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

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

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

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

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

References

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

Chapter 9

Factor Analysis and Inference for Structured Covariance Matrices

9.1 Introduction

Factor analysis has provoked rather turbulent controversy throughout its history. Its modern beginnings lie in the early-20th-century attempts of Karl Pearson, Charles Spearman, and others to define and measure intelligence. Because of this early association with constructs such as intelligence, factor analysis was nurtured and developed primarily by scientists interested in psychometrics. Arguments over the psychological interpretations of several early studies and the lack of powerful computing facilities impeded its initial development as a statistical method. The advent of high-speed computers has generated a renewed interest in the theoretical and computational aspects of factor analysis. Most of the original techniques have been abandoned and early controversies resolved in the wake of recent developments. It is still true, however, that each application of the technique must be examined on its own merits to determine its success.

The essential purpose of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying, but unobservable, random quantities called *factors*. Basically, the factor model is motivated by the following argument: Suppose variables can be grouped by their correlations. That is, suppose all variables within a particular group are highly correlated among themselves, but have relatively small correlations with variables in a different group. Then it is conceivable that each group of variables represents a single underlying construct, or factor, that is responsible for the observed correlations. For example, correlations from the group of test scores in classics, French, English, mathematics, and music collected by Spearman suggested an underlying "intelligence" factor. A second group of variables, representing physical-fitness scores, if available, might correspond to another factor. It is this type of structure that factor analysis seeks to confirm.

Factor analysis can be considered an extension of principal component analysis. Both can be viewed as attempts to approximate the covariance matrix $\mathbf{\Sigma}$. However, the approximation based on the factor analysis model is more elaborate. The primary question in factor analysis is whether the data are consistent with a prescribed structure.

9.2 The Orthogonal Factor Model

The observable random vector $\mathbf{X}$, with p components, has mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$. The factor model postulates that $\mathbf{X}$ is linearly dependent upon a few unobservable random variables $F_1, F_2, \ldots, F_m$, called *common factors*, and p additional sources of variation $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$, called *errors* or, sometimes, *specific factors*.[1] In particular, the factor analysis model is

$$\begin{aligned} X_1 - \mu_1 &= \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 &= \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \varepsilon_2 \\ &\vdots \\ X_p - \mu_p &= \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \varepsilon_p \end{aligned} \tag{9-1}$$

or, in matrix notation,

$$\underset{(p\times 1)}{\mathbf{X} - \boldsymbol{\mu}} = \underset{(p\times m)}{\mathbf{L}}\underset{(m\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\varepsilon}} \tag{9-2}$$

The coefficient ℓ_{ij} is called the *loading* of the ith variable on the jth factor, so the matrix $\mathbf{L}$ is the *matrix of factor loadings*. Note that the ith specific factor ε_i is associated only with the ith response X_i. The p deviations $X_1 - \mu_1, X_2 - \mu_2, \ldots, X_p - \mu_p$ are expressed in terms of $p + m$ random variables $F_1, F_2, \ldots, F_m, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$ which are *unobservable*. This distinguishes the factor model of (9-2) from the multivariate regression model in (7-23), in which the independent variables [whose position is occupied by $\mathbf{F}$ in (9-2)] can be observed.

With so many unobservable quantities, a direct verification of the factor model from observations on $X_1, X_2, \ldots, X_p$ is hopeless. However, with some additional assumptions about the random vectors $\mathbf{F}$ and $\boldsymbol{\varepsilon}$, the model in (9-2) implies certain covariance relationships, which can be checked.

We assume that

$$E(\mathbf{F}) = \underset{(m\times 1)}{\mathbf{0}}, \qquad \text{Cov}(\mathbf{F}) = E[\mathbf{F}\mathbf{F}'] = \underset{(m\times m)}{\mathbf{I}}$$

$$E(\boldsymbol{\varepsilon}) = \underset{(p\times 1)}{\mathbf{0}}, \qquad \text{Cov}(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \underset{(p\times p)}{\mathbf{\Psi}} = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix} \tag{9-3}$$

[1] As Maxwell [12] points out, in many investigations the ε_i tend to be combinations of measurement error and factors that are uniquely associated with the individual variables.

and that $\mathbf{F}$ and $\boldsymbol{\varepsilon}$ are independent, so

$$\text{Cov}(\boldsymbol{\varepsilon}, \mathbf{F}) = E(\boldsymbol{\varepsilon}\mathbf{F}') = \underset{(p\times m)}{\mathbf{0}}$$

These assumptions and the relation in (9-2) constitute the *orthogonal factor model*.[2]

Orthogonal Factor Model with m Common Factors

$$\begin{aligned} \underset{(p\times 1)}{\mathbf{X}} &= \underset{(p\times 1)}{\boldsymbol{\mu}} + \underset{(p\times m)}{\mathbf{L}}\underset{(m\times 1)}{\mathbf{F}} + \underset{(p\times 1)}{\boldsymbol{\varepsilon}} \\ \mu_i &= \textit{mean} \text{ of variable } i \\ \varepsilon_i &= i\text{th } \textit{specific factor} \\ F_j &= j\text{th } \textit{common factor} \\ \ell_{ij} &= \textit{loading} \text{ of the } i\text{th variable on the } j\text{th factor} \end{aligned} \tag{9-4}$$

The unobservable random vectors $\mathbf{F}$ and $\boldsymbol{\varepsilon}$ satisfy the following conditions:

$\mathbf{F}$ and $\boldsymbol{\varepsilon}$ are independent

$E(\mathbf{F}) = \mathbf{0}$, $\text{Cov}(\mathbf{F}) = \mathbf{I}$

$E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\text{Cov}(\boldsymbol{\varepsilon}) = \mathbf{\Psi}$, where $\mathbf{\Psi}$ is a diagonal matrix

The orthogonal factor model implies a covariance structure for $\mathbf{X}$. From the model in (9-4),

$$\begin{aligned} (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' &= (\mathbf{LF} + \boldsymbol{\varepsilon})(\mathbf{LF} + \boldsymbol{\varepsilon})' \\ &= (\mathbf{LF} + \boldsymbol{\varepsilon})((\mathbf{LF})' + \boldsymbol{\varepsilon}') \\ &= \mathbf{LF}(\mathbf{LF})' + \boldsymbol{\varepsilon}(\mathbf{LF})' + \mathbf{LF}\boldsymbol{\varepsilon}' + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \end{aligned}$$

so that

$$\begin{aligned} \mathbf{\Sigma} = \text{Cov}(\mathbf{X}) &= E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' \\ &= \mathbf{L}E(\mathbf{FF}')\mathbf{L}' + E(\boldsymbol{\varepsilon}\mathbf{F}')\mathbf{L}' + \mathbf{L}E(\mathbf{F}\boldsymbol{\varepsilon}') + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\ &= \mathbf{LL}' + \mathbf{\Psi} \end{aligned}$$

according to (9-3). Also by independence, $\text{Cov}(\boldsymbol{\varepsilon}, \mathbf{F}) = E(\boldsymbol{\varepsilon}, \mathbf{F}') = \mathbf{0}$

Also, by the model in (9-4), $(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}' = (\mathbf{LF} + \boldsymbol{\varepsilon})\mathbf{F}' = \mathbf{LFF}' + \boldsymbol{\varepsilon}\mathbf{F}'$. $\text{Cov}(\mathbf{X}, \mathbf{F}) = E(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}' = \mathbf{L}E(\mathbf{FF}') + E(\boldsymbol{\varepsilon}\mathbf{F}') = \mathbf{L}$.

[2] Allowing the factors $\mathbf{F}$ to be correlated so that $\text{Cov}(\mathbf{F})$ is *not* diagonal gi model. The oblique model presents some additional estimation difficulties ap this book. (See [10].)

