smaller in number than the number of observed variables) contribute to the covariation among the observed variables.

Review of Factor Analysis

- Given these postulates and the properties of linear systems, it is possible to identify exactly the underlying factor pattern from the examination of the resulting covariance structure, provided that the underlying pattern is relatively simple and that it satisfies the requirements of simple factor structure.
- In Figure 1, two-common factor model can be recovered from the error-free correlation matrix shown in in the lower triangle Table I.

Review of Factor Analysis Basics

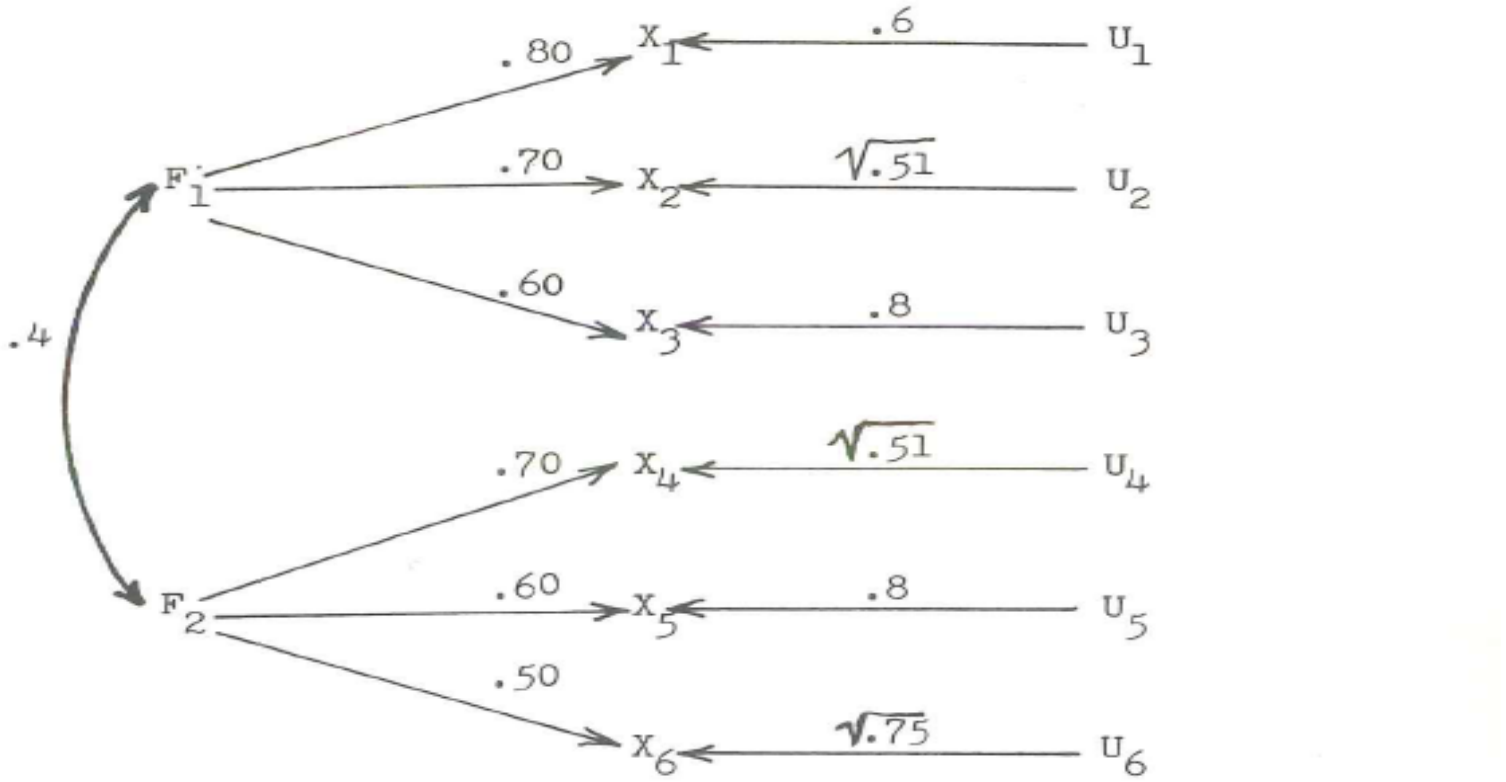


Figure 1: Path Model for Six-Variable, Two Oblique Factor Model Example, where the observed variables represent opinions on:

- X_1 = whether government should spend more money on schools,
- X_2 = whether government should spend more money to reduce unemployment,
- X_3 = whether government should control big business,
- X_4 = whether government should expedite desegregation, through busing,
- X_5 = whether government sees to it that minorities get their respective quota in jobs,
- X_6 = whether government should expand the headstart program.

Review of Factor Analysis Basics

TABLE 1

Correlations for the Population (in the lower triangle) and for a Simulated Sample of 100 Cases (in the upper triangle), Pertaining to the Two-Common Factor Model Represented in Figure 1[a]

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	--	.6008	.4984	.1920	.1959	.3466
X_2	.560	--	.4749	.2196	.1912	.2979
X_3	.480	.420	--	.2079	.2010	.2445
X_4	.224	.196	.168	--	.4334	.3197
X_5	.192	.168	.144	.420	--	.4207
X_6	.160	.140	.120	.350	.300	--

a. Reproduced from Tables 8 and 13 in University Paper 07-013.

Review of Factor Analysis

- There are three steps a researcher usually employs in obtaining solutions to exploratory factor analysis: (1) the preparation of an appropriate covariance matrix; (2) extraction of initial (orthogonal) factorial; (3) and rotation to a terminal solution.

Basic strategies and Methods to be Covered

- The uses of factor analysis are mainly exploratory or confirmatory depending on the major objectives of the researcher.
- In both applications, the three basic steps-of preparing the relevant covariance matrix, extracting initial factors, and rotating to a terminal solution-are implicitly involved.

Basic strategies and Methods to be Covered

- In the initial factoring step, there is the common factor model, serving as model of reference, and principal components, where the underlying rationale is different from the common factor analysis. Both methods are effective, and widely used, means of exploring “interdependence” among variables.
- Principal components are a certain mathematical function of the observed variables while common factors are not expressible by the combination of the observed variables.

Basic strategies and Methods to be Covered

- The rotation step involves two major options—the orthogonal rotation and the oblique rotation. The oblique rotation can be further subdivided into those which are based on the direct simplification of loadings in the factor pattern matrix and the indirect simplification in the factor pattern matrix.
- How many factors to extract and retain will also be examined.

Methods of Extracting Initial Factors

- The main goal is to determine the minimum number of common factors that would satisfactorily produce the correlations among the observed variables.
- Two major types of solution that follow faithfully the common factor models: the maximum likelihood method and the least squares method, whose variants include principal axis factoring with iterated communalities.

Methods of Extracting Initial Factors

- Determine the minimum number of common factors that would satisfactorily produce the correlations among the observed variables.
- Two major types of solution that follow faithfully the common factor models: the maximum likelihood method and the least squares method, whose variants include principal axis factoring with iterated communalities (numbers between 0 and 1 i.e. the sum of the squared loadings for variables across factors).

Principle Components, Eigenvalues, and Vectors

- Principal components analysis is a method of transforming a given set of observed variables into another set of variables.
- First axis is more informative in describing cases as the association between X and Y.
- In extreme cases, the first principal components will contain all the information necessary to describe each case.

