CHAPTER 13

EXPLORATORY FACTOR ANALYSIS

13.1 INTRODUCTION

In factor analysis we represent the variables $y_1, y_2, \ldots, y_p$ as linear combinations of a few random variables $f_1, f_2, \ldots, f_m$ $(m < p)$ called *factors*. The factors are underlying *constructs* or *latent* variables that "generate" the y's. Like the original variables, the factors vary from individual to individual; but unlike the variables, the factors cannot be measured or observed. The existence of these hypothetical variables is therefore open to question.

If the original variables $y_1, y_2, \ldots, y_p$ are at least moderately correlated, the basic dimensionality of the system is less than p. The goal of factor analysis is to reduce the redundancy among the variables by using a smaller number of factors.

Suppose the pattern of the high and low correlations in the correlation matrix is such that the variables in a particular subset have high correlations among themselves but low correlations with all the other variables. Then there may be a single underlying factor that gave rise to the variables in the subset. If the other variables can be similarly grouped into subsets with a like pattern of correlations, then a few factors can represent these groups of variables. In this case the pattern in the correlation ma-

Methods of Multivariate Analysis, Third Edition. By Alvin C. Rencher and William F. Christensen

trix corresponds directly to the factors. For example, suppose the correlation matrix has the form

$$\begin{pmatrix} 1 & .9 & .05 & .05 & .05 \\ .9 & 1 & .05 & .05 & .05 \\ .05 & .05 & 1 & .9 & .9 \\ .05 & .05 & .9 & 1 & .9 \\ .05 & .05 & .9 & .9 & 1 \end{pmatrix}.$$

Then variables 1 and 2 correspond to a factor and variables 3, 4, and 5 correspond to another factor. In some cases where the correlation matrix does not have such a simple pattern, factor analysis will still partition the variables into clusters.

Factor analysis is related to principal component analysis in that both seek a simpler structure in a set of variables, but they differ in many respects (see Section 13.8). For example, two differences in basic approach are as follows:

1. Principal components are defined as linear combinations of the original variables. In factor analysis, the original variables are expressed as linear combinations of the factors.

2. In principal component analysis, we explain a large part of the total variance of the variables, $\sum_i s_{ii}$. In factor analysis, we seek to account for the covariances or correlations among the variables.

Although we have thus far compared principal component analysis with "factor analysis," we can draw further distinctions between the *exploratory factor analysis* described in this chapter and the *confirmatory factor analysis* that will be discussed in Chapter 14. Exploratory factor analysis (or EFA) is most often used to explore multivariate data to identify possible latent structure. In EFA, the number of latent factors is not determined before the analysis. In contrast, confirmatory factor analysis (or CFA) allows the researcher to hypothesize the number of latent factors and the specific nature of the latent structure in the data, and then test the hypotheses that have been formulated. Thus the objectives and methods of EFA are more closely related to those of principal component analysis, while CFA uses the more traditional statistical notions of model formulation, parameter estimation, model evaluation, and statistical inference. Because of the subjectivity and relative lack of formal statistical inference associated with EFA, it is considered by some to be controversial or of limited utility. Regardless, EFA has proven to be a useful tool in many settings, particularly as a technique for simplifying complex multivariate data and formulating hypotheses that can be later verified by using a CFA on an independently obtained data set (see Gerbing and Hamilton, 1996).

For either type of factor analysis, there will be some data sets for which the factor analysis model does not provide a satisfactory fit. Sometimes a few easily interpretable factors emerge, but for other data sets neither the number of factors nor the interpretation is clear. Some possible reasons for these failures are discussed in Section 13.7.

In this chapter, we describe the methods associated with exploratory factor analysis. Throughout this chapter, we will use "exploratory factor analysis" and "factor analysis" interchangeably.

13.2 ORTHOGONAL FACTOR MODEL

13.2.1 Model Definition and Assumptions

Factor analysis is basically a one-sample procedure [for possible applications to data with groups, see Rencher (1998, Section 10.8)]. We assume a random sample $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ from a homogeneous population with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

The factor analysis model expresses each variable as a linear combination of underlying *common factors* $f_1, f_2, \ldots, f_m$, with an accompanying error term to account for that part of the variable that is unique (not in common with the other variables). For $y_1, y_2, \ldots, y_p$ in any observation vector $\mathbf{y}$, the model is as follows:

$$\begin{aligned} y_1 - \mu_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1m} f_m + \varepsilon_1 \\ y_2 - \mu_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2m} f_m + \varepsilon_2 \\ &\vdots \\ y_p - \mu_p &= \lambda_{p1} f_1 + \lambda_{p2} f_2 + \cdots + \lambda_{pm} f_m + \varepsilon_p. \end{aligned} \tag{13.1}$$

Ideally, m should be substantially smaller than p; otherwise we have not achieved a parsimonious description of the variables as functions of a few underlying factors. We might regard the f's in (13.1) as random variables that engender the y's. The coefficients λ_{ij} are called *loadings* and serve as weights, showing how each y_i individually depends on the f's. (In this chapter, we defer to common usage in the factor analysis literature and use the notation λ_{ij} for loadings rather than eigenvalues.) With appropriate assumptions, λ_{ij} indicates the importance of the jth factor f_j to the ith variable y_i and can be used in interpretation of f_j. We describe or interpret f_2, for example, by examining its coefficients, $\lambda_{12}, \lambda_{22}, \ldots, \lambda_{p2}$. The larger loadings relate f_2 to the corresponding y's. From these y's, we infer a meaning or description of f_2. After estimating the λ_{ij}'s (and rotating them, see Sections 13.2.2 and 13.5), they will hopefully partition the variables into groups corresponding to factors.

The system of equations (13.1) bears a superficial resemblance to the multiple regression model (10.1), but there are fundamental differences. For example, (1) the f's are unobserved and (2) the model in (13.1) represents only one observation vector, while (10.1) depicts all n observations.

It is assumed that for $j = 1, 2, \ldots, m$, $E(f_j) = 0$, $\text{var}(f_j) = 1$, and $\text{cov}(f_j, f_k) = 0$, $j \neq k$. The assumptions for $\varepsilon_i, i = 1, 2, \ldots, p$, are similar, except that we must allow each ε_i to have a different variance, since it shows the residual part of y_i that is not in common with the other variables. Thus we assume that $E(\varepsilon_i) = 0$, $\text{var}(\varepsilon_i) = \psi_i$, and $\text{cov}(\varepsilon_i, \varepsilon_k) = 0$, $i \neq k$. In addition, we assume that $\text{cov}(\varepsilon_i, f_j) = 0$ for all i and j. We refer to ψ_i as the *specific variance*.

These assumptions are natural consequences of the basic model (13.1) and the goals of factor analysis. Since $E(y_i - \mu_i) = 0$, we need $E(f_j) = 0, j = 1, 2, \ldots, m$. The assumption $\text{cov}(f_j, f_k) = 0$ is made for parsimony in expressing the y's as functions of as few factors as possible. The assumptions $\text{var}(f_j) = 1$, $\text{var}(\varepsilon_i) = \psi_i$, $\text{cov}(f_j, f_k) = 0$, and $\text{cov}(\varepsilon_i, f_j) = 0$ yield a simple expression for the variance of y_i,

$$\text{var}(y_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2 + \psi_i, \tag{13.2}$$

which plays an important role in our development. The assumption $\text{cov}(\varepsilon_i, \varepsilon_k) = 0$ implies that the factors account for all the correlations among the y's, that is, all that the y's have in common. Thus the emphasis in factor analysis is on modeling the covariances or correlations among the y's.

Model (13.1) can be written in matrix notation as

$$\mathbf{y} - \boldsymbol{\mu} = \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\varepsilon}, \tag{13.3}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_p)'$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)'$, $\mathbf{f} = (f_1, f_2, \ldots, f_m)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p)'$, and

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pm} \end{pmatrix}. \tag{13.4}$$

We illustrate the model in (13.1) and (13.3) with $p = 5$ and $m = 2$. The model for each variable in (13.1) becomes

$$\begin{aligned} y_1 - \mu_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \varepsilon_1 \\ y_2 - \mu_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \varepsilon_2 \\ y_3 - \mu_3 &= \lambda_{31} f_1 + \lambda_{32} f_2 + \varepsilon_3 \\ y_4 - \mu_4 &= \lambda_{41} f_1 + \lambda_{42} f_2 + \varepsilon_4 \\ y_5 - \mu_5 &= \lambda_{51} f_1 + \lambda_{52} f_2 + \varepsilon_5. \end{aligned}$$

In matrix notation as in (13.3), this becomes

$$\begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ y_3 - \mu_3 \\ y_4 - \mu_4 \\ y_5 - \mu_5 \end{pmatrix} = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ \lambda_{41} & \lambda_{42} \\ \lambda_{51} & \lambda_{52} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}, \tag{13.5}$$

or $\mathbf{y} - \boldsymbol{\mu} = \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\varepsilon}$.

The assumptions described above between (13.1) and (13.2) can be expressed concisely using vector and matrix notation:

$E(f_j) = 0, j = 1, 2, \ldots, m$, becomes

$$E(\mathbf{f}) = \mathbf{0}, \tag{13.6}$$

$\text{var}(f_j) = 1, j = 1, 2, \ldots, m$, and $\text{cov}(f_j, f_k) = 0, j \neq k$, become

$$\text{cov}(\mathbf{f}) = \mathbf{I}, \tag{13.7}$$

$E(\varepsilon_i) = 0, i = 1, 2, \ldots, p$, becomes

$$E(\boldsymbol{\varepsilon}) = \mathbf{0}, \tag{13.8}$$

$\text{var}(\varepsilon_i) = \psi_i, i = 1, 2, \ldots, p$, and $\text{cov}(\varepsilon_i, \varepsilon_k) = 0, i \neq k$, become

$$\text{cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi} = \begin{pmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \psi_p \end{pmatrix}, \tag{13.9}$$

and $\text{cov}(\varepsilon_i, f_j) = 0$ for all i and j becomes

$$\text{cov}(\mathbf{f}, \boldsymbol{\varepsilon}) = \mathbf{O}. \tag{13.10}$$

The notation $\text{cov}(\mathbf{f}, \boldsymbol{\varepsilon})$ indicates a rectangular matrix containing the covariances of the f's with the ε's:

$$\text{cov}(\mathbf{f}, \boldsymbol{\varepsilon}) = \begin{pmatrix} \sigma_{f_1\varepsilon_1} & \sigma_{f_1\varepsilon_2} & \cdots & \sigma_{f_1\varepsilon_p} \\ \sigma_{f_2\varepsilon_1} & \sigma_{f_2\varepsilon_2} & \cdots & \sigma_{f_2\varepsilon_p} \\ \vdots & \vdots & & \vdots \\ \sigma_{f_m\varepsilon_1} & \sigma_{f_m\varepsilon_2} & \cdots & \sigma_{f_m\varepsilon_p} \end{pmatrix}.$$

It was noted following (13.2) that the emphasis in factor analysis is on modeling the covariances among the y's. We wish to express the $\frac{1}{2}p(p-1)$ covariances (and the p variances) of the variables $y_1, y_2, \ldots, y_p$ in terms of a simplified structure involving the pm loadings λ_{ij} and the p specific variances ψ_i; that is, we wish to express $\boldsymbol{\Sigma}$ in terms of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$. We can do this using the model (13.3) and the assumptions (13.7), (13.9), and (13.10). Since $\boldsymbol{\mu}$ does not affect variances and covariances of $\mathbf{y}$, we have from (13.3)

$$\boldsymbol{\Sigma} = \text{cov}(\mathbf{y}) = \text{cov}(\boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\varepsilon}).$$

By (13.10), $\boldsymbol{\Lambda}\mathbf{f}$ and $\boldsymbol{\varepsilon}$ are uncorrelated; therefore, the covariance matrix of their sum is the sum of their covariance matrices:

$$\begin{aligned} \boldsymbol{\Sigma} &= \text{cov}(\boldsymbol{\Lambda}\mathbf{f}) + \text{cov}(\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\Lambda}\text{cov}(\mathbf{f})\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \quad [\text{by (3.74) and (13.9)}] \\ &= \boldsymbol{\Lambda}\mathbf{I}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \quad [\text{by (13.7)}] \\ &= \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \end{aligned} \tag{13.11}$$

If $\boldsymbol{\Lambda}$ has only a few columns, say two or three, then $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ in (13.11) represents a simplified structure for $\boldsymbol{\Sigma}$, in which the covariances are modeled by the

λ_{ij}'s alone since $\boldsymbol{\Psi}$ is diagonal. For example, in the above illustration in (13.5) with $m = 2$ factors, σ_{12} would be the product of the first two rows of $\boldsymbol{\Lambda}$, that is,

$$\sigma_{12} = \text{cov}(y_1, y_2) = \lambda_{11}\lambda_{21} + \lambda_{12}\lambda_{22},$$

where $(\lambda_{11}, \lambda_{12})$ is the first row of $\boldsymbol{\Lambda}$ and $(\lambda_{21}, \lambda_{22})$ is the second row of $\boldsymbol{\Lambda}$. If y_1 and y_2 have a great deal in common, they will have similar loadings on the common factors f_1 and f_2; that is, $(\lambda_{11}, \lambda_{12})$ will be similar to $(\lambda_{21}, \lambda_{22})$. In this case, either $\lambda_{11}\lambda_{21}$ or $\lambda_{12}\lambda_{22}$ is likely to be high. On the other hand, if y_1 and y_2 have little in common, then their loadings λ_{11} and λ_{21} on f_1 will be different and their loadings λ_{12} and λ_{22} on f_2 will likewise differ. In this case, the products $\lambda_{11}\lambda_{21}$ and $\lambda_{12}\lambda_{22}$ will tend to be small.