Covariance Structure for the Orthogonal Factor Model

1. $\text{Cov}(\mathbf{X}) = \mathbf{LL}' + \boldsymbol{\Psi}$

   or

$$\text{Var}(X_i) = \ell_{i1}^2 + \cdots + \ell_{im}^2 + \psi_i$$

$$\text{Cov}(X_i, X_k) = \ell_{i1}\ell_{k1} + \cdots + \ell_{im}\ell_{km} \qquad (9\text{-}5)$$

2. $\text{Cov}(\mathbf{X}, \mathbf{F}) = \mathbf{L}$

   or

$$\text{Cov}(X_i, F_j) = \ell_{ij}$$

The model $\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon}$ is *linear* in the common factors. If the p responses $\mathbf{X}$ are, in fact, related to underlying factors, but the relationship is nonlinear, such as in $X_1 - \mu_1 = \ell_{11}F_1F_3 + \varepsilon_1$, $X_2 - \mu_2 = \ell_{21}F_2F_3 + \varepsilon_2$, and so forth, then the covariance structure $\mathbf{LL}' + \boldsymbol{\Psi}$ given by (9-5) may not be adequate. The very important assumption of linearity is inherent in the formulation of the traditional factor model.

That portion of the variance of the ith variable contributed by the m common factors is called the ith *communality*. That portion of $\text{Var}(X_i) = \sigma_{ii}$ due to the specific factor is often called the *uniqueness*, or *specific variance*. Denoting the ith communality by h_i^2, we see from (9-5) that

$$\underbrace{\sigma_{ii}}_{\text{Var}(X_i)} = \underbrace{\ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2}_{\text{communality}} + \underbrace{\psi_i}_{\text{specific variance}}$$

or

$$h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2 \qquad (9\text{-}6)$$

and

$$\sigma_{ii} = h_i^2 + \psi_i, \qquad i = 1, 2, \ldots, p$$

The ith communality is the sum of squares of the loadings of the ith variable on the m common factors.

Example 9.1 (Verifying the relation $\boldsymbol{\Sigma} = \mathbf{LL}' + \boldsymbol{\Psi}$ for two factors) Consider the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{bmatrix}$$

The equality

$$\begin{bmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} 4 & 7 & -1 & 1 \\ 1 & 2 & 6 & 8 \end{bmatrix} + \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

or

$$\boldsymbol{\Sigma} = \mathbf{LL}' + \boldsymbol{\Psi}$$

may be verified by matrix algebra. Therefore, $\boldsymbol{\Sigma}$ has the structure produced by an $m = 2$ orthogonal factor model. Since

$$\mathbf{L} = \begin{bmatrix} \ell_{11} & \ell_{12} \\ \ell_{21} & \ell_{22} \\ \ell_{31} & \ell_{32} \\ \ell_{41} & \ell_{42} \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix},$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & 0 & 0 & 0 \\ 0 & \psi_2 & 0 & 0 \\ 0 & 0 & \psi_3 & 0 \\ 0 & 0 & 0 & \psi_4 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

the communality of X_1 is, from (9-6),

$$h_1^2 = \ell_{11}^2 + \ell_{12}^2 = 4^2 + 1^2 = 17$$

and the variance of X_1 can be decomposed as

$$\sigma_{11} = (\ell_{11}^2 + \ell_{12}^2) + \psi_1 = h_1^2 + \psi_1$$

or

$$\underbrace{19}_{\text{variance}} = \underbrace{4^2 + 1^2}_{\text{communality}} + \underbrace{2}_{\text{specific variance}} = 17 + 2$$

A similar breakdown occurs for the other variables. ■

The factor model assumes that the $p + p(p-1)/2 = p(p+1)/2$ variances and covariances for $\mathbf{X}$ can be reproduced from the pm factor loadings ℓ_{ij} and the p specific variances ψ_i. When $m = p$, any covariance matrix $\boldsymbol{\Sigma}$ can be reproduced exactly as $\mathbf{LL}'$ [see (9-11)], so $\boldsymbol{\Psi}$ can be the zero matrix. However, it is when m is small relative to p that factor analysis is most useful. In this case, the factor model provides a "[illegible]ple" explanation of the covariation in $\mathbf{X}$ with fewer parameters than the $p(p$ [illegible] parameters in $\boldsymbol{\Sigma}$. For example, if $\mathbf{X}$ contains $p = 12$ variables, and the fact[illegible] (9-4) with $m = 2$ is appropriate, then the $p(p+1)/2 = 12(13)/2 =$ [illegible] $\boldsymbol{\Sigma}$ are described in terms of the $mp + p = 12(2) + 12 = 36$ param[illegible] the factor model.

Unfortunately for the factor analyst, most covariance matrices cannot be factored as $\mathbf{LL}' + \mathbf{\Psi}$, where the number of factors m is much less than p. The following example demonstrates one of the problems that can arise when attempting to determine the parameters ℓ_{ij} and ψ_i from the variances and covariances of the observable variables.

Example 9.2 (Nonexistence of a proper solution) Let $p = 3$ and $m = 1$, and suppose the random variables X_1, X_2, and X_3 have the positive definite covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & .9 & .7 \\ .9 & 1 & .4 \\ .7 & .4 & 1 \end{bmatrix}$$

Using the factor model in (9-4), we obtain

$$X_1 - \mu_1 = \ell_{11}F_1 + \varepsilon_1$$
$$X_2 - \mu_2 = \ell_{21}F_1 + \varepsilon_2$$
$$X_3 - \mu_3 = \ell_{31}F_1 + \varepsilon_3$$

The covariance structure in (9-5) implies that

$$\mathbf{\Sigma} = \mathbf{LL}' + \mathbf{\Psi}$$

or

$$1 = \ell_{11}^2 + \psi_1 \qquad .90 = \ell_{11}\ell_{21} \qquad .70 = \ell_{11}\ell_{31}$$
$$1 = \ell_{21}^2 + \psi_2 \qquad .40 = \ell_{21}\ell_{31}$$
$$1 = \ell_{31}^2 + \psi_3$$

The pair of equations

$$.70 = \ell_{11}\ell_{31}$$
$$.40 = \ell_{21}\ell_{31}$$

implies that

$$\ell_{21} = \left(\frac{.40}{.70}\right)\ell_{11}$$

Substituting this result for ℓ_{21} in the equation

$$.90 = \ell_{11}\ell_{21}$$

yields $\ell_{11}^2 = 1.575$, or $\ell_{11} = \pm 1.255$. Since $\text{Var}(F_1) = 1$ (by assumption) and $\text{Var}(X_1) = 1$, $\ell_{11} = \text{Cov}(X_1, F_1) = \text{Corr}(X_1, F_1)$. Now, a correlation coefficient cannot be greater than unity (in absolute value), so, from this point of view, $|\ell_{11}| = 1.255$ is too large. Also, the equation

$$1 = \ell_{11}^2 + \psi_1, \quad \text{or} \quad \psi_1 = 1 - \ell_{11}^2$$

gives

$$\psi_1 = 1 - 1.575 = -.575$$

which is unsatisfactory, since it gives a negative value for $\text{Var}(\varepsilon_1) = \psi_1$.

