4
Exploratory Factor Analysis

4.1 Introduction

In many areas of psychology and other disciplines in the behavioural sciences, it is often not possible to measure directly the concepts of primary interest. Two obvious examples are *intelligence* and *social class*. In such cases the researcher is forced to examine the concepts *indirectly* by collecting information on variables that *can* be measured or observed directly, and which can also realistically be assumed to be indicators, in some sense, of the concepts of real interest. The psychologist who is interested in an individual's "intelligence," for example, may record examination scores in a variety of different subjects in the expectation that these scores are related in some way to what is widely regarded as "intelligence." And a sociologist, say, concerned with people's "social class," might pose questions about a person's occupation, educational background, home ownership, etc., on the assumption that these do reflect the concept he or she is really interested in.

Both "intelligence" and "social class" are what are generally referred to as *latent variables*; i.e., concepts that cannot be measured directly but can be assumed to relate to a number of measurable or *manifest* variables. The method of analysis most generally used to help uncover the relationships between the assumed latent variables and the manifest variables is *exploratory factor analysis*. The model on which the method is based is essentially that of multiple regression, except now the manifest variables are regressed on the unobservable latent variables (often referred to in this context as *common factors*), so that direct estimation of the corresponding regression coefficients (*factor loadings*) is not possible.

4.2 The Factor Analysis Model

The basis of factor analysis is a regression model linking the manifest variables to a set of unobserved (and unobservable) latent variables. In essence the model assumes that the observed relationships between the manifest variables (as measured by their covariances or correlations) are a result of the relationships of these variables to the latent variables.

(Since it is the covariances or correlations of the manifest variables that that are central to factor analysis we can, in the description of the mathematics of the method given in Display 4.1, assume that the manifest variables all have zero mean.)

Display 4.1
Mathematics of the Factor Analysis Model

- We assume that we have a set of observed or manifest variables, $\mathbf{x}' = [x_1, x_2, \ldots, x_q]$, assumed to be linked to a smaller number of unobserved latent variables, $f_1, f_2, \ldots, f_k$ where $k < q$, by a regression model of the form

$$\begin{aligned} x_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1k} f_k + u_1, \\ x_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2k} f_k + u_2, \\ &\vdots \\ x_q &= \lambda_{q1} f_1 + \lambda_{q2} f_2 + \cdots + \lambda_{qk} f_k + u_q. \end{aligned}$$

- The λ_{ij}'s are weights showing how each x_i depends on the common factors.
- The λ_{ij}'s are used in the interpretation of the factors, i.e., larger values relate a factor to the corresponding observed variables and from these we infer a meaningful description of each factor.
- The equations above may be written more concisely as

$$\mathbf{x} = \Lambda \mathbf{f} + \mathbf{u},$$

where

$$\Lambda = \begin{pmatrix} \lambda_{11} & L & \lambda_{1k} \\ \vdots & \vdots & \vdots \\ \lambda_{q1} & L & \lambda_{qk} \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} f_1 \\ \vdots \\ f_k \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_q \end{pmatrix}.$$

- We assume that the "residual" terms $u_1, \ldots, u_q$ are uncorrelated with each other and with the factors $f_1, \ldots, f_k$. The elements of u are specific to each x_i and hence are known as *specific variates*.
- The two assumptions above imply that, given the values of the factors, the manifest variables are independent, that is, the correlations of the observed variables arise from their relationships with the factors. In factor analysis the regression coefficients in Λ are more usually known as *factor loadings*.
- Since the factors are unobserved we can fix their location and scale arbitrarily. We shall assume they occur in standardized form with mean zero and standard deviation one. We shall also assume, initially at least, that the factors are uncorrelated with one another, in which case the factor loadings are the correlations of the manifest variables and the factors.

- With these additional assumptions about the factors, the factor analysis model implies that the variance of variable x_i, σ_i^2, is given by

$$\sigma_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_i,$$

  where ψ_i is the variance of u_i.
- So the factor analysis model implies that the variance of each observed variable can be split into two parts. The first, h_i^2, given by

$$h_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2,$$

  is known as the *communality* of the variable and represents the variance shared with the other variables via the common factors. The second part, ψ_i, is called the *specific* or *unique* variance, and relates to the variability in x_i not shared with other variables.
- In addition, the factor model leads to the following expression for the covariance of variables x_i and x_j:

$$\sigma_{ij} = \sum_{l=1}^{k} \lambda_{il}\lambda_{jl}.$$

- The covariances are not dependent on the specific variates in any way; the common factors above account for the relationships between the manifest variables.
- So the factor analysis model implies that the population covariance matrix, $\mathbf{\Sigma}$, of the observed variables has the form

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi},$$

  where

$$\mathbf{\Psi} = \text{diag}(\psi_i).$$

- The converse also holds: If $\mathbf{\Sigma}$ can be decomposed into the form given above, then the k-factor model holds for $\mathbf{x}$.
- In practice, $\mathbf{\Sigma}$ will be estimated by the sample covariance matrix $\mathbf{S}$ (alternatively, the model will be applied to the correlation matrix $\mathbf{R}$), and we will need to obtain estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ so that the observed covariance matrix takes the form required by the model (see later in the chapter for an account of estimation methods).
- We will also need to determine the value of k, the number of factors, so that the model provides an adequate fit to $\mathbf{S}$ or $\mathbf{R}$.

To apply the factor analysis model outlined in Display 4.1 to a sample of multivariate observations we need to estimate the parameters of the model in some way. The estimation problem in factor analysis is essentially that of finding $\hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{\Psi}}$ for which

$$\mathbf{S} \approx \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}.$$

(If the $x_i s$ are standardized, then $\mathbf{S}$ is replaced by $\mathbf{R}$.)

There are two main methods of estimation leading to what are known as *principal factor analysis* and *maximum likelihood factor analysis*, both of which are now briefly described.

4.2.1 *Principal Factor Analysis*

Principal factor analysis is an eigenvalue and eigenvector technique similar in many respects to principal components analysis (see Chapter 3), but operating not directly on $\mathbf{S}$ (or $\mathbf{R}$), but on what is known as the *reduced covariance matrix*, $\mathbf{S}^*$, defined as

$$\mathbf{S}^* = \mathbf{S} - \hat{\mathbf{\Psi}},$$

where $\hat{\mathbf{\Psi}}$ is a diagonal matrix containing estimates of the ψ_i.

The diagonal elements of $\mathbf{S}^*$ contain estimated *communalities*—the parts of the variance of each observed variable that can be explained by the common factors. Unlike principal components analysis, factor analysis does not try to account for *all* observed variance only that shared through the common factors. Of more concern in factor analysis is to account for the covariances or correlations between the manifest variables.

To calculate $\mathbf{S}^*$ (or with $\mathbf{R}$ replacing $\mathbf{S}$, $\mathbf{R}^*$) we need values for the communalities. Clearly we cannot calculate them on the basis of factor loadings as described in Display 4.1 since these loadings still have to be estimated. To get round this seemingly "chicken and egg" situation we need to find a sensible way of finding initial values for the communalities that does not depend on knowing the factor loadings. When the factor analysis is based on the correlation matrix of the manifest variables two frequently used methods are the following:

- Take the communality of a variable x_i as the square of the multiple correlation coefficient of x_i with the other observed variables.
- Take the communality of x_i as the largest of the absolute values of the correlation coefficients between x_i and one of the other variables.

