3

The normal linear factor model

3.1 The model

We have already introduced the normal linear factor (NLFM) model in Chapter 1 as an example of our general approach. Its main properties arise as special cases of the GLLVM given in Chapter 2. The emphasis in this chapter is on the implementation of the model and, especially, its interpretation.

The NLFM is the oldest and most widely used latent variable model. The analyses based on the model may be motivated in various ways and not all of them require the normality assumption. We shall therefore include sections at appropriate points which show what can be done if we do not assume normality. The distinctive feature of our presentation is that its development flows naturally from the framework provided in Chapter 2.

The NLFM assumes that

$$\mathbf{x} \,|\, \mathbf{y} \sim N_p(\boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{y},\ \boldsymbol{\Psi}) \tag{3.1}$$

and

$$\mathbf{y} \sim N_q(\mathbf{0},\ \mathbf{I}), \tag{3.2}$$

where $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix of *specific variances* and $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of *factor loadings*. We have already noted that there is no loss of generality in assuming zero means and unit variances for the ys. For some purposes, to be considered later, we may wish to allow correlations among the ys, but for the moment this is excluded. Note that there are two assumptions of normality involved.

Latent Variable Models and Factor Analysis: A Unified Approach, Third Edition.
David Bartholomew, Martin Knott and Irini Moustaki.
 Published 2011 by John Wiley & Sons, Ltd.

An equivalent, and more common, way of writing the model is as a linear equation in normal random variables. Thus

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{y} + \mathbf{e}, \tag{3.3}$$

where $\mathbf{y} \sim N_q(\mathbf{0},\ \mathbf{I})$, $\mathbf{e} \sim N_p(\mathbf{0},\ \boldsymbol{\Psi})$ and $\mathbf{y}$ is independent of $\mathbf{e}$.

The difference between the two versions lies in the manner in which the manifest variables are supposed to have been generated. In the first version this takes place in two stages. First we select the individual from the population; the values of $\mathbf{x}$ for that individual are then sampled from the conditional distribution of $\mathbf{x}$. In the second version $\mathbf{y}$ and $\mathbf{e}$ are selected simultaneously and independently and are then combined according to (3.3) to give $\mathbf{x}$. Whether we write the model in terms of probability distributions, as in (3.1) and (3.2), or as an equation in random variables, as in (3.3), is perhaps more a matter of taste than of substance. However, the blurring of the distinction between random variables and their realised values which (3.3) encourages can lead to confusion when we come to the question of what can be said about $\mathbf{y}$ after $\mathbf{x}$ has been observed.

Since the conditional distribution of $\mathbf{x}$ in (3.1) belongs to the exponential family there is a sufficient statistic which we gave by way of example in Section 2.6. In matrix notation it is

$$\mathbf{X} = \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\mathbf{x}, \tag{3.4}$$

in the parameterisation of the present section.

3.2 Some distributional properties

The two main properties have already been given in (1.12) and (1.13), namely

$$\mathbf{x} \sim N_p(\boldsymbol{\mu},\ \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}) \tag{3.5}$$

and

$$\mathbf{y} \mid \mathbf{x} \sim N_q(\boldsymbol{\Lambda}'(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1}(\mathbf{x} - \boldsymbol{\mu}),\ (\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} + \mathbf{I})^{-1}). \tag{3.6}$$

The first result is required for fitting the model by maximum likelihood; the second for making inferences about the latent variables on the basis of the observed *x*s.

The link between the conditional expectation of $\mathbf{y}$ and the components given in (3.4) is established by noting that

$$\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}) = (\mathbf{I} + \boldsymbol{\Gamma})\boldsymbol{\Lambda}', \tag{3.7}$$

where $\boldsymbol{\Gamma} = \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$. From this it follows that

$$\boldsymbol{\Lambda}'(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1} = (\mathbf{I} + \boldsymbol{\Gamma})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1} \tag{3.8}$$

and hence that

$$E(\mathbf{y} \mid \mathbf{x}) = (\mathbf{I} + \boldsymbol{\Gamma})^{-1}(\mathbf{X} - E(\mathbf{X})). \tag{3.9}$$

The matrix $\boldsymbol{\Gamma}$ plays a key role in what follows. If it is diagonal we see that, given $\mathbf{x}$, the ys are independent and each posterior mean is proportional to the corresponding component. Other consequences will emerge below.