We can also find the covariances of the y's with the f's in terms of the λ's. Consider, for example, $\text{cov}(y_1, f_2)$. By (13.1), $y_1 - \mu_1 = \lambda_{11}f_1 + \lambda_{12}f_2 + \cdots + \lambda_{1m}f_m + \varepsilon_1$. From (13.7), f_2 is uncorrelated with all other f_j's, and by (13.10), f_2 is uncorrelated with ε_1. Thus

$$\begin{aligned}
\text{cov}(y_1, f_2) &= E[(y_1 - \mu_1)(f_2 - \mu_{f_2})] \\
&= E[(\lambda_{11}f_1 + \lambda_{12}f_2 + \cdots + \lambda_{1m}f_m)f_2] \\
&= E(\lambda_{11}f_1f_2 + \lambda_{12}f_2^2 + \cdots + \lambda_{1m}f_mf_2) \\
&= \lambda_{11}\text{cov}(f_1, f_2) + \lambda_{12}\text{var}(f_2) + \cdots + \lambda_{1m}\text{cov}(f_m, f_2) \\
&= \lambda_{12},
\end{aligned}$$

since $\text{var}(f_2) = 1$. Hence the loadings themselves represent covariances of the variables with the factors. In general,

$$\text{cov}(y_i, f_j) = \lambda_{ij}, \quad i = 1, 2, \ldots, p, \quad j = 1, 2, \ldots, m. \tag{13.12}$$

Since λ_{ij} is the (ij)th element of $\boldsymbol{\Lambda}$, we can write (13.12) in the form

$$\text{cov}(\mathbf{y}, \mathbf{f}) = \boldsymbol{\Lambda}. \tag{13.13}$$

If standardized variables are used, (13.11) is replaced by $\mathbf{P}_\rho = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$, and the loadings become correlations:

$$\text{corr}(y_i, f_j) = \lambda_{ij}. \tag{13.14}$$

In (13.2), we have a partitioning of the variance of y_i into a component due to the common factors, called the *communality*, and a component unique to y_i, called the *specific variance*:

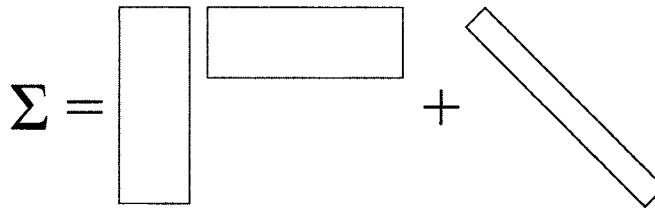
$$\begin{aligned}
\sigma_{ii} = \text{var}(y_i) &= (\lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2) + \psi_i \\
&= h_i^2 + \psi_i \\
&= \text{communality} + \text{specific variance},
\end{aligned}$$

where

$$\begin{aligned} \text{Communality} &= h_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2, \\ \text{Specific variance} &= \psi_i. \end{aligned} \tag{13.15}$$

The communality h_i^2 is also referred to as *common variance*, and the specific variance ψ_i has been called *specificity, unique variance*, or *residual variance.*

Assumptions (13.6)–(13.10) lead to the simple covariance structure of (13.11), $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$, which is an essential part of the factor analysis model. In schematic form, $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ has the following appearance:



The diagonal elements of $\boldsymbol{\Sigma}$ can be easily modeled by adjusting the diagonal elements of $\boldsymbol{\Psi}$, but $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ is a simplified configuration for the off-diagonal elements. Hence the critical aspect of the model involves the covariances, and this is the major emphasis of factor analysis, as noted in Section 13.1 and in comments following (13.2) and (13.10).

It is a rare population covariance matrix $\boldsymbol{\Sigma}$ that can be expressed exactly as $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}$ is diagonal and $\boldsymbol{\Lambda}$ is $p \times m$, with m relatively small. In practice, many sample covariance matrices do not come satisfactorily close to this ideal pattern. However, we do not relax the assumptions because the structure $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ is essential for estimation of $\boldsymbol{\Lambda}$.

One advantage of the factor analysis model is that when it does not fit the data, the estimate of $\boldsymbol{\Lambda}$ clearly reflects this failure. In such cases, there are two problems in the estimates: (1) it is unclear how many factors there should be, and (2) it is unclear what the factors are. In other statistical procedures, failure of assumptions may not lead to such obvious consequences in the estimates or tests. In factor analysis, the assumptions are essentially self-checking, whereas in other procedures, we typically have to check the assumptions with residual plots, tests, and so on.

13.2.2 Nonuniqueness of Factor Loadings

The loadings in the model (13.3) can be multiplied by an orthogonal matrix without impairing their ability to reproduce the covariance matrix in $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. To see this, let $\mathbf{T}$ be an arbitrary orthogonal matrix. Then by (2.102), $\mathbf{T}\mathbf{T}' = \mathbf{I}$, and we can insert $\mathbf{T}\mathbf{T}'$ into the basic model (13.3) to obtain

$$\mathbf{y} - \boldsymbol{\mu} = \boldsymbol{\Lambda}\mathbf{T}\mathbf{T}'\mathbf{f} + \boldsymbol{\varepsilon}.$$

We then associate $\mathbf{T}$ with $\boldsymbol{\Lambda}$ and associate $\mathbf{T}'$ with $\mathbf{f}$ so that the model becomes

$$\mathbf{y} - \boldsymbol{\mu} = \boldsymbol{\Lambda}^*\mathbf{f}^* + \boldsymbol{\varepsilon}, \tag{13.16}$$

where

$$\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\mathbf{T}, \tag{13.17}$$

$$\mathbf{f}^* = \mathbf{T}'\mathbf{f}. \tag{13.18}$$

If $\boldsymbol{\Lambda}$ in $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ is replaced by $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\mathbf{T}$, we have

$$\begin{aligned}\boldsymbol{\Sigma} &= \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*'} + \boldsymbol{\Psi} = \boldsymbol{\Lambda}\mathbf{T}(\boldsymbol{\Lambda}\mathbf{T})' + \boldsymbol{\Psi} \\ &= \boldsymbol{\Lambda}\mathbf{T}\mathbf{T}'\boldsymbol{\Lambda}' + \boldsymbol{\Psi} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi},\end{aligned}$$

since $\mathbf{T}\mathbf{T}' = \mathbf{I}$. Thus the new loadings $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\mathbf{T}$ in (13.17) reproduce the covariance matrix, just as $\boldsymbol{\Lambda}$ does in (13.11):

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*'} + \boldsymbol{\Psi} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \tag{13.19}$$

The new factors $\mathbf{f}^* = \mathbf{T}'\mathbf{f}$ in (13.18) satisfy the assumptions (13.6), (13.7), and (13.10); that is, $E(\mathbf{f}^*) = \mathbf{0}$, $\text{cov}(\mathbf{f}^*) = \mathbf{I}$, and $\text{cov}(\mathbf{f}^*, \boldsymbol{\varepsilon}) = \mathbf{O}$.

The communalities $h_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2$, $i = 1, 2, \ldots, p$, as defined in (13.15), are also unaffected by the transformation $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\mathbf{T}$. This can be seen as follows. The communality h_i^2 is the sum of squares of the ith row of $\boldsymbol{\Lambda}$. If we denote the ith row of $\boldsymbol{\Lambda}$ by $\boldsymbol{\lambda}_i'$, then the sum of squares in vector notation is $h_i^2 = \boldsymbol{\lambda}_i'\boldsymbol{\lambda}_i$. The ith row of $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\mathbf{T}$ is $\boldsymbol{\lambda}_i^{*'} = \boldsymbol{\lambda}_i'\mathbf{T}$, and the corresponding communality is

$$h_i^{*2} = \boldsymbol{\lambda}_i^{*'}\boldsymbol{\lambda}_i^* = \boldsymbol{\lambda}_i'\mathbf{T}\mathbf{T}'\boldsymbol{\lambda}_i = \boldsymbol{\lambda}_i'\boldsymbol{\lambda}_i = h_i^2.$$

Thus the communalities remain the same for the new loadings. Note that $h_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2 = \boldsymbol{\lambda}_i'\boldsymbol{\lambda}_i$ is the distance from the origin to the point $\boldsymbol{\lambda}_i' = (\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{im})$ in the m-dimensional space of the factor loadings. Since the distance $\boldsymbol{\lambda}_i'\boldsymbol{\lambda}_i$ is the same as $\boldsymbol{\lambda}_i^{*'}\boldsymbol{\lambda}_i^*$, the points $\boldsymbol{\lambda}_i^*$ are rotated from the points $\boldsymbol{\lambda}_i$. [This also follows because $\boldsymbol{\lambda}_i^{*'} = \boldsymbol{\lambda}_i'\mathbf{T}$, where $\mathbf{T}$ is orthogonal. Multiplication of a vector by an orthogonal matrix is equivalent to a rotation of axes; see (2.103).]

The inherent potential to rotate the loadings to a new frame of reference without affecting any assumptions or properties is very useful in interpretation of the factors and will be exploited in Section 13.5.

Note that the coefficients (loadings) in (13.1) are applied to the factors, not to the variables, as they are in discriminant functions and principal components. Thus in factor analysis the observed variables are not involved in the rotation as they are in discriminant functions and principal components.

13.3 ESTIMATION OF LOADINGS AND COMMUNALITIES

In Sections 13.3.1–13.3.4, we discuss four approaches to estimation of loadings and communalities.

13.3.1 Principal Component Method

The first technique we consider is commonly called the *principal component* method. This name is perhaps unfortunate in that it adds to the confusion between factor analysis and principal component analysis. In the principal component method for estimation of loadings, we do not actually calculate any principal components. The reason for the name is given following (13.25).

From a random sample $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, we obtain the sample covariance matrix $\mathbf{S}$ and then attempt to find an estimator $\hat{\mathbf{\Lambda}}$ that will approximate the fundamental expression (13.11) with $\mathbf{S}$ in place of $\mathbf{\Sigma}$:

$$\mathbf{S} \cong \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}. \tag{13.20}$$

In the principal component approach, we neglect $\hat{\mathbf{\Psi}}$ and factor $\mathbf{S}$ into $\mathbf{S} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'$.

In order to factor $\mathbf{S}$, we use the spectral decomposition in (2.109),

$$\mathbf{S} = \mathbf{CDC}', \tag{13.21}$$

where $\mathbf{C}$ is an orthogonal matrix constructed with normalized eigenvectors ($\mathbf{c}_i'\mathbf{c}_i = 1$) of $\mathbf{S}$ as columns and $\mathbf{D}$ is a diagonal matrix with the eigenvalues $\theta_1, \theta_2, \ldots, \theta_p$ of $\mathbf{S}$ on the diagonal:

$$\mathbf{D} = \begin{pmatrix} \theta_1 & 0 & \cdots & 0 \\ 0 & \theta_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \theta_p \end{pmatrix}. \tag{13.22}$$

We use the notation θ_i for eigenvalues instead of the usual λ_i in order to avoid confusion with the notation λ_{ij} used for the loadings.

To finish factoring $\mathbf{CDC}'$ in (13.21) into the form $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'$, we observe that since the eigenvalues θ_i of the positive semidefinite matrix $\mathbf{S}$ are all positive or zero, we can factor $\mathbf{D}$ into

$$\mathbf{D} = \mathbf{D}^{1/2}\mathbf{D}^{1/2},$$

where

$$\mathbf{D}^{1/2} = \begin{pmatrix} \sqrt{\theta_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\theta_2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sqrt{\theta_p} \end{pmatrix}.$$

With this factoring of $\mathbf{D}$, (13.21) becomes,

$$\begin{aligned} \mathbf{S} &= \mathbf{CDC}' = \mathbf{CD}^{1/2}\mathbf{D}^{1/2}\mathbf{C}' \\ &= (\mathbf{CD}^{1/2})(\mathbf{CD}^{1/2})'. \end{aligned} \tag{13.23}$$

This is of the form $\mathbf{S} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'$, but we do not define $\hat{\mathbf{\Lambda}}$ to be $\mathbf{CD}^{1/2}$ because $\mathbf{CD}^{1/2}$ is $p \times p$, and we are seeking a $\hat{\mathbf{\Lambda}}$ that is $p \times m$ with $m < p$. We therefore define

$\mathbf{D}_1 = \text{diag}(\theta_1, \theta_2, \ldots, \theta_m)$ with the m largest eigenvalues $\theta_1 > \theta_2 > \cdots > \theta_m$ and $\mathbf{C}_1 = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_m)$ containing the corresponding eigenvectors. We then estimate $\mathbf{\Lambda}$ by the first m columns of $\mathbf{CD}^{1/2}$,

$$\hat{\mathbf{\Lambda}} = \mathbf{C}_1\mathbf{D}_1^{1/2} = (\sqrt{\theta_1}\mathbf{c}_1, \sqrt{\theta_2}\mathbf{c}_2, \ldots, \sqrt{\theta_m}\mathbf{c}_m) \tag{13.24}$$

by (2.56), where $\hat{\mathbf{\Lambda}}$ is $p \times m$, $\mathbf{C}_1$ is $p \times m$, and $\mathbf{D}_1^{1/2}$ is $m \times m$.

We illustrate the structure of the $\hat{\lambda}_{ij}$ in (13.24) for $p = 5$ and $m = 2$:

$$\begin{aligned}
\begin{pmatrix} \hat{\lambda}_{11} & \hat{\lambda}_{12} \\ \hat{\lambda}_{21} & \hat{\lambda}_{22} \\ \hat{\lambda}_{31} & \hat{\lambda}_{32} \\ \hat{\lambda}_{41} & \hat{\lambda}_{42} \\ \hat{\lambda}_{51} & \hat{\lambda}_{52} \end{pmatrix} &= \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \\ c_{41} & c_{42} \\ c_{51} & c_{52} \end{pmatrix} \begin{pmatrix} \sqrt{\theta_1} & 0 \\ 0 & \sqrt{\theta_2} \end{pmatrix} \\
&= \begin{pmatrix} \sqrt{\theta_1}c_{11} & \sqrt{\theta_2}c_{12} \\ \sqrt{\theta_1}c_{21} & \sqrt{\theta_2}c_{22} \\ \sqrt{\theta_1}c_{31} & \sqrt{\theta_2}c_{32} \\ \sqrt{\theta_1}c_{41} & \sqrt{\theta_2}c_{42} \\ \sqrt{\theta_1}c_{51} & \sqrt{\theta_2}c_{52} \end{pmatrix} \quad \text{[by (2.56)]}.
\end{aligned} \tag{13.25}$$

We can see in (13.25) the source of the term *principal component* solution. The columns of $\hat{\mathbf{\Lambda}}$ are proportional to the eigenvectors of $\mathbf{S}$, so that the loadings on the jth factor are proportional to coefficients in the jth principal component. The factors are thus related to the first m principal components, and it would seem that interpretation would be the same as for principal components. But after rotation of the loadings, the interpretation of the factors is usually different. The researcher will ordinarily prefer the rotated factors for reasons to be treated in Section 13.5.

By (2.52), the ith diagonal element of $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'$ is the sum of squares of the ith row of $\hat{\mathbf{\Lambda}}$, or $\hat{\boldsymbol{\lambda}}_i'\hat{\boldsymbol{\lambda}}_i = \sum_{j=1}^m \hat{\lambda}_{ij}^2$. Hence to complete the approximation of $\mathbf{S}$ in (13.20), we define

$$\hat{\psi}_i = s_{ii} - \sum_{j=1}^m \hat{\lambda}_{ij}^2 \tag{13.26}$$

and write

$$\mathbf{S} \cong \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}, \tag{13.27}$$

where $\hat{\mathbf{\Psi}} = \text{diag}(\hat{\psi}_1, \hat{\psi}_2, \ldots, \hat{\psi}_p)$. Thus in (13.27) the variances on the diagonal of $\mathbf{S}$ are modeled exactly, but the off-diagonal covariances are only approximate. Again, this is the challenge of factor analysis.