Each of these possibilities will lead to higher values for the initial communality when x_i is highly correlated with at least some of the other manifest variables, which is essentially what is required.

Given initial communality values, a principal components analysis is performed on $\mathbf{S}^*$, and the first k eigenvectors used to provide the estimates of the loadings in the k-factor model. The estimation process can stop here or the loadings obtained at this stage ($\hat{\lambda}_{ij}$) can provide revised communality estimates calculated as $\sum_{j=1}^{k} \hat{\lambda}_{ij}^2$. The

procedure is then repeated until some convergence criterion is satisfied. Difficulties can sometimes arise with this iterative approach if at any time a communality estimate exceeds the variance of the corresponding manifest variable, resulting in a negative estimate of the variable's specific variance. Such a result is known as a *Heywood case* (Heywood, 1931) and is clearly unacceptable since we cannot have a negative specific variance.

4.2.2 *Maximum Likelihood Factor Analysis*

Maximum likelihood is regarded, by statisticians at least, as perhaps the most respectable method of estimating the parameters in the factor analysis model. The essence of this approach is to define a type of "distance" measure, F, between the observed covariance matrix and the predicted value of this matrix from the factor analysis model. The measure F is defined as

$$F = \ln|\mathbf{\Lambda\Lambda}' + \mathbf{\Psi}| + \text{trace}(\mathbf{S}|\mathbf{\Lambda\Lambda}' + \mathbf{\Psi}|^{-1}) - \ln|\mathbf{S}| - q.$$

The function F takes the value zero if $\mathbf{\Lambda\Lambda}' + \mathbf{\Psi}$ is equal to $\mathbf{S}$ and values greater than zero otherwise. Estimates of the loadings and the specific variances are found by minimizing F; details are given in Lawley and Maxwell (1971), Mardia et al. (1979), and Everitt (1984, 1987).

Minimizing F is equivalent to maximizing L, the likelihood function for the k-factor model, under the assumption of multivariate normality of the data, since L equals $-\frac{1}{2}nF$ plus a function of the observations. As with iterated principal factor analysis, the maximum likelihood approach can also experience difficulties with Heywood cases.

4.3 Estimating the Numbers of Factors

The decision over how many factors, k, are needed to give an adequate representation of the observed covariances or correlations is generally critical when fitting an exploratory factor analysis model. A k and $k + 1$ solution will often produce quite different factors and factor loadings for all factors, unlike a principal component analysis in which the first k components will be identical in each solution. And as pointed out by Jolliffe (1989), with too few factors there will be too many high loadings, and with too many factors, factors may be fragmented and difficult to interpret convincingly.

Choosing k might be done by examining solutions corresponding to different values of k and deciding subjectively which can be given the most convincing interpretation. Another possibility is to use the scree diagram approach described in Chapter 3, although the usefulness of this rule is not so clear in factor analysis since the eigenvalues represent variances of principal components not factors.

An advantage of the maximum likelihood approach is that it has an associated formal hypothesis testing procedure for the number of factors. The test statistic is

$$U = n' \min(F),$$

where $n' = n + 1 - \frac{1}{6}(2q + 5) - \frac{2}{3}k$. If k common factors are adequate to account for the observed covariances or correlations of the manifest variables, then U has, asymptotically, a chi-squared distribution with ν degrees of freedom, where

$$\nu = \frac{1}{2}(q - k)^2 - \frac{1}{2}(q + k).$$

In most exploratory studies k cannot be specified in advance and so a sequential procedure is used. Starting with some small value for k (usually $k = 1$), the parameters in the corresponding factor analysis model are estimated by maximum likelihood. If U is not significant the current value of k is accepted, otherwise k is increased by one and the process repeated. If at any stage the degrees of freedom of the test become zero, then either no nontrivial solution is appropriate or alternatively the factor model itself with its assumption of linearity between observed and latent variables is questionable.

4.4 A Simple Example of Factor Analysis

The estimation procedures outlined in the previous section are needed in practical applications of factor analysis where invariably there are fewer parameters in the model than there are independent elements in **S** or **R** from which these parameters are to be estimated. Consequently the fitted model represents a genuinely parsimonious description of the data. But it is of some interest to consider a simple example in which the number of parameters is equal to the number of independent elements in **R** so that an exact solution is possible.

Spearman (1904) considered a sample of children's examination marks in three subjects—Classics (x_1), French (x_2), and English (x_3)—from which he calculated the following correlation matrix for a sample of children:

$$\mathbf{R} = \begin{matrix} \text{Classics} \\ \text{French} \\ \text{English} \end{matrix} \begin{pmatrix} 1.00 & & \\ 0.83 & 1.00 & \\ 0.78 & 0.67 & 1.00 \end{pmatrix}.$$

If we assume a single factor, then the appropriate factor analysis model is

$$\begin{aligned} x_1 &= \lambda_1 f + u_1, \\ x_2 &= \lambda_2 f + u_2, \\ x_3 &= \lambda_3 f + u_3. \end{aligned}$$

In this example the common factor, f, might be equated with intelligence or general intellectual ability, and the specific variates, u_1, u_2, u_3 will have small

variances if their associated observed variable is closely related to f. Here the number of parameters in the model (6) is equal to the number of independent elements in **R**, and so by equating elements of the observed correlation matrix to the corresponding values predicted by the single-factor model we will be able to find estimates of λ_1, λ_2, λ_3, ψ_1, ψ_2, and ψ_3 such that the model fits exactly. The six equations derived from the matrix equality implied by the factor analysis model, namely

$$\mathbf{R} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} \begin{bmatrix} \lambda_1 & \lambda_2 & \lambda_3 \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ 0 & 0 & \psi_3 \end{bmatrix}$$

are

$$\begin{aligned} \hat{\lambda}_1\lambda_2 &= 0.83, \\ \hat{\lambda}_1\lambda_3 &= 0.78, \\ \hat{\lambda}_1\lambda_4 &= 0.67, \\ \hat{\psi}_1 &= 1.0 - \hat{\lambda}_1^2, \\ \hat{\psi}_2 &= 1.0 - \hat{\lambda}_2^2, \\ \hat{\psi}_3 &= 1.0 - \hat{\lambda}_3^2. \end{aligned}$$

The solutions of these equations are

$$\begin{aligned} &\hat{\lambda}_1 = 0.99, \quad \hat{\lambda}_2 = 0.84, \quad \hat{\lambda}_3 = 0.79, \\ &\hat{\psi}_1 = 0.02, \quad \hat{\psi}_2 = 0.30, \quad \hat{\psi}_3 = 0.38. \end{aligned}$$

Suppose now that the observed correlations had been

$$\mathbf{R} = \begin{matrix} \text{Classics} \\ \text{French} \\ \text{English} \end{matrix} \begin{pmatrix} 1.00 & & \\ 0.84 & 1.00 & \\ 0.60 & 0.35 & 1.00 \end{pmatrix}.$$

In this case the solution for the parameters of a single factor model is

$$\begin{aligned} &\hat{\lambda}_1 = 1.2, \qquad \hat{\lambda}_2 = 0.7, \qquad \hat{\lambda}_3 = 0.5, \\ &\hat{\psi}_1 = -0.44, \quad \hat{\psi}_2 = 0.51, \quad \hat{\psi}_3 = 0.75. \end{aligned}$$

Clearly this solution is unacceptable because of the negative estimate for the first specific variance.