It follows from (3.5) that

$$\text{var}(x_i) = \sum_{j=1}^{q} \lambda_{ij}^2 + \psi_i \quad (i = 1, 2, \ldots, p). \tag{3.10}$$

The variance of x_i is thus composed of two parts. The first, $\sum \lambda_{ij}^2$, arises from what is common to all xs. For this reason it is known as the *communality*. The complementary part, ψ_i, is the variance specific to that particular x_i.

Other distributional properties, which are useful for interpreting the model, concern the relationship between the factors on the one hand and the observables, $\mathbf{x}$ and $\mathbf{X}$, on the other.

(a) The covariance between the manifest variables and the factor is given by

$$\begin{aligned} E(\mathbf{x} - \boldsymbol{\mu})\mathbf{y}' &= E\left[E(\mathbf{x} - \boldsymbol{\mu})\mathbf{y}' \mid \mathbf{y}\right] \\ &= E\left[E\{(\mathbf{x} - \boldsymbol{\mu}) \mid \mathbf{y}\}\mathbf{y}'\right] \\ &= E(\boldsymbol{\Lambda}\mathbf{y}\mathbf{y}') = \boldsymbol{\Lambda}. \end{aligned}$$

The factor loadings can therefore be interpreted as covariances between individual manifest variables and factors. The correlations are given by

$$\{\text{diag } \boldsymbol{\Sigma}\}^{-1/2}\, \boldsymbol{\Lambda}.$$

(b) The covariance between the components and the factors is

$$\begin{aligned} E(\mathbf{X} - E(\mathbf{X}))\mathbf{y}' &= \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}E[(\mathbf{x} - \boldsymbol{\mu})\mathbf{y}'] \\ &= \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} = \boldsymbol{\Gamma}. \end{aligned}$$

If $\boldsymbol{\Gamma}$ is diagonal there are no cross-correlations between components and factors. For the correlations we require the covariance matrix of $\mathbf{X}$ which is

$$\begin{aligned} \text{var}(\mathbf{X}) &= E\left[\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right] \\ &= \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi})\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} \\ &= \boldsymbol{\Gamma}^2 + \boldsymbol{\Gamma}. \end{aligned}$$

The required correlation matrix is thus

$$\boldsymbol{\Gamma}\{\text{diag}(\boldsymbol{\Gamma}^2 + \boldsymbol{\Gamma})\}^{-1/2}.$$

So if $\boldsymbol{\Gamma}$ is diagonal the correlation between X_j and y_j is $(1 + \Gamma_j^{-1})^{-1/2}$, where Γ_j is the jth diagonal element of $\boldsymbol{\Gamma}$. The larger Γ_j, the larger will the correlation be. Recalling the role which $\boldsymbol{\Gamma}$ plays in the posterior distribution of $\mathbf{y}$, we can say that the smaller the variance of y_j (i.e. the more precisely it is determined by the data) the larger will Γ_j be and hence the more closely related will X_j and y_j be.

3.3 Constraints on the model

There are circumstances in which we may wish to place constraints on the parameters. For example, we have seen above that if the matrix $\boldsymbol{\Gamma} = \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ is diagonal the ys will be independent *a posteriori* and the relationship between the ys and the components is particularly simple. This may help to facilitate the interpretation of the model. This particular constraint also removes the freedom to arbitrarily rotate $\boldsymbol{\Lambda}$ (unless some elements of $\boldsymbol{\Gamma}$ are equal).

In confirmatory factor analysis, to which we come in Chapter 8, there may be grounds for specifying the pattern of the elements in $\boldsymbol{\Lambda}$. This usually takes the form of setting the values of certain λs equal to zero. This too will usually remove the rotational freedom.