In this method of estimation, the sums of squares of the rows and columns of $\hat{\mathbf{\Lambda}}$ are equal to communalities and eigenvalues, respectively. This is easily shown. By (13.26) and by analogy with (13.15), the ith communality is estimated by

$$\hat{h}_i^2 = \sum_{j=1}^m \hat{\lambda}_{ij}^2, \tag{13.28}$$

which is the sum of squares of the ith row of $\hat{\mathbf{\Lambda}}$. The sum of squares of the jth column of $\hat{\mathbf{\Lambda}}$ is the jth eigenvalue of $\mathbf{S}$:

$$\begin{aligned}\sum_{i=1}^{p} \hat{\lambda}_{ij}^2 &= \sum_{i=1}^{p} (\sqrt{\theta_j} c_{ij})^2 \qquad \text{[by (13.25)]} \\ &= \theta_j \sum_{i=1}^{p} c_{ij}^2 \\ &= \theta_j, \end{aligned} \tag{13.29}$$

since the normalized eigenvectors (columns of $\mathbf{C}$) have length 1.

By (13.26) and (13.28), the variance of the ith variable is partitioned into a part due to the factors and a part due uniquely to the variable:

$$\begin{aligned} s_{ii} &= \hat{h}_i^2 + \hat{\psi}_i \\ &= \hat{\lambda}_{i1}^2 + \hat{\lambda}_{i2}^2 + \cdots + \hat{\lambda}_{im}^2 + \hat{\psi}_i. \end{aligned} \tag{13.30}$$

Thus the jth factor contributes $\hat{\lambda}_{ij}^2$ to s_{ii}. The contribution of the jth factor to the total sample variance, $\text{tr}(\mathbf{S}) = s_{11} + s_{22} + \cdots + s_{pp}$, is therefore

$$\text{variance due to } j\text{th factor} = \sum_{i=1}^{p} \hat{\lambda}_{ij}^2 = \hat{\lambda}_{1j}^2 + \hat{\lambda}_{2j}^2 + \cdots + \hat{\lambda}_{pj}^2, \tag{13.31}$$

which is the sum of squares of loadings in the jth column of $\hat{\mathbf{\Lambda}}$. By (13.29), this is equal to the jth eigenvalue, θ_j. The proportion of total sample variance due to the jth factor is

$$\frac{\sum_{i=1}^{p} \hat{\lambda}_{ij}^2}{\text{tr}(\mathbf{S})} = \frac{\theta_j}{\text{tr}(\mathbf{S})}. \tag{13.32}$$

If the variables are not commensurate, we can use standardized variables and work with the correlation matrix $\mathbf{R}$. The eigenvalues and eigenvectors of $\mathbf{R}$ are then used in place of those of $\mathbf{S}$ in (13.24) to obtain estimates of the loadings. In practice, $\mathbf{R}$ is used more often than $\mathbf{S}$ and is the default in most software packages. Since the emphasis in factor analysis is on reproducing the covariances or correlations rather than the variances, use of $\mathbf{R}$ is more appropriate in factor analysis than in principal components. In applications, $\mathbf{R}$ often gives better results than $\mathbf{S}$.

If we are factoring $\mathbf{R}$, the proportion corresponding to (13.32) is

$$\frac{\sum_{i=1}^{p} \hat{\lambda}_{ij}^2}{\text{tr}(\mathbf{R})} = \frac{\theta_j}{p}, \tag{13.33}$$

where p is the number of variables.

We can assess the fit of the factor analysis model by comparing the left and right sides of (13.27). The error matrix

$$\mathbf{E} = \mathbf{S} - (\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}})$$

Table 13.1 Perception Data: Ratings on Five Adjectives for Seven People

People	Kind	Intelligent	Happy	Likeable	Just
FSM1[a]	1	5	5	1	1
SISTER	8	9	7	9	8
FSM2	9	8	9	9	8
FATHER	9	9	9	9	9
TEACHER	1	9	1	1	9
MSM[b]	9	7	7	9	9
FSM3	9	7	9	9	7

[a]Female schoolmate 1.
[b]Male schoolmate.

has zeros on the diagonal but nonzero off-diagonal elements. The following inequality gives a bound on the size of the elements in $\mathbf{E} = \mathbf{S} - (\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}})$:

$$\sum_{ij} e_{ij}^2 \leq \theta_{m+1}^2 + \theta_{m+2}^2 + \cdots + \theta_p^2; \tag{13.34}$$

that is, the sum of squared entries in the matrix $\mathbf{E} = \mathbf{S} - (\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}})$ is at most equal to the sum of squares of the deleted eigenvalues of $\mathbf{S}$. If the eigenvalues are small, the residuals in the error matrix $\mathbf{S} - (\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}})$ are small and the fit is good.

■ **EXAMPLE 13.3.1**

To illustrate the principal component method of estimation, we use a simple data set collected by Brown et al. (1984). A 12-year-old girl made five ratings on a nine-point semantic differential scale for each of seven of her acquaintances. The ratings were based on the five adjectives "kind," "intelligent," "happy," "likeable," and "just." Her ratings are given in Table 13.1.

The correlation matrix for the five variables (adjectives) is as follows, with the larger values bolded:

$$\mathbf{R} = \begin{pmatrix} 1.000 & .296 & \mathbf{.881} & \mathbf{.995} & .545 \\ .296 & 1.000 & -.022 & .326 & \mathbf{.837} \\ \mathbf{.881} & -.022 & 1.000 & \mathbf{.867} & .130 \\ \mathbf{.995} & .326 & \mathbf{.867} & 1.000 & .544 \\ .545 & \mathbf{.837} & .130 & .544 & 1.000 \end{pmatrix}. \tag{13.35}$$

The boldface values indicate two groups of variables: $\{1, 3, 4\}$ and $\{2, 5\}$. We would therefore expect that the correlations among the variables can be explained fairly well by two factors.

The eigenvalues of $\mathbf{R}$ are 3.263, 1.538, .168, .031, and 0. Thus $\mathbf{R}$ is singular, which is possible in a situation such as this with only seven observations on five variables recorded in a single-digit scale. The multicollinearity among

Table 13.2 Factor Loadings by the Principal Component Method for the Perception Data of Table 13.1

Variables	Loadings $\hat{\lambda}_{1j}$	Loadings $\hat{\lambda}_{2j}$	Communalities, $\hat{h}_i^2$	Specific Variances, $\hat{\psi}_i$
Kind	.969	−.231	.993	.007
Intelligent	.519	.807	.921	.079
Happy	.785	−.587	.960	.040
Likeable	.971	−.210	.987	.013
Just	.704	.667	.940	.060
Variance accounted for	3.263	1.538	4.802	
Proportion of total variance	.653	.308	.960	
Cumulative proportion	.653	.960	.960	

the variables induced by the fifth eigenvalue, 0, could be ascertained from the corresponding eigenvector, as noted in Section 12.7 (see Problem 13.6).

By (13.33), the first two factors account for $(3.263 + 1.538)/5 = .96$ of the total sample variance. We therefore extract two factors. The first two eigenvectors are

$$\mathbf{c}_1 = \begin{pmatrix} .537 \\ .288 \\ .434 \\ .537 \\ .390 \end{pmatrix} \quad \text{and} \quad \mathbf{c}_2 = \begin{pmatrix} -.186 \\ .651 \\ -.473 \\ -.169 \\ .538 \end{pmatrix}.$$

When these are multiplied by the square roots of the respective eigenvalues 3.263 and 1.538 as in (13.25), we obtain the loadings in Table 13.2.

The communalities in Table 13.2 are obtained from the sum of squares of the rows of the loadings, as in (13.28). The first one, for example, is $(.969)^2 + (-.231)^2 = .993$. The specific variances are obtained from (13.26) as $\hat{\psi}_i = 1 - \hat{h}_i^2$ using 1 in place of s_{ii} because we are factoring $\mathbf{R}$ rather than $\mathbf{S}$. The variance accounted for by each factor is the sum of squares of the corresponding column of the loadings, as in (13.31). By (13.29), the variance accounted for is also equal to the eigenvalue in each case. Note that the variance accounted for by the two factors adds to the sum of the communalities, since the latter is the sum of all squared loadings. By (13.33), the proportion of total variance for each factor is the variance accounted for divided by 5.

The two factors account for 96% of the total variance and therefore represent the five variables very well. To see how well the two-factor model reproduces the correlation matrix, we examine

$$\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}} = \begin{pmatrix} .969 & -.231 \\ .519 & .807 \\ .785 & -.587 \\ .971 & -.210 \\ .704 & .667 \end{pmatrix} \begin{pmatrix} .969 & .519 & .785 & .971 & .704 \\ -.231 & .807 & -.587 & -.210 & .667 \end{pmatrix}$$
$$+ \begin{pmatrix} .007 & 0 & 0 & 0 & 0 \\ 0 & .079 & 0 & 0 & 0 \\ 0 & 0 & .040 & 0 & 0 \\ 0 & 0 & 0 & .013 & 0 \\ 0 & 0 & 0 & 0 & .060 \end{pmatrix}$$
$$= \begin{pmatrix} 1.000 & .317 & .896 & .990 & .528 \\ .317 & 1.000 & -.066 & .335 & .904 \\ .896 & -.066 & 1.000 & .885 & .161 \\ .990 & .335 & .885 & 1.000 & .543 \\ .528 & .904 & .161 & .543 & 1.000 \end{pmatrix},$$

which is very close to the original $\mathbf{R}$. We will not attempt to interpret the factors at this point but will wait until they have been rotated in Section 13.5.2. □

13.3.2 Principal Factor Method

In the principal component approach to estimation of the loadings, we neglected $\mathbf{\Psi}$ and factored $\mathbf{S}$ or $\mathbf{R}$. The *principal factor* method (also called the *principal axis* method) uses an initial estimate $\hat{\mathbf{\Psi}}$ and factors $\mathbf{S} - \hat{\mathbf{\Psi}}$ or $\mathbf{R} - \hat{\mathbf{\Psi}}$ to obtain

$$\mathbf{S} - \hat{\mathbf{\Psi}} \cong \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}', \tag{13.36}$$
$$\mathbf{R} - \hat{\mathbf{\Psi}} \cong \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}', \tag{13.37}$$

where $\hat{\mathbf{\Lambda}}$ is $p \times m$ and is calculated as in (13.24) using eigenvalues and eigenvectors of $\mathbf{S} - \hat{\mathbf{\Psi}}$ or $\mathbf{R} - \hat{\mathbf{\Psi}}$.

The ith diagonal element of $\mathbf{S} - \hat{\mathbf{\Psi}}$ is given by $s_{ii} - \hat{\psi}_i$, which is the ith communality, $\hat{h}_i^2 = s_{ii} - \hat{\psi}_i$ [see (13.30)]. Likewise, the diagonal elements of $\mathbf{R} - \hat{\mathbf{\Psi}}$ are the communalities $\hat{h}_i^2 = 1 - \hat{\psi}_i$. (Obviously, $\hat{\psi}_i$ and $\hat{h}_i^2$ have different values for $\mathbf{S}$

than for $\mathbf{R}$.) With these diagonal values, $\mathbf{S} - \hat{\mathbf{\Psi}}$ and $\mathbf{R} - \hat{\mathbf{\Psi}}$ have the form

$$\mathbf{S} - \hat{\mathbf{\Psi}} = \begin{pmatrix} \hat{h}_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & \hat{h}_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & \hat{h}_p^2 \end{pmatrix}, \tag{13.38}$$

$$\mathbf{R} - \hat{\mathbf{\Psi}} = \begin{pmatrix} \hat{h}_1^2 & r_{12} & \cdots & r_{1p} \\ r_{21} & \hat{h}_2^2 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & \hat{h}_p^2 \end{pmatrix}. \tag{13.39}$$

A popular initial estimate for a communality in $\mathbf{R} - \hat{\mathbf{\Psi}}$ is $\hat{h}_i^2 = R_i^2$, the squared multiple correlation between y_i and the other $p - 1$ variables. This can be found as

$$\hat{h}_i^2 = R_i^2 = 1 - \frac{1}{r^{ii}}, \tag{13.40}$$

where r^{ii} is the ith diagonal element of $\mathbf{R}^{-1}$.

For $\mathbf{S} - \hat{\mathbf{\Psi}}$, an initial estimate of communality analogous to (13.40) is

$$\hat{h}_i^2 = s_{ii} - \frac{1}{s^{ii}}, \tag{13.41}$$

where s_{ii} is the ith diagonal element of $\mathbf{S}$ and s^{ii} is the ith diagonal element of $\mathbf{S}^{-1}$. It can be shown that (13.41) is equivalent to

$$\hat{h}_i^2 = s_{ii} - \frac{1}{s^{ii}} = s_{ii} R_i^2, \tag{13.42}$$

which is a reasonable estimate of the amount of variance that y_i has in common with the other y's.

To use (13.40) or (13.41), $\mathbf{R}$ or $\mathbf{S}$ must be nonsingular. If $\mathbf{R}$ is singular we can use the absolute value or the square of the largest correlation in the ith row of $\mathbf{R}$ as an estimate of communality.

After obtaining communality estimates, we calculate eigenvalues and eigenvectors of $\mathbf{S} - \hat{\mathbf{\Psi}}$ or $\mathbf{R} - \hat{\mathbf{\Psi}}$ and use (13.24) to obtain estimates of factor loadings, $\hat{\mathbf{\Lambda}}$. Then the columns and rows of $\hat{\mathbf{\Lambda}}$ can be used to obtain new eigenvalues (variance explained) and communalities, respectively. The sum of squares of the jth column of $\hat{\mathbf{\Lambda}}$ is the jth eigenvalue of $\mathbf{S} - \hat{\mathbf{\Psi}}$ or $\mathbf{R} - \hat{\mathbf{\Psi}}$, and the sum of squares of the ith row of $\hat{\mathbf{\Lambda}}$ is the communality of y_i. The proportion of variance explained by the jth factor is

$$\frac{\theta_j}{\text{tr}(\mathbf{S} - \hat{\mathbf{\Psi}})} = \frac{\theta_j}{\sum_{i=1}^{p} \theta_i}$$

or

$$\frac{\theta_j}{\text{tr}(\mathbf{R} - \hat{\mathbf{\Psi}})} = \frac{\theta_j}{\sum_{i=1}^{p} \theta_i},$$

where θ_j is the jth eigenvalue of $\mathbf{S} - \hat{\boldsymbol{\Psi}}$ or $\mathbf{R} - \hat{\boldsymbol{\Psi}}$. The matrices $\mathbf{S} - \hat{\boldsymbol{\Psi}}$ and $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ are not necessarily positive semidefinite and will often have some small negative eigenvalues. In such a case, the cumulative proportion of variance will exceed 1 and then decline to 1 as the negative eigenvalues are added. [Note that loadings cannot be obtained by (13.24) for the negative eigenvalues.]