4.5 Factor Rotation

Until now we have ignored one problematic feature of the factor analysis model, namely that as formulated in Display 4.1, there is no unique solution for the factor

loading matrix. We can see that this is so by introducing an orthogonal matrix $\mathbf{M}$ of order $k \times k$, and rewriting the basic regression equation linking the observed and latent variables as

$$\mathbf{x} = (\mathbf{\Lambda M})(\mathbf{M}'\mathbf{f}) + \mathbf{u}.$$

This "new" model satisfies all the requirements of a k-factor model as previously outlined with new factors $\mathbf{f}^* = \mathbf{M}'\mathbf{f}$ and the new factor loadings $\mathbf{\Lambda M}$. This model implies that the covariance matrix of the observed variables is

$$\mathbf{\Sigma} = (\mathbf{\Lambda M})(\mathbf{\Lambda M})' + \mathbf{\Psi},$$

which, since $\mathbf{MM}' = \mathbf{I}$, reduces to $\mathbf{\Sigma} = \mathbf{\Lambda\Lambda}' + \mathbf{\Psi}$ as before. Consequently factors $\mathbf{f}$ with loadings $\mathbf{\Lambda}$ and factors $\mathbf{f}^*$ with loadings $\mathbf{\Lambda M}$ are, for any orthogonal matrix $\mathbf{M}$, equivalent for explaining the covariance matrix of the observed variables. Essentially then there are an infinite number of solutions to the factor analysis model as previously formulated.

The problem is generally solved by introducing some constraints in the original model. One possibility is to require the matrix $\mathbf{G}$ given by

$$\mathbf{G} = \mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}$$

to be diagonal, with its element arranged in descending order of magnitude. Such a requirement sets the first factor to have maximal contribution to the common variance of the observed variables, the second has maximal contribution to this variance subject to being uncorrelated with the first, and so on (cf. principal components analysis in Chapter 3).

The constraints on the factor loadings imposed by a condition such as that given above need to be introduced to make the parameter estimates in the factor analysis model unique. These conditions lead to orthogonal factors that are arranged in descending order of importance and enable an initial factor analysis solution to be found. The properties are not, however, inherent in the factor model, and merely considering such a solution may lead to difficulties of interpretation. For example, two consequences of these properties of a factor solution are as follows:

- The factorial complexity of variables is likely to be greater than one regardless of the underlying true model; consequently variables may have substantial loadings on more than one factor.
- Except for the first factor, the remaining factors are often *bipolar*, that is, they have a mixture of positive and negative loadings.

It may be that a more interpretable solution can be achieved using the equivalent model with loadings $\mathbf{\Lambda}^* = \mathbf{\Lambda M}$ for some particular orthogonal matrix, $\mathbf{M}$. Such a process is generally known as *factor rotation*, but before we consider how to choose $\mathbf{M}$, that is, how to "rotate" the factors, we need to address the question "Is factor rotation an acceptable process?"

Certainly in the past, factor analysis has been the subject of severe criticism because of the possibility of rotating factors. Critics have suggested that this apparently allows the investigator to impose on the data whatever type of solution they

are looking for. Some have even gone so far as to suggest that factor analysis has become popular in some areas precisely because it *does* enable users to impose their preconceived ideas of the structure behind the observed correlations (Blackith and Reyment, 1971). But, on the whole, such suspicions are not justified and factor rotation can be a useful procedure for simplifying an exploratory factor analysis. Factor rotation merely allows the fitted factor analysis model to be described as simply as possible; rotation does not alter the overall structure of a solution but only how the solution is described.

Rotation is a process by which a solution is made more interpretable without changing its underlying mathematical properties. Initial factor solutions with variables loading on several factors and with bipolar factors can be difficult to interpret. Interpretation is more straightforward if each variable is highly loaded on at most one factor, and if all factor loadings are either large and positive, or near zero, with few intermediate values. The variables are thus split into disjoint sets, each of which is associated with a single factor. This aim is essentially what Thurstone (1931) referred to as *simple structure*. In more detail such structure has the following properties:

- Each row or the factor-loading matrix should contain at least one zero.
- Each column of the loading matrix should contain at least k zeros.
- Every pair of columns of the loading matrix should contain several variables whose loadings vanish in one column but not in the other.
- If the number of factors is four or more, every pair of columns should contain a large number of variables with zero loadings in both columns.
- Conversely for every pair of columns of the loading matrix only a small number of variables should have nonzero loadings in both columns.

When simple structure is achieved the observed variables will fall into mutually exclusive groups whose loadings are high on single factors, perhaps moderate to low on a few factors, and of negligible size on the remaining factors.

The search for simple structure or something close to it begins after an initial factoring has determined the number of common factors necessary and the communalties of each observed variable. The factor loadings are then transformed by post multiplication by a suitably chosen orthogonal matrix. Such a transformation is equivalent to a rigid rotation of the axes of the originally identified factor space. For a two-factor model the process of rotation can be performed graphically. As an example, consider the following correlation matrix for six school subjects:

$$\mathbf{R} = \begin{array}{c} \text{French} \\ \text{English} \\ \text{History} \\ \text{Arithmetic} \\ \text{Algebra} \\ \text{Geometry} \end{array} \begin{pmatrix} 1.00 & & & & & \\ 0.44 & 1.00 & & & & \\ 0.41 & 0.35 & 1.00 & & & \\ 0.29 & 0.35 & 0.16 & 1.00 & & \\ 0.33 & 0.32 & 0.19 & 0.59 & 1.00 & \\ 0.25 & 0.33 & 0.18 & 0.47 & 0.46 & 1.00 \end{pmatrix}.$$

The initial factor loadings are plotted in Figure 4.1. By referring each variable to the new axes shown, which correspond to a rotation of the original axes through about

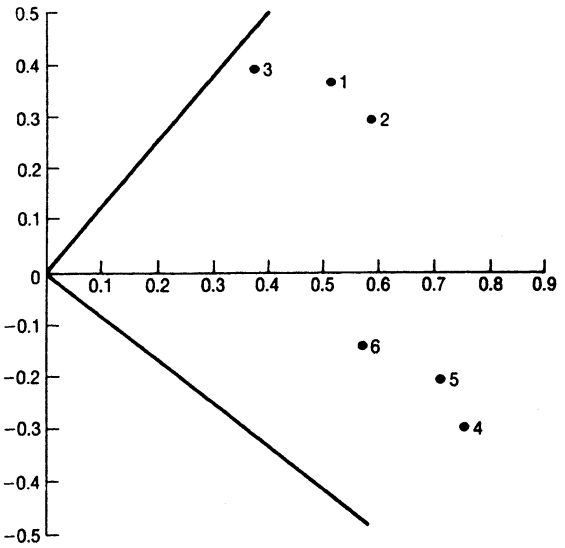

Figure 4.1 Plot of factor loadings showing a rotation of original axis.