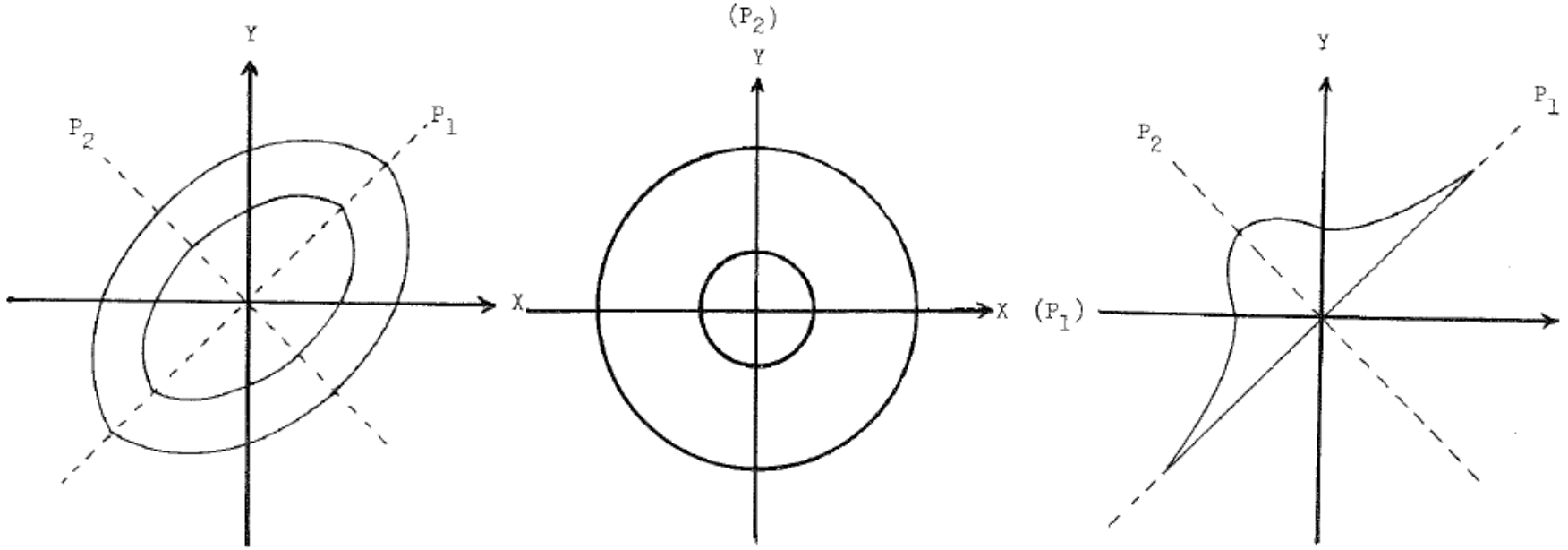
Principle Components, Eigenvalues, and Vectors

- If X and Y are independent, there will be no principal axis and the use of principal components analysis will not provide any economy.

Principle Components, Eigenvalues, and Vectors

- A bivariate normal relationship is illustrated in the next figure by the use of contour maps. These maps show because of the positive relationship between x and y, the data points cluster such that higher values of X tend to be associated with higher values of Y (and vice versa). Therefore, more cases are contained within the 1^{st} and 3^{rd} quadrants than along the 2^{nd} and 4^{th}.
- The principal axes (P1) runs along the line on which the most data points are located; the second axis runs along the line on which the fewest data points are located.

Example of Principal Axes for Bivariate Distributions



a. Some correlation between X and Y

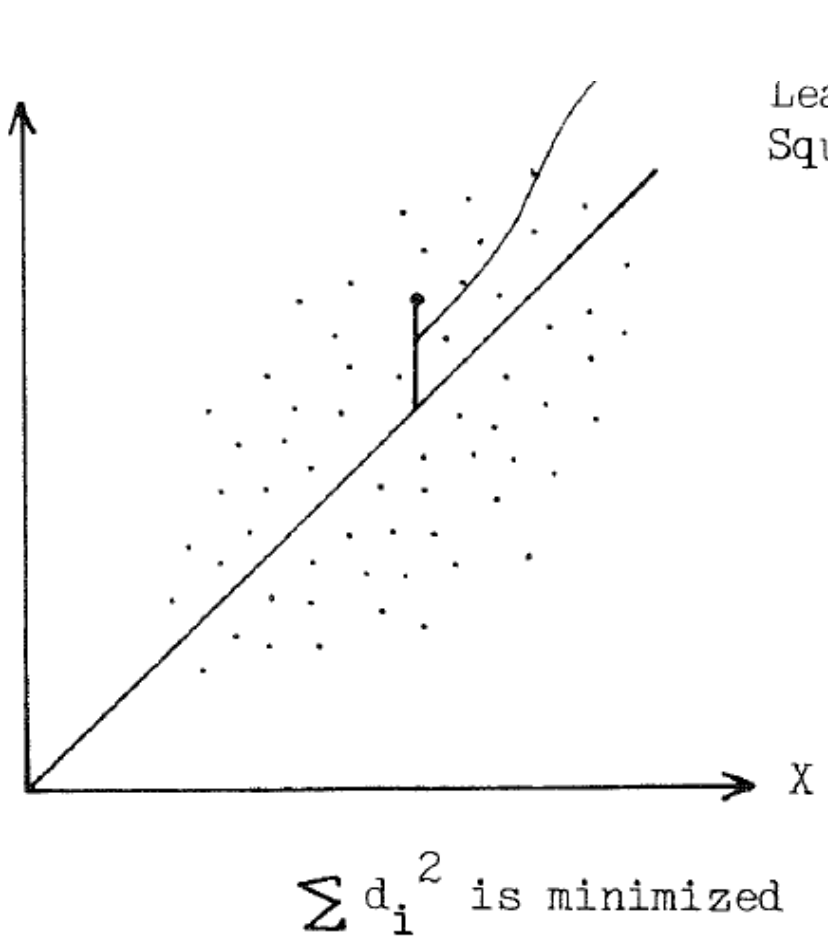
b. No correlation between X and Y

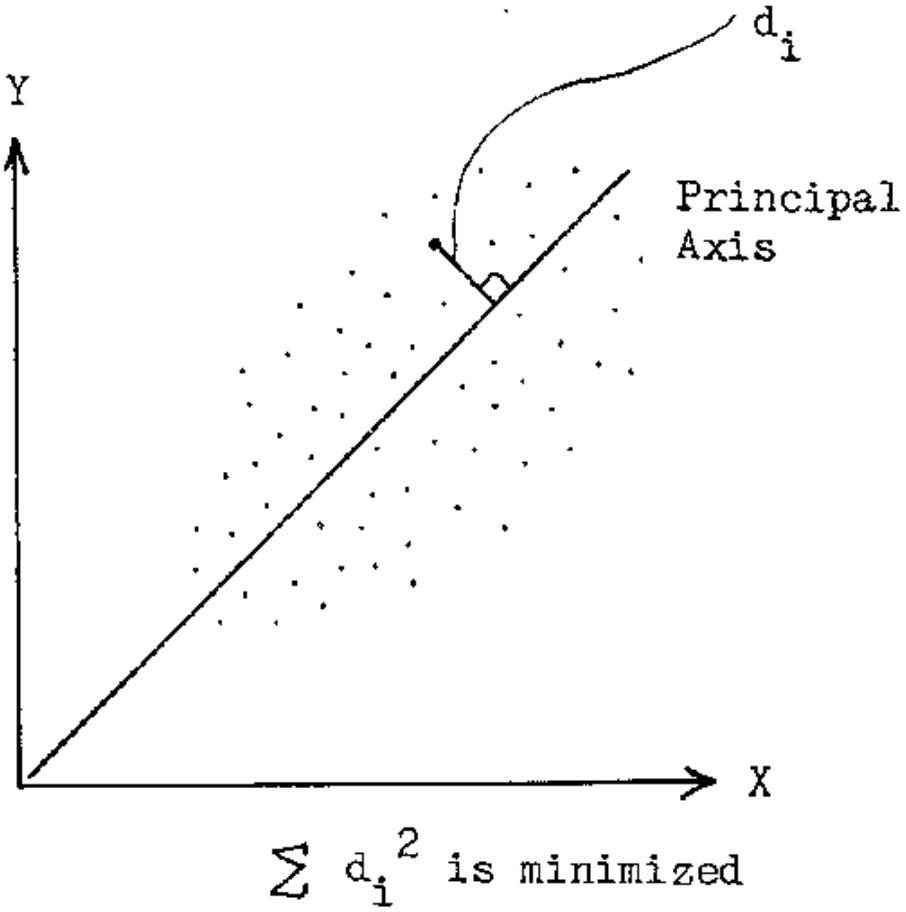
c. Perfect correlation

Principle Components, Eigenvalues, and Vectors

- In general, the principal axis is given by a line from which the sum of the squared distances from each point is a minimum value.
- In finding a least squares regression line Y=mx+b or Y = a + Bx), the sum of the squared distances Y and Y-hat are minimized. Distance is measured by a line parallel to the y-axis and perpendicular to the x-axis.
- In finding a principal axis, the perpendicular distance between the data point and the axis is minimized (the distance is from the point perpendicular to the principal axis not to x.). Next figure depicts in detail.