Thus, for this example with $m = 1$, it is possible to get a unique numerical solution to the equations $\mathbf{\Sigma} = \mathbf{LL}' + \mathbf{\Psi}$. However, the solution is not consistent with the statistical interpretation of the coefficients, so it is not a proper solution. ■

When $m > 1$, there is always some inherent ambiguity associated with the factor model. To see this, let $\mathbf{T}$ be any $m \times m$ orthogonal matrix, so that $\mathbf{TT}' = \mathbf{T}'\mathbf{T} = \mathbf{I}$. Then the expression in (9-2) can be written

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon} = \mathbf{LTT}'\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{L}^*\mathbf{F}^* + \boldsymbol{\varepsilon} \tag{9-7}$$

where

$$\mathbf{L}^* = \mathbf{LT} \quad \text{and} \quad \mathbf{F}^* = \mathbf{T}'\mathbf{F}$$

Since

$$E(\mathbf{F}^*) = \mathbf{T}'E(\mathbf{F}) = \mathbf{0}$$

and

$$\text{Cov}(\mathbf{F}^*) = \mathbf{T}'\text{Cov}(\mathbf{F})\mathbf{T} = \mathbf{T}'\mathbf{T} = \underset{(m\times m)}{\mathbf{I}}$$

it is impossible, on the basis of observations on $\mathbf{X}$, to distinguish the loadings $\mathbf{L}$ from the loadings $\mathbf{L}^*$. That is, the factors $\mathbf{F}$ and $\mathbf{F}^* = \mathbf{T}'\mathbf{F}$ have the same statistical properties, and even though the loadings $\mathbf{L}^*$ are, in general, different from the loadings $\mathbf{L}$, they both generate the same covariance matrix $\mathbf{\Sigma}$. That is,

$$\mathbf{\Sigma} = \mathbf{LL}' + \mathbf{\Psi} = \mathbf{LTT}'\mathbf{L}' + \mathbf{\Psi} = (\mathbf{L}^*)(\mathbf{L}^*)' + \mathbf{\Psi} \tag{9-8}$$

This ambiguity provides the rationale for "factor rotation," since orthogonal matrices correspond to rotations (and reflections) of the coordinate system for $\mathbf{X}$.

Factor loadings $\mathbf{L}$ are determined only up to an orthogonal matrix $\mathbf{T}$. Thus, the loadings

$$\mathbf{L}^* = \mathbf{LT} \quad \text{and} \quad \mathbf{L} \tag{9-9}$$

both give the same representation. The communalities, given by the diagonal elements of $\mathbf{LL}' = (\mathbf{L}^*)(\mathbf{L}^*)'$ are also unaffected by the choice of $\mathbf{T}$.

The analysis of the factor model proceeds by imposing conditions that allow one to uniquely estimate $\mathbf{L}$ and $\mathbf{\Psi}$. The loading matrix is then rotated (multiplied by an orthogonal matrix), where the rotation is determined by some "ease-of-interpretation" criterion. Once the loadings and specific variances are obtained, factors are identified, and estimated values for the factors themselves (called *factor scores*) are frequently constructed.

9.3 Methods of Estimation

Given observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ on p generally correlated variables, factor analysis seeks to answer the question, Does the factor model of (9-4), with a small number of factors, adequately represent the data? In essence, we tackle this statistical model-building problem by trying to verify the covariance relationship in (9-5).

The sample covariance matrix $\mathbf{S}$ is an estimator of the unknown population covariance matrix $\boldsymbol{\Sigma}$. If the off-diagonal elements of $\mathbf{S}$ are small or those of the sample correlation matrix $\mathbf{R}$ essentially zero, the variables are not related, and a factor analysis will not prove useful. In these circumstances, the *specific* factors play the dominant role, whereas the major aim of factor analysis is to determine a few important common factors.

If $\boldsymbol{\Sigma}$ appears to deviate significantly from a diagonal matrix, then a factor model can be entertained, and the initial problem is one of estimating the factor loadings ℓ_{ij} and specific variances ψ_i. We shall consider two of the most popular methods of parameter estimation, the *principal component* (and the related *principal factor*) *method* and the *maximum likelihood method*. The solution from either method can be rotated in order to simplify the interpretation of factors, as described in Section 9.4. It is always prudent to try more than one method of solution; if the factor model is appropriate for the problem at hand, the solutions should be consistent with one another.

Current estimation and rotation methods require iterative calculations that must be done on a computer. Several computer programs are now available for this purpose.

The Principal Component (and Principal Factor) Method

The spectral decomposition of (2-16) provides us with one factoring of the covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{\Sigma}$ have eigenvalue–eigenvector pairs $(\lambda_i, \mathbf{e}_i)$ with $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_p \ge 0$. Then

$$\boldsymbol{\Sigma} = \lambda_1 \mathbf{e}_1\mathbf{e}_1' + \lambda_2 \mathbf{e}_2\mathbf{e}_2' + \cdots + \lambda_p \mathbf{e}_p\mathbf{e}_p'$$

$$= \left[\sqrt{\lambda_1}\,\mathbf{e}_1 \;\vdots\; \sqrt{\lambda_2}\,\mathbf{e}_2 \;\vdots\; \cdots \;\vdots\; \sqrt{\lambda_p}\,\mathbf{e}_p\right] \begin{bmatrix} \sqrt{\lambda_1}\,\mathbf{e}_1' \\ \hdashline \sqrt{\lambda_2}\,\mathbf{e}_2' \\ \hdashline \vdots \\ \hdashline \sqrt{\lambda_p}\,\mathbf{e}_p' \end{bmatrix} \tag{9-10}$$

This fits the prescribed covariance structure for the factor analysis model having as many factors as variables $(m = p)$ and specific variances $\psi_i = 0$ for all i. The loading matrix has jth column given by $\sqrt{\lambda_j}\,\mathbf{e}_j$. That is, we can write

$$\underset{(p\times p)}{\boldsymbol{\Sigma}} = \underset{(p\times p)}{\mathbf{L}}\underset{(p\times p)}{\mathbf{L}'} + \underset{(p\times p)}{\mathbf{0}} = \mathbf{L}\mathbf{L}' \tag{9-11}$$

Apart from the scale factor $\sqrt{\lambda_j}$, the factor loadings on the jth factor are the coefficients for the jth principal component of the population.

Although the factor analysis representation of $\boldsymbol{\Sigma}$ in (9-11) is exact, it is not particularly useful: It employs as many common factors as there are variables and does not allow for any variation in the specific factors $\boldsymbol{\varepsilon}$ in (9-4). We prefer models that explain the covariance structure in terms of just a few common factors. One

approach, when the last $p - m$ eigenvalues are small, is to neglect the contribution of $\lambda_{m+1}\mathbf{e}_{m+1}\mathbf{e}_{m+1}' + \cdots + \lambda_p\mathbf{e}_p\mathbf{e}_p'$ to $\boldsymbol{\Sigma}$ in (9-10). Neglecting this contribution, we obtain the approximation

$$\boldsymbol{\Sigma} \doteq \left[\sqrt{\lambda_1}\,\mathbf{e}_1 \;\vdots\; \sqrt{\lambda_2}\,\mathbf{e}_2 \;\vdots\; \cdots \;\vdots\; \sqrt{\lambda_m}\,\mathbf{e}_m\right] \begin{bmatrix} \sqrt{\lambda_1}\,\mathbf{e}_1' \\ \hdashline \sqrt{\lambda_2}\,\mathbf{e}_2' \\ \hdashline \vdots \\ \hdashline \sqrt{\lambda_m}\,\mathbf{e}_m' \end{bmatrix} = \underset{(p\times m)}{\mathbf{L}}\underset{(m\times p)}{\mathbf{L}'} \tag{9-12}$$

The approximate representation in (9-12) assumes that the specific factors $\boldsymbol{\varepsilon}$ in (9-4) are of minor importance and can also be ignored in the factoring of $\boldsymbol{\Sigma}$. If specific factors are included in the model, their variances may be taken to be the diagonal elements of $\boldsymbol{\Sigma} - \mathbf{L}\mathbf{L}'$, where $\mathbf{L}\mathbf{L}'$ is as defined in (9-12).

Allowing for specific factors, we find that the approximation becomes

$$\boldsymbol{\Sigma} \doteq \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$$

$$= \left[\sqrt{\lambda_1}\,\mathbf{e}_1 \;\vdots\; \sqrt{\lambda_2}\,\mathbf{e}_2 \;\vdots\; \cdots \;\vdots\; \sqrt{\lambda_m}\,\mathbf{e}_m\right] \begin{bmatrix} \sqrt{\lambda_1}\,\mathbf{e}_1' \\ \hdashline \sqrt{\lambda_2}\,\mathbf{e}_2' \\ \hdashline \vdots \\ \hdashline \sqrt{\lambda_m}\,\mathbf{e}_m' \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix} \tag{9-13}$$

where $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} \ell_{ij}^2$ for $i = 1, 2, \ldots, p$.