40 degrees, a new set of loadings that give an improved description of the fitted model can be obtained. The two sets of loadings are given explicitly in Table 4.1

When there are more than two factors, more formal methods of rotation are needed. And during the rotation phase we might choose to abandon one of the assumptions made previously, namely that factors are orthogonal, that is, independent (the condition was assumed initially simply for convenience in describing the factor analysis model). Consequently two types of rotation are possible:

- *Orthogonal rotation*: methods restrict the rotated factors to being uncorrelated.
- *Oblique rotation*: methods allow correlated factors.

So the first question that needs to be considered when rotating factors is whether or not we should use an orthogonal or oblique rotation? As for many questions posed

Table 4.1 Two-Factor Solution for Correlations of Six School Subjects

	Unrotated loadings		Rotated loadings	
Variable	1	2	1	2
French	0.55	0.43	0.20	0.62
English	0.57	0.29	0.30	0.52
History	0.39	0.45	0.05	0.55
Arithmetic	0.74	−0.27	0.75	0.15
Algebra	0.72	−0.21	0.65	0.18
Geometry	0.59	−0.13	0.50	0.20

in data analysis, there is no universal answer to this question. There are advantages and disadvantages to using either type of rotation procedure. As a general rule, if a researcher is primarily concerned with getting results that "best fit" his/her data, then the researcher should rotate the factors obliquely. If, on the other hand, the researcher is more interested in the generalizability of his/her results, then orthogonal rotation is probably to be preferred.

One major advantage of an orthogonal rotation is simplicity since the loadings represent correlations between factors and manifest variables. This is *not* the case with an oblique rotation because of the correlations between the factors. Here there are two parts of the solution to consider:

- *Factor pattern coefficients*: regression coefficients that multiply with factors to produce measured variables according to the common factor model.
- *Factor structure coefficients*: correlation coefficients between manifest variables and the factors.

Additionally there is a matrix of factor correlations to consider. In many cases where these correlations are relatively small, researchers may prefer to return to an orthogonal solution.

There are a variety of rotation techniques although only relatively few are in general use. For orthogonal rotation the two most commonly used techniques are know as *varimax* and *quartimax*:

- *Varimax rotation*: originally proposed by Kaiser (1958), this has as its rationale the aim of factors with a few large loadings and as many near-zero loadings as possible. This is achieved by iterative maximization of a quadratic function of the loadings; details are given in Marda et al. (1979). This produces factors that have high correlations with one small set of variables and little or no correlation with other sets. There is a tendency for any general factor to disappear because the factor variance is redistributed.
- *Quartimax rotation*: originally suggested by Carroll (1953) this approach forces a given variable to correlate highly on one factor and either not at all or very low on other factors. Far less popular than varimax.

For oblique rotation the two methods most often used are *oblimin* and *promax*.

- *Oblimin rotation*: invented by Jennrich and Sampson (1966) this method attempts to find simple structure with regard to the factor pattern matrix through a parameter that is used to control the degree of correlation between the factors. Fixing a value for this parameter is not straightforward, but Lackey and Sullivan (2003) suggest that values between about -0.5 and 0.5 are sensible for many applications.
- *Promax rotation*: a method due to Hendrickson and White (1964) that operates by raising the loadings in an orthogonal solution (generally a varimax rotation) to some power. The goal is to obtain a solution that provides the best structure using the lowest possible power loadings and the lowest correlation between the factors.

As mentioned earlier, factor rotation is often regarded as controversial since it apparently allows the investigator to impose on the data whatever type of solution is required. But this is clearly *not* the case since although the axes may be rotated about their origin, or may be allowed to become oblique, *the distribution of the points will remain invariant*. Rotation is simply a procedure that allows new axes to be chosen so that the positions of the points can be described as simply as possible.

It should be noted that rotation techniques are also often applied to the results from a principal components analysis in the hope that it will aid in their interpretability. Although in some cases this may be acceptable, it does have several disadvantages which are listed by Jolliffe (1989). The main problem is that the defining property of principal components, namely that of accounting for maximal proportions of the total variation in the observed variables, is lost after rotation.

4.6 Estimating Factor Scores

In most applications an exploratory factor analysis will consist of the estimation of the parameters in the model and the rotation of the factors, followed by an (often heroic) attempt to interpret the fitted model. There are occasions, however, when the investigator would like to find factor scores for each individual in the sample. Such scores, like those derived in a principal components analysis (see Chapter 3), might be useful in a variety of ways. But the calculation of factor scores is not as straightforward as the calculation of principal components scores. In the original equation defining the factor analysis model, the variables are expressed in terms of the factors, whereas to calculate scores we require the relationship to be in the opposite direction. Bartholomew (1987) makes the point that to talk about "estimating" factor score is essentially misleading since they are random variables, and the issue is really one of prediction.

But if we make the assumption of normality, the conditional distribution of $\mathbf{f}$ given $\mathbf{x}$ can be found. It is

$$\mathbf{N}[\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\mathbf{x}, (\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} + \mathbf{I})^{-1}].$$

Consequently, one plausible way of calculating factor scores would be to use the sample version of the mean of this distribution, namely

$$\hat{\mathbf{f}} = \hat{\boldsymbol{\Lambda}}'\mathbf{S}^{-1}\mathbf{x},$$

where the vector of scores for an individual, $\mathbf{x}$, is assumed to have mean zero, that is, sample means for each variable have already been subtracted. Other possible methods for deriving factor scores are described in Rencher (1995). In many respects the most damaging problem with factor analysis is not the rotational indeterminacy of the loadings, but the indeterminacy of the factor scores.

4.7 Two Examples of Exploratory Factor Analysis

4.7.1 *Expectations of Life*

The data in Table 4.2 show life expectancy in years by country, age, and sex. The data come from Keyfitz and Flieger (1971) and relate to life expectancies in the 1960s.

We will use the formal test for number of factors incorporated into the maximum likelihood approach. We can apply this test to the data, assumed to be contained in the dataframe `life` with the country names labelling the rows and variables names as given in parentheses in Table 4.2, using the following R and S-PLUS code:

```
life.fa1<-factanal(life,factors=1,method="mle")
life.fa1
life.fa2<-factanal(life,factors=2,method="mle")
life.fa2
life.fa3<-factanal(life,factors=3,method="mle")
life.fa3
```

The results from the test are shown in Table 4.3. These results indicate that a three-factor solution is adequate for the data, although it has to be remembered that with only 31 countries, use of an asymptotic test result may be rather suspect. (The numerical results from R and S-PLUS® may differ a little.)

To find the details of the three-factor solution given by maximum likelihood we use the single R instruction

```
life.fa3
```

(In S-PLUS `summary(life.fa3)` is needed.)

The results, shown in Table 4.4, correspond to a varimax-rotated solution (the default for the `factanal` function). For interest we might also compare this with results from the quartimax rotation technique. The necessary S-PLUS code to find this solution is

```
life.fa3<-factanal(life,factors=3,method="mle",
  rotation="quartimax")summary(life.fa3)
```

(R does not have the quartimax option in `factanal`.) The results are shown in Table 4.5.

The first two factors from both varimax and quartimax are similar. The first factor is dominated by life expectancy at birth for both males and females and the second reflects life expectancies at older ages. The third factor from the varimax rotation has its highest loadings for the life expectancies of men aged 50 and 75.