Comparison Between Least Squares Regression Line and Principal Axis





Principal Components, Eigenvalues and Vectors

The primary mathematical tool by which hierarchical decomposition or transformations are arrived at is referred to as the characteristic equation or eigenequation, where R is the matrix for which a solution is ought, V is the eigenvector to be found and lambda is an eigenvalue.

Principal Components, Eigenvalues and Vectors

- RV = λV
- Det (R - Iλ) = 0 = $Det\begin{pmatrix} 1-\lambda & r_{12} \\ r_{12} & 1-\lambda \end{pmatrix}$
- = (1 - λ)(1 - λ) – $r_{12}(r12)$
- = λ - 2 λ + (1 – r_{12}^2)
- Proportion explained by a given component = (corresponding eigenvalue)/m

Principal Components, Eigenvalues and Vectors

- The eigenvalues can now be obtained if you remember how to solve an equation $ax^2 + bx + c = 0$. The eigenvalues for a bivariate correlation matrix are lambda $_1 = 1 + r_{12}$ and lambda $_2 = 1 - r_{12.}$
- Note: if the correlation between the two variables I perfect, one of the eigenvalues will be 2 and other zero; and that if correlation is zero, both eigenvalues will be 1.

Principal Components, Eigenvalues and Vectors

- Note: Most important is the largest eigenvalue represents the amount of variance explained by the first principal axis; the second largest eigenvalue represents the amount of the variance explained by the second axis, etc.
- Since the sum of all the eigenvalues is equal to the number of variables in the analysis (when the correlation matrix is used), by dividing the first eigenvalue by the m (number of variables), the proportion of variance obtained can also be explained by a given axis or component.

First Two Principal Components of the Correlation Matrix in the Lower Triangle

Variables	Principal Components F_1	F_2	$h^{2\,a}$
X_1	.749	-.395	.713
X_2	.706	-.405	.666
X_3	.651	-.417	.597
X_4	.595	.579	.623
X_5	.548	.529	.581
X_6	.488	.526	.514
Eigenvalues	2.372	1.323	Sum = 3.695
Percent of Variance Explained	39.5	22.1	
Cumulative Percent of Variance Explained	39.5	61.6	

Variants in the Common Factor Model

- Historically speaking, most of the earlier expository treatments of factor analysis identified the common factor model by principal axis factor procedure, which uses the decomposition strategies of PCA as applied to the adjusted correlation matrix whose diagonal elements of 1 are replaced by corresponding estimates of communalities. After inserting communality estimates in the main diagonal of the correlation matrix, factors are extracted like PCA.
- Det (R_1 - Iλ)
- Where R_1 is the correlation matrix with communality estimates in the main diagonal. While this method is still used, it is gradually being replaced by least-squares approach.

Least Squares Approach

- The principal behind the least squares approach to common factor analysis is to minimize the number of factors, and to assess the degree of fit between the reproduced correlations under the model and the observed correlations (the squared differences are examined).
- Procedure is as follows: (1) assume that k factors account for the observed correlations: (2)obtain initial estimates of communalities. (3) extract k factors that can best reproduce the observed correlation matrix (according to the least squares principle).

Solutions Based on the Maximum Likelihood Procedure

- The overall objective of the maximum likelihood solution is the same as the least squares solution: the find the factor solution which would best fit the observed correlations.
- The observed data comprise a sample from a population where k-common factor model exactly applies, and where the distribution of the variables (including the factors) is multivariate normal.

Principal Axis Factoring with Iterated Communalities: Political Opinion Example

Variables	F_1	F_2	h^2
X_1	.731	-.320	.637
X_2	.642	-.282	.492
X_3	.550	-.241	.360
X_4	.513	.473	.487
X_5	.441	.409	.362
X_6	.367	.340	.251
Eigenvalues	1.842	.746	
Percentage Explained	30.7	12.4	

Solutions Based on the Maximum Likelihood Procedure

- Is the previous solution a one- or two-factor model?

Solutions Based on the Maximum Likelihood Procedure

- Alternative solution by Joreskog (1967) not different from the eigenequation solutions is the determinant of R:
- Det $(R_2 - I\lambda) = 0$, [10]
- where $R_2 = U^{-1}(R - U^2)U^{-1}$ [11]
- $= U^{-1}R_1U^{-1}$ [12]
- Where U^2 is the estimate of unique variance at each stage. R_2 is adjusted at every stage in such a way that greater weight is given to correlations involving less unique variance.

Maximum Likelihood Two-Common Factor Solution (Table I Data)

- Sums of squares are equivalent to eigenvalues in the un-rotated solution and this value divided by m gives the proportion of variance explained by that factor. In an obliquely rotated solution, they represent merely might be called a direct contribution of each factor. The joint contribution (including that due to the correlations between the factors) is still equivalent to the sum of eigenvalues in the unrotated solution.

Maximum Likelihood Two-Common Factor Solution Applied to Data in the Upper Triangle

Variables	Unrotated		Communality	Rotated Using Direct Oblimin Criterion	
	F_1	F_2		F_1	F_2
X_1	.747	-.300	.648	.817	-.027
X_2	.701	-.266	.562	.754	-.009
X_3	.599	-.176	.389	.602	.046
X_4	.428	.362	.314	.027	.547
X_5	.505	.605	.621	-.113	.833
X_6	.534	.248	.367	.202	.468
Sum of Squares[a]	2.132	.749		1.652	1.215
χ^2 with 4 degrees of freedom =	.825				

Significance Test Chi-Square

- As expected, the significance test indicates that the fit is adequate. The exact formula for calculating chi-square value is presented to show that the value is dependent on the sample size, while the degrees of freedom are independent of the sample size.
- What is important is for a fixed correlation matrix, the U_k value increases directly proportional to N.

X^2 statistic

$$U_k = N\left\{\ln |C| - \ln |R| + \mathrm{tr}\,(RC^{-1}) - n\right\}$$

ln = natural logarithm, and tr = trace of a matrix
N = the sample size;
n = number of variables;
R = the covariance matrix;
$C = FF' + U^2$, where
F = Factor loadings and U^2, unique variance.

Methods of Rotation

- The initial factoring step usually determines the minimum number of factors that can adequately account for observed correlations, and in the process determines the communalities of each variables. The next step in factor analysis involves finding simpler and more easily interpreted factors through rotations, while keep the number of factors and communalities of each variable fixed.

Methods of Rotation

- There are three different approaches: (1) examine the pattern of variables graphically and then rotate the axis or define new axes in such a way that best satisfy one's criterion for simple and meaningful structure; (2) rely on analytic rotation method, i.e. orthogonal rotation and the other oblique rotation. (3) define a target matrix or configuration before actual rotation. The objective is find the factor patterns that are closest to the given target matrix, presumption based on hypothesis about the nature of the factor structure, approaching confirmatory factor analysis.

Graphics Rotation, Simple Structure, and Reference Axes

- The goal of all rotations is to achieve the simplest possible factor structure. Using criterion from Mulaik (1972):
- (1) Each row of the reference structure matrix V should have at least one zero.
  (2) For each column of K of the reference-structure matrix V, there should be at least r linearly independent observed variables whose correlations (as found in the kth column of V) with the kth reference-axis variable are zero. This criterion is needed to overdetermine the corresponding reference axis.