■ **EXAMPLE 13.3.2**

To illustrate the principal factor method, we use the perception data from Table 13.1. The correlation matrix as given in Example 13.3.1 is singular. Hence in place of multiple correlations as communality estimates, we use the (absolute value of) the largest correlation in each row of $\mathbf{R}$. [The multiple correlation of y with several variables is greater than the simple correlation of y with any of the individual variables; see, for example, Rencher and Schaalje (2008, p. 257).] The diagonal elements of $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ as given by (13.39) are therefore .995, .837, .881, .995, and .837, which are obtained from $\mathbf{R}$ in (13.35). The eigenvalues of $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ are 3.202, 1.395, .030, −.0002, and −.080, whose sum is 4.546. The first two eigenvectors of $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ are

$$\mathbf{c}_1 = \begin{pmatrix} .548 \\ .272 \\ .431 \\ .549 \\ .373 \end{pmatrix} \quad \text{and} \quad \mathbf{c}_2 = \begin{pmatrix} -.178 \\ .656 \\ -.460 \\ -.159 \\ .549 \end{pmatrix}.$$

When these are multiplied by the square roots of the respective eigenvalues, we obtain the principal factor loadings. In Table 13.3, these are compared with the loadings obtained by the principal component method in Example 13.3.1. The two sets of loadings are very similar, as we would have expected because of the large size of the communalities. The communalities in Table 13.3 are for the principal factor loadings, as noted above. The proportion of variance in each case for the principal factor loadings is obtained by dividing the variance accounted for (eigenvalue) by the sum of the eigenvalues, 4.546; for example, $3.202/4.546 = .704$. □

13.3.3 Iterated Principal Factor Method

The principal factor method can easily be iterated to improve the estimates of communality. After obtaining $\hat{\boldsymbol{\Lambda}}$ from $\mathbf{S} - \hat{\boldsymbol{\Psi}}$ or $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ in (13.36) or (13.37) using initial communality estimates, we can obtain new communality estimates from the loadings in $\hat{\boldsymbol{\Lambda}}$ using (13.28),

$$\hat{h}_i^2 = \sum_{j=1}^{m} \hat{\lambda}_{ij}^2.$$

These values of $\hat{h}_i^2$ are substituted into the diagonal of $\mathbf{S} - \hat{\boldsymbol{\Psi}}$ or $\mathbf{R} - \hat{\boldsymbol{\Psi}}$, from which we obtain a new value of $\hat{\boldsymbol{\Lambda}}$ using (13.24). This process is continued until the

Table 13.3 Loadings Obtained by Two Different Methods for Perception Data of Table 13.1

	Principal Component Loadings		Principal Factor Loadings		
Variables	f_1	f_2	f_1	f_2	Communalities
Kind	.969	−.231	.981	−.210	.995
Intelligent	.519	.807	.487	.774	.837
Happy	.785	−.587	.771	−.544	.881
Likeable	.971	−.210	.982	−.188	.995
Just	.704	.667	.667	.648	.837
Variance accounted for	3.263	1.538	3.202	1.395	
Proportion of total variance	.653	.308	.704	.307	
Cumulative proportion	.653	.960	.704	1.01	

communality estimates converge. (For some data sets, the iterative procedure does not converge.) Then the eigenvalues and eigenvectors of the final version of $\mathbf{S} - \hat{\mathbf{\Psi}}$ or $\mathbf{R} - \hat{\mathbf{\Psi}}$ are used in (13.24) to obtain the loadings.

The principal factor method and the iterated principal factor method will typically yield results very close to those from the principal component method when *either* of the following is true.

1. The correlations are fairly large, with a resulting small value of m.
2. The number of variables, p, is large.

A shortcoming of the iterative approach is that sometimes it leads to a communality estimate $\hat{h}_i^2$ exceeding 1 (when factoring $\mathbf{R}$). Such a result is known as a *Heywood case* (Heywood 1931). If $\hat{h}_i^2 > 1$, then $\hat{\psi}_i < 0$ by (13.26) and (13.28), which is clearly improper, since we cannot have a negative specific variance. Thus when a communality exceeds 1, the iterative process should stop, with the program reporting that a solution cannot be reached. Some software programs have an option of continuing the iterations by setting the communality equal to 1 in all subsequent iterations. The resulting solution with $\hat{\psi}_i = 0$ is somewhat questionable because it implies exact dependence of a variable on the factors, a possible but unlikely outcome.

■ **EXAMPLE 13.3.3**

We illustrate the iterated principal factor method using the Seishu data in Table 7.1. The correlation matrix is given below:

$$\mathbf{R} = \begin{pmatrix} 1.00 & .56 & .22 & .10 & .20 & -.04 & .13 & .03 & -.07 & .09 \\ .56 & 1.00 & -.09 & .13 & .20 & -.17 & .17 & .24 & .16 & .06 \\ .22 & -.09 & 1.00 & .16 & .70 & -.31 & -.45 & -.34 & -.11 & .68 \\ .10 & .13 & .16 & 1.00 & .49 & -.03 & -.16 & .01 & .42 & .37 \\ .20 & .20 & .70 & .49 & 1.00 & -.32 & -.34 & -.19 & .30 & .87 \\ -.04 & -.17 & -.31 & -.03 & -.32 & 1.00 & -.42 & -.57 & -.11 & -.26 \\ .13 & .17 & -.45 & -.16 & -.34 & -.42 & 1.00 & .82 & .23 & -.30 \\ .03 & .24 & -.34 & .01 & -.19 & -.57 & .82 & 1.00 & .45 & -.17 \\ -.07 & .16 & -.11 & .42 & .30 & -.11 & .23 & .45 & 1.00 & .29 \\ .09 & .06 & .68 & .37 & .87 & -.26 & -.30 & -.17 & .29 & 1.00 \end{pmatrix}.$$

The eigenvalues of $\mathbf{R}$ are 3.17, 2.56, 1.43, 1.28, .54, .47, .25, .12, .10, and .06. There is a notable gap between 1.28 and .54, and we therefore extract four factors (see Section 13.4). The first four eigenvalues account for a proportion

$$\frac{3.17 + 2.56 + 1.43 + 1.28}{10} = .84$$

of tr($\mathbf{R}$).

For initial communality estimates, we use the squared multiple correlation between each variable and the other nine variables. These are given in Table 13.4, along with the final communalities after iteration. We multiply the first four eigenvectors of the final iterated version of $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ by the square roots of the respective eigenvalues, as in (13.24), to obtain the factor loadings given in Table 13.4. We will not attempt to interpret the factors until after they have been rotated in Example 13.5.2b(b). □

13.3.4 Maximum Likelihood Method

If we assume that the observations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ constitute a random sample from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ can be estimated by the method of maximum likelihood. It can be shown that the estimates $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Psi}}$ satisfy the following:

$$\mathbf{S}\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Lambda}} = \hat{\boldsymbol{\Lambda}}(\mathbf{I} + \hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}}), \tag{13.43}$$

$$\hat{\boldsymbol{\Psi}} = \text{diag}\,(\mathbf{S} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'), \tag{13.44}$$

$$\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Psi}}^{-1}\hat{\boldsymbol{\Lambda}} \quad \text{is diagonal.} \tag{13.45}$$

These equations must be solved iteratively, and in practice the procedure may fail to converge or may yield a Heywood case (Section 13.3.3).

We note that the proportion of variance accounted for by the factors, as given by (13.32) or (13.33), will not necessarily be in descending order for maximum likelihood factors, as it is for factors obtained from the principal component or principal factor method.

Table 13.4 Iterated Principal Factor Loadings and Communalities for the Seishu Data

Variables	Loadings				Initial Communalities	Final Communalities
	f_1	f_2	f_3	f_4		
Taste	.22	.31	.92	.12	.57	1.00
Odor	.07	.40	.43	−.20	.54	.38
pH	.80	.04	.05	−.40	.78	.79
Acidity 1	.41	.22	−.11	.37	.40	.36
Acidity 2	.94	.28	−.07	.05	.88	.98
Sake meter	−.13	−.67	.10	.56	.77	.79
Reducing sugar	−.55	.66	.03	−.11	.79	.75
Total sugar	−.45	.88	−.14	−.07	.87	.99
Alcohol	.13	.54	−.37	.54	.66	.74
Formyl-nitrogen	.84	.21	−.17	−.02	.80	.78
Variance accounted for	3.00	2.37	1.25	.96	7.06	7.57

■ **EXAMPLE 13.3.4**

We illustrate the maximum likelihood method with the Seishu data of Table 7.1. The correlation matrix and its eigenvalues were given in Example 13.3.3. We extract four factors, as in Example 13.3.3. The iterative solution of (13.43), (13.44), and (13.45) yielded the loadings and communalities given in Table 13.5.

The pattern of the loadings is different from that obtained using the iterated principal factor method in Example 13.3.3, but we will not compare them until after rotation in Example 13.5.2b(b). Note that the four values of variance accounted for are not in descending order. □

13.4 CHOOSING THE NUMBER OF FACTORS, m

Several criteria have been proposed for choosing m, the number of factors. We consider four criteria, which are similar to those given in Section 12.6 for choosing the number of principal components to retain.

1. Choose m equal to the number of factors necessary for the variance accounted for to achieve a predetermined percentage, say 80%, of the total variance tr($\mathbf{S}$) or tr($\mathbf{R}$).

2. Choose m equal to the number of eigenvalues greater than the average eigenvalue. For $\mathbf{R}$ the average is 1; for $\mathbf{S}$ it is $\sum_{j=1}^{p} \theta_j / p$.

3. Use the scree test based on a plot of the eigenvalues of $\mathbf{S}$ or $\mathbf{R}$. If the graph drops sharply, followed by a straight line with much smaller slope, choose m equal to the number of eigenvalues before the straight line begins.

Table 13.5 Maximum Likelihood Loadings and Communalities for the Seishu Data

Variables	Loadings				Communalities
	f_1	f_2	f_3	f_4	
Taste	1.00	0	0	0	1.00
Odor	.45	−.05	.22	.19	.29
pH	.22	.68	−.20	−.40	.71
Acidity 1	.10	.47	.10	.37	.38
Acidity 2	.20	.98	.02	.00	1.00
Sake meter	−.04	−.31	−.68	.55	.86
Reducing sugar	.13	−.39	.76	−.02	.75
Total sugar	.03	−.22	.96	.02	.98
Alcohol	−.07	.31	.52	.60	.72
Formyl-nitrogen	.02	.79	−.05	−.10	.63
Variance accounted for	1.33	2.66	2.34	1.00	7.32

4. Test the hypothesis that m is the correct number of factors, H_0: $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$, where $\boldsymbol{\Lambda}$ is $p \times m$.

Method 1 applies particularly to the principal component method. By (13.32), the proportion of total sample variance (variance accounted for) due to the jth factor from $\mathbf{S}$ is $\sum_{i=1}^{p} \hat{\lambda}_{ij}^2 / \text{tr}(\mathbf{S})$. The corresponding proportion from $\mathbf{R}$ is $\sum_{i=1}^{p} \hat{\lambda}_{ij}^2 / p$, as in (13.33). The contribution of all m factors to tr($\mathbf{S}$) or p is therefore $\sum_{i=1}^{p} \sum_{j=1}^{m} \hat{\lambda}_{ij}^2$, which is the sum of squares of all elements of $\hat{\boldsymbol{\Lambda}}$. For the principal component method, we see by (13.28) and (13.29) that this sum is also equal to the sum of the first m eigenvalues or to the sum of all p communalities:

$$\sum_{i=1}^{p} \sum_{j=1}^{m} \hat{\lambda}_{ij}^2 = \sum_{i=1}^{p} \hat{h}_i^2 = \sum_{j=1}^{m} \theta_j. \tag{13.46}$$

Thus we choose m sufficiently large so that the sum of the communalities or the sum of the eigenvalues (variance accounted for) constitutes a relatively large portion of tr($\mathbf{S}$) or p.

Method 1 can be extended to the principal factor method, where prior estimates of communalities are used to form $\mathbf{S} - \hat{\boldsymbol{\Psi}}$ or $\mathbf{R} - \hat{\boldsymbol{\Psi}}$. However, $\mathbf{S} - \hat{\boldsymbol{\Psi}}$ or $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ will often have some negative eigenvalues. Therefore, as values of m range from 1 to p, the cumulative proportion of eigenvalues, $\sum_{j=1}^{m} \theta_j / \sum_{j=1}^{p} \theta_j$, will exceed 1.0 and then reduce to 1.0 as the negative eigenvalues are added. Hence a percentage such as 80% will be reached for a lower value of m than would be the case for $\mathbf{S}$ or $\mathbf{R}$, and a better strategy might be to choose m equal to the value for which the percentage first exceeds 100%.

In the iterated principal factor method, m is specified before iteration, and $\sum_i \hat{h}_i^2$ is obtained after iteration as $\sum_i \hat{h}_i^2 = \text{tr}(\mathbf{S} - \hat{\boldsymbol{\Psi}})$. To choose m before iterating, one could use a priori considerations or the eigenvalues of $\mathbf{S}$ or $\mathbf{R}$, as in the principal component method.

Method 2 is a popular criterion of long standing and is the default in many software packages. Although heuristically based, it often works well in practice. A variation to method 2 that has been suggested for use with $\mathbf{R} - \hat{\boldsymbol{\Psi}}$ is to let m equal the number of positive eigenvalues. (There will typically be some negative eigenvalues of $\mathbf{R} - \hat{\boldsymbol{\Psi}}$.) However, this criterion will often result in too many factors, since the sum of the positive eigenvalues will exceed the sum of the communalities.

The scree test in method 3 was named after the geological term *scree*, referring to the debris at the bottom of a rocky cliff. It also performs well in practice.

In method 4 we wish to test

$$H_0: \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \quad \text{vs.} \quad H_1: \boldsymbol{\Sigma} \neq \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi},$$

where $\boldsymbol{\Lambda}$ is $p \times m$. The test statistic, a function of the likelihood ratio, is

$$\left(n - \frac{2p + 4m + 11}{6}\right) \ln\left(\frac{|\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}}|}{|\mathbf{S}|}\right), \tag{13.47}$$

which is approximately χ_ν^2 when H_0 is true, where $\nu = \frac{1}{2}[(p-m)^2 - p - m]$ and $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Psi}}$ are the maximum likelihood estimators. Rejection of H_0 implies that m is too small and more factors are needed.

In practice, when n is large, the test in method 4 often shows more factors to be significant than do the other three methods. We may therefore consider the value of m indicated by the test to be an upper bound on the number of factors with practical importance.