If using S-PLUS the estimated factor scores are already available in `life.fa3$scores`. In R the scores have to be requested as follows;

```
scores<-factanal(life,factors=3,method="mle",
scores="regression")$scores
```

Table 4.2 Life Expectancies for Different Countries by Age and Sex

	Male				Female			
Age	0 (m0)	25 (m25)	50 (m50)	75 (m75)	0 (w0)	25 (w25)	50 (w50)	75 (w75)
Algeria	63	51	30	13	67	54	34	15
Cameroon	34	29	13	5	38	32	17	6
Madagascar	38	30	17	7	38	34	20	7
Mauritius	59	42	20	6	64	46	25	8
Reunion	56	38	18	7	62	46	25	10
Seychelles	62	44	24	7	69	50	28	14
South Africa (B)	50	39	20	7	55	43	23	8
South Africa (W)	65	44	22	7	72	50	27	9
Tunisia	56	46	24	11	63	54	33	19
Canada	69	47	24	8	75	53	29	10
Cost Rica	65	48	26	9	68	50	27	10
Dominican Republic	64	50	28	11	66	51	29	11
El Salvador	56	44	25	10	61	48	27	12
Greenland	60	44	22	6	65	45	25	9
Grenada	61	45	22	8	65	49	27	10
Guatemala	49	40	22	9	51	41	23	8
Honduras	59	42	22	6	61	43	22	7
Jamaica	63	44	23	8	67	48	26	9
Mexico	59	44	24	8	63	46	25	8
Nicaragua	65	48	28	14	68	51	29	13
Panama	65	48	26	9	67	49	27	10
Trinidad (62)	64	63	21	7	68	47	25	9
Trinidad (67)	64	43	21	6	68	47	24	8
United States (66)	67	45	23	8	74	51	28	10
United States (NW66)	61	40	21	10	67	46	25	11
United States (W66)	68	46	23	8	75	52	29	10
United States (67)	67	45	23	8	74	51	28	10
Argentina	65	46	24	9	71	51	28	10
Chile	59	43	23	10	66	49	27	12
Colombia	58	44	24	9	62	47	25	10
Ecuador	57	46	28	9	60	49	28	11

Table 4.3 Results from Test for Number of Factors on the Data in Table 4.2 Using R

1. Test of the hypothesis that one factor is sufficient versus the alternative that more are required:

   The chi square statistic is 163.11 on 20 degrees of freedom. The p-value is <0.0001.

2. Test of the hypothesis that two factors are sufficient versus the alternative that more are required:

   The chi square statistic is 45.24 on 13 degrees of freedom. The p-value is <0.0001.

3. Test of the hypothesis that three factors are sufficient versus the alternative that more are required:

   The chi square statistic is 6.73 on 7 degrees of freedom. The p-value is 0.458.

Table 4.4 Maximum Likelihood Three-Factor Solution for Life Expectancy Data After Varimax Rotation Using R

Importance of factors:

	Factor 1	Factor 2	Factor 3
SS loadings	3.38	2.08	1.64
Proportion Var	0.42	0.26	0.21
Cumulative Var	0.42	0.68	0.89

The degrees of freedom for the model is 7.

Uniquenesses:

M0	M25	M50	M75	W0	W25	W50	W75
0.005	0.362	0.066	0.288	0.005	0.011	0.020	0.146

Loadings:

	Factor 1	Factor 2	Factor 3	Communality
M0	0.97	0.12	0.23	0.9999
M25	0.65	0.17	0.44	0.6491
M50	0.43	0.35	0.79	0.9018
M75	—	0.53	0.66	0.7077
W0	0.97	0.22	—	0.9951
W25	0.76	0.56	0.31	0.9890
W50	0.54	0.73	0.40	0.9793
W75	0.16	0.87	0.28	0.8513

The factor scores are shown in Table 4.6 (again the scores from R and S-PLUS may differ a little). We can use the scores to provide a 3-D plot of the data by first creating a new dataframe

```
#if using S-PLUS we need scores<-life.fa3$scores
lifex<-data.frame(life,scores)
attach(lifex)
```

Table 4.5 Three-Factor Solution for Life Expectancy Data After Quartimax Rotation Using S-PLUS

	Factor 1	Factor 2	Factor 3
SS loadings	4.57	2.13	0.37
Proportion Var	0.57	0.26	0.04
Cumulative Var	0.57	0.84	0.88

The degrees of freedom for the model is 7.

Uniquenesses:

M0	M25	M50	M75	W0	W25	W50	W75
0.0000876	0.3508347	0.09818739	0.2923066	0.004925743	0.01100307	0.02074596	0.1486658

Loadings:

	Factor 1	Factor 2	Factor 3	Communality
M0	0.99	—	—	0.9999
M25	0.76	0.18	0.21	0.6491
M50	0.66	0.57	0.37	0.9018
M75	0.33	0.74	0.23	0.7077
W0	0.98	—	−0.16	0.9951
W25	0.90	0.39	−0.14	0.9890
W50	0.74	0.65	−0.14	0.9793
W75	0.37	0.80	−0.28	0.8513

Table 4.6 Factor Scores from the Three-Factor Solution for the Life Expectancy Data

	Factor 1	Factor 2	Factor 3
Algeria	−0.26	1.92	1.96
Cameroon	−2.84	−0.69	−1.98
Madagascar	−2.82	−1.03	0.29
Mauritius	0.15	−0.36	−0.77
Reunion	−0.19	0.35	−1.39
Seychelles	0.38	0.90	−0.71
South Africa (B)	−1.07	0.06	−0.87
South Africa (W)	0.95	0.12	−1.02
Tunisia	−0.87	3.52	−0.21
Canada	1.27	0.26	−0.22
Cost Rica	0.52	−0.52	1.06
Dominican Republic	0.11	−0.01	1.94
El Salvador	−0.64	0.82	0.25
Greenland	0.24	−0.67	−0.45
Grenada	0.15	0.11	0.08
Guatemala	−1.48	−0.64	0.62
Honduras	0.07	−1.93	0.38

(*Continued*)

Table 4.6 (*Continued*)

Jamaica	0.48	−0.58	0.17
Mexico	−0.07	−0.60	0.26
Nicaragua	0.28	0.08	1.77
Panama	0.47	−0.84	1.43
Trinidad (62)	0.72	−1.07	−0.00
Trinidad (67)	0.82	−1.24	−0.36
United States (66)	1.14	0.20	−0.75
United States (NW66)	0.41	−0.39	−0.74
United States (W66)	1.23	0.40	−0.68
United States (67)	1.14	0.20	−0.75
Argentina	0.73	0.31	−0.21
Chile	−0.02	0.91	−0.73
Colombia	−0.26	−0.19	0.28
Ecuador	−0.75	0.62	1.36

and then using the S-PLUS GUI as described in Chapter 2. The resulting diagram is shown in Figure 4.2.

Ordering along the first axis reflects life expectancy at birth ranging from Cameroon and Madagascar to countries such as the United States. And on the third axis Algeria is prominent because it has high life expectancy amongst men at higher ages with Cameroon at the lower end of the scale with a low life expectancy for men over 50.