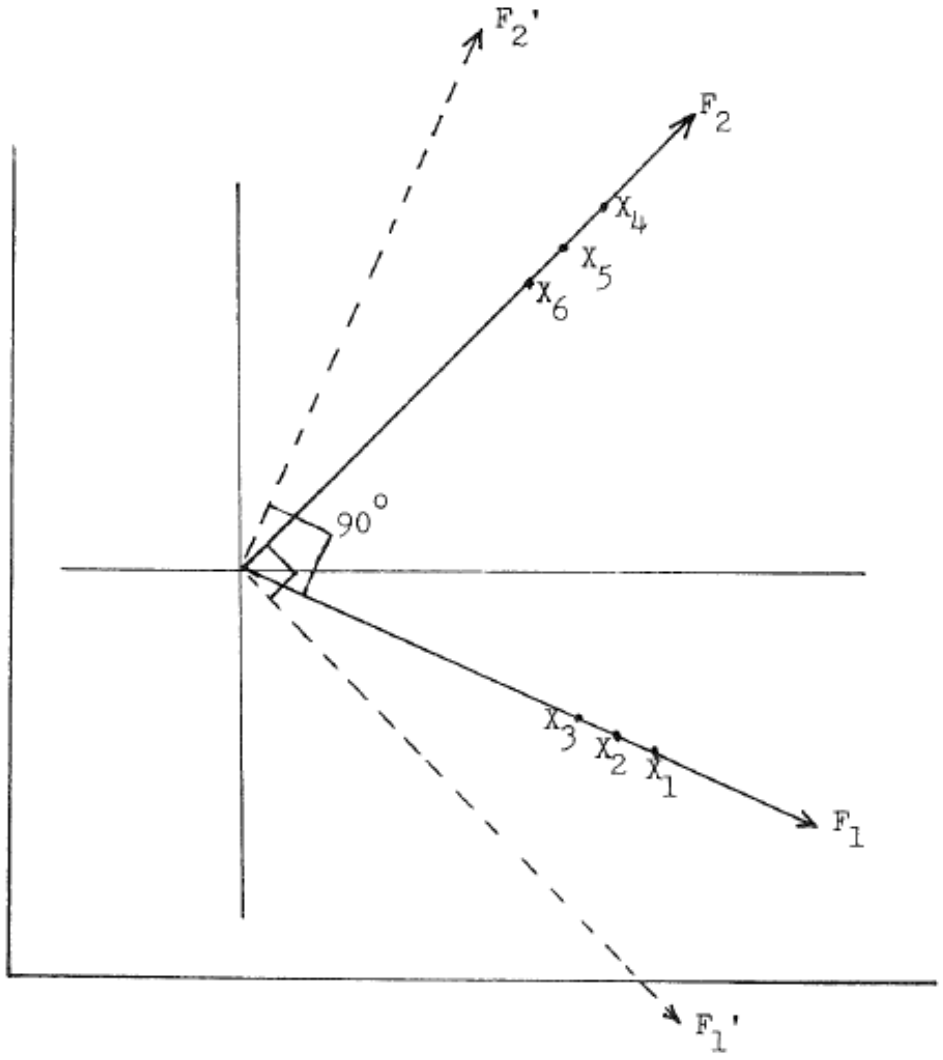
Graphics Rotation, Simple Structure, and Reference Axes

- (3) For every pair of columns of V there should be several zero entries in one column corresponding to nonzero entries in the other. This requirement assures the distinctness of the reference axes and their corresponding sub-spaces of r-1 dimensions of the common-factor space.
- (4) When four or more common factors are obtained, each pair of columns V should have a proportion of corresponding zero entries. This requirement assures that each reference axis is pertinent to only a few of the observed variables and thus guarantees a separation of observed variables in distinct clusters..
- (5) For every pair of columns V there should be only a small number of corresponding entries in both columns which do not vanish. This criterion further ensures the simplicity of the variables.

Graphics Rotation, Simple Structure, and Reference Axes

- Note that in the orthogonal case, the notion of simply structure implies that one set of points will have zero loadings (zero projection) on the other axis or factor. The zero projection is precluded if the angle between the clusters is not orthogonal (i.e. not 90 degrees). Given such an oblique angle, a new procedure is to set up another reference axis that is perpendicular to the hyper-plane (in this two factor model it is simply a line) that passes through the cluster of points which one considers to be a primary factor axis.
- Thus, Factor 1 and Factor 2 are primary oblique factors and F_1 prime and F_2 prime are corresponding reference axes. The projections of X_1, X_2 and X_3 will be zero loadings on F_2 and the projections of X_4, X_5, and X_6 will be zero on F_1

Primary Factor Axis Example



Methods of Orthogonal Rotation: Quartimax, and Varimax

- Factorial complexity of a variable I is expressed where r is the number of columns in a pattern matrix, bij is the factor loading of variable I on the factor j, and b-bar ij squared is the mean of the squared factor loadings for the rows. The equation can be re-written to isolate qi. Once an initial factor solution is given, both r and the communality of each variable are fixed. Hence, the term after the minus sign remains fixed because it is an orthogonal solution. Then the overall measure of simplicity can be obtained by summing q for all the variables. Application of the quartimax criterion results in rotating axes in such a way that factor loadings maximize q.

Methods of Orthogonal Rotation: Quartimax and Varimax

Factorial Complexity of a Variable i

$$= \frac{1}{r} \sum_{j=1}^{r} (b_{ij}^2 - \bar{b}_{ij}^2)^2, \qquad [19]$$

$$q_i = \frac{\sum_{j=1}^{r} (b_{ij}^4) - \left(\sum_{j=1}^{r} b_{ij}^2 \right)^2}{r^2}. \qquad [20]$$

$$\sum_{j=1}^{r} b_{ij}^2 = h_i^2$$

$$q = \sum_{i=1}^{n} q_i = \sum_{i=1}^{n} \frac{\sum_{j=1}^{r} (b_{ij}^4) - \left(\sum_{j=1}^{r} b_{ij}^2 \right)^2}{r^2}. \qquad [21]$$

Methods of Orthogonal Rotation: Quartimax and Varimax

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{r} b_{ij}^4,$$

- Application of the quartimax criterion results in rotating axes in such a way that the factor loadings maximize q. Maximization of q is equivalent to the above equation, because the remaining terms are all constants. Hence, name quartimax.

Methods of Orthogonal Rotation: Quartimax and Varimax

- The varimax rotation uses a slightly different criterion which simplifies each column of the factor matrix.

$$v_j = \frac{n \sum_{i=1}^{n} b_{ij}^4 - \left(\sum_{i=1}^{n} b_{ij}^2 \right)^2}{n^2} \qquad [23]$$

$$\left(\sum_{j=1}^{n} b_{ij}^2 \right)$$

$$V = \sum_{j=1}^{r} v_j = \frac{\sum_{j=1}^{r} n \sum_{i=1}^{n} b_{ij}^4 - \left(\sum_{i=1}^{n} b_{ij}^2 \right)^2}{n^2}, \qquad [24]$$

Methods of Orthogonal Rotation: Quartimax and Varimax

- Instead of maximizing the variance of squared loadings for each variable, it maximizes the variance of the squared loadings for each factor. The quantity to maximize—the index of simplicity of a factor j is then, i.e row varimax criterion.

$$v_j = \frac{n \sum_{i=1}^{n} b_{ij}^4 - \left(\sum_{i=1}^{n} b_{ij}^2 \right)^2}{n^2} \qquad [23]$$

$$\left(\sum_{j=1}^{n} b_{ij}^2 \right)$$

$$V = \sum_{j=1}^{r} v_j = \frac{\sum_{j=1}^{r} n \sum_{i=1}^{n} b_{ij}^4 - \left(\sum_{i=1}^{n} b_{ij}^2 \right)^2}{n^2}, \qquad [24]$$

Varimax and Quartimax Rotations Applied to the Same Pattern Matrix

Variables	Varimax Rotation		Quartimax Rotation	
	F_1	F_2	F_1	F_2
X_1	.787	.167	.793	.133
X_2	.730	.170	.736	.143
X_3	.595	.187	.602	.166
X_4	.154	.539	.173	.533
X_5	.083	.783	.111	.780
X_6	.306	.503	.324	.492

Methods of Oblique Rotation

- An oblique rotation is more general than an orthogonal rotation in that it does not arbitrarily impose the restriction that factors be uncorrelated. Its advantage over orthogonal rotations is that, after making oblique rotations, if the resulting factors are orthogonal, one can be sure that the orthogonality is not an artifact of the method of rotation. Further, high-order factorial causation may need to be assumed to explain correlations among the factors.
- There are two different types of oblique rotation—one uses reference axes and the other uses the primary pattern matrix.

Solutions Based on Reference Axes

- All solutions included are based on the fact that if there are definable clusters of variables representing separate dimensions, and if these clusters are correctly identified by primary factors, each cluster of variables will have near-zero projections on all reference axes but one. The quartimin criterion is as follows:

Solutions Based on Reference Axes

- Where a_{ij} and a_{ji} are projections on the j^{th} and k^{th} reference axes. This value will be zero if all the variables are uni-factorial. But the goal is to minimize N. (In the orthogonal rotations, this criterion is equivalent to the quartimax.)
- Parallel to the Varimax modification of the quartimax criterion in orthogonal rotations, there is a covarimin or biquartimin criterion. The minimized value in this case is the covariance of the squared elements of the projections on the reference axes.