For many data sets, the choice of m will not be obvious. This indeterminacy leaves many statisticians skeptical as to the validity of factor analysis. A researcher may begin with one of the methods (say method 2) for an initial choice of m, will inspect the resulting percent of tr($\mathbf{R}$) or tr($\mathbf{S}$), and will then examine the rotated loadings for interpretability. If the percent of variance or interpretation does not seem satisfactory, the experimenter will try other values of m in a search for an acceptable compromise between percent of tr($\mathbf{R}$) and interpretability of the factors. Admittedly, this is a subjective procedure, and for such data sets one could well question the outcome (see Section 13.7).

When a data set is successfully fitted by a factor analysis model, the first three methods will almost always give the same value of m, and there will be little question as to what this value should be. Thus for a "good" data set, the entire procedure becomes much more objective.

■ **EXAMPLE 13.4(a)**

We compare the four methods of choosing m for the perception data used in Examples 13.3.1 and 13.3.2.

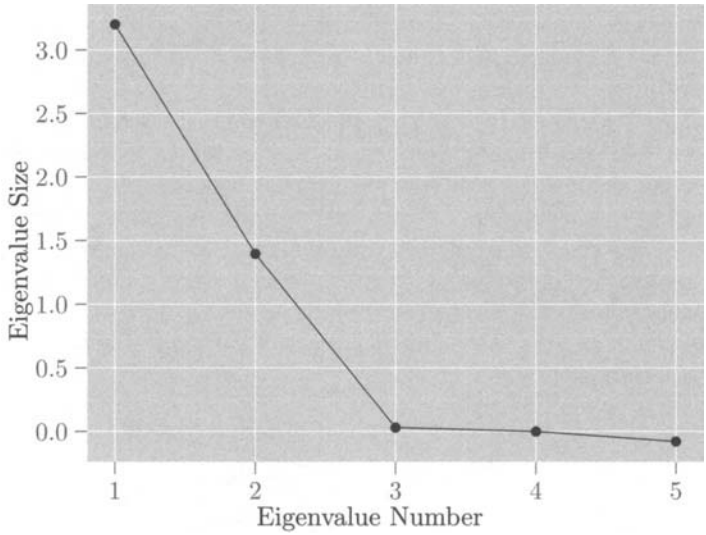


Figure 13.1 Scree graph for the perception data.

Method 1 gives $m = 2$, because one eigenvalue accounts for 65% of tr($\mathbf{R}$), while two eigenvalues account for 96%.

Method 2 gives $m = 2$, since $\lambda_2 = 1.54$ and $\lambda_3 = .17$.

For method 3, we examine the scree plot in Figure 13.1. It is clear that $m = 2$ is indicated.

Method 4 is not available for the perception data because $\mathbf{R}$ is singular (fifth eigenvalue is zero), and the test involves $|\mathbf{R}|$.

Hence for the perception data, all three available methods agree on $m = 2$. □

■ **EXAMPLE 13.4(b)**

We compare the four methods of choosing m for the Seishu data used in Examples 13.3.3 and 13.3.4.

Method 1 gives $m = 4$ for the principal component method, because four eigenvalues of $\mathbf{R}$ account for 82% of tr($\mathbf{R}$). For the principal factor method with initial communality estimates R_i^2, the eigenvalues of $\mathbf{R} - \hat{\mathbf{\Psi}}$ and corresponding proportions are as follows:

Eigenvalues	2.86	2.17	.94	.88	.12	.08	.01	−.06	−.13	−.22
Proportions	.43	.33	.14	.16	.02	.01	.00	−.01	−.02	−.03
Cumulative proportions	.43	.76	.90	1.03	1.05	1.06	1.06	1.06	1.03	1.00

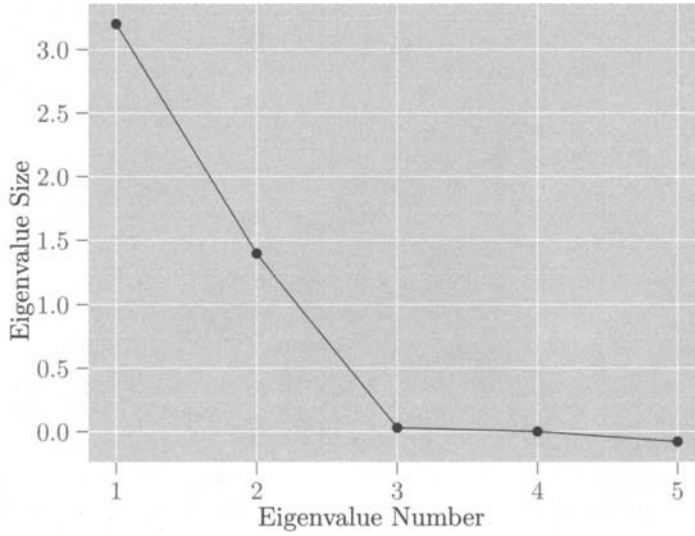


Figure 13.2 Scree graph for the Seishu data.

The proportions are obtained by dividing the eigenvalues by their sum, 6.63. Thus the cumulative proportion first exceeds 1.00 for $m = 4$.

Method 2 gives $m = 4$, since $\lambda_4 = 1.31$ and $\lambda_5 = .61$, where λ_4 and λ_5 are eigenvalues of $\mathbf{R}$.

For method 3, we examine the scree plot in Figure 13.2. There is a discernible bend in slope at the fifth eigenvalue.

For method 4, we use $m = 4$ in the approximate chi-squared statistic in (13.47) and obtain $\chi^2 = 9.039$, with degrees of freedom

$$\nu = \tfrac{1}{2}[(p-m)^2 - p - m] = \tfrac{1}{2}[(10-4)^2 - 10 - 4] = 11.$$

Since $9.039 < \chi^2_{.05,11} = 19.68$, we do not reject the hypothesis that four factors are adequate.

Thus for the Seishu data, all four methods agree on $m = 4$. □

13.5 ROTATION

13.5.1 Introduction

As noted in Section 13.2.2, the factor loadings (rows of $\mathbf{\Lambda}$) in the population model are unique only up to multiplication by an orthogonal matrix that rotates the loadings. The rotated loadings preserve the essential properties of the original loadings; they reproduce the covariance matrix and satisfy all basic assumptions. The estimated loading matrix $\hat{\mathbf{\Lambda}}$ can likewise be rotated to obtain $\hat{\mathbf{\Lambda}}^* = \hat{\mathbf{\Lambda}}\mathbf{T}$, where $\mathbf{T}$ is orthogonal.

Since $\mathbf{TT}' = \mathbf{I}$ by (2.102), the rotated loadings provide the same estimate of the covariance matrix as before:

$$\mathbf{S} \cong \hat{\mathbf{\Lambda}}^{*}\hat{\mathbf{\Lambda}}^{*\prime} + \hat{\mathbf{\Psi}} = \hat{\mathbf{\Lambda}}\mathbf{T}\mathbf{T}'\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}. \tag{13.48}$$

Geometrically, the loadings in the ith row of $\hat{\mathbf{\Lambda}}$ constitute the coordinates of a point in the loading space corresponding to y_i. Rotation of the p points gives their coordinates with respect to new axes (factors) but otherwise leaves their basic geometric configuration intact. We hope to find a new frame of reference in which the factors are more interpretable. To this end, the goal of rotation is to place the axes close to as many points as possible. If there are clusters of points (corresponding to groupings of y's), we seek to move the axes so as to pass through or near these clusters. This would associate each group of variables with a factor (axis) and make interpretation more objective. The resulting axes then represent the natural factors.

If we can achieve a rotation in which every point is close to an axis, then each variable loads highly on the factor corresponding to the axis and has small loadings on the remaining factors. In this case, there is no ambiguity. Such a happy state of affairs is called *simple structure*, and interpretation is greatly simplified. We merely observe which variables are associated with each factor, and the factor is defined or named accordingly.

In order to identify the natural groupings of variables, we seek a rotation to an interpretable pattern for the loadings, in which the variables load highly on only one factor. The number of factors on which a variable has moderate or high loadings is called the *complexity* of the variable. In the ideal situation referred to above as simple structure, the variables all have a complexity of 1. In this case, the variables have been clearly clustered into groups corresponding to the factors.

We consider two basic types of rotation: *orthogonal* and *oblique*. The rotation in (13.48) involving an orthogonal matrix is an orthogonal rotation; the original perpendicular axes are rotated rigidly and remain perpendicular. In an orthogonal rotation, angles and distances are preserved, communalities are unchanged, and the basic configuration of the points remains the same. Only the reference axes differ. In an oblique "rotation" (transformation), the axes are not required to remain perpendicular and are thus free to pass closer to clusters of points.

In Sections 13.5.2 and 13.5.3, we discuss orthogonal and oblique rotations, followed by some guidelines for interpretation in Section 13.5.4.

13.5.2 Orthogonal Rotation

It was noted above in Section 13.5.1 that orthogonal rotations preserve communalities. This is because the rows of $\hat{\mathbf{\Lambda}}$ are rotated, and the distance to the origin is unchanged, which, by (13.28), is the communality. However, the variance accounted for by each factor as given in (13.31) will change, as will the corresponding proportion in (13.32) or (13.33). The proportions due to the rotated loadings will not necessarily be in descending order.

In Sections 13.5.2a and 13.5.2b, we consider two approaches to orthogonal rotation.

13.5.2a Graphical Approach

If there are only two factors ($m = 2$), we can use a *graphical* rotation based on a visual inspection of a plot of factor loadings. In this case, the rows of $\hat{\mathbf{\Lambda}}$ are pairs of loadings, $(\hat{\lambda}_{i1}, \hat{\lambda}_{i2}), i = 1, 2, \ldots, p$, corresponding to $y_1, y_2, \ldots, y_p$. We choose an angle ϕ through which the axes can be rotated to move them closer to groupings of points. The new rotated loadings $(\hat{\lambda}^*_{i1}, \hat{\lambda}^*_{i2})$ can be measured directly on the graph as coordinates of the axes or calculated from $\hat{\mathbf{\Lambda}}^* = \hat{\mathbf{\Lambda}}\mathbf{T}$ using

$$\mathbf{T} = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}. \tag{13.49}$$

■ **EXAMPLE 13.5.2a**

In Example 13.3.1, the initial factor loadings for the perception data did not provide an interpretation consistent with the two groupings of variables apparent in the pattern of correlations in $\mathbf{R}$. The five pairs of loadings $(\hat{\lambda}_{i1}, \hat{\lambda}_{i2})$ corresponding to the five variables are plotted in Figure 13.3. An orthogonal rotation through $-35°$ would bring the axes (factors) closer to the two clusters of points (variables) identified in Example 13.3.1. With the rotation, each cluster of variables corresponds much more closely to a factor. Using $\hat{\mathbf{\Lambda}}$ from Example 13.3.1 and $-35°$ in $\mathbf{T}$ as given in (13.49), we obtain the following rotated loadings:

$$\hat{\mathbf{\Lambda}}^* = \hat{\mathbf{\Lambda}}\mathbf{T} = \begin{pmatrix} .969 & -.231 \\ .519 & .807 \\ .785 & -.587 \\ .971 & -.210 \\ .704 & .667 \end{pmatrix} \begin{pmatrix} .819 & .574 \\ -.574 & .819 \end{pmatrix}$$

$$= \begin{pmatrix} .927 & .367 \\ -.037 & .959 \\ .980 & -.031 \\ .916 & .385 \\ .194 & .950 \end{pmatrix}.$$

In Table 13.6, we compare the rotated loadings in $\hat{\mathbf{\Lambda}}^*$ with the original loadings in $\hat{\mathbf{\Lambda}}$.

The interpretation of the rotated loadings is clear. As indicated by the boldface loadings in Table 13.6, the first factor is associated with variables 1, 3, and 4: kind, happy, and likeable. The second factor is associated with the other two variables: intelligent and just. This same grouping of variables is indicated by the pattern in the correlation matrix in (13.35) and can also be seen in the two clusters of points in Figure 13.3. The first factor might be described as representing a person's perceived humanity or amiability, while the second involves more logical or rational practices.

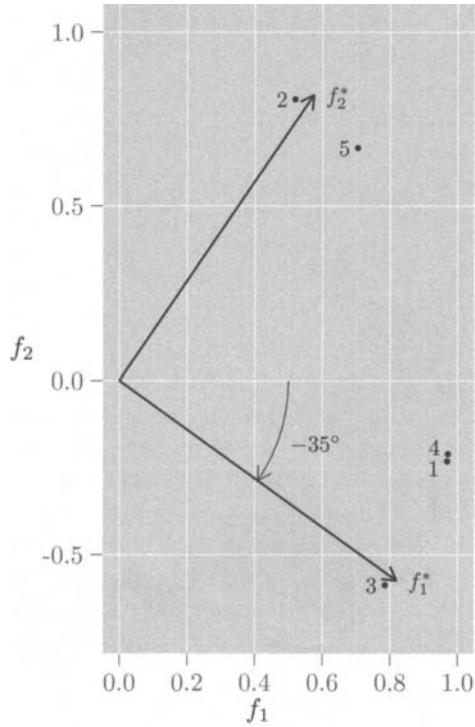


Figure 13.3 Plot of the two loadings for each of the five variables in the perception data of Table 13.1.

Note that if the angle between the rotated axes was allowed to be less than 90° (an oblique rotation), the lower axis representing f_1^* could come closer to the points corresponding to variables 1 and 4 so that the coordinates on f_2^*, .367 and .385, could be reduced. However, the basic interpretation would not change; variables 1 and 4 would still be associated with f_1^*. □

13.5.2b Varimax Rotation

The graphical approach to rotation is generally limited to $m = 2$. For $m > 2$, various analytical methods have been proposed. The most popular of these is the *varimax* technique, which seeks rotated loadings that maximize the variance of the squared loadings in each column of $\hat{\mathbf{\Lambda}}^*$. If the loadings in a column were nearly equal, the variance would be close to 0. As the squared loadings approach 0 and 1 (for factoring $\mathbf{R}$), the variance will approach a maximum. Thus the varimax method attempts to make the loadings either large or small to facilitate interpretation.

The varimax procedure cannot guarantee that all variables will load highly on only one factor. In fact, no procedure could do this for all possible data sets. The configuration of the points in the loading space remains fixed; we merely rotate the

Table 13.6 Graphically Rotated Loadings for the Perception Data of Table 13.1

Variables	Principal Component Loadings		Graphically Rotated Loadings		Communalities,
	f_1	f_2	f_1	f_2	$\hat{h}_i^2$
Kind	.969	−.231	**.927**	.367	.993
Intelligent	.519	.807	−.037	**.959**	.921
Happy	.785	−.587	**.980**	−.031	.960
Likeable	.971	−.210	**.916**	.385	.987
Just	.704	.667	.194	**.950**	.940
Variance accounted for	3.263	1.538	2.696	2.106	4.802
Proportion of total variance	.653	.308	.539	.421	.960
Cumulative proportion	.653	.960	.539	.960	.960

axes to be as close to as many points as possible. In many cases, the points are not well clustered, and the axes simply cannot be rotated so as to be near all of them. This problem is compounded by having to choose m. If m is changed, the coordinates $(\hat{\lambda}_{i1}, \hat{\lambda}_{i2}, \ldots, \hat{\lambda}_{im})$ change, and the relative position of the points is altered.