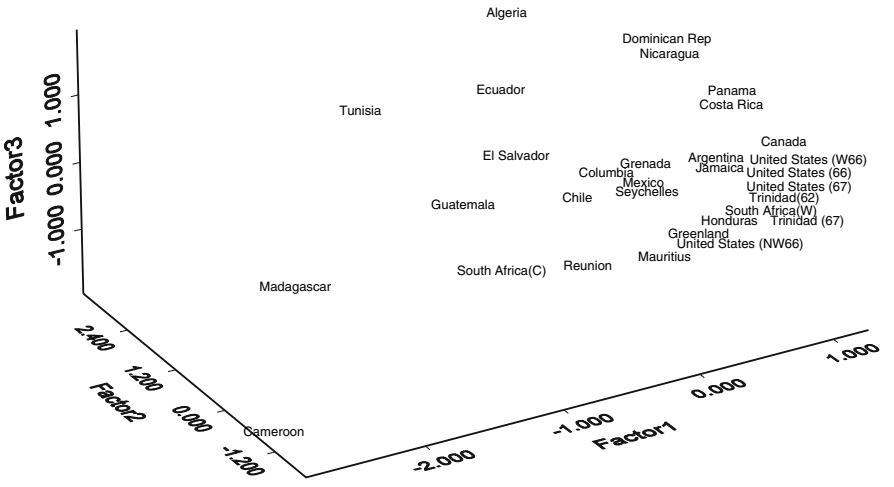

Figure 4.2 Plot of three-factor scores for life expectancy data.

4.7.2 *Drug Usage by American College Students*

The majority of adult and adolescent Americans regularly use psychoactive substances during an increasing proportion of their lifetime. Various forms of licit and illicit psychoactive substances use are prevalent, suggesting that patterns of psychoactive substance taking are a major part of the individual's behavioural repertory and have pervasive implications for the performance of other behaviors. In an investigation of these phenomena, Huba et al. (1981) collected data on drug usage rates for 1634 students in the seventh to ninth grades in 11 schools in the greater metropolitan area of Los Angeles. Each participant completed a questionnaire about the number of times a particular substance had ever been used. The substances asked about were as follows:

X1. cigarettes
X2. beer
X3. wine
X4. liquor
X5. cocaine
X6. tranquillizers
X7. drug store medications used to get high
X8. heroin and other opiates
X9. marijuana
X10. hashish
X11. inhalants (glue, gasoline, etc.)
X12. hallucinogenics (LSD, mescaline, etc.)
X13. amphetamine, stimulants

Responses were recorded on a five-point scale;

1. never tried
2. only once
3. a few times
4. many times
5. regularly

The correlations between the usage rates of the 13 substances are shown in Table 4.7. We first try to determine the number of factors using the maximum likelihood test. Here the S-PLUS code needs to accommodate the use of the correlation matrix rather than the raw data. We assume the correlation matrix is available as the data frame `druguse.cor`. The R code for finding the results of the test for number of factors here is

R

```
druguse.fa<-lapply(1:6,function(nf)
factanal(covmat=druguse.cor,factors=nf,method="mle",
  n.obs=1634)
```

Table 4.7 Correlation Matrix for Drug Usage Data

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
X1	1												
X2	0.447	1											
X3	0.442	0.619	1										
X4	0.435	0.604	0.583	1									
X5	0.114	0.068	0.053	0.115	1								
X6	0.203	0.146	0.139	0.258	0.349	1							
X7	0.091	0.103	0.110	0.122	0.209	0.221	1						
X8	0.082	0.063	0.066	0.097	0.321	0.355	0.201	1					
X9	0.513	0.445	0.365	0.482	0.186	0.315	0.150	0.154	1				
X10	0.304	0.318	0.240	0.368	0.303	0.377	0.163	0.219	0.534	1			
X11	0.245	0.203	0.183	0.255	0.272	0.323	0.310	0.288	0.301	0.302	1		
X12	0.101	0.088	0.074	0.139	0.279	0.367	0.232	0.320	0.204	0.368	0.304	1	
X13	0.245	0.199	0.184	0.293	0.278	0.545	0.232	0.314	0.394	0.467	0.392	0.511	1

The S-PLUS code here is a little different

S-PLUS

```
druguse.list<-list(cov=druguse.cor,center=rep(0,13),
  n.obs=1634)
druguse.fa<-lapply(1:6,function(nf)
  factanal(covlist=druguse.list,factors=nf,method="mle"))
```

The results from the test of number of factors are shown in Table 4.8. The test suggests that a six-factor model is needed. The results from the six-factor varimax solution are obtained from

R: `druguse.fa[[6]]`

S-PLUS: `summary(druguse.fa[[6]])`

and are shown in Table 4.9. The first factor involves cigarettes, beer, wine, liquor, and marijuana and we might label it "social/soft drug usage." The second factor has high loadings on cocaine, tranquillizers, and heroin. The obvious label for the factor is "hard drug usage." Factor three is essentially simply amphetamine use, and factor four is hashish use. We will not try to interpret the last two factors even though the formal test for number of factors indicated that a six-factor solution was necessary. It may be that we should not take the results of the formal test too literally. Rather, it may be a better strategy to consider the value of k indicated by the test to be an upper bound on the number of factors with practical importance. Certainly a six-factor solution for a data set with only 13 manifest variables might be regarded as not entirely satisfactory, and clearly we would have some difficulties interpreting all the factors.

Table 4.8 Results of Formal Test for Number of Factors on Drug Usage Data from R

1. Test of the hypothesis that one factor is sufficient versus the alternative that more are required:

   The chi square statistic is 2278.25 on 65 degrees of freedom. The p-value is <0.00001.

2. Test of the hypothesis that two factor is sufficient versus the alternative that more are required:

   The chi square statistic is 477.37 on 53 degrees of freedom. The p-value is <0.00001.

3. Test of the hypothesis that three factors are sufficient versus the alternative that more are required:

   The chi square statistic is 231.95 on 42 degrees of freedom. The p-value is <0.00001.

4. Test of the hypothesis that four factors are sufficient versus the alternative that more are required:

   The chi square statistic is 113.42 on 32 degrees of freedom. The p-value is <0.00001.

5. Test of the hypothesis that five factors are sufficient versus the alternative that more are required:

   The chi square statistic is 60.57 on 23 degrees of freedom. The p-value is <0.00001.

6. Test of the hypothesis that six factors are sufficient versus the alternative that more are required:

   The chi square statistic is 23.97 on 15 degrees of freedom. The p-value is 0.066.

One of the problems is that with the large sample size in this example, even small discrepancies between the correlation matrix predicted by a proposed model and the observed correlation matrix may lead to rejection of the model. One way to investigate this possibility is to simply look at the differences between the observed and predicted correlations. We shall do this first for the six-factor model using the following R and S-PLUS code:

```
pred<-druguse.fa[[6]]$loadings%*%t(druguse.fa[[6]]
   $loadings)+
diag(druguse.fa[[6]]$uniquenesses)
druguse.cor-pred
```

The resulting matrix of differences is shown in Table 4.10. The differences are all very small, underlining that the six-factor model does describe the data very well.