To apply this approach to a data set $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, it is customary first to center the observations by subtracting the sample mean $\bar{\mathbf{x}}$. The centered observations

$$\mathbf{x}_j - \bar{\mathbf{x}} = \begin{bmatrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jp} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{bmatrix} x_{j1} - \bar{x}_1 \\ x_{j2} - \bar{x}_2 \\ \vdots \\ x_{jp} - \bar{x}_p \end{bmatrix} \qquad j = 1, 2, \ldots, n \tag{9-14}$$

have the same sample covariance matrix $\mathbf{S}$ as the original observations.

In cases in which the units of the variables are not commensurate, it is usually desirable to work with the standardized variables

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

whose sample covariance matrix is the sample correlation matrix $\mathbf{R}$ of the observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Standardization avoids the problems of having one variable with large variance unduly influencing the determination of factor loadings.

The representation in (9-13), when applied to the sample covariance matrix **S** or the sample correlation matrix **R**, is known as the *principal component solution*. The name follows from the fact that the factor loadings are the scaled coefficients of the first few sample principal components. (See Chapter 8.)

Principal Component Solution of the Factor Model

The principal component factor analysis of the sample covariance matrix **S** is specified in terms of its eigenvalue–eigenvector pairs $(\hat{\lambda}_1, \hat{\mathbf{e}}_1)$, $(\hat{\lambda}_2, \hat{\mathbf{e}}_2), \ldots, (\hat{\lambda}_p, \hat{\mathbf{e}}_p)$, where $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \cdots \ge \hat{\lambda}_p$. Let $m < p$ be the number of common factors. Then the matrix of estimated factor loadings $\{\tilde{\ell}_{ij}\}$ is given by

$$\tilde{\mathbf{L}} = \left[\sqrt{\hat{\lambda}_1}\,\hat{\mathbf{e}}_1 \;\vdots\; \sqrt{\hat{\lambda}_2}\,\hat{\mathbf{e}}_2 \;\vdots\; \cdots \;\vdots\; \sqrt{\hat{\lambda}_m}\,\hat{\mathbf{e}}_m\right] \tag{9-15}$$

The estimated specific variances are provided by the diagonal elements of the matrix $\mathbf{S} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}'$, so

$$\tilde{\boldsymbol{\Psi}} = \begin{bmatrix} \tilde{\psi}_1 & 0 & \cdots & 0 \\ 0 & \tilde{\psi}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\psi}_p \end{bmatrix} \quad \text{with} \quad \tilde{\psi}_i = s_{ii} - \sum_{j=1}^{m} \tilde{\ell}_{ij}^2 \tag{9-16}$$

Communalities are estimated as

$$\tilde{h}_i^2 = \tilde{\ell}_{i1}^2 + \tilde{\ell}_{i2}^2 + \cdots + \tilde{\ell}_{im}^2 \tag{9-17}$$

The principal component factor analysis of the sample correlation matrix is obtained by starting with **R** in place of **S**.

For the principal component solution, the estimated loadings for a given factor do not change as the number of factors is increased. For example, if $m = 1$, $\tilde{\mathbf{L}} = \left[\sqrt{\hat{\lambda}_1}\,\hat{\mathbf{e}}_1\right]$, and if $m = 2$, $\tilde{\mathbf{L}} = \left[\sqrt{\hat{\lambda}_1}\,\hat{\mathbf{e}}_1 \;\vdots\; \sqrt{\hat{\lambda}_2}\,\hat{\mathbf{e}}_2\right]$, where $(\hat{\lambda}_1, \hat{\mathbf{e}}_1)$ and $(\hat{\lambda}_2, \hat{\mathbf{e}}_2)$ are the first two eigenvalue–eigenvector pairs for **S** (or **R**).

By the definition of $\tilde{\psi}_i$, the diagonal elements of **S** are equal to the diagonal elements of $\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\boldsymbol{\Psi}}$. However, the off-diagonal elements of **S** are not usually reproduced by $\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\boldsymbol{\Psi}}$. How, then, do we select the number of factors m?

If the number of common factors is not determined by a priori considerations, such as by theory or the work of other researchers, the choice of m can be based on the estimated eigenvalues in much the same manner as with principal components. Consider the *residual matrix*

$$\mathbf{S} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\boldsymbol{\Psi}}) \tag{9-18}$$

resulting from the approximation of **S** by the principal component solution. The diagonal elements are zero, and if the other elements are also small, we may subjectively take the m factor model to be appropriate. Analytically, we have (see Exercise 9.5)

$$\text{Sum of squared entries of } (\mathbf{S} - (\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\boldsymbol{\Psi}})) \le \hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_p^2 \tag{9-19}$$

Consequently, a small value for the sum of the squares of the neglected eigenvalues implies a small value for the sum of the squared errors of approximation.

Ideally, the contributions of the first few factors to the sample variances of the variables should be large. The contribution to the sample variance s_{ii} from the first common factor is $\tilde{\ell}_{i1}^2$. The contribution to the *total* sample variance, $s_{11} + s_{22} + \cdots + s_{pp} = \text{tr}(\mathbf{S})$, from the first common factor is then

$$\tilde{\ell}_{11}^2 + \tilde{\ell}_{21}^2 + \cdots + \tilde{\ell}_{p1}^2 = \left(\sqrt{\hat{\lambda}_1}\,\hat{\mathbf{e}}_1\right)'\left(\sqrt{\hat{\lambda}_1}\,\hat{\mathbf{e}}_1\right) = \hat{\lambda}_1$$

since the eigenvector $\hat{\mathbf{e}}_1$ has unit length. In general,

$$\begin{pmatrix}\text{Proportion of total} \\ \text{sample variance} \\ \text{due to } j\text{th factor}\end{pmatrix} = \begin{cases} \dfrac{\hat{\lambda}_j}{s_{11} + s_{22} + \cdots + s_{pp}} & \text{for a factor analysis of } \mathbf{S} \\[2ex] \dfrac{\hat{\lambda}_j}{p} & \text{for a factor analysis of } \mathbf{R} \end{cases} \tag{9-20}$$

Criterion (9-20) is frequently used as a heuristic device for determining the appropriate number of common factors. The number of common factors retained in the model is increased until a "suitable proportion" of the total sample variance has been explained.

Another convention, frequently encountered in packaged computer programs, is to set m equal to the number of eigenvalues of **R** greater than one if the sample correlation matrix is factored, or equal to the number of positive eigenvalues of **S** if the sample covariance matrix is factored. These rules of thumb should not be applied indiscriminately. For example, $m = p$ if the rule for **S** is obeyed, since all the eigenvalues are expected to be positive for large sample sizes. The best approach is to retain few rather than many factors, assuming that they provide a satisfactory interpretation of the data and yield a satisfactory fit to **S** or **R**.

Example 9.3 (Factor analysis of consumer-preference data) In a consumer-preference study, a random sample of customers were asked to rate several attributes of a new product. The responses, on a 7-point semantic differential scale, were tabulated and the attribute correlation matrix constructed. The correlation matrix is presented next:

Attribute (Variable)		1	2	3	4	5
Taste	1	1.00	.02	(.96)	.42	.01
Good buy for money	2	.02	1.00	.13	.71	(.85)
Flavor	3	.96	.13	1.00	.50	.11
Suitable for snack	4	.42	.71	.50	1.00	(.79)
Provides lots of energy	5	.01	.85	.11	.79	1.00

It is clear from the circled entries in the correlation matrix that variables 1 and 3 and variables 2 and 5 form groups. Variable 4 is "closer" to the (2, 5) group than the (1, 3) group. Given these results and the small number of variables, we might expect that the apparent linear relationships between the variables can be explained in terms of, at most, two or three common factors.