A third type of constraint arises from the arbitrariness of the scale of the manifest variables in many applications, especially in social sciences. Any change of scale in an x will be reflected in the covariances and hence in the parameter estimates and their interpretation. We therefore need a parameterisation which is unaffected by arbitrary changes of scale. This can be done by expressing the xs in units of their standard deviation. The effect of this is to make all the diagonal elements of $\boldsymbol{\Sigma}$ equal to 1, that is,

$$\sum_{j=1}^{q} \lambda_{ij}^2 + \Psi_i = 1 \quad (i = 1, 2, \ldots, p).$$

The implications of this constraint for maximum likelihood estimation will be explained in Section 3.12.2.

3.4 Maximum likelihood estimation

Estimation by maximum likelihood is not easy. The treatment below follows somewhat the same lines as Lawley and Maxwell (1971), with some help from matrix differentiation results in Magnus and Neudecker (1988). There is an attempt to fill some of the gaps in previously available treatments, including that of the second edition of the present volume (Bartholomew and Knott 1999), by including Heywood cases, and being more careful about the choice of stationary values. The derivation here is, as far as we know, different from any already published.

If $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then the likelihood function for observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ may be written

$$l(\boldsymbol{\Sigma}) = \text{constant} + \frac{n}{2}[\ln \mid \boldsymbol{\Sigma}^{-1} \mid - \text{trace}[\boldsymbol{\Sigma}^{-1}\mathbf{S}]] \tag{3.11}$$

where $\mathbf{S} = \sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'/n$. The first step is to maximise with respect to $\boldsymbol{\mu}$. This is a standard problem and it is easily shown that $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$. Henceforth we shall suppose this estimate to have been substituted into $\mathbf{S}$. The novelty in the second stage is that we wish to maximise with respect to $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$, where

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \tag{3.12}$$

and where $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of rank q, and $\boldsymbol{\Psi}$ is diagonal.

It is convenient to use matrix differentiation results; see, for instance, Rao (1973), Magnus and Neudecker (1988). One can safely ignore the symmetry constraints. We have

$$\mathrm{d}l(\boldsymbol{\Sigma}) = \frac{n}{2}[-\text{trace}(\boldsymbol{\Sigma}^{-1}\mathrm{d}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}^{-1}(\mathrm{d}\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\mathbf{S})] \tag{3.13}$$

$$= -\frac{n}{2}\text{trace}[(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1})\mathrm{d}\boldsymbol{\Sigma}]. \tag{3.14}$$

From (3.14)

$$\mathrm{d}l(\boldsymbol{\Sigma}) = -\frac{n}{2}\text{trace}[(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1})((\mathrm{d}\boldsymbol{\Lambda})\boldsymbol{\Lambda}' + \boldsymbol{\Lambda}(\mathrm{d}\boldsymbol{\Lambda})')] \tag{3.15}$$

$$= -n \text{ trace}[\boldsymbol{\Lambda}'(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1})\mathrm{d}\boldsymbol{\Lambda}]. \tag{3.16}$$

So the stationarity conditions for $\boldsymbol{\Lambda}$ can be written

$$(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1})\boldsymbol{\Lambda} = \mathbf{0}. \tag{3.17}$$

Similarly, one can see that

$$\mathrm{d}l(\boldsymbol{\Sigma}) = -\frac{n}{2}\text{trace}[(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1})\mathrm{d}\boldsymbol{\Psi}], \tag{3.18}$$

so that stationarity conditions for $\boldsymbol{\Psi}$ are

$$\text{diag}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{S}\boldsymbol{\Sigma}^{-1}) = \mathbf{0}, \tag{3.19}$$

where diag $\mathbf{A}$ is the matrix with diagonal elements as for $\mathbf{A}$ and zeros elsewhere. Pre- and post-multiplying the matrices in (3.19) by the diagonal matrix $\boldsymbol{\Psi} = \boldsymbol{\Sigma} - \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ and using (3.17) gives the stationarity conditions for $\boldsymbol{\Psi}$ in the form

$$\text{diag}(\boldsymbol{\Sigma} - \mathbf{S}) = \mathbf{0}. \tag{3.20}$$

Now we follow Lawley and Maxwell in spirit, but unlike them allow $\boldsymbol{\Psi}$ to be singular to accommodate the Heywood cases that arise in practice (see Section 3.12.3). The

form of the stationary values for $\mathbf{\Lambda}$ can be made more transparent. Pre-multiplying (3.17) by $\mathbf{S}^{-1/2}\mathbf{\Sigma}\mathbf{S}^{-1}\mathbf{\Sigma}$ leads to

$$[\mathbf{S}^{-1/2}\mathbf{\Sigma}\mathbf{S}^{-1/2} - \mathbf{I}]\mathbf{S}^{-1/2}\mathbf{\Lambda} = \mathbf{0}. \tag{3.21}$$