Now let us look at the corresponding matrices for the three- and four-factor solutions found in a similar way; see Table 4.11. Again in both cases the residuals are all relatively small, suggesting perhaps that use of the formal test for number of

Table 4.9 Maximum Likelihood of Six-Factor Solution for Drug Usage Data—Varimax Rotation

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
SS loadings	2.30	1.43	1.13	0.95	0.68	0.61
Proportion Var	0.18	0.11	0.09	0.07	0.05	0.05
Cumulative Var	0.80	0.29	0.37	0.45	0.50	0.55

The degrees of freedom for the model is 15.

Uniquenesses:

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
0.560	0.368	0.374	0.411	0.681	0.526	0.748	0.665	0.324	0.025	0.597	0.630	4r-010

Loadings:

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
X1	0.49	—	—	—	0.41	—
X2	0.78	—	—	0.10	0.11	—
X3	0.79	—	—	—	—	—
X4	0.72	0.12	0.10	0.12	0.16	—
X5	—	0.52	—	0.13	—	0.16
X6	0.13	0.56	0.32	0.10	0.14	—
X7	—	0.24	—	—	—	0.42
X8	—	0.54	0.10	—	—	0.19
X9	0.43	0.16	0.15	0.26	0.60	0.10
X10	0.24	0.28	0.19	0.87	0.20	—
X11	0.17	0.32	0.16	—	0.15	0.47
X12	—	0.39	0.34	0.19	—	0.26
X13	0.15	0.34	0.89	0.14	0.14	0.17

factors leads, in this case, to overfitting. The three-factor model appears to provide a perfectly adequate fit for these data.

4.8 Comparison of Factor Analysis and Principal Components Analysis

Factor analysis, like principal components analysis, is an attempt to explain a set of multivariate data using a smaller number of dimensions than one begins with, but the procedures used to achieve this goal are essentially quite different in the two approaches. Some differences between the two are as follows:

- Factor analysis tries to explain the covariances or correlations of the observed variables by means of a few common factors. Principal components analysis is primarily concerned with explaining the variance of the observed variables.
- If the number of retained components is increased, say, from m to $m + 1$, the first m components are unchanged. This is not the case in factor analysis, where there can be substantial changes in *all* factors if the number of factors is unchanged.
- The calculation of principal component scores is straightforward. The calculation of factor scores is more complex, and a variety of methods have been suggested.

Table 4.10 Differences Between Observed and Predicted Correlations for Six-Factor Model Fitted to Drug Usage Data

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
X1	0.000	−0.001	0.015	−0.018	0.010	0.000	−0.020	−0.005	0.002	0	0.013	−0.003	0
X2	−0.001	0.000	−0.002	0.004	0.004	−0.011	−0.002	0.007	0.002	0	−0.004	0.005	0
X3	0.015	−0.002	0.000	−0.001	0.000	−0.005	0.007	0.008	−0.004	0	−0.008	−0.001	0
X4	−0.018	0.004	−0.001	0.000	−0.07	0.020	−0.002	−0.018	0.004	0	0.013	−0.004	0
X5	0.010	0.004	0.000	−0.007	0.000	0.002	0.005	0.003	−0.004	0	−0.002	−0.008	0
X6	0.000	−0.011	−0.005	0.020	0.002	−0.001	0.011	−0.004	−0.003	0	−0.002	−0.008	0
X7	−0.020	−0.002	0.007	−0.002	0.005	0.011	0.002	−0.018	0.007	0	0.005	−0.003	0
X8	−0.005	0.007	0.008	−0.018	0.003	−0.004	−0.018	0.001	0.006	0	0.002	0.021	0
X9	0.002	0.002	−0.004	0.004	−0.004	−0.003	0.007	0.006	0.000	0	−0.007	0.002	0
X10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0.000	0.000	0
X11	0.013	−0.004	−0.008	0.013	−0.002	−0.002	0.005	0.002	−0.007	0	−0.003	0.017	0
X12	−0.003	0.005	−0.001	−0.004	−0.008	−0.008	−0.003	0.021	0.002	0	−0.019	−0.003	0
X13	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0	0.000	0.000	0

Table 4.11 Differences Between Observed and Predicted Correlations for Three- and Four-Factor Models Fitted to Drug Usage Data

1. Three factor

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
X1	0.000	−0.001	0.009	−0.013	0.011	0.009	−0.011	−0.004	0.002	−0.026	0.039	−0.017	0.002
X2	−0.001	0.000	−0.001	0.002	0.002	−0.014	0.000	0.005	−0.001	0.019	−0.003	0.009	−0.007
X3	0.009	−0.001	0.000	0.000	−0.002	−0.004	0.012	0.013	0.001	−0.017	−0.007	0.004	0.002
X4	−0.013	0.002	0.000	0.000	−0.008	0.023	−0.018	−0.020	−0.001	0.014	−0.002	−0.015	0.005
X5	0.011	0.002	−0.002	−0.008	0.000	0.030	0.038	0.082	−0.002	0.041	0.026	−0.027	−0.076
X6	0.009	−0.014	−0.004	0.023	0.030	0.002	−0.022	0.024	−0.001	−0.017	−0.035	−0.057	0.040
X7	−0.011	0.000	0.012	−0.018	0.038	−0.022	0.000	0.021	0.006	−0.040	0.116	0.002	−0.038
X8	−0.004	0.005	0.013	−0.020	0.082	0.024	0.021	0.000	0.006	−0.035	0.034	−0.003	−0.050
X9	0.002	−0.001	0.001	−0.001	−0.002	−0.001	0.006	0.006	0.000	0.001	0.002	−0.002	−0.002
X10	−0.026	0.019	−0.017	0.014	0.041	−0.017	−0.040	−0.035	0.001	0.000	−0.033	0.036	0.010
X11	0.039	−0.003	−0.007	−0.002	0.026	−0.035	0.116	0.034	0.002	−0.033	−0.004	0.014	−0.012
X12	−0.017	0.009	0.004	−0.015	−0.027	−0.057	0.002	−0.003	−0.002	0.036	−0.022	−0.002	0.044
X13	0.002	−0.007	0.002	0.005	−0.076	0.040	−0.038	−0.050	−0.002	0.010	−0.012	0.044	0.001

2. Four factor

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
X1	0.000	−0.001	0.008	−0.012	0.010	0.008	−0.014	−0.007	0.001	−0.023	0.037	−0.020	0
X2	−0.001	0.000	−0.001	0.001	0.001	−0.016	−0.002	0.003	−0.001	0.018	−0.005	0.006	0
X3	0.008	−0.001	0.000	0.000	−0.001	−0.005	0.012	0.014	0.001	−0.020	−0.008	0.001	0
X4	−0.012	0.001	0.000	0.000	−0.005	0.029	−0.015	−0.016	−0.001	0.018	0.001	−0.009	0
X5	0.010	0.001	−0.001	−0.005	0.002	0.020	−0.014	0.004	−0.003	0.033	−0.018	−0.026	0
X6	0.008	−0.016	−0.005	0.029	0.020	0.001	−0.021	0.024	−0.001	0.001	−0.029	−0.025	0
X7	−0.014	−0.002	0.012	−0.015	−0.014	−0.021	−0.001	−0.018	0.003	−0.042	0.095	0.011	0
X8	−0.007	0.003	0.014	−0.016	0.004	0.024	−0.018	0.001	0.003	−0.038	0.003	0.008	0
X9	0.001	−0.001	0.001	−0.001	−0.003	−0.001	0.003	0.003	0.000	0.000	0.001	−0.002	0
X10	−0.023	0.018	−0.020	0.018	0.033	0.001	−0.042	−0.038	0.000	0.000	−0.029	0.057	0
X11	0.037	−0.005	−0.008	0.001	−0.018	−0.029	0.095	0.003	0.001	−0.029	−0.006	0.028	0
X12	−0.020	0.006	0.001	−0.009	−0.026	−0.025	0.011	0.008	−0.002	0.057	−0.008	−0.002	0
X13	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0

- There is usually no relationship between the principal components of the sample correlation matrix and the sample covariance matrix. For maximum likelihood factor analysis, however, the results of analyzing either matrix are essentially equivalent (this is not true of principal factor analysis).