The varimax rotation is available in virtually all factor analysis software programs. The output typically includes the rotated loading matrix $\hat{\mathbf{\Lambda}}^*$, the variance accounted for (sum of squares of each column of $\hat{\mathbf{\Lambda}}^*$), the communalities (sum of squares of each row of $\hat{\mathbf{\Lambda}}^*$), and the orthogonal matrix $\mathbf{T}$ used to obtain $\hat{\mathbf{\Lambda}}^* = \hat{\mathbf{\Lambda}}\mathbf{T}$.

■ **EXAMPLE 13.5.2b(a)**

In Example 13.5.2a, a graphical rotation was devised visually to achieve interpretable loadings for the perception data of Table 13.1. As we would expect, the varimax method yields a similar result. The varimax rotated loadings are given in Table 13.7. For comparison, we have included the original unrotated loadings from Table 13.3 and the graphically rotated loadings from Table 13.6.

The orthogonal matrix $\mathbf{T}$ for the varimax rotation is

$$\mathbf{T} = \begin{pmatrix} .859 & .512 \\ -.512 & .859 \end{pmatrix}.$$

By (13.49), $-\sin\phi = .512$, and the angle of rotation is $\phi = -\sin^{-1}(.512) = -30.8°$. Thus the varimax rotation chose an angle of rotation of $-30.8°$ as compared to the $-35°$ we selected visually, but the results are very close and the interpretation is exactly the same. □

Table 13.7 Varimax Rotated Factor Loadings for the Perception Data of Table 13.1

Variables	Principal Component Loadings		Graphically Rotated Loadings		Varimax Rotated Loadings		Communalities,
	f_1	f_2	f_1	f_2	f_1	f_2	$\hat{h}_i^2$
Kind	.969	−.231	**.927**	.367	**.951**	.298	.993
Intelligent	.519	.807	−.037	**.959**	.033	**.959**	.921
Happy	.785	−.587	**.980**	−.031	**.975**	−.103	.960
Likeable	.971	−.210	**.916**	.385	**.941**	.317	.987
Just	.704	.667	.194	**.950**	.263	**.933**	.940
Variance accounted for	3.263	1.538	2.696	2.106	2.811	1.991	4.802
Proportion of total variance	.653	.308	.539	.421	.562	.398	.960
Cumulative proportion	.653	.960	.539	.960	.562	.960	.960

■ **EXAMPLE 13.5.2b(b)**

In Examples 13.3.3 and 13.3.4, we obtained the iterated principal factor loadings and maximum likelihood loadings for the Seishu data. In Table 13.8, we show the varimax rotation of these two sets of loadings. The similarities in the two sets of rotated loadings are striking. The interpretation in each case is the same. The variances accounted for are virtually identical.

The rotation in each case has achieved a satisfactory simple structure, and most variables show a complexity of 1. The boldface loadings indicate the variables associated with each factor for interpretation purposes. These may be meaningful to the researcher. For example, factor 2 is associated with sake meter, reducing sugar, and total sugar, while factor 3 is aligned with taste and odor. □

13.5.3 Oblique Rotation

The term *oblique rotation* refers to a transformation in which the axes do not remain perpendicular. Technically, the term oblique *rotation* is a misnomer, since rotation implies an orthogonal transformation that preserves distances. A more accurate characterization would be oblique *transformation*, but the term oblique rotation is well established in the literature.

Instead of the orthogonal transformation matrix $\mathbf{T}$ used in (13.16), (13.17), and (13.18), an oblique rotation uses a general nonsingular transformation matrix $\mathbf{Q}$ to obtain $\mathbf{f}^* = \mathbf{Q}'\mathbf{f}$, and by (3.74),

$$\text{cov}(\mathbf{f}^*) = \mathbf{Q}'\mathbf{I}\mathbf{Q} = \mathbf{Q}'\mathbf{Q} \neq \mathbf{I}. \tag{13.50}$$

Table 13.8 Varimax Rotated Loadings for the Seishu Data

Variables	Iterated Principal Factor Rotated Loadings				Maximum Likelihood Rotated Loadings			
	f_1	f_2	f_3	f_4	f_1	f_2	f_3	f_4
Taste	.16	−.01	**.99**	−.09	.16	−.00	**.98**	−.10
Odor	−.11	.14	**.48**	.14	−.07	.14	**.49**	.17
pH	**.88**	−.12	.02	−.13	**.82**	−.10	.08	−.15
Acidity 1	.26	−.09	.09	**.54**	.29	−.08	.11	**.53**
Acidity 2	**.89**	−.06	.10	.43	**.91**	−.06	.10	.39
Sake meter	−.43	−**.76**	.01	.07	−.46	−**.80**	.04	.10
Reducing sugar	−.37	**.76**	.18	.03	−.37	**.75**	.20	.08
Total sugar	−.26	**.92**	.10	.25	−.27	**.91**	.11	.26
Alcohol	−.01	.25	.00	**.80**	−.00	.25	.01	**.81**
Formyl-nitrogen	**.74**	−.07	−.08	.20	**.76**	−.07	−.08	.22
Variance accounted for	2.62	2.12	1.27	1.27	2.61	2.14	1.29	1.28

Thus the new factors are correlated. Since distances and angles are not preserved, the communalities for $\mathbf{f}^*$ are different from those for $\mathbf{f}$. Some program packages report communalities obtained from the original loadings, rather than the oblique loadings.

When the axes are not required to be perpendicular, they can more easily pass through the major clusters of points in the loading space (assuming there are such clusters). For example, in Figure 13.4, we have plotted the varimax rotated loadings for two factors extracted from the sons data of Table 3.8 (see Example 13.5.3 at the end of this section). Oblique axes with an angle of 38° would pass much closer to the points, and the resulting loadings would be very close to 0 and 1. However, the interpretation would not change, since the same points (variables) would be associated with the oblique axes as with the orthogonal axes.

Various analytical methods for achieving oblique rotations have been proposed and are available in program packages. Typically, the output of one of these procedures includes a *pattern matrix*, a *structure matrix*, and a matrix of correlations among the oblique factors. For interpretation, we would usually prefer the pattern matrix rather than the structure matrix. The loadings in a row of the pattern matrix are the natural coordinates of the point (variable) on the oblique axes and serve as coefficients in the model relating the variable to the factors.

One use for an oblique rotation is to check on the orthogonality of the factors. The orthogonality in the original factors is imposed by the model and maintained by an orthogonal rotation. If an oblique rotation produces a correlation matrix that is nearly diagonal, we can be more confident that the factors are indeed orthogonal.

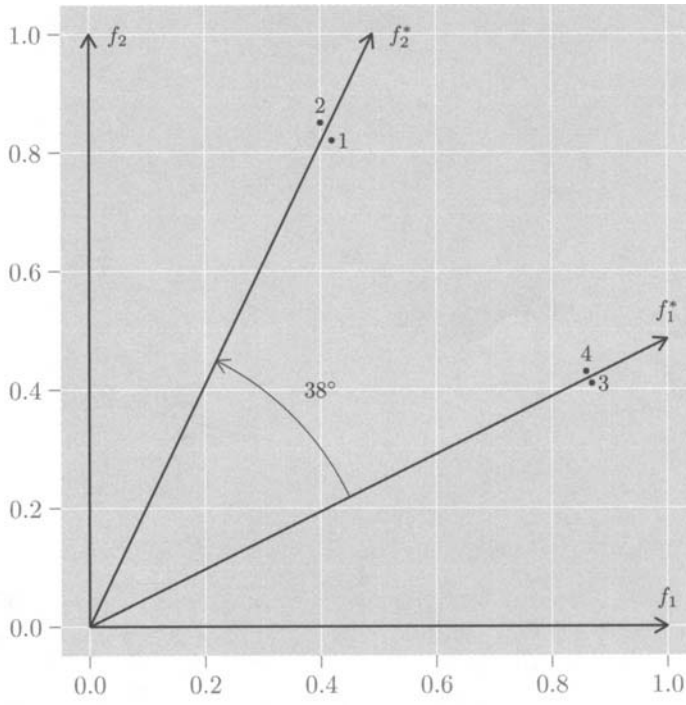


Figure 13.4 Orthogonal and oblique rotations for the sons data.

■ EXAMPLE 13.5.3

The correlation matrix for the sons data of Table 3.8 is given below:

$$\mathbf{R} = \begin{pmatrix} 1.000 & .735 & .711 & .704 \\ .735 & 1.000 & .693 & .709 \\ .711 & .693 & 1.000 & .839 \\ .704 & .709 & .839 & 1.000 \end{pmatrix}.$$

The varimax rotated loadings for two factors obtained by the principal component method are given in Table 13.9 and plotted in Figure 13.4. An analytical oblique rotation (Harris–Kaiser orthoblique method in SAS) produced oblique axes with an angle of 38°, the same as obtained by a graphical approach. The correlation between the two factors is .79 [obtained from $\mathbf{Q}'\mathbf{Q}$ in (13.50)], which is related to the angle by (3.15), $.79 = \cos 38°$. The pattern loadings are given in Table 13.9.

The oblique loadings give a much cleaner simple structure than the varimax loadings, but the interpretation is essentially the same if we neglect loadings below .45 on the varimax rotation.

Table 13.9 Varimax and Orthoblique Loadings for the Sons Data

	Varimax Loadings		Orthoblique Pattern Matrix	
Variable	f_1	f_2	f_1	f_2
1	.42	**.82**	.03	**.90**
2	.40	**.85**	−.03	**.96**
3	**.87**	.41	**.97**	−.01
4	**.86**	.43	**.95**	.01

In Figure 13.4, it is evident that a single factor would be adequate since the angle between axes is less than 45°. The suggestion to let $m = 1$ is also supported by the first three criteria in Section 13.4: the eigenvalues of $\mathbf{R}$ are 3.20, .38, .27, and .16. The first accounts for 80%, the second for an additional 9%. The large correlation, .79, between the two oblique factors constitutes additional evidence that a single factor model would suffice here. In fact, the pattern in $\mathbf{R}$ itself indicates the presence of only one factor. The four variables form only one cluster, since all are highly correlated. There are no small correlations between groupings of variables. □

13.5.4 Interpretation

In Sections 13.5.1, 13.5.2, and 13.5.3, we have discussed the usefulness of rotation as an aid to interpretation. Our goal is to achieve a simple structure in which each variable loads highly on only one factor, with small loadings on all other factors. In practice, we often fail to achieve this goal, but rotation usually produces loadings that are closer to the desired simple structure.

We now suggest general guidelines for interpreting the factors by examination of the matrix of rotated factor loadings. Moving horizontally from left to right across the m loadings in each row, identify the highest loading (in absolute value). If the highest loading is of a significant size (a subjective determination, see the next paragraph), circle or underline it. This is done for each of the p variables. There may be other significant loadings in a row besides the one circled. If these are considered, the interpretation is less simple. On the other hand, there may be variables with such small communalities that no significant loading appears on any factor. In this case, the researcher may wish to increase the number of factors and run the program again so that these variables might associate with a new factor.

To assess significance of factor loadings $\hat{\lambda}_{ij}$ obtained from $\mathbf{R}$, a threshold value of .3 has been advocated by many writers. For most successful applications, however, a critical value of .3 is too low and will result in variables of complexity greater than 1. A target value of .5 or .6 is typically more useful. The .3 criterion is loosely based on the critical value for significance of an ordinary correlation coefficient, r. However, the distribution of the sample loadings is not the same as that of r arising

from the bivariate normal. In addition, the critical value should be increased because mp values of $\hat{\lambda}_{ij}$ are being tested. On the other hand, if m is large, the critical value might possibly need to be reduced somewhat. Since $\hat{h}_i^2 = \sum_{j=1}^{m} \hat{\lambda}_{ij}^2$ is bounded by 1, an increase in m reduces the average squared loading in a row.

After identifying potentially significant loadings, the experimenter then attempts to discover some meaning in the factors and, ideally, to label or name them. This can readily be done if the group of variables associated with each factor makes sense to the researcher. But in many situations, the groupings are not so logical, and a revision can be tried, such as adjusting the size of loading deemed to be important, changing m, using a different method of estimating the loadings, or employing another type of rotation.

13.6 FACTOR SCORES

In many applications, the researcher wishes only to ascertain whether a factor analysis model fits the data and to identify the factors. In other applications, the experimenter wishes to obtain *factor scores* $\hat{\mathbf{f}}_i = (\hat{f}_{i1}, \hat{f}_{i2}, \ldots, \hat{f}_{im})'$, $i = 1, 2, \ldots, n$, which are defined as estimates of the underlying factor values for each observation. There are two potential uses for such scores: (1) the behavior of the observations in terms of the factors may be of interest and (2) we may wish to use the factor scores as input to another analysis, such as MANOVA. The latter usage resembles a similar application of principal components.

Since the f's are not observed, we must estimate them as functions of the observed y's. The most popular approach to estimating the factors is based on regression (Thompson 1951). We will discuss this method and also briefly describe an informal technique that can be used when $\mathbf{R}$ (or $\mathbf{S}$) is singular. For other approaches see Harman (1976, Chapter 16).

Since $E(f_i) = 0$, we relate the f's to the y's by a centered regression model

$$\begin{aligned} f_1 &= \beta_{11}(y_1 - \overline{y}_1) + \beta_{12}(y_2 - \overline{y}_2) + \cdots + \beta_{1p}(y_p - \overline{y}_p) + \epsilon_1 \\ f_2 &= \beta_{21}(y_1 - \overline{y}_1) + \beta_{22}(y_2 - \overline{y}_2) + \cdots + \beta_{2p}(y_p - \overline{y}_p) + \epsilon_2 \\ &\vdots \\ f_m &= \beta_{m1}(y_1 - \overline{y}_1) + \beta_{m2}(y_2 - \overline{y}_2) + \cdots + \beta_{mp}(y_p - \overline{y}_p) + \epsilon_m, \end{aligned} \tag{13.51}$$

which can be written in matrix form as

$$\mathbf{f} = \mathbf{B}_1'(\mathbf{y} - \overline{\mathbf{y}}) + \boldsymbol{\epsilon}. \tag{13.52}$$

We have used the notation $\boldsymbol{\epsilon}$ to distinguish this error from $\boldsymbol{\varepsilon}$ in the original factor model $\mathbf{y} - \boldsymbol{\mu} = \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\varepsilon}$ given in (13.3). Our approach is to estimate $\mathbf{B}_1$ and use the predicted value $\hat{\mathbf{f}} = \hat{\mathbf{B}}_1'(\mathbf{y} - \overline{\mathbf{y}})$ to estimate $\mathbf{f}$.