So at stationary values of $\mathbf{\Lambda}$, the columns of $\mathbf{S}^{-1/2}\mathbf{\Lambda}$ are eigenvectors of $\mathbf{S}^{-1/2}\mathbf{\Sigma}\mathbf{S}^{-1/2}$ with eigenvalues 1. As might be expected, one can rotate the columns of $\mathbf{\Lambda}$ by post-multiplying by an orthogonal matrix, and (3.12) is still satisfied. Replacing $\mathbf{\Sigma}$ by $\mathbf{\Psi} + \mathbf{\Lambda}\mathbf{\Lambda}'$ gives

$$[\mathbf{S}^{-1/2}\mathbf{\Psi}\mathbf{S}^{-1/2}]\mathbf{S}^{-1/2}\mathbf{\Lambda} = \mathbf{S}^{-1/2}\mathbf{\Lambda}[\mathbf{I} - \mathbf{\Lambda}'\mathbf{S}^{-1}\mathbf{\Lambda}]. \tag{3.22}$$

So, if the columns of $\mathbf{\Lambda}$ are scaled to make $\mathbf{\Lambda}'\mathbf{S}^{-1}\mathbf{\Lambda}$ a diagonal matrix, one can see that, at stationary values of $\mathbf{\Lambda}$, we then have $\mathbf{S}^{-1/2}\mathbf{\Lambda}$ with columns which are eigenvectors of $\mathbf{S}^{-1/2}\mathbf{\Psi}\mathbf{S}^{-1/2}$ with eigenvalues no greater than 1. It is always possible to choose $\mathbf{\Psi}$ so that there are q eigenvalues no greater than 1. The scaling is always possible. It is easy to verify directly that if $\mathbf{V}$ is a $p \times q$ set of orthogonal eigenvectors of $\mathbf{S}^{-1/2}\mathbf{\Psi}\mathbf{S}^{-1/2}$ with eigenvalues (all no greater than 1) on the diagonal of the $q \times q$ matrix $\mathbf{\Delta}$, one can take

$$\mathbf{\Lambda} = \mathbf{S}^{1/2}\mathbf{V}(\mathbf{I} - \mathbf{\Delta})^{1/2}. \tag{3.23}$$

It is interesting to note that $\mathbf{\Lambda}$ from (3.23) cannot be arbitrarily rotated while still keeping $\mathbf{\Lambda}'\mathbf{S}^{-1}\mathbf{\Lambda}$ diagonal. A rotation is chosen by the way in which we solve the likelihood equations.

The choice of a particular stationary value for $\mathbf{\Lambda}$ can be made looking to see which one maximises the likelihood. To maximise the log-likelihood is to minimise

$$-\ln|\mathbf{\Sigma}^{-1}\mathbf{S}| + \text{trace}[\mathbf{\Sigma}^{-1}\mathbf{S}] = -\ln|\mathbf{S}^{1/2}\mathbf{\Sigma}^{-1}\mathbf{S}^{1/2}| + \text{trace}[\mathbf{S}^{1/2}\mathbf{\Sigma}^{-1}\mathbf{S}^{1/2}]. \tag{3.24}$$

So, if θ_i are the eigenvalues of $\mathbf{S}^{-1/2}\mathbf{\Sigma}\mathbf{S}^{-1/2}$, one must minimise

$$\sum_{i=1}^{n}[-\ln\theta_i + 1/\theta_i]. \tag{3.25}$$

As θ varies, $-\ln\theta + 1/\theta$ can be seen to take a minimum value of 1 when $\theta = 1$. For any stationary value of $\mathbf{\Lambda}$, q of the $\{\theta_i\}$ are equal to 1, and the other $p - q$ of the $\{\theta_i\}$ are also eigenvalues of $\mathbf{S}^{-1/2}\mathbf{\Psi}\mathbf{S}^{-1/2}$, and are less than 1. So the log-likelihood is maximised by choosing the columns of $\mathbf{S}^{-1/2}\mathbf{\Lambda}$ to be the eigenvectors corresponding to the q smallest of the eigenvalues of $\mathbf{S}^{-1/2}\mathbf{\Psi}\mathbf{S}^{-1/2}$.