The first two eigenvalues, $\hat{\lambda}_1 = 2.85$ and $\hat{\lambda}_2 = 1.81$, of **R** are the only eigenvalues greater than unity. Moreover, $m = 2$ common factors will account for a cumulative proportion

$$\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{p} = \frac{2.85 + 1.81}{5} = .93$$

of the total (standardized) sample variance. The estimated factor loadings, communalities, and specific variances, obtained using (9-15), (9-16), and (9-17), are given in Table 9.1.

Table 9.1

Variable	Estimated factor loadings $\tilde{\ell}_{ij} = \sqrt{\hat{\lambda}_i}\hat{e}_{ij}$ F_1	F_2	Communalities $\tilde{h}_i^2$	Specific variances $\tilde{\psi}_i = 1 - \tilde{h}_i^2$
1. Taste	.56	.82	.98	.02
2. Good buy for money	.78	−.53	.88	.12
3. Flavor	.65	.75	.98	.02
4. Suitable for snack	.94	−.10	.89	.11
5. Provides lots of energy	.80	−.54	.93	.07
Eigenvalues	2.85	1.81		
Cumulative proportion of total (standardized) sample variance	.571	.932		

Now,

$$\tilde{\mathbf{L}}\tilde{\mathbf{L}}' + \tilde{\mathbf{\Psi}} = \begin{bmatrix} .56 & .82 \\ .78 & -.53 \\ .65 & .75 \\ .94 & -.10 \\ .80 & -.54 \end{bmatrix} \begin{bmatrix} .56 & .78 & .65 & .94 & .80 \\ .82 & -.53 & .75 & -.10 & -.54 \end{bmatrix}$$

$$+ \begin{bmatrix} .02 & 0 & 0 & 0 & 0 \\ 0 & .12 & 0 & 0 & 0 \\ 0 & 0 & .02 & 0 & 0 \\ 0 & 0 & 0 & .11 & 0 \\ 0 & 0 & 0 & 0 & .07 \end{bmatrix} = \begin{bmatrix} 1.00 & .01 & .97 & .44 & .00 \\ & 1.00 & .11 & .79 & .91 \\ & & 1.00 & .53 & .11 \\ & & & 1.00 & .81 \\ & & & & 1.00 \end{bmatrix}$$

nearly reproduces the correlation matrix **R**. Thus, on a purely descriptive basis, we would judge a two-factor model with the factor loadings displayed in Table 9.1 as providing a good fit to the data. The communalities (.98, .88, .98, .89, .93) indicate that the two factors account for a large percentage of the sample variance of each variable.

We shall not interpret the factors at this point. As we noted in Section 9.2, the factors (and loadings) are unique up to an orthogonal rotation. A rotation of the factors often reveals a simple structure and aids interpretation. We shall consider this example again (see Example 9.9 and Panel 9.1) after factor rotation has been discussed. ■

Example 9.4 (Factor analysis of stock-price data) Stock-price data consisting of $n = 103$ weekly rates of return on $p = 5$ stocks were introduced in Example 8.5. In that example, the first two sample principal components were obtained from **R**. Taking $m = 1$ and $m = 2$, we can easily obtain principal component solutions to the orthogonal factor model. Specifically, the estimated factor loadings are the sample principal component coefficients (eigenvectors of **R**), scaled by the square root of the corresponding eigenvalues. The estimated factor loadings, communalities, specific variances, and proportion of total (standardized) sample variance explained by each factor for the $m = 1$ and $m = 2$ factor solutions are available in Table 9.2. The communalities are given by (9-17). So, for example, with $m = 2$, $\tilde{h}_1^2 = \tilde{\ell}_{11}^2 + \tilde{\ell}_{12}^2 = (.732)^2 + (-.437)^2 = .73$.

Table 9.2

	One-factor solution		Two-factor solution		
Variable	Estimated factor loadings F_1	Specific variances $\tilde{\psi}_i = 1 - \tilde{h}_i^2$	Estimated factor loadings F_1	F_2	Specific variances $\tilde{\psi}_i = 1 - \tilde{h}_i^2$
1. J P Morgan	.732	.46	.732	−.437	.27
2. Citibank	.831	.31	.831	−.280	.23
3. Wells Fargo	.726	.47	.726	−.374	.33
4. Royal Dutch Shell	.605	.63	.605	.694	.15
5. ExxonMobil	.563	.68	.563	.719	.17
Cumulative proportion of total (standardized) sample variance explained	.487		.487	.769	

The residual matrix corresponding to the solution for $m = 2$ factors is

$$\mathbf{R} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}' - \tilde{\mathbf{\Psi}} = \begin{bmatrix} 0 & -.099 & -.185 & -.025 & .056 \\ -.099 & 0 & -.134 & .014 & -.054 \\ -.185 & -.134 & 0 & .003 & .006 \\ -.025 & .014 & .003 & 0 & -.156 \\ .056 & -.054 & .006 & -.156 & 0 \end{bmatrix}$$

The proportion of the total variance explained by the two-factor solution is appreciably larger than that for the one-factor solution. However, for $m = 2$, $\widetilde{\mathbf{L}}\widetilde{\mathbf{L}}'$ produces numbers that are, in general, larger than the sample correlations. This is particularly true for r_{13}.

It seems fairly clear that the first factor, F_1, represents general economic conditions and might be called a *market factor*. All of the stocks load highly on this factor, and the loadings are about equal. The second factor contrasts the banking stocks with the oil stocks. (The banks have relatively large negative loadings, and the oils have large positive loadings, on the factor.) Thus, F_2 seems to differentiate stocks in different industries and might be called an *industry factor*. To summarize, rates of return appear to be determined by general market conditions and activities that are unique to the different industries, as well as a residual or firm specific factor. This is essentially the conclusion reached by an examination of the sample principal components in Example 8.5. ■

A Modified Approach—the Principal Factor Solution

A modification of the principal component approach is sometimes considered. We describe the reasoning in terms of a factor analysis of **R**, although the procedure is also appropriate for **S**. If the factor model $\boldsymbol{\rho} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$ is correctly specified, the *m common* factors should account for the *off-diagonal* elements of $\boldsymbol{\rho}$, as well as the *communality portions* of the diagonal elements

$$\rho_{ii} = 1 = h_i^2 + \psi_i$$

If the specific factor contribution ψ_i is removed from the diagonal or, equivalently, the 1 replaced by h_i^2, the resulting matrix is $\boldsymbol{\rho} - \boldsymbol{\Psi} = \mathbf{L}\mathbf{L}'$.

Suppose, now, that initial estimates ψ_i^* of the specific variances are available. Then replacing the ith diagonal element of **R** by $h_i^{*2} = 1 - \psi_i^*$, we obtain a "reduced" sample correlation matrix

$$\mathbf{R}_r = \begin{bmatrix} h_1^{*2} & r_{12} & \cdots & r_{1p} \\ r_{12} & h_2^{*2} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & h_p^{*2} \end{bmatrix}$$

Now, apart from sampling variation, all of the elements of the reduced sample correlation matrix $\mathbf{R}_r$ should be accounted for by the m common factors. In particular, $\mathbf{R}_r$ is factored as

$$\mathbf{R}_r \doteq \mathbf{L}_r^* \mathbf{L}_r^{*\prime} \tag{9-21}$$

where $\mathbf{L}_r^* = \{\ell_{ij}^*\}$ are the estimated loadings.