The model (13.52) holds for each observation:

$$\mathbf{f}_i = \mathbf{B}_1'(\mathbf{y}_i - \overline{\mathbf{y}}) + \boldsymbol{\epsilon}_i, \quad i = 1, 2, \ldots, n.$$

In transposed form, the model becomes

$$\mathbf{f}_i' = (\mathbf{y}_i - \bar{\mathbf{y}})'\mathbf{B}_1 + \boldsymbol{\epsilon}_i', \quad i = 1, 2, \ldots, n,$$

and these n equations can be combined into a single model,

$$\begin{aligned}\mathbf{F} = \begin{pmatrix} \mathbf{f}_1' \\ \mathbf{f}_2' \\ \vdots \\ \mathbf{f}_n' \end{pmatrix} &= \begin{pmatrix} (\mathbf{y}_1 - \bar{\mathbf{y}})'\mathbf{B}_1 \\ (\mathbf{y}_2 - \bar{\mathbf{y}})'\mathbf{B}_1 \\ \vdots \\ (\mathbf{y}_n - \bar{\mathbf{y}})'\mathbf{B}_1 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1' \\ \boldsymbol{\epsilon}_2' \\ \vdots \\ \boldsymbol{\epsilon}_n' \end{pmatrix} \\ &= \begin{pmatrix} (\mathbf{y}_1 - \bar{\mathbf{y}})' \\ (\mathbf{y}_2 - \bar{\mathbf{y}})' \\ \vdots \\ (\mathbf{y}_n - \bar{\mathbf{y}})' \end{pmatrix} \mathbf{B}_1 + \boldsymbol{\Xi} \\ &= \mathbf{Y}_c\mathbf{B}_1 + \boldsymbol{\Xi} \qquad \text{[by (10.11)]}. \end{aligned} \tag{13.53}$$

The model (13.53) has the appearance of a centered multivariate multiple regression model as in Section 10.4.5, with $\mathbf{Y}_c$ in place of $\mathbf{X}_c$. By (10.55), the estimate for $\mathbf{B}_1$ would be

$$\hat{\mathbf{B}}_1 = (\mathbf{Y}_c'\mathbf{Y}_c)^{-1}\mathbf{Y}_c'\mathbf{F}. \tag{13.54}$$

However, $\mathbf{F}$ is unobserved. To evaluate $\hat{\mathbf{B}}_1$ in spite of this, we first use (10.57) to rewrite (13.54) in terms of covariance matrices,

$$\hat{\mathbf{B}}_1 = \mathbf{S}_{yy}^{-1}\mathbf{S}_{yf}. \tag{13.55}$$

In the notation of the present chapter, $\mathbf{S}_{yy}$ is represented by $\mathbf{S}$; for $\mathbf{S}_{yf}$ we use $\hat{\boldsymbol{\Lambda}}$, since $\hat{\boldsymbol{\Lambda}}$ estimates $\text{cov}(\mathbf{y}, \mathbf{f}) = \boldsymbol{\Lambda}$ in (13.13). Thus, based on the assumptions in Section 13.2.1, we can write (13.55) as

$$\hat{\mathbf{B}}_1 = \mathbf{S}^{-1}\hat{\boldsymbol{\Lambda}}. \tag{13.56}$$

Then from model (13.53), the estimated (predicted) $\mathbf{f}_i$ values are given by

$$\begin{aligned}\hat{\mathbf{F}} = \begin{pmatrix} \hat{\mathbf{f}}_1' \\ \hat{\mathbf{f}}_2' \\ \vdots \\ \hat{\mathbf{f}}_n' \end{pmatrix} &= \mathbf{Y}_c\hat{\mathbf{B}}_1 \\ &= \mathbf{Y}_c\mathbf{S}^{-1}\hat{\boldsymbol{\Lambda}}. \end{aligned} \tag{13.57}$$

If $\mathbf{R}$ is factored instead of $\mathbf{S}$, (13.56) and (13.57) become

$$\hat{\mathbf{B}}_1 = \mathbf{R}^{-1}\boldsymbol{\Lambda}, \tag{13.58}$$

$$\hat{\mathbf{F}} = \mathbf{Y}_s\mathbf{R}^{-1}\hat{\boldsymbol{\Lambda}}, \tag{13.59}$$

respectively, where $\mathbf{Y}_s$ is the observed matrix of standardized variables, $(y_{ij} - \bar{y}_j)/s_j$.

We would ordinarily obtain factor scores for the rotated factors rather than the original factors. Thus $\hat{\mathbf{\Lambda}}$ in (13.57) or (13.59) would be replaced by $\hat{\mathbf{\Lambda}}^*$.

In order to obtain factor scores by (13.57) or (13.59), $\mathbf{S}$ or $\mathbf{R}$ must be nonsingular. When $\mathbf{R}$ (or $\mathbf{S}$) is singular, we can obtain factor scores by a simple method based directly on the rotated loadings. We cluster the variables into groups (factors) according to the loadings and find a score for each factor by averaging the variables associated with the factor. If the variables are not commensurate, the variables should be standardized before averaging. An alternative approach would be to weight the variables by their loadings when averaging.

■ EXAMPLE 13.6

The speaking rate of four voices was artificially manipulated by means of a rate changer without altering the pitch (Brown, Strong, and Rencher 1973). There were five rates for each voice:

FF = 45% faster
F = 25% faster
N = normal rate
S = 22% slower
SS = 42% slower.

The resulting 20 voices were played to 30 judges who rated them on 15 paired-opposite adjectives (variables) with a 14-point scale between poles. The following adjectives were used: intelligent, ambitious, polite, active, confident, happy, just, likeable, kind, sincere, dependable, religious, good-looking, sociable, and strong. The results were averaged over the 30 judges to produce 20 observation vectors of 15 variables each. The averaging produced very reliable data so that even though there were only 20 observations on 15 variables, the factor analysis model fit very well. The correlation matrix is as follows:

$$\mathbf{R} = \begin{pmatrix}
1.00 & .90 & -.17 & .88 & .92 & .88 & .15 & .39 & -.02 & -.16 & .52 & -.15 & -.79 & -.78 & .73 \\
.90 & 1.00 & -.46 & .93 & .87 & .79 & -.16 & .10 & -.35 & -.42 & .25 & -.40 & .68 & -.60 & .62 \\
-.17 & -.46 & 1.00 & -.56 & -.13 & .07 & .85 & .75 & .88 & .91 & .68 & .88 & .21 & .31 & .25 \\
.88 & .93 & -.56 & 1.00 & .85 & .73 & -.25 & -.02 & -.45 & -.57 & .10 & -.53 & .58 & .84 & .50 \\
.92 & .87 & -.13 & .85 & 1.00 & .91 & .20 & .39 & -.09 & -.16 & .49 & -.10 & .85 & .80 & .81 \\
.88 & .79 & .07 & .73 & .91 & 1.00 & .27 & .53 & .12 & .06 & .66 & .08 & .90 & .85 & .78 \\
.15 & -.16 & .85 & -.25 & .20 & .27 & 1.00 & .85 & .81 & .79 & .79 & .81 & .43 & .54 & .53 \\
.39 & .10 & .75 & -.02 & .39 & .53 & .85 & 1.00 & .84 & .79 & .93 & .77 & .71 & .69 & .76 \\
-.02 & -.35 & .88 & -.45 & -.09 & .12 & .81 & .84 & 1.00 & .91 & .76 & .85 & .28 & .36 & .35 \\
-.16 & -.42 & .91 & -.57 & -.16 & .06 & .79 & .79 & .91 & 1.00 & .72 & .96 & .26 & .28 & .29 \\
.52 & .25 & .67 & .10 & .49 & .66 & .79 & .93 & .76 & .72 & 1.00 & .72 & .75 & .77 & .78 \\
-.15 & -.40 & .88 & -.53 & -.10 & .08 & .81 & .77 & .85 & .96 & .72 & 1.00 & .33 & .32 & .34 \\
.79 & .68 & .21 & .58 & .85 & .90 & .43 & .71 & .28 & .26 & .75 & .33 & 1.00 & .86 & .92 \\
.78 & .60 & .31 & .54 & .80 & .85 & .54 & .69 & .36 & .28 & .77 & .32 & .86 & 1.00 & .82 \\
.73 & .62 & .25 & .50 & .81 & .78 & .53 & .76 & .35 & .29 & .78 & .34 & .92 & .82 & 1.00
\end{pmatrix}$$

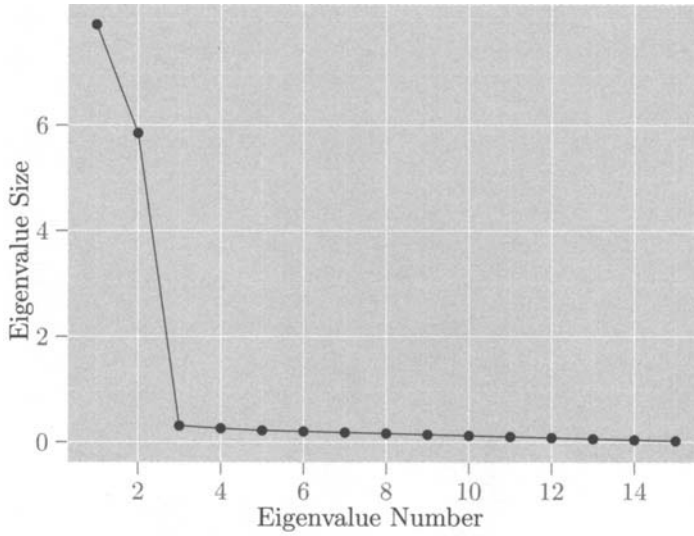


Figure 13.5 Scree graph for voice data.

The eigenvalues of $\mathbf{R}$ are $7.91, 5.85, .31, .26, \cdots, .002$, with the scree plot in Figure 13.5. Clearly, by any criterion for choosing m, there are two factors.

All four major methods of factor extraction discussed in Section 13.3 produced nearly identical results (after rotation). We give the initial and rotated loadings obtained from the principal component method in Table 13.10.

The two rotated factors were labeled "competence" and "benevolence." The same two factors emerged consistently in similar studies with different voices and different judges.

The two groupings of variables can also be seen in the correlation matrix. For example, in the first row, the large correlations correspond to the boldface rotated loadings for f_1, while in the third row, the large correlations correspond to the boldface rotated loadings for f_2.

The factor scores were of primary interest in this study. The goal was to ascertain the effect of the rate manipulations on the two factors, that is, to determine the perceived change in competence and benevolence when the speaking rate is increased or decreased.

The two factor scores were obtained for each of the 20 voices; these are plotted in Figure 13.6, where a consistent effect of the manipulation of speaking rate on all four voices can clearly be seen. Decreasing the speaking rate causes the speaker to be rated less competent; increasing the rate causes the speaker to be rated less benevolent. The mean vectors (centroids) are also given in Figure 13.6 for the four speakers. □

Table 13.10 Initial and Varimax Rotated Loadings for the Voice Data

	Initial Loadings		Rotated Loadings		
Variable	f_1	f_2	f_1	f_2	Communalities
Intelligent	.71	−.65	**.96**	−.06	.93
Ambitious	.48	−.84	**.90**	−.36	.94
Polite	.50	.81	−.12	**.95**	.92
Active	.37	−.91	**.86**	−.48	.97
Confident	.73	−.64	**.97**	−.04	.95
Happy	.83	−.47	**.94**	.15	.91
Just	.71	.58	.20	**.89**	.84
Likeable	.89	.39	.45	**.87**	.95
Kind	.58	.75	−.02	**.95**	.89
Sincere	.52	.82	−.11	**.97**	.95
Dependable	.93	.27	.56	**.79**	.94
Religious	.55	.79	−.07	**.96**	.92
Good-looking	.91	−.29	**.89**	.35	.91
Sociable	.91	−.22	**.84**	.40	.87
Strong	.91	−.21	**.84**	.41	.86
Variance accounted for	7.91	5.85	7.11	6.65	13.76
Proportion of total variance	.53	.39	.47	.44	.92
Cumulative proportion	.53	.92	.47	.92	.92

13.7 VALIDITY OF THE FACTOR ANALYSIS MODEL

For many statisticians, factor analysis is controversial and does not belong in a tool kit of legitimate multivariate techniques. The reasons for this mistrust include the following: the difficulty in choosing m, the many methods of extracting factors, the many rotation techniques, and the subjectivity in interpretation. Some statisticians also criticize factor analysis because of the indeterminacy of the factor loading matrix $\mathbf{\Lambda}$ or $\hat{\mathbf{\Lambda}}$, first noted in Section 13.2.2. However, it is the ability to rotate that gives factor analysis its utility, if not its charm.

The basic question is whether the factors really exist. The model (13.11) for the covariance matrix is $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$ or $\mathbf{\Sigma} - \mathbf{\Psi} = \mathbf{\Lambda}\mathbf{\Lambda}'$, where $\mathbf{\Lambda}\mathbf{\Lambda}'$ is of rank m. Many populations have covariance matrices that do not approach this pattern unless m is large. Thus the model will not fit data from such a population when we try to impose a small value of m. On the other hand, for a population in which $\mathbf{\Sigma}$ is reasonably close to $\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$ for small m, the sampling procedure leading to $\mathbf{S}$ may obscure this pattern. The researcher may believe there are underlying factors but has difficulty collecting data that will reveal them. In many cases, the basic problem is

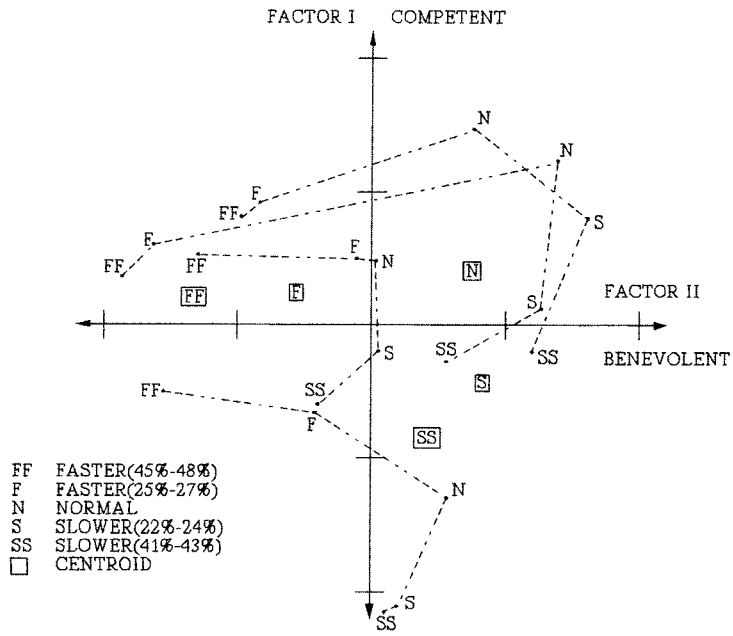


Figure 13.6 Factor scores of adjective rating of voices with five levels of manipulated rate.

that $\mathbf{S}$ (or $\mathbf{R}$) contains both structure and error, and the methods of factor analysis cannot separate the two.