Having found $\mathbf{\Lambda}$, one can use (3.20) to find a new value for $\mathbf{\Psi}$ and then iterate. As Jöreskog (1967) first showed, the hard part of maximum likelihood estimation is the determination of $\mathbf{\Psi}$, while given $\mathbf{\Psi}$ the maximisation over $\mathbf{\Lambda}$ is easy. One can see from the presentation above that the maximum likelihood estimation of $\mathbf{\Psi}$ chooses it so that the largest $p - q$ eigenvalues $\{\theta_i\}$, all less than 1, of $\mathbf{S}^{-1/2}\mathbf{\Psi}\mathbf{S}^{-1/2}$ minimise $\sum(-\ln\theta_i + 1/\theta_i)$.

It is interesting to see what happens if there is a Heywood case (see also Section 3.12.3). The simplest way this arises is if the maximum likelihood estimate of one of the diagonal elements of $\boldsymbol{\Psi}$, say the first one, is zero. Then $\mathbf{S}^{-1/2}\boldsymbol{\Psi}\mathbf{S}^{-1/2}$ is singular, so one of the $\{\theta_i\}$ is zero. The eigenvector associated with that zero eigenvalue will therefore be one of the columns of the maximum likelihood estimate of $\mathbf{S}^{-1/2}\boldsymbol{\Lambda}$. It is easy to check directly that a suitable column of $\boldsymbol{\Lambda}$ is $\frac{1}{s_{11}}\mathbf{S}_1$, where $\mathbf{S}_1$ is the first column of $\mathbf{S}$, and s_{11} is its first element. The maximum likelihood estimates for the remaining factors, from (3.22), are found by using the sub-matrix of $\mathbf{S}^{-1/2}$ obtained by omitting its first row and first column. That is the same as using the conditional covariance matrix for variables $2, \ldots, p$ given variable 1. Jöreskog (1967) has this result, but no obvious way of integrating it into his analytic treatment of maximum likelihood estimation.

3.5 Maximum likelihood estimation by the E-M algorithm

Another way of carrying out the maximum likelihood estimation is to use the E-M method, which is an iterative technique well suited to maximum likelihood estimation for models where there is missing information. The missing pieces of information here are the values of the latent variables. The E-M approach was introduced in Dempster *et al.* (1977), and applied directly to the normal factor analysis model in Rubin and Thayer (1982).

Starting with some initial values for the parameters, the procedure is to write down the joint likelihood of $(\mathbf{x}_i, \mathbf{y}_i)$ for $i = 1, \ldots, n$. The log-likelihood is replaced by its expected value (E-step) conditional on the $\mathbf{x}_i$. The expected values are worked out with the current values of the parameters at that iteration.

This modified likelihood is maximised (M-step) to give new values for the parameters, and the whole procedure iterated until convergence. Although convergence to the global maximum is not guaranteed, the marginal likelihood of $\mathbf{x}_i$ for $i = 1, \ldots, n$ will never decrease with each change in the parameters. This follows from use of the maximising property of the M-step, and a well-known information theory inequality on the conditional distributions of $\mathbf{y}$ given $\mathbf{x}$; see Dempster *et al.* (1977, equation (3.10)).

In practice it is easier to set the conditional expected value of the score function from the joint likelihood of $(\mathbf{x}_i, \mathbf{y}_i)$ given $\mathbf{x}_i$ equal to zero. Notice that this is in fact the score function from the marginal likelihood of the $\mathbf{x}_i$ (McLachlan and Krishnan 1997, (3.42), for instance).