The *principal factor method* of factor analysis employs the estimates

$$\mathbf{L}_r^* = \left[\sqrt{\hat{\lambda}_1^*}\,\hat{\mathbf{e}}_1^* \;\vdots\; \sqrt{\hat{\lambda}_2^*}\,\hat{\mathbf{e}}_2^* \;\vdots\; \cdots \;\vdots\; \sqrt{\hat{\lambda}_m^*}\,\hat{\mathbf{e}}_m^*\right]$$
$$\psi_i^* = 1 - \sum_{j=1}^{m} \ell_{ij}^{*2} \tag{9-22}$$

where $(\hat{\lambda}_i^*, \hat{\mathbf{e}}_i^*)$, $i = 1, 2, \ldots, m$ are the (largest) eigenvalue-eigenvector pairs determined from $\mathbf{R}_r$. In turn, the communalities would then be (re)estimated by

$$\tilde{h}_i^{*2} = \sum_{j=1}^{m} \ell_{ij}^{*2} \tag{9-23}$$

The principal factor solution can be obtained iteratively, with the communality estimates of (9-23) becoming the initial estimates for the next stage.

In the spirit of the principal component solution, consideration of the estimated eigenvalues $\hat{\lambda}_1^*, \hat{\lambda}_2^*, \ldots, \hat{\lambda}_p^*$ helps determine the number of common factors to retain. An added complication is that now some of the eigenvalues may be negative, due to the use of initial communality estimates. Ideally, we should take the number of common factors equal to the rank of the reduced *population* matrix. Unfortunately, this rank is not always well determined from $\mathbf{R}_r$, and some judgment is necessary.

Although there are many choices for initial estimates of specific variances, the most popular choice, when one is working with a correlation matrix, is $\psi_i^* = 1/r^{ii}$, where r^{ii} is the ith diagonal element of $\mathbf{R}^{-1}$. The initial communality estimates then become

$$h_i^{*2} = 1 - \psi_i^* = 1 - \frac{1}{r^{ii}} \tag{9-24}$$

which is equal to the square of the multiple correlation coefficient between X_i and the other $p - 1$ variables. The relation to the multiple correlation coefficient means that h_i^{*2} can be calculated even when **R** is not of full rank. For factoring **S**, the initial specific variance estimates use s^{ii}, the diagonal elements of $\mathbf{S}^{-1}$. Further discussion of these and other initial estimates is contained in [6].

Although the principal component method for **R** can be regarded as a principal factor method with *initial* communality estimates of unity, or specific variances equal to zero, the two are philosophically and geometrically different. (See [6].) In practice, however, the two frequently produce comparable factor loadings if the number of variables is large and the number of common factors is small.

We do not pursue the principal factor solution, since, to our minds, the solution methods that have the most to recommend them are the principal component method and the maximum likelihood method, which we discuss next.

The Maximum Likelihood Method

If the common factors **F** and the specific factors $\boldsymbol{\varepsilon}$ can be assumed to be normally distributed, then maximum likelihood estimates of the factor loadings and specific variances may be obtained. When $\mathbf{F}_j$ and $\boldsymbol{\varepsilon}_j$ are jointly normal, the observations $\mathbf{X}_j - \boldsymbol{\mu} = \mathbf{L}\mathbf{F}_j + \boldsymbol{\varepsilon}_j$ are then normal, and from (4-16), the likelihood is

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} e^{-\left(\frac{1}{2}\right)\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})' + n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'\right)\right]} \\ &= (2\pi)^{-\frac{(n-1)p}{2}} |\boldsymbol{\Sigma}|^{-\frac{(n-1)}{2}} e^{-\left(\frac{1}{2}\right)\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'\right)\right]} \\ &\quad \times (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\left(\frac{n}{2}\right)(\bar{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})} \end{aligned} \tag{9-25}$$

which depends on $\mathbf{L}$ and $\boldsymbol{\Psi}$ through $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$. This model is still not well defined, because of the multiplicity of choices for $\mathbf{L}$ made possible by orthogonal transformations. It is desirable to make $\mathbf{L}$ well defined by imposing the computationally convenient *uniqueness condition*

$$\mathbf{L}'\boldsymbol{\Psi}^{-1}\mathbf{L} = \boldsymbol{\Delta} \quad \text{a diagonal matrix} \tag{9-26}$$

The maximum likelihood estimates $\hat{\mathbf{L}}$ and $\hat{\boldsymbol{\Psi}}$ must be obtained by numerical maximization of (9-25). Fortunately, efficient computer programs now exist that enable one to get these estimates rather easily.

We summarize some facts about maximum likelihood estimators and, for now, rely on a computer to perform the numerical details.

Result 9.1. Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}$ is the covariance matrix for the m common factor model of (9-4). The maximum likelihood estimators $\hat{\mathbf{L}}, \hat{\boldsymbol{\Psi}}$, and $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ maximize (9-25) subject to $\hat{\mathbf{L}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\mathbf{L}}$ being diagonal.

The maximum likelihood estimates of the communalities are

$$\hat{h}_i^2 = \hat{\ell}_{i1}^2 + \hat{\ell}_{i2}^2 + \cdots + \hat{\ell}_{im}^2 \quad \text{for } i = 1, 2, \ldots, p \tag{9-27}$$

so

$$\begin{pmatrix} \text{Proportion of total sample} \\ \text{variance due to } j\text{th factor} \end{pmatrix} = \frac{\hat{\ell}_{1j}^2 + \hat{\ell}_{2j}^2 + \cdots + \hat{\ell}_{pj}^2}{s_{11} + s_{22} + \cdots + s_{pp}} \tag{9-28}$$

Proof. By the invariance property of maximum likelihood estimates (see Section 4.3), functions of $\mathbf{L}$ and $\boldsymbol{\Psi}$ are estimated by the same functions of $\hat{\mathbf{L}}$ and $\hat{\boldsymbol{\Psi}}$. In particular, the communalities $h_i^2 = \ell_{i1}^2 + \cdots + \ell_{im}^2$ have maximum likelihood estimates $\hat{h}_i^2 = \hat{\ell}_{i1}^2 + \cdots + \hat{\ell}_{im}^2$. ■

If, as in (8-10), the variables are standardized so that $\mathbf{Z} = \mathbf{V}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$, then the covariance matrix $\boldsymbol{\rho}$ of $\mathbf{Z}$ has the representation

$$\boldsymbol{\rho} = \mathbf{V}^{-1/2}\boldsymbol{\Sigma}\mathbf{V}^{-1/2} = (\mathbf{V}^{-1/2}\mathbf{L})(\mathbf{V}^{-1/2}\mathbf{L})' + \mathbf{V}^{-1/2}\boldsymbol{\Psi}\mathbf{V}^{-1/2} \tag{9-29}$$

Thus, $\boldsymbol{\rho}$ has a factorization analogous to (9-5) with loading matrix $\mathbf{L}_z = \mathbf{V}^{-1/2}\mathbf{L}$ and specific variance matrix $\boldsymbol{\Psi}_z = \mathbf{V}^{-1/2}\boldsymbol{\Psi}\mathbf{V}^{-1/2}$. By the invariance property of maximum likelihood estimators, the maximum likelihood estimator of $\boldsymbol{\rho}$ is

$$\begin{aligned} \hat{\boldsymbol{\rho}} &= (\hat{\mathbf{V}}^{-1/2}\hat{\mathbf{L}})(\hat{\mathbf{V}}^{-1/2}\hat{\mathbf{L}})' + \hat{\mathbf{V}}^{-1/2}\hat{\boldsymbol{\Psi}}\hat{\mathbf{V}}^{-1/2} \\ &= \hat{\mathbf{L}}_z\hat{\mathbf{L}}_z' + \hat{\boldsymbol{\Psi}}_z \end{aligned} \tag{9-30}$$

where $\hat{\mathbf{V}}^{-1/2}$ and $\hat{\mathbf{L}}$ are the maximum likelihood estimators of $\mathbf{V}^{-1/2}$ and $\mathbf{L}$, respectively. (See Supplement 9A.)