Solutions Based on Reference Axes

- The covariance criterion tends to produce fewer oblique factors while the quartimin criterion produces more oblique factors. Given the opposite tendencies shown by both these two criteria, they are combined.
- Then by multiplying the combined equation by n, combining the terms and setting lambda = Beta (alpha + Beta), the oblimin criterion is yielded as follows:

Solutions Based on Reference Axes

$$N = \sum_{i=1}^{n} \sum_{j<k=1}^{r} a_{ij}^2 \, a_{ik}^2 , \qquad [27]$$

$$C = \sum_{j<k=1}^{r} \left(n \sum_{i=1}^{n} a_{ij}^2 \, a_{ik}^2 - \sum_{i=1}^{n} a_{ij}^2 \sum_{i=1}^{n} a_{ik}^2 \right). \qquad [28]$$

$B = \alpha N + \beta C / n$ = minimum where α and β are weights to be assigned and N and C are given above. [29]

$$B = \sum_{j<k=1}^{r} \left(n \sum_{i=1}^{n} a_{ij}^2 \, a_{ik}^2 - \gamma \sum_{i=1}^{n} a_{ij}^2 \sum_{i=1}^{n} a_{ik}^2 \right). \qquad [30]$$

Solutions Based on Reference Axes

This general criterion reduces to:

Quartimin when $\gamma = 0$ (most oblique)
Biquartimin when $\gamma = .5$ (less oblique)
Covarimin where $\gamma = 1$ (least oblique).

- An attempt to objectify the choice of lambda in the preceding equation as a means of correcting for too oblique bias of quartimin and too orthogonal of covarimin is binormamin. Compared to biquartimin, which adopts ½ for lambda, the binormamin is reported to be more satisfactory if the data are particularly simple or particularly complex.

Number of Factors Problem Revisited

- Reasons for dissatisfaction with the number of factors: (1) initial solutions not satisfactory; (2) imperfect fit between the factor analytic model and the data.
- Rules for questioning the number of factors: (1) significance tests associated with the maximum likelihood and least squares solutions; (2) varieties of eigenvalue criterion; (3) substantive importance; (4) Scree Test; (5) criterion of interpretability and invariance.

Significance Tests

- The large sample Chi-square test associated with the maximum likelihood solution is the most satisfactory solution.
- Maximum likelihood criterion is most appropriate when applied to known population models without substantively insignificant minor factors.
- The method may be too costly (time consuming, expensive) on a large set of variables. Therefore, consider decreasing the number of variables for testing (based on theoretical construct).

Eigenvalue Specification

- One of the most popular criteria for addressing the number of factors is to retain factors with eigenvalues greater than one, when the correlational matrix is decomposed.
- Another related eigenvalue criterion is that of retaining vectors with eigenvalues greater than zero, when the reduced correlation matrix is decomposed. The rationale applies to a population correlation matrix, due to large N, but not hold the sample correlation matrix. When the communalities are estimated and inserted in the main diagonal, this eigenvalue criterion can be applied.

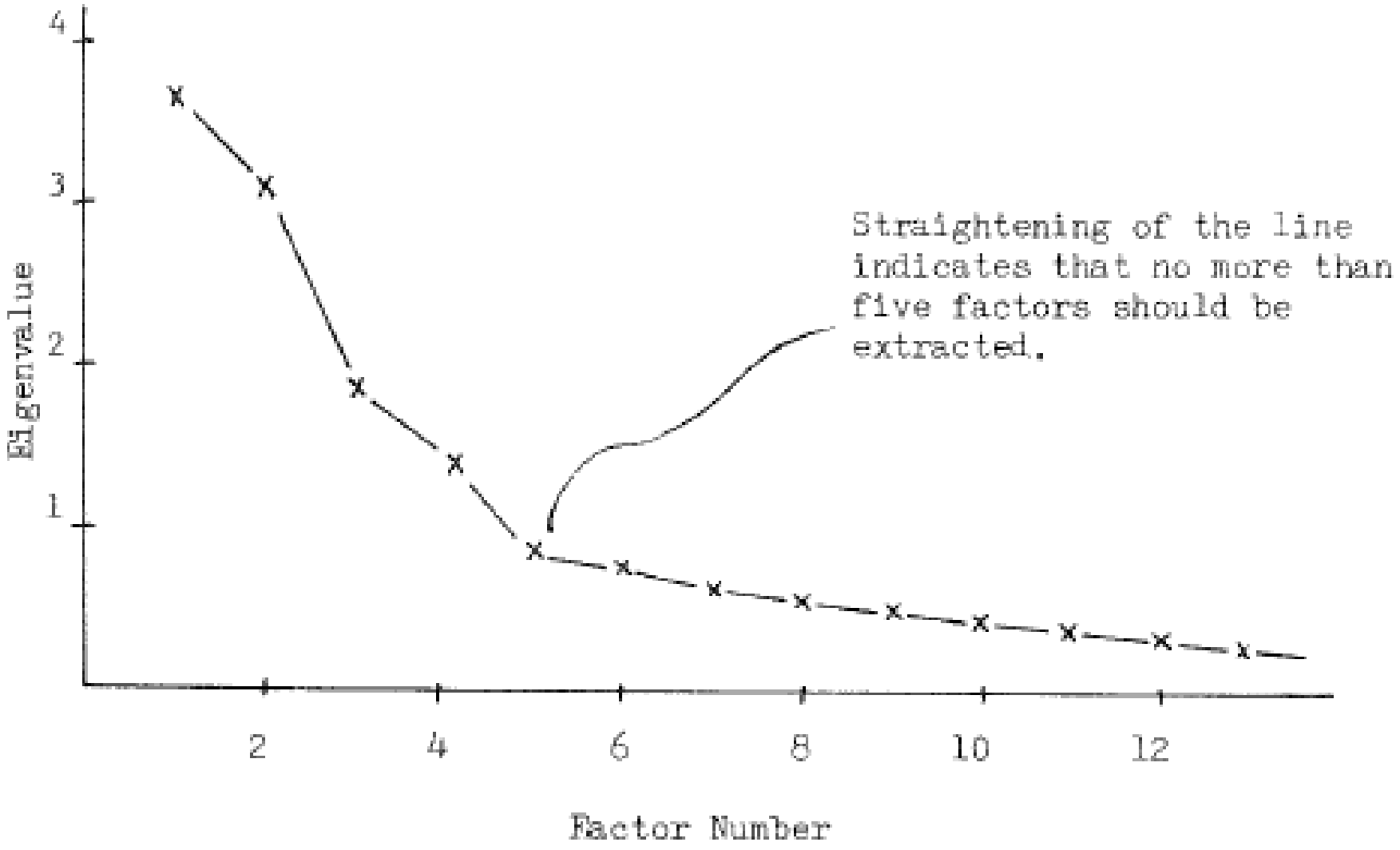
Criterion of Substantive Importance

- Considering that the significance tests focus on sampling variability and the eigenvalue criterion focuses on some abstract properties of a matrix, a third alternative is to focus directly on what should be considered a minimum contribution by a factor to be evaluated as substantively significant.
- This initial factoring criterion is based on the decomposition of the unaltered correlation matrix; the proportion to specify is the proportion of the total variance (which is the number of variables) to be explained by the last factor to be retained.

Scree-Test

- This test advocated by Cattell (1965), directs one to examine the graph of the eigenvalues, and stop factoring at the point where the eigenvalues (or characteristic roots) begin to level forming a straight line with an almost horizontal slope.
- Beyond this point, Cattell describes the smooth slope as “factorial litter or scree” (the geological term referring to debris that collects on the lower part of a rocky slope).

Illustration of Scree-Test



Criteria of Interpretability and Invariance

- As a way to protect oneself from accepting results which are dubious, a general rule of thumb is to try to combine various rules, and accept only those conclusions that are supported by several criteria.
- Employ the standards of scholarship in the field. Refer to the literature for precedent.