The log-likelihood of the $(\mathbf{x}_i, \mathbf{y}_i)$ for $i = 1, \ldots, n$ is

$$\text{constant} - \frac{n}{2}\log \mid \boldsymbol{\Psi} \mid - \frac{n}{2}\text{trace}\, \boldsymbol{\Psi}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda}\mathbf{y}_i)(\mathbf{x}_i - \boldsymbol{\mu} - \boldsymbol{\Lambda}\mathbf{y}_i)'\right]$$
$$- \frac{n}{2}\text{trace}\, \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i\mathbf{y}_i'),$$

which gives a score function for $\boldsymbol{\mu}$,

$$n\boldsymbol{\Psi}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu} - \boldsymbol{\Lambda}\bar{\mathbf{y}}), \tag{3.26}$$

for $\boldsymbol{\Lambda}$,

$$n\boldsymbol{\Psi}^{-1}(\mathbf{S}'_{xy} - \boldsymbol{\mu}\bar{\mathbf{y}}' - \boldsymbol{\Lambda}\mathbf{S}'_{yy}), \tag{3.27}$$

and for $\boldsymbol{\Psi}$ the diagonal elements of

$$\begin{aligned} -\frac{n}{2}\boldsymbol{\Psi}^{-1} + \frac{n}{2}\boldsymbol{\Psi}^{-1}[&\mathbf{S}'_{xx} - \boldsymbol{\mu}\bar{\mathbf{x}}' - \bar{\mathbf{x}}\boldsymbol{\mu}' - \mathbf{S}'_{xy}\boldsymbol{\Lambda}' - \boldsymbol{\Lambda}\mathbf{S}'_{yx} \\ &+ \boldsymbol{\mu}\bar{\mathbf{y}}'\boldsymbol{\Lambda}' + \boldsymbol{\Lambda}\bar{\mathbf{y}}\boldsymbol{\mu}' + \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\Lambda}\mathbf{S}'_{yy}\boldsymbol{\Lambda}']\boldsymbol{\Psi}^{-1}, \end{aligned} \tag{3.28}$$

where

$$\mathbf{S}'_{xx} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}'_i.$$

To find the conditional expected values of the score functions, it is enough here to do so for the sufficient statistics:

$$\begin{aligned} \bar{\mathbf{y}} &= \frac{1}{n}\sum \mathbf{y}_i, \\ \mathbf{S}'_{xy} &= \frac{1}{n}\sum \mathbf{x}_i\mathbf{y}'_i, \\ \mathbf{S}'_{yy} &= \frac{1}{n}\sum \mathbf{y}_i\mathbf{y}'_i. \end{aligned}$$

Now,

$$E[\bar{\mathbf{y}} \mid \mathbf{x}_i, i = 1, \ldots, n] = \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) = \hat{\bar{\mathbf{y}}}, \tag{3.29}$$

$$\begin{aligned} E[\mathbf{S}'_{xy} \mid \mathbf{x}_i, i = 1, \ldots, n] &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i E[\mathbf{y}'_i \mid \mathbf{x}_i] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}))' \\ &= (\mathbf{S}'_{xx} - \bar{\mathbf{x}}\boldsymbol{\mu}')\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} = \hat{\mathbf{S}}'_{xy}, \end{aligned} \tag{3.30}$$

$$\begin{aligned} E[\mathbf{S}'_{yy} \mid \mathbf{x}_i, i = 1, \ldots, n] &= (\mathbf{I}_q + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1} + \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}(\mathbf{S}'_{xx} - \boldsymbol{\mu}\bar{\mathbf{x}}' - \bar{\mathbf{x}}\boldsymbol{\mu}' + \boldsymbol{\mu}\boldsymbol{\mu}')\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} \\ &= \hat{\mathbf{S}}'_{yy}. \end{aligned} \tag{3.31}$$

These are calculated replacing $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda}$ by their current estimates and then substituted into the score functions of (3.26), (3.27), (3.28) above. Setting the score functions equal to zero and solving,