A statistical consultant in a university setting or elsewhere all too often sees the following scenario. A researcher designs a long questionnaire, with answers to be given in, say, a five-point semantic differential scale or Likert scale. The respondents, who vary in attitude from uninterested to resentful, hurriedly mark answers that in many cases are not even good subjective responses to the questions. Then the researcher submits the results to a handy factor analysis program. Being disappointed in the results, he or she appeals to a statistician for help. They attempt to improve the results by trying different methods of extraction, different rotations, different values of m, and so on. But it is all to no avail. The scree plot looks more like foothills than a steep cliff with gently sloping debris at the bottom. There is no clear value of m. They have to extract 10 or 12 factors to account for, say, 60% of the variance, and interpretation of this large number of factors is hopeless. If a few underlying dimensions exist, they are totally obscured by both systematic and random errors in marking the questionnaire. A factor analysis model simply does not fit such a data set, unless a large value of m is used, which gives useless results.

It is not necessarily the "discreteness" of the data that causes the problem, but the "noisiness" of the data. The specified variables are not measured accurately. In some cases, discrete variables yield satisfactory results, such as in Examples 13.3.1, 13.3.2, 13.5.2a, and 13.5.2b(a), where a 12-year-old girl, responding carefully to a semantic

differential scale, produced data leading to an unambiguous factor analysis. On the other hand, continuous variables do not guarantee good results [see Example 13.7(a) below].

In cases in which some factors are found that provide a satisfactory fit to the data, we should still be tentative in interpretation until we can independently establish the existence of the factors. If the same factors emerge in repeated sampling from the same population or a similar one, then we can have confidence that application of the model has uncovered some real factors. Thus it is good practice to repeat the experiment to check the stability of the factors. If the data set is large enough, it could be split in half and a factor analysis performed on each half. The two solutions could be compared with each other and with that for the complete set.

If there is replication in the data set, it may be helpful to average over the replications. This was done to great advantage in Example 13.6, where several judges rated the same voices. Averaging over the judges produced variables that apparently possessed very low noise. Similar experimentation with different judges always produced the same factors. Unfortunately, replication of this type is unavailable in most situations.

As with other techniques in this book, factor analysis assumes that the variables are at least approximately linearly related to each other. We could make bivariate scatterplots to check this assumption.

A basic prerequisite for a factor analysis application is that the variables not be independent. To check this requirement, we could test $H_0\colon \mathbf{P}_\rho = \mathbf{I}$ by using the test in Section 7.4.3.

Some writers have suggested that $\mathbf{R}^{-1}$ should be a near-diagonal matrix in order to successfully fit a factor analysis model. To assess how close $\mathbf{R}^{-1}$ is to a diagonal matrix, Kaiser (1970) proposed a *measure of sampling adequacy*,

$$\text{MSA} = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} q_{ij}^2}, \tag{13.60}$$

where r_{ij}^2 is the square of an element from $\mathbf{R}$ and q_{ij}^2 is the square of an element from $\mathbf{Q} = \mathbf{D}\mathbf{R}^{-1}\mathbf{D}$, with $\mathbf{D} = [(\text{diag}\, \mathbf{R}^{-1})^{1/2}]^{-1}$. As $\mathbf{R}^{-1}$ approaches a diagonal matrix, MSA approaches 1. Kaiser and Rice (1974) suggest that MSA should exceed .8 for satisfactory results to be expected. We show some results for MSA in Example 13.7(b).

In summary, there are many data sets to which factor analysis should not be applied. One indication that $\mathbf{R}$ is inappropriate for factoring is the failure of the methods in Section 13.4 to clearly and rather objectively choose a value for m. If the scree plot does not have a pronounced bend or the eigenvalues do not show a large gap around 1, then $\mathbf{R}$ is likely to be unsuitable for factoring. In addition, the communality estimates after factoring should be fairly large.

To balance the "good" examples in this chapter, we now give an example involving a data set that cannot be successfully modeled by factor analysis. Likewise, the problems at the end of the chapter include both "good" and "bad" data sets.

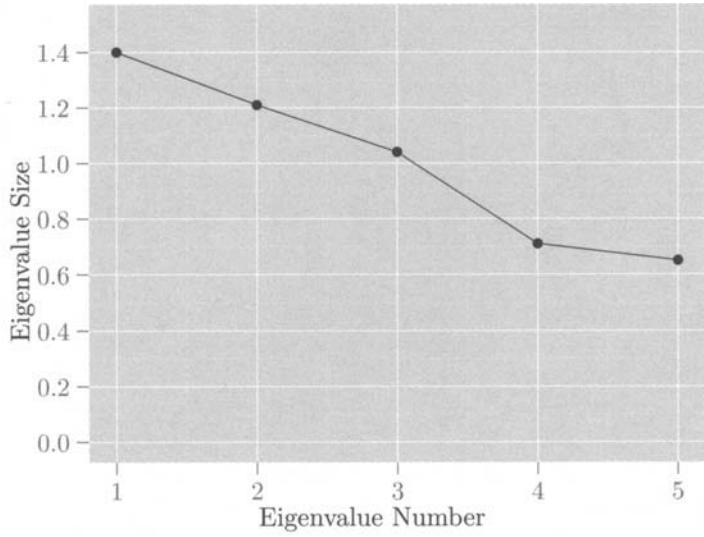


Figure 13.7 Scree graph for diabetes data.

■ **EXAMPLE 13.7(a)**

As an illustration of an application of factor analysis that is less successful than previous examples in this chapter, we consider the diabetes data of Table 3.4. The correlation matrix for the five variables is as follows:

$$\mathbf{R} = \begin{pmatrix} 1.00 & .05 & -.13 & .07 & .21 \\ .05 & 1.00 & -0.1 & .01 & -.10 \\ -.13 & -.01 & 1.00 & .29 & .05 \\ .07 & .01 & .29 & 1.00 & .21 \\ .21 & -.10 & .05 & .21 & 1.00 \end{pmatrix}.$$

The correlations are all small, and the variables do not appear to have much in common. The MSA value is .49. The eigenvalues are 1.40, 1.21, 1.04, .71, and .65. Three factors would be required to account for 73% of the variance and four factors to reach 87%. This is not a useful reduction in dimensionality. The eigenvalues are plotted in a scree graph in Figure 13.7. The lack of a clear value of m is apparent.

It is evident from the small correlations in $\mathbf{R}$ that the communalities of the variables will not be large. The principal component method, which essentially estimates the initial communalities as 1, gave very different final communality estimates than did the iterated principal factor method:

	Communalities				
Principal component method	.71	.91	.71	.67	.64
Iterated principal factor method	.31	.16	.35	.37	.33

Table 13.11 Varimax Rotated Factor Loadings for Iterated Principal Factors from the Diabetes Data

Variable	Rotated Loadings f_1	f_2	f_3	Communalities
1	−.08	**.54**	.12	.31
2	.01	.01	**.40**	.16
3	**.57**	−.15	−.03	.35
4	**.57**	.22	.02	.37
5	.19	**.47**	−.27	.33
Variance accounted for	.69	.59	.24	1.52

The communalities obtained by the iterated approach reflect more accurately the small correlations among the variables.

The varimax rotated factor loadings for three factors extracted by the iterated principal factor method are given in Table 13.11. The first factor is associated with variables 3 and 4, the second factor with variables 1 and 5, and the third with variable 2. This clustering of variables can be seen in $\mathbf{R}$, where variables 1 and 5 have a correlation of .21, variables 3 and 4 have a correlation of .29, and variable 2 has very low correlations with all other variables. However, these correlations (.21 and .29) are small, and in this case the collapsing of five variables to three factors is not a useful reduction in dimensionality, especially since the first three eigenvalues account for only 73% of tr($\mathbf{R}$). The 73% is not convincingly greater than 60%, which we would expect from three original variables picked at random. This conclusion is borne out by a test of H_0: $\mathbf{P}_\rho = \mathbf{I}$. Using (7.37) and (7.38), we obtain

$$u = |\mathbf{R}| = .80276, \quad \nu = 20 - 1 = 19, \quad p = 5,$$

$$u' = -[\nu - \tfrac{1}{6}(2p+5)] \ln u = -\left(19 - \frac{15}{6}\right)(-.2197) = 3.625.$$

With $\frac{1}{2}p(p-1) = 10$ degrees of freedom, the .05 critical value for this approximate χ^2 test is 18.31, and we have no basis to question the independence of the five variables. Thus the three factors we obtained are very likely an artifact of the present sample and would not reappear in another sample from the same population. □

■ **EXAMPLE 13.7(b)**

For data sets used in previous examples in this chapter, the values of MSA from (13.60) are calculated as follows:

Seishu data: MSA = .53
Sons data: MSA = .82
Voice data: MSA = .73
Diabetes data: MSA = .49.

The MSA value cannot be computed for the perception data, because $\mathbf{R}$ is singular.

These results do not suggest great confidence in the MSA index as a sole guide to the suitability of $\mathbf{R}$ for factoring. We see a wide disparity in the MSA values for the first three data sets. Yet all three yielded successful factor analyses. These three MSA values seem to be inversely related to the number of factors: In the sons data, there were indications that one factor would suffice; the voice data clearly had two factors; and for the Seishu data, there were four factors.

The MSA for the diabetes data is close to that of the Seishu data. Yet the diabetes data are totally unsuitable for factor analysis, while the factor analysis of the Seishu data is very convincing. □

13.8 RELATIONSHIP OF FACTOR ANALYSIS TO PRINCIPAL COMPONENT ANALYSIS

Both factor analysis and principal component analysis have the goal of reducing dimensionality. Because the objectives are similar, many authors discuss principal component analysis as another type of factor analysis. This can be confusing, and we wish to underscore the distinguishing characteristics of the two techniques.

Two of the differences between factor analysis and principal component analysis were mentioned in Section 13.1: (1) In factor analysis, the variables are expressed as linear combinations of the factors, whereas the principal components are linear functions of the variables, and (2) in principal component analysis, the emphasis is on explaining the total variance $\sum_i s_{ii}$, as contrasted with the attempt to explain the covariances in factor analysis.

Additional differences are that (3) principal component analysis requires essentially no assumptions, while factor analysis makes several key assumptions; (4) the principal components are unique (assuming distinct eigenvalues of $\mathbf{S}$), whereas the factors are subject to an arbitrary rotation; and (5) if we change the number of factors, the (estimated) factors change. This does not happen in principal components.

The ability to rotate to improve interpretability is one of the advantages of factor analysis over principal components. If finding and describing some underlying factors is the goal, factor analysis may prove more useful than principal components; we would prefer factor analysis if the factor model fits the data well and we like the interpretation of the rotated factors. On the other hand, if we wish to define a smaller number of variables for input into another analysis, we would ordinarily pre-

fer principal components, although this can sometimes be accomplished with factor scores. Occasionally, principal components are interpretable, as in the size and shape components in Example 12.8.1.

PROBLEMS

13.1 Show that the assumptions lead to (13.2), $\text{var}(y_i) = \lambda_{i1}^2 + \lambda_{i2}^2 + \cdots + \lambda_{im}^2 + \psi_i$.

13.2 Verify directly that $\text{cov}(\mathbf{y}, \mathbf{f}) = \boldsymbol{\Lambda}$ as in (13.13).

13.3 Show that $\mathbf{f}^* = \mathbf{T}'\mathbf{f}$ in (13.18) satisfies the assumptions (13.6) and (13.7), $E(\mathbf{f}^*) = \mathbf{0}$ and $\text{cov}(\mathbf{f}^*) = \mathbf{I}$.

13.4 Show that $\sum_{ij} e_{ij}^2 \leq \theta_{m+1}^2 + \theta_{m+2}^2 + \cdots + \theta_p^2$ as in (13.34), where the e_{ij}'s are the elements of $\mathbf{E} = \mathbf{S} - (\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}})$ and the θ_i's are eigenvalues of $\mathbf{S}$.

13.5 Show that $\sum_{i=1}^{p} \sum_{j=1}^{m} \hat{\lambda}_{ij}^2$ is equal to the sum of the first m eigenvalues and also equal to the sum of all p communalities, as in (13.46).

13.6 In Example 13.3.2, the correlation matrix for the perception data was shown to have an eigenvalue equal to 0. Find the multicollinearity among the five variables that this implies.

13.7 Use the words data of Table 5.9.

- **(a)** Obtain principal component loadings for two factors.
- **(b)** Do a graphical rotation of the two factors.
- **(c)** Do a varimax rotation and compare the results with those in part (b).

13.8 Use the ramus bone data of Table 3.7.

- **(a)** Extract loadings by the principal component method and do a varimax rotation. Use two factors.
- **(b)** Do all variables have a complexity of 1? Carry out an oblique rotation to improve the loadings.
- **(c)** What is the angle between the oblique axes? Would a single factor ($m = 1$) be more appropriate here?

13.9 Carry out a factor analysis of the rootstock data of Table 6.2. Combine the six groups into a single sample.

- **(a)** Estimate the loadings for two factors by the principal component method and do a varimax rotation.
- **(b)** Did the rotation improve the loadings?

13.10 Use the fish data of Table 6.17. Combine the three groups into a single sample.

(a) Obtain loadings on two factors by the principal component method and do a varimax rotation.

(b) Note the similarity of loadings for y_1 and y_2. Is there any indication in the correlation matrix as to why this is so?

(c) Compute factor scores.

(d) Using the factor scores, carry out a MANOVA comparing the three groups.

13.11 Carry out a factor analysis of the flea data in Table 5.5. Combine the two groups into a single sample.

(a) From an examination of the eigenvalues greater than 1, the scree plot, and the percentages, is there a clear choice of m?

(b) Extract two factors by the principal component method and carry out a varimax rotation.

(c) Is the rotation an improvement? Try an oblique rotation.

13.12 Use the engineer data of Table 5.6. Combine the two groups into a single sample.

(a) Using a scree plot, the number of eigenvalues greater than 1, and the percentages, is there a clear choice of m?

(b) Extract three factors by the principal component method and carry out a varimax rotation.

(c) Extract three factors by the principal factor method and carry out a varimax rotation.

(d) Compare the results of parts (b) and (c).

13.13 Use the probe word data of Table 3.6.

(a) Obtain loadings for two factors by the principal component method and carry out a varimax rotation.

(b) Note the near duplication of loadings for y_2 and y_4. Is there any indication in the correlation matrix as to why this is so?

(c) Is the rotation satisfactory? Try an oblique rotation.