Principal Components: Variance-covariance Structure for Data Reduction and Interpretation

PRINCIPAL COMPONENTS

8.1 Introduction

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

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

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

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

Principal Components: Linear Combinations of p Random Variables Selecting a New Coordinate System By Rotating Axes

8.2 Population Principal Components

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

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

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

Consider the linear combinations

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

Principal Components Uncorrelated Linear Combinations With Maximized Variance

Then, using (2-45), we obtain

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

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

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

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

First principal component = linear combination $\mathbf{a}_1'\mathbf{X}$ that maximizes $\text{Var}(\mathbf{a}_1'\mathbf{X})$ subject to $\mathbf{a}_1'\mathbf{a}_1 = 1$

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

Let ∑ Be the Covariance Matrix Associated with Random Vector X' With Eigenvalue-eigenvector pairs

At the ith step,

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

Result 8.1. Let $\boldsymbol{\Sigma}$ be the covariance matrix associated with the random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$. Let $\boldsymbol{\Sigma}$ 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

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

With these choices,

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

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

Principal Components Uncorrelated With Variances Equal to Eigenvalues of ∑

From Result 8.1, the principal components are uncorrelated and have variances equal to the eigenvalues of $\mathbf{\Sigma}$.

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

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

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

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

Thus,

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

Proportion of Total Population Variance Due to the kth Principal Component

Result 8.2 says that

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

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

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

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

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

Principal Components Obtained from the Covariance Matrix ∑

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

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

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

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

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

Correlations Only Explain Univariate Contribution to Help Interpret the Components; Only the Coefficients e_{ik} Be Used to Interpret Components

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

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

Example: Calculating Population Principal Components

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

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

It may be verified that the eigenvalue–eigenvector pairs are

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

Therefore, the principal components become

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

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

Example: Calculating Population Principal Components

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

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

It is also readily apparent that

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

It is also readily apparent that

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

Variance Accounted for By Each Component: Sum of the Eigenvalue-Eigenvector pairs Divided by Sum of the Covariance Matrix Diagonal

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

Next, using (8-8), we obtain

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

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

Evaluating the Weight of the Principal Components

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

Finally,

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

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

Principal Components Derived from Multivariate Normal Random Variables μ = 0

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

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

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

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

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

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

Principal Components Lie in Direction of the Axes of a Constant Density Ellipsoid

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

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

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

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

Constant Density Ellipse and Principal Components for Bivariate Normal Vector μ = 0 and p = .75

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

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

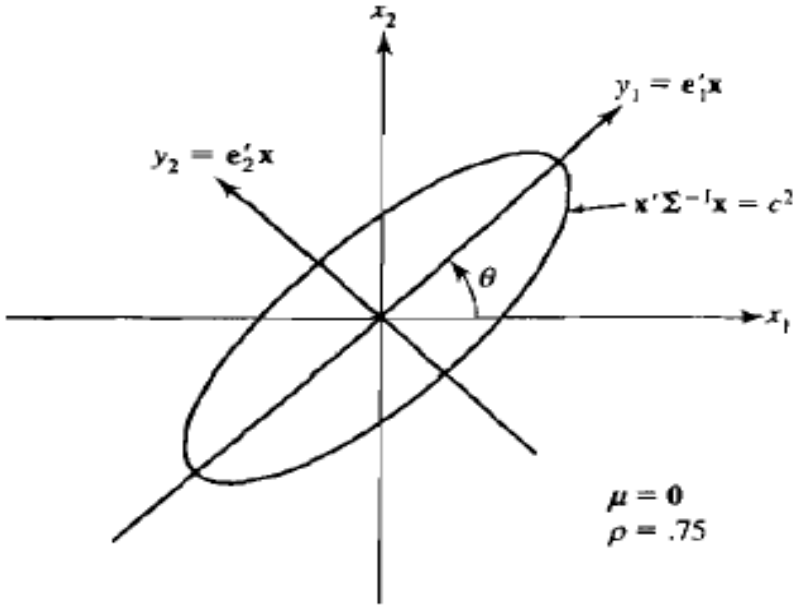


Figure 8.1 The constant density ellipse $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} = c^2$ and the principal components y_1, y_2 for a bivariate normal random vector $\mathbf{X}$ having mean $\mathbf{0}$.

Principal Components Obtained from Standardized Variables i.e. Z-Scores

Principal Components Obtained from Standardized Variables

Principal components may also be obtained for the standardized variables

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

Principal Components Obtained from Eigenvectors of Correlation Matrix p of X

In matrix notation,

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

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

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

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

Principal Component – Eigenvalue-Eigenvector Paris for p

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

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

Moreover,

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

and

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

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

Proportion of Standardized Population Variance Due to kth Principal Component = $\frac{\lambda_k}{p}$

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

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

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

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

Example: Principal Components Obtained from Covariance & Correlation Matrices Are Different

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

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

and the derived correlation matrix

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

The eigenvalue–eigenvector pairs from Σ are

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

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

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

Principal Components: Variables That Dominate a Component

The respective principal components become

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

and

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

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

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

of the total population variance.

When Variables Standardized, Resulting Variables Contribute Equally to Principal Components

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

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

and

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

In this case, the first principal component explains a proportion

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

of the total (standardized) population variance.

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

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

Preceding Example Shows Principal Components Derived from ∑ Different From Those Derived from p or Correlation

The preceding example demonstrates that the principal components derived from Σ are different from those derived from ρ. Furthermore, one set of principal components is not a simple function of the other. This suggests that the standardization is not inconsequential.

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

Principal Components for Covariance Matrices with Special Structures

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 Σ is the diagonal matrix

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

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

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

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.

If Variable X is Normal, Contours of Constant Density Are Ellipsoids Whose Axes Lie in Direction of Maximum Variation; Thus No Need to Rotate the Coordinate System

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

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

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

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

The resulting correlation matrix

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

P Eigenvalues of Correlation Matrix Divided into Two Groups

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

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

with associated eigenvector

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

The remaining $p - 1$ eigenvalues are

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

and one choice for their eigenvectors is

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

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

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

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

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

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

Standardized Variables Multivariate Normal Distribution With Covariance Matrix, Then Ellipsoids of Constant Density Are Cigar Shaped With Major Axis Proportional to 1st Principal Component

The first principal component

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

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

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

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

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

Summarizing Sample Variation by Principal Components

8.3 Summarizing Sample Variation by Principal Components

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

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

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

Recall that the n values of any linear combination

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

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

Sample Principal Component Defined As Linear Combination Maximum Sample Variance

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

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

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

At the ith step, we have

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

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

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

Correlation Equal to Eigenvector x Square Root of Eigenvalue Divided by the Square Root of the Variance

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

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

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

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

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

In addition,

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

and

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

Sample Mean of Each Principal Component Equal to Zero

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 **S** or **R**.[2] The components constructed from **S** and **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 **S** and gives the ith principal component

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

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

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

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

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

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

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

Summarizing Sample Variability With Two Sample Principal Components

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

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

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?

Principal Component Coefficients

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. ■

Examine Both Eigenvectors and Correlations To Interpret and Decide How Many Principal Components to Retain