As a consequence of the factorization of (9-30), whenever the maximum likelihood analysis pertains to the correlation matrix, we call

$$\hat{h}_i^2 = \hat{\ell}_{i1}^2 + \hat{\ell}_{i2}^2 + \cdots + \hat{\ell}_{im}^2 \quad i = 1, 2, \ldots, p \tag{9-31}$$

the maximum likelihood estimates of the communalities, and we evaluate the importance of the factors on the basis of

$$\begin{pmatrix} \text{Proportion of total (standardized)} \\ \text{sample variance due to } j\text{th factor} \end{pmatrix} = \frac{\hat{\ell}_{1j}^2 + \hat{\ell}_{2j}^2 + \cdots + \hat{\ell}_{pj}^2}{p} \tag{9-32}$$

To avoid more tedious notations, the preceding $\hat{\ell}_{ij}$'s denote the elements of $\hat{\mathbf{L}}_z$.

Comment. Ordinarily, the observations are standardized, and a sample correlation matrix is factor analyzed. The sample correlation matrix $\mathbf{R}$ is inserted for $[(n-1)/n]\mathbf{S}$ in the likelihood function of (9-25), and the maximum likelihood estimates $\hat{\mathbf{L}}_z$ and $\hat{\boldsymbol{\Psi}}_z$ are obtained using a computer. Although the likelihood in (9-25) is appropriate for $\mathbf{S}$, not $\mathbf{R}$, surprisingly, this practice is equivalent to obtaining the maximum likelihood estimates $\hat{\mathbf{L}}$ and $\hat{\boldsymbol{\Psi}}$ based on the sample covariance matrix $\mathbf{S}$, setting $\hat{\mathbf{L}}_z = \hat{\mathbf{V}}^{-1/2}\hat{\mathbf{L}}$ and $\hat{\boldsymbol{\Psi}}_z = \hat{\mathbf{V}}^{-1/2}\hat{\boldsymbol{\Psi}}\hat{\mathbf{V}}^{-1/2}$. Here $\hat{\mathbf{V}}^{-1/2}$ is the diagonal matrix with the reciprocal of the sample standard deviations (computed with the divisor $\sqrt{n}$) on the main diagonal.

Going in the other direction, given the estimated loadings $\hat{\mathbf{L}}_z$ and specific variances $\hat{\boldsymbol{\Psi}}_z$ obtained from $\mathbf{R}$, we find that the resulting maximum likelihood estimates for a factor analysis of the covariance matrix $[(n-1)/n]\mathbf{S}$ are $\hat{\mathbf{L}} = \hat{\mathbf{V}}^{1/2}\hat{\mathbf{L}}_z$ and $\hat{\boldsymbol{\Psi}} = \hat{\mathbf{V}}^{1/2}\hat{\boldsymbol{\Psi}}_z\hat{\mathbf{V}}^{1/2}$, or

$$\hat{\ell}_{ij} = \hat{\ell}_{z,ij}\sqrt{\hat{\sigma}_{ii}} \quad \text{and} \quad \hat{\psi}_i = \hat{\psi}_{z,i}\hat{\sigma}_{ii}$$

where $\hat{\sigma}_{ii}$ is the sample variance computed with divisor n. The distinction between divisors can be ignored with principal component solutions. ■

The equivalency between factoring $\mathbf{S}$ and $\mathbf{R}$ has apparently been confused in many published discussions of factor analysis. (See Supplement 9A.)

Example 9.5 (Factor analysis of stock-price data using the maximum likelihood method) The stock-price data of Examples 8.5 and 9.4 were reanalyzed assuming an $m = 2$ factor model and using the *maximum likelihood method*. The estimated factor loadings, communalities, specific variances, and proportion of total (standardized) sample variance explained by each factor are in Table 9.3.[3] The corresponding figures for the $m = 2$ factor solution obtained by the *principal component method* (see Example 9.4) are also provided. The communalities corresponding to the maximum likelihood factoring of $\mathbf{R}$ are of the form [see (9-31)] $\hat{h}_i^2 = \hat{\ell}_{i1}^2 + \hat{\ell}_{i2}^2$.

So, for example,

$$\hat{h}_1^2 = (.115)^2 + (.765)^2 = .58$$

[3] The maximum likelihood solution leads to a *Heywood case*. For this example, the solution of the likelihood equations give estimated loadings such that a specific variance is negative. The software program obtains a feasible solution by slightly adjusting the loadings so that all specific variance estimates are nonnegative. A Heywood case is suggested here by the .00 value for the specific variance of Royal Dutch Shell.

Table 9.3

Variable	Maximum likelihood			Principal components		
	Estimated factor loadings		Specific variances	Estimated factor loadings		Specific variances
	F_1	F_2	$\hat{\psi}_i = 1 - \hat{h}_i^2$	F_1	F_2	$\tilde{\psi}_i = 1 - \tilde{h}_i^2$
1. J P Morgan	.115	.755	.42	.732	−.437	.27
2. Citibank	.322	.788	.27	.831	−.280	.23
3. Wells Fargo	.182	.652	.54	.726	−.374	.33
4. Royal Dutch Shell	1.000	−.000	.00	.605	.694	.15
5. Texaco	.683	−.032	.53	.563	.719	.17
Cumulative proportion of total (standardized) sample variance explained	.323	.647		.487	.769	

The residual matrix is

$$\mathbf{R} - \hat{\mathbf{L}}\hat{\mathbf{L}}' - \hat{\mathbf{\Psi}} = \begin{bmatrix} 0 & .001 & -.002 & .000 & .052 \\ .001 & 0 & .002 & .000 & -.033 \\ -.002 & .002 & 0 & .000 & .001 \\ .000 & .000 & .000 & 0 & .000 \\ .052 & -.033 & .001 & .000 & 0 \end{bmatrix}$$

The elements of $\mathbf{R} - \hat{\mathbf{L}}\hat{\mathbf{L}}' - \hat{\mathbf{\Psi}}$ are much smaller than those of the residual matrix corresponding to the principal component factoring of $\mathbf{R}$ presented in Example 9.4. On this basis, we prefer the maximum likelihood approach and typically feature it in subsequent examples.

The cumulative proportion of the total sample variance explained by the factors is larger for principal component factoring than for maximum likelihood factoring. It is not surprising that this criterion typically favors principal component factoring. Loadings obtained by a principal component factor analysis are related to the principal components, which have, by design, a variance optimizing property. [See the discussion preceding (8-19).]

Focusing attention on the maximum likelihood solution, we see that all variables have positive loadings on F_1. We call this factor the *market factor*, as we did in the principal component solution. The interpretation of the second factor is not as clear as it appeared to be in the principal component solution. The bank stocks have large positive loadings and the oil stocks have negligible loadings on the second factor F_2. From this perspective, the second factor differentiaties the bank stocks from the oil stocks and might be called an *industry factor*. Alternatively, the second factor might be simply called a *banking factor*.

The patterns of the initial factor loadings for the maximum likelihood solution are constrained by the uniqueness condition that $\hat{\mathbf{L}}'\hat{\mathbf{\Psi}}^{-1}\hat{\mathbf{L}}$ be a diagonal matrix. Therefore, useful factor patterns are often not revealed until the factors are rotated (see Section 9.4). ■