Despite these differences, the results from both types of analysis are frequently very similar. Certainly if the specific variances are small we would expect both forms of analysis to give similar results. However, if the specific variances are large they will be absorbed into all the principal components, both retained and rejected, whereas factor analysis makes special provision for them.

Lastly, it should be remembered that both principal components analysis and factor analysis are similar in one important respect—they are both pointless if the observed variables are almost uncorrelated. In this case factor analysis has nothing to explain and principal components analysis will simply lead to components which are similar to the original variables.

4.9 Confirmatory Factor Analysis

The methods described in this chapter have been those of exploratory factor analysis. In such models no constraints are placed on which of the manifest variables load on the common factors. But there is an alternative approach known as *confirmatory factor analysis* in which specific constraints *are* introduced, for example, that particular manifest variables are related to only one of the common factors with their loadings on other factors set a priori to be zero. These constraints may be suggested by theoretical considerations or perhaps from earlier exploratory factor analyses on similar data sets. Fitting confirmatory factor analysis models requires specialized software and readers are referred to Dunn et al. (1993) and Muthen and Muthen (1998).

4.10 Summary

Factor analysis has probably attracted more critical comments than any other statistical technique. Hills (1977), for example, has gone so far as to suggest that factor analysis is not worth the time necessary to understand it and carry it out. And Chatfield and Collins (1980) recommend that factor analysis should not be used in most practical situations. The reasons that these authors and others are so openly sceptical about factor analysis arises firstly from the central role of latent variables in the factor analysis model and secondly from the lack of uniqueness of the factor loadings in the model that gives rise to the possibility of rotating factors. It certainly is the case that since the common factors cannot be measured or observed, the existence of these hypothetical variables is open to question. A factor is a construct operationally defined by its factor loadings, and overly enthusiastic reification is not to be recommended.

It is the case that given one factor-loading matrix, there are an infinite number of factor-loading matrices that could equally well (or equally badly) account for the variances and covariances of the manifest variables. Rotation methods are designed to find an easily interpretable solution from among this infinitely large set of alternatives by finding a solution that exhibits the best simple structure.

Factor analysis can be a useful tool for investigating particular features of the structure of multivariate data. Of course, like many models used in data analysis, the one used in factor analysis may be only a very idealized approximation to the truth. Such an approximation may, however, prove a valuable starting point for further investigations.

Exercises

4.1 Show how the result $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$ arises from the assumptions of uncorrelated factors, independence of the specific variates, and independence of common factors and specific variances. What form does $\boldsymbol{\Sigma}$ take if the factors are allowed to be correlated?

4.2 Show that the communalities in a factor analysis model are unaffected by the transformation $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\mathbf{M}$.

4.3 Give a formula for the proportion of variance explained by the jth factor estimated by the principal factor approach.

4.4 Apply the factor analysis model separately to the life expectancies of men and women and compare the results.

4.5 Apply principal factor analysis to the drug usage data and compare the results with those given in the text from maximum likelihood factor analysis. Investigate the use of oblique rotation for these data.

4.6 The correlation matrix given below arises from the scores of 220 boys in six school subjects: (1) French, (2) English, (3) history, (4) arithmetic, (5) algebra, and (6) geometry. The two-factor solution from a maximum likelihood factor analysis is shown in Table 4.12. By plotting the derived loadings, find an

Table 4.12 Maximum Likelihood Factor Analysis for School Subjects Data

	Factor loadings		
Subject	F1	F2	Communality
1. French	0.55	0.43	0.49
2. English	0.57	0.29	0.41
3. History	0.39	0.45	0.36
4. Arithmetic	0.74	−0.27	0.62
5. Algebra	0.72	−0.21	0.57
6. Geometry	0.60	−0.13	0.37

orthogonal rotation that allows easier interpretation of the results.

$$\mathbf{R} = \begin{array}{c} \text{French} \\ \text{English} \\ \text{History} \\ \text{Arithmetic} \\ \text{Algebra} \\ \text{Geometry} \end{array} \begin{pmatrix} 1.00 & & & & & \\ 0.44 & 1.00 & & & & \\ 0.41 & 0.35 & 1.00 & & & \\ 0.29 & 0.35 & 0.16 & 1.00 & & \\ 0.33 & 0.32 & 0.19 & 0.59 & 1.00 & \\ 0.25 & 0.33 & 0.18 & 0.47 & 0.46 & 1.00 \end{pmatrix}.$$

4.7 The matrix below shows the correlations between ratings on nine statements about pain made by 123 people suffering from extreme pain. Each statement was scored on a scale from 1 to 6 ranging from agreement to disagreement. The nine pain statements were as follows:

1. Whether or not I am in pain in the future depends on the skills of the doctors.
2. Whenever I am in pain, it is usually because of something I have done or not done.
3. Whether or not I am in pain depends on what the doctors do for me.
4. I cannot get any help for my pain unless I go to seek medical advice.
5. When I am in pain I know that it is because I have not been taking proper exercise or eating the right food.
6. People's pain results from their own carelessness.
7. I am directly responsible for my pain.
8. Relief from pain is chiefly controlled by the doctors.
9. People who are never in pain are just plain lucky.

$$\mathbf{R} = \begin{pmatrix} 1.00 & & & & & & & & \\ -0.04 & 1.00 & & & & & & & \\ 0.61 & -0.07 & 1.00 & & & & & & \\ 0.45 & -0.12 & 0.59 & 1.00 & & & & & \\ 0.03 & 0.49 & 0.03 & -0.08 & 1.00 & & & & \\ -0.29 & 0.43 & -0.13 & -0.21 & 0.47 & 1.00 & & & \\ -0.30 & 0.30 & -0.24 & -0.19 & 0.41 & 0.63 & 1.00 & & \\ 0.45 & -0.31 & 0.59 & 0.63 & -0.14 & -0.13 & -0.26 & 1.00 & \\ 0.30 & -0.17 & 0.32 & 0.37 & -0.24 & -0.15 & -0.29 & 0.40 & 1.00 \end{pmatrix}$$

(a) Perform a principal components analysis on these data and examine the associated scree plot to decide on the appropriate number of components.
(b) Apply maximum likelihood factor analysis and use the test described in the chapter to select the necessary number of common factors.
(c) Rotate the factor solution selected using both an orthogonal and an oblique procedure, and interpret the results.