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

The Number of Principal Components

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

Scree Plot: Visual Aid To Determine How Many Principal Components

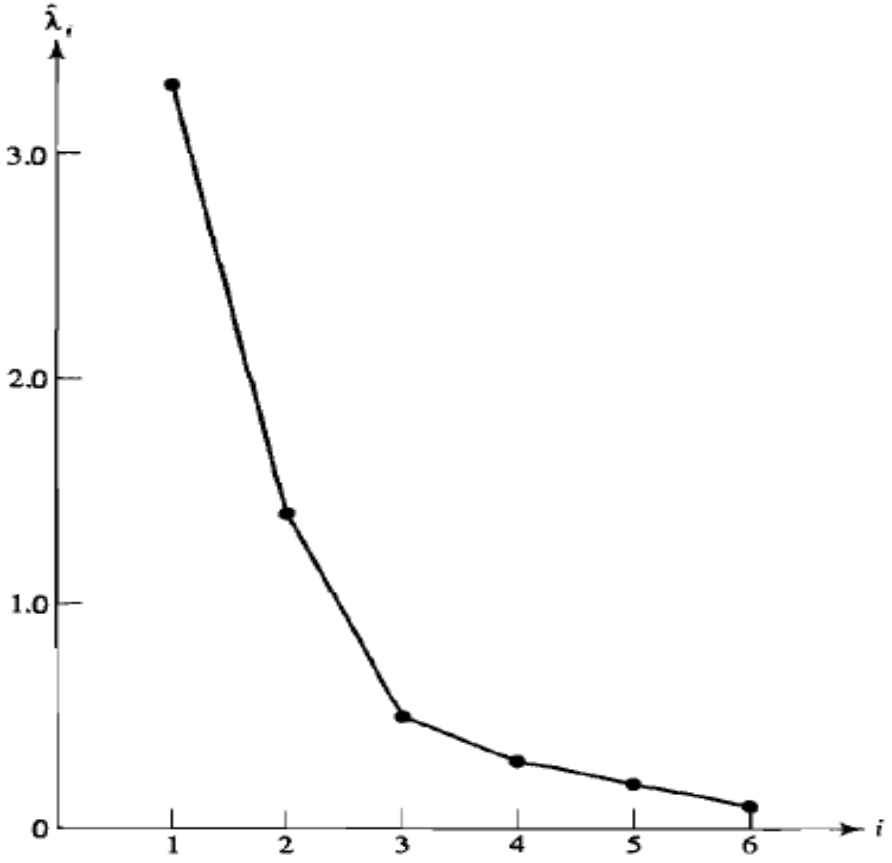


Figure 8.2 A scree plot.

A useful visual aid to determining an appropriate number of principal components is a *scree plot*.[3] With the eigenvalues ordered from largest to smallest, a scree plot is a plot of $\hat{\lambda}_i$ versus i—the magnitude of an eigenvalue versus its number. To determine the appropriate number of components, we look for an

Scree Plot: Eigenvalues Ordered From Largest to Smallest

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

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

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

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

Summarizing Sample Variability with One Principal Component

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

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

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

Coefficients for the Principal Components
(Correlation Coefficients in Parentheses)

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

Scree Plot – Detects How Many Principal Components

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

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

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

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. ■

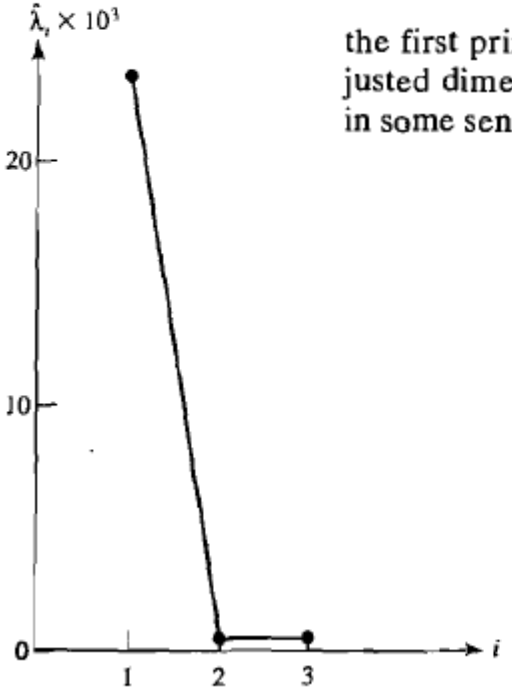


Figure 8.3 A scree plot for the turtle data.

Sample SAS Output to Generate Principal Components

PANEL 8.1 SAS ANALYSIS FOR EXAMPLE 8.4 USING PROC PRINCOMP.

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

PROGRAM COMMANDS

Principal Components Analysis

24 Observations
3 Variables

OUTPUT

Simple Statistics

	X1	X2	X3
Mean	4.725443647	4.477573765	3.703185794
StD	0.105223590	0.080104466	0.082296771

Covariance Matrix

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

Total Variance = 0.024261488

Sample SAS Output to Generate Principal Components

Total Variance = 0.024261488

Eigenvalues of the Covariance Matrix

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

Eigenvectors

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

Interpretation of Sample Principal Components Using Contours on Scatter Plot to Indicate Normality

Interpretation of the Sample Principal Components

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

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

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

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

Sample Principal Components Lie Along the Axes of Hyperellipsoid and Their Absolute Values Are Lengths of the Projections of $x - \bar{x}$ in the Directions of the Eigenvector

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

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

Principal Components and Ellipses of Constant Distance

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

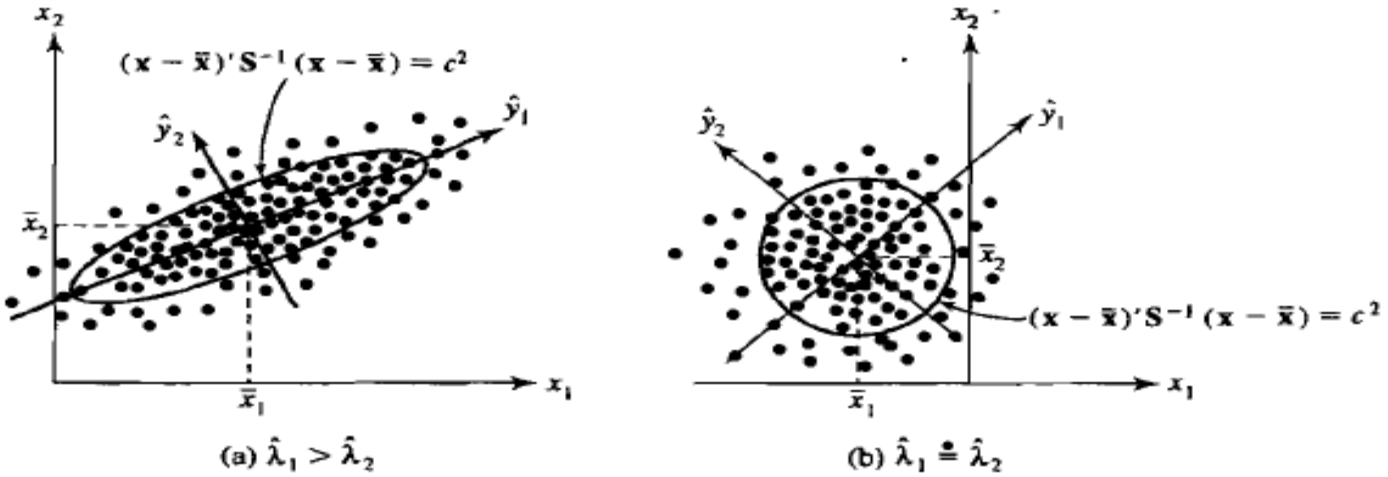


Figure 8.4 Sample principal components and ellipses of constant distance.

directions of the original coordinate axes. Similarly, the sample principal components

When Contours of Constant Distance Nearly Circular or When Eigenvalues of S Are Nearly Equal, Sample Variation Homogenous in All Directions

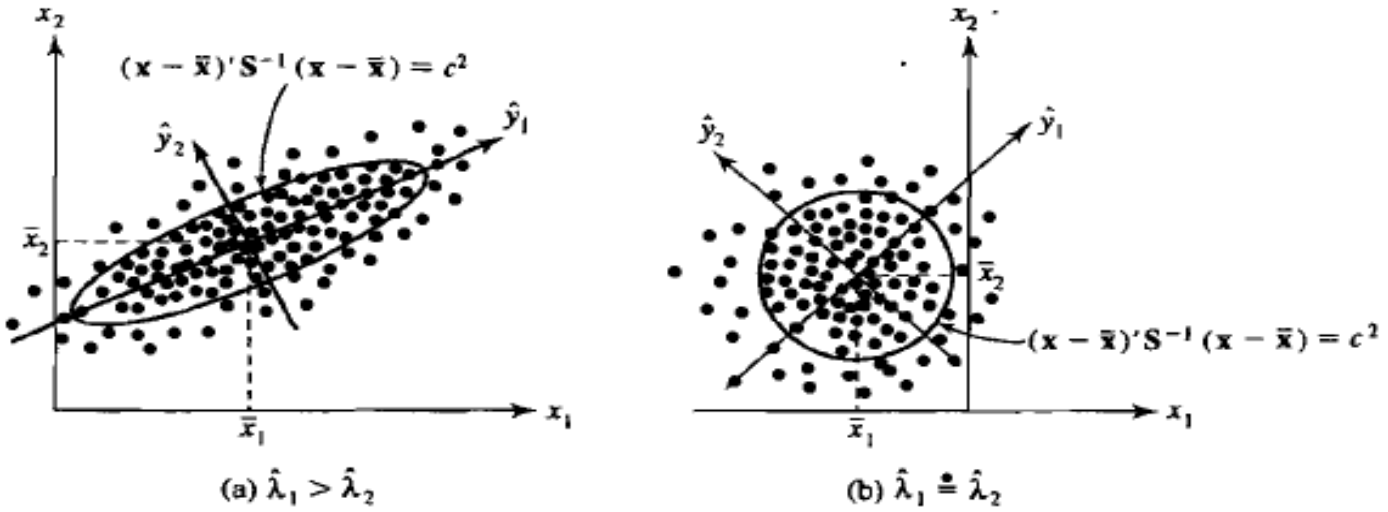


Figure 8.4 Sample principal components and ellipses of constant distance.