Example 9.6 (Factor analysis of Olympic decathlon data) Linden [11] originally conducted a factor analytic study of Olympic decathlon results for all 160 complete starts from the end of World War II until the mid-seventies. Following his approach we examine the $n = 280$ complete starts from 1960 through 2004. The recorded values for each event were standardized and the signs of the timed events changed so that large scores are good for all events. We, too, analyze the correlation matrix, which based on all 280 cases, is

$$\mathbf{R} = \begin{bmatrix} 1.000 & .6386 & .4752 & .3227 & .5520 & .3262 & .3509 & .4008 & .1821 & -.0352 \\ .6386 & 1.0000 & .4953 & .5668 & .4706 & .3520 & .3998 & .5167 & .3102 & .1012 \\ .4752 & .4953 & 1.0000 & .4357 & .2539 & .2812 & .7926 & .4728 & .4682 & -.0120 \\ .3227 & .5668 & .4357 & 1.0000 & .3449 & .3503 & .3657 & .6040 & .2344 & .2380 \\ .5520 & .4706 & .2539 & .3449 & 1.0000 & .1546 & .2100 & .4213 & .2116 & .4125 \\ .3262 & .3520 & .2812 & .3503 & .1546 & 1.0000 & .2553 & .4163 & .1712 & .0002 \\ .3509 & .3998 & .7926 & .3657 & .2100 & .2553 & 1.0000 & .4036 & .4179 & .0109 \\ .4008 & .5167 & .4728 & .6040 & .4213 & .4163 & .4036 & 1.0000 & .3151 & .2395 \\ .1821 & .3102 & .4682 & .2344 & .2116 & .1712 & .4179 & .3151 & 1.0000 & .0983 \\ -.0352 & .1012 & -.0120 & .2380 & .4125 & .0002 & .0109 & .2395 & .0983 & 1.0000 \end{bmatrix}$$

From a principal component factor analysis perspective, the first four eigenvalues, 4.21, 1.39, 1.06, .92, of $\mathbf{R}$ suggest a factor solution with $m = 3$ or $m = 4$. A subsequent interpretation, much like Linden's original analysis, reinforces the choice $m = 4$.

In this case, the two solution methods produced very different results. For the principal component factorization, all events except the 1,500-meter run have large positive loading on the first factor. This factor might be labeled *general athletic ability*. Factor 2, which loads heavily on the 400-meter run and 1,500-meter run might be called a *running endurance* factor. The remaining factors cannot be easily interpreted to our minds.

For the maximum likelihood method, the first factor appears to be a *general athletic ability factor* but the loading pattern is not as strong as with principal component factor solution. The second factor is primarily a *strength* factor because shot put and discus load highly on this factor. The third factor is *running endurance* since the 400-meter run and 1,500-meter run have large loadings. Again, the fourth factor is not easily identified, although it may have something to do with jumping ability or *leg strength*. We shall return to an interpretation of the factors in Example 9.11 after a discussion of factor rotation.

The four-factor principal component solution accounts for much of the total (standardized) sample variance, although the estimated specific variances are large in some cases (for example, the javelin). This suggests that some events might require *unique* or specific attributes not required for the other events. The four-factor maximum likelihood solution accounts for less of the total sample

Table 9.4

Variable	Principal component: Estimated factor loadings F_1	F_2	F_3	F_4	Specific variances $\tilde{\psi}_i = 1 - \tilde{h}_i^2$	Maximum likelihood: Estimated factor loadings F_1	F_2	F_3	F_4	Specific variances $\hat{\psi}_i = 1 - \hat{h}_i^2$
1. 100-m run	.696	.022	−.468	−.416	.12	.993	−.069	−.021	.002	.01
2. Long jump	.793	.075	−.255	−.115	.29	.665	.252	.239	.220	.39
3. Shot put	.771	−.434	.197	−.112	.17	.530	.777	−.141	−.079	.09
4. High jump	.711	.181	.005	.367	.33	.363	.428	.421	.424	.33
5. 400-m run	.605	.549	−.045	−.397	.17	.571	.019	.620	−.305	.20
6. 100 m hurdles	.513	−.083	−.372	.561	.28	.343	.189	.090	.323	.73
7. Discus	.690	−.456	.289	−.078	.23	.402	.718	−.102	−.095	.30
8. Pole vault	.761	.162	.018	.304	.30	.440	.407	.390	.263	.42
9. Javelin	.518	−.252	.519	−.074	.39	.218	.461	.084	−.085	.73
10. 1500-m run	.220	.746	.493	.085	.15	−.016	.091	.609	−.145	.60
Cumulative proportion of total variance explained	.42	.56	.67	.76		.27	.45	.57	.62	

variance, but, as the following residual matrices indicate, the maximum likelihood estimates $\hat{\mathbf{L}}$ and $\hat{\mathbf{\Psi}}$ do a better job of reproducing $\mathbf{R}$ than the principal component estimates $\tilde{\mathbf{L}}$ and $\tilde{\mathbf{\Psi}}$.

Principal component:

$$\mathbf{R} - \tilde{\mathbf{L}}\tilde{\mathbf{L}}' - \tilde{\mathbf{\Psi}} = \begin{bmatrix} 0 & -.082 & -.006 & -.021 & -.068 & .031 & -.016 & .003 & .039 & .062 \\ -.082 & 0 & -.046 & .033 & -.107 & -.078 & -.048 & -.059 & .042 & .006 \\ -.006 & -.046 & 0 & .006 & -.010 & -.014 & -.003 & -.013 & -.151 & .055 \\ -.021 & .033 & .006 & 0 & -.038 & -.204 & -.015 & -.078 & -.064 & -.086 \\ -.068 & -.107 & -.010 & -.038 & 0 & .096 & .025 & -.006 & .030 & -.074 \\ .031 & -.078 & -.014 & -.204 & .096 & 0 & .015 & -.124 & .119 & .085 \\ -.016 & -.048 & -.003 & -.015 & .025 & .015 & 0 & -.029 & -.210 & .064 \\ .003 & -.059 & -.013 & -.078 & -.006 & -.124 & -.029 & 0 & -.026 & -.084 \\ .039 & .042 & -.151 & -.064 & .030 & .119 & -.210 & -.026 & 0 & -.078 \\ .062 & .006 & .055 & -.086 & -.074 & .085 & .064 & -.084 & -.078 & 0 \end{bmatrix}$$

Maximum likelihood:

$$\mathbf{R} - \hat{\mathbf{L}}\hat{\mathbf{L}}' - \hat{\mathbf{\Psi}} = \begin{bmatrix} 0 & .000 & .000 & -.000 & -.000 & .000 & -.000 & .000 & -.001 & 000 \\ .000 & 0 & -.002 & .023 & .005 & .017 & -.003 & -.030 & .047 & -.024 \\ .000 & -.002 & 0 & .004 & -.000 & -.009 & .000 & -.001 & -.001 & .000 \\ -.000 & .023 & .004 & 0 & -.002 & -.030 & -.004 & -.006 & -.042 & .010 \\ -.000 & .005 & -.001 & -.002 & 0 & -.002 & .001 & .001 & .000 & -.001 \\ .000 & -.017 & -.009 & -.030 & -.002 & 0 & .022 & .069 & .029 & -.019 \\ -.000 & -.003 & .000 & -.004 & .001 & .022 & 0 & -.000 & -.000 & .000 \\ .000 & -.030 & -.001 & -.006 & .001 & .069 & -.000 & 0 & .021 & .011 \\ -.001 & .047 & -.001 & -.042 & .001 & .029 & -.000 & .021 & 0 & -.003 \\ .000 & -.024 & .000 & .010 & -.001 & -.019 & .000 & .011 & -.003 & 0 \end{bmatrix}$$

■

A Large Sample Test for the Number of Common Factors

The assumption of a normal population leads directly to a test of the adequacy of the model. Suppose the m common factor model holds. In this case $\mathbf{\Sigma} = \mathbf{LL}' + \mathbf{\Psi}$, and testing the adequacy of the m common factor model is equivalent to testing

$$H_0: \underset{(p\times p)}{\mathbf{\Sigma}} = \underset{(p\times m)}{\mathbf{L}} \underset{(m\times p)}{\mathbf{L}'} + \underset{(p\times p)}{\mathbf{\Psi}} \tag{9-33}$$

versus H_1: $\mathbf{\Sigma}$ any other positive definite matrix. When $\mathbf{\Sigma}$ does not have any special form, the maximum of the likelihood function [see (4-18) and Result 4.11 with $\hat{\mathbf{\Sigma}} = ((n-1)/n)\mathbf{S} = \mathbf{S}_n$] is proportional to

$$|\mathbf{S}_n|^{-n/2} e^{-np/2} \tag{9-34}$$