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

Standardizing the Principal Components

Standardizing the Sample Principal Components

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

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

Standardizing the Principal Components

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

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

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

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

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

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

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

Sample Principal Components of Standardized Observations With Covariance Matrix R in Place of S

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

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

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

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

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

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

Correlation Equal to Eigenvector x Square Root of Variance

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

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

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

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

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

In addition, (8-29)

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

and

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

Proportion of Total Sample Variance Explained by ith Sample Principal Component or = Variance / Proportion

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

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

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

Sample Principal Components from Standardized Data

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

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

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

Eigenvalues and Eigenvectors Calculated for R

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

$$\hat{\lambda}_1 = 2.437, \quad \hat{\mathbf{e}}_1' = [\ .469,\ .532,\ .465,\ .387,\ .361]$$

$$\hat{\lambda}_2 = 1.407, \quad \hat{\mathbf{e}}_2' = [-.368, -.236, -.315,\ .585,\ .606]$$

$$\hat{\lambda}_3 = .501, \quad \hat{\mathbf{e}}_3' = [-.604, -.136,\ .772,\ .093, -.109]$$

$$\hat{\lambda}_4 = .400, \quad \hat{\mathbf{e}}_4' = [\ .363, -.629,\ .289, -.381,\ .493]$$

$$\hat{\lambda}_5 = .255, \quad \hat{\mathbf{e}}_5' = [\ .384, -.496,\ .071,\ .595, -.498]$$

Interpreting The Components From Standardized Variables

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

$$\hat{y}_1 = \hat{\mathbf{e}}_1'\mathbf{z} = .469z_1 + .532z_2 + .465z_3 + .387z_4 + .361z_5$$

$$\hat{y}_2 = \hat{\mathbf{e}}_2'\mathbf{z} = -.368z_1 - .236z_2 - .315z_3 + .585z_4 + .606z_5$$

These components, which account for

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

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

The second component represents a contrast between the banking stocks (JP Morgan, Citibank, Wells Fargo) and the oil stocks (Royal Dutch Shell, 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 behavior also has been suggested by King [12].

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

Components From Correlation Matrix With Special Structure

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

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

and

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

The eigenvalues of this matrix are

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

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

First Principal Component Vs. Large Eigenvalues and Corresponding Eigenvector Close to Zero May Point Out linear Dependencies in the Data

The first principal component

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

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

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

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

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

Graphing Principal Components: Scatter Diagram and Q-Q Plots

8.4 Graphing the Principal Components

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

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

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

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

The following statements summarize these ideas.

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

Example: Plotting the Principal Components for the Turtle Data

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

$$\hat{y}_1 = .683(x_1 - 4.725) + .510(x_2 - 4.478) + .523(x_3 - 3.703)$$

$$\hat{y}_2 = -.159(x_1 - 4.725) - .594(x_2 - 4.478) + .788(x_3 - 3.703)$$

$$\hat{y}_3 = -.713(x_1 - 4.725) + .622(x_2 - 4.478) + .324(x_3 - 3.703)$$

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

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

Q-Q Plot for 2nd Principal Component and Scatter Plot for 1st and 2nd

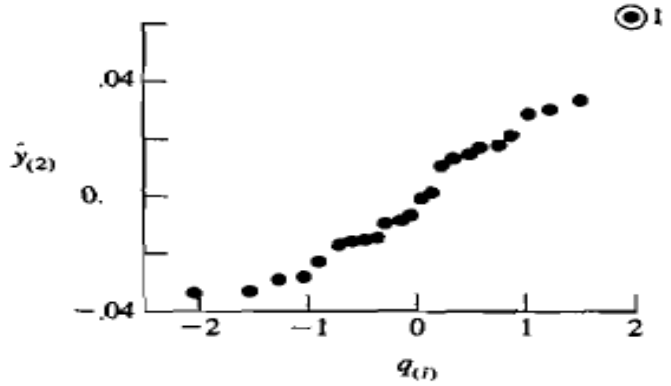


Figure 8.5 A Q–Q plot for the second principal component $\hat{y}_2$ from the data on male turtles.

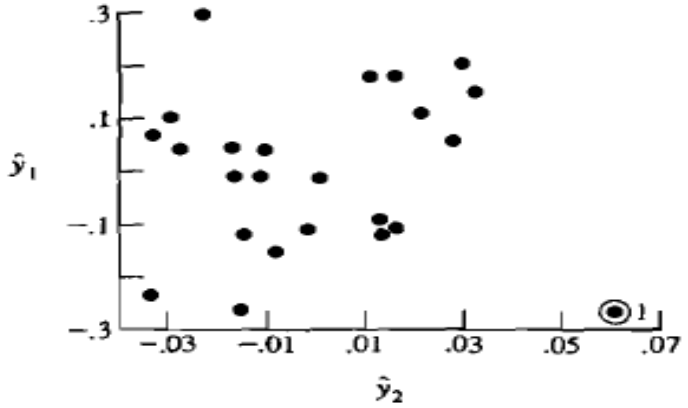


Figure 8.6 Scatter plot of the principal components $\hat{y}_1$ and $\hat{y}_2$ of the data on male turtles.

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

Residual Vector = Observation Vector Minus the Vector of Predicted Values

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

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

or

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

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

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

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

Eigenvectors Determine Direction of Maximum Variability; Eigenvalues Specify Variances, Thus Principal Component Quality Based on Pair

8.5 Large Sample Inferences

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

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

Large Sample Properties of Eigenvalue, Eigenvector Pairs

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 $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Sigma}$ has a special structure that yields equal eigenvalues. Even when the normal assumption is violated, the confidence intervals obtained in this manner still provide some indication of the uncertainty in $\hat{\lambda}_i$ and $\hat{\mathbf{e}}_i$.

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

1. Let $\boldsymbol{\Lambda}$ 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 $N_p(\mathbf{0}, \mathbf{E}_i)$.
3. Each $\hat{\lambda}_i$ is distributed independently of the elements of the associated $\hat{\mathbf{e}}_i$.

Eigenvalues Have Approximate Normal Distribution

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

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

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

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

Constructing Confidence Interval for Eigenvector, Variance 1st Principal Component

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

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

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

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

Testing for the Equal Correlation Structure

Testing for the Equal Correlation Structure

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

To test for this structure, let

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

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 [14] has demonstrated that an equivalent test procedure can be constructed from the off-diagonal elements of $\mathbf{R}$.

Use Chi-Square As Critical Value

Lawley's procedure requires the quantities

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

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

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

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

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

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

Example: Testing the Equi-correlation Structure

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

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

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

Here $p = 4$, and we set

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

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

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

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

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

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

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

Example: Testing the Equi-correlation Structure Noting Eigenvalues Slightly Different. Thus, Differences in Equal Differences Emerge as Statistically Significant

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

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

and

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

Since $(p + 1)(p - 2)/2 = 5(2)/2 = 5$, the 5% critical value for the test in (8-35) is $\chi^2_5(.05) = 11.07$. 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 in Example 8.6, the smallest eigenvalues $\hat{\lambda}_2$, $\hat{\lambda}_3$, and $\hat{\lambda}_4$ are slightly different, with $\hat{\lambda}_4$ being somewhat smaller than the other two. Consequently, with the large sample size in this problem, small differences from the equal correlation structure show up as statistically significant. ■

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