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} - \boldsymbol{\Lambda}\hat{\bar{\mathbf{y}}}, \tag{3.32}$$

$$\hat{\boldsymbol{\Lambda}} = (\hat{\mathbf{S}}'_{xy} - \bar{\mathbf{x}}\hat{\bar{\mathbf{y}}}')(\hat{\mathbf{S}}'_{yy} - \hat{\bar{\mathbf{y}}}\hat{\bar{\mathbf{y}}}')^{-1}, \tag{3.33}$$

$$\begin{aligned}\hat{\boldsymbol{\Psi}} = \operatorname{diag}(&\mathbf{S}'_{xx} - \hat{\boldsymbol{\mu}}\bar{\mathbf{x}}' - \bar{\mathbf{x}}\hat{\boldsymbol{\mu}}' - \hat{\mathbf{S}}_{xy}\hat{\boldsymbol{\Lambda}}' - \hat{\boldsymbol{\Lambda}}\hat{\mathbf{S}}_{yx} \\ &+\hat{\boldsymbol{\mu}}\hat{\bar{\mathbf{y}}}'\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Lambda}}\hat{\bar{\mathbf{y}}}\hat{\boldsymbol{\mu}}' + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}' + \hat{\boldsymbol{\Lambda}}\hat{\mathbf{S}}'_{yy}\hat{\boldsymbol{\Lambda}}').\end{aligned} \tag{3.34}$$

The procedure is iterated until it converges. It is easier, as Rubin and Thayer noticed, to leave out $\boldsymbol{\mu}$, which has maximum likelihood estimator $\bar{\mathbf{x}}$. The natural way to allow for this would seem to be to substitute $\boldsymbol{\mu} = \bar{\mathbf{x}}$ into (3.29), (3.30) and (3.31) to give

$$\begin{aligned}\hat{\bar{\mathbf{y}}} &= 0, \\ \hat{\mathbf{S}}'_{xy} &= \mathbf{S}_{xx}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}, \\ \hat{\mathbf{S}}'_{yy} &= (\mathbf{I}_q + \boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1} + \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\mathbf{S}_{xx}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}.\end{aligned}$$

These are then used in (3.33) and (3.34) to provide new updating equations. The updating equation for $\hat{\boldsymbol{\Psi}}$ that results from this approach is not that obtained by Rubin and Thayer who, in effect, substitute the result of updating $\hat{\boldsymbol{\Lambda}}$ into the updating equation for $\hat{\boldsymbol{\Psi}}$.

3.6 Sampling variation of estimators

The lack of information about the sampling behaviour of parameter estimates in the factor model when used in exploratory mode is a serious defect of the standard software packages. Standard errors, whether asymptotic or exact, are only crude indicators of uncertainty in problems with many parameters, especially if they turn out to be highly correlated. Nevertheless, without some indication of the precision of estimators the interpretation of factors, which depends critically on those values, must be suspect. The need for such information was recognised from an early stage, as the Cudeck and O'Dell (1994) account shows, and the lack of adequate theory has led to the adoption of a variety of rules of thumb which bear no relation to the sampling variability of the estimators themselves. For example, it is common to regard factor loadings greater than 0.3 as 'significant' and it is not unusual to find published tables of loadings with the smaller loadings deleted. This may make the interpretation easier but it owes nothing to the statistical significance of the loadings themselves.

In the absence of exact sampling theory we can have recourse to the asymptotic theory of maximum likelihood. For the normal factor model this was first provided by Lawley (1967) for the case of unstandardised variables and extended to standardised

variables by Lawley and Maxwell (1971) who also gave a numerical illustration. Subsequently, Jennrich and Thayer (1973) provided a small correction to Lawley's results. As we pointed out in Section 2.11, the estimation process yields a set of estimates rather than a single point in the parameter space. What the Lawley (1967) method did was to give the sampling behaviour of the estimator given by the standard maximum likelihood routine, namely that for which $\mathbf{\Lambda}'\mathbf{\Psi}^{-1}\mathbf{\Lambda}$ is diagonal. Estimated standard errors for the loadings of rotated factors can be obtained by the usual 'delta' methods, and this was done for orthogonal rotation by Archer and Jennrich (1973) and for oblique rotations by Jennrich (1973). An account of the history of the subject is given in Cudeck and O'Dell (1994) who go on to extend the theory in various ways, including simultaneous inference for sets of parameters based on confidence intervals. Further results can be found in Ogasawara (1998).

The heavy computations required may have inhibited the implementation of these methods in a routine fashion but they are well within the scope of modern desktop computers. The Mplus software (Muthén and Muthén 2010) and the CEFA computer program (Browne *et al.* 1998) fit the NLFM and give standard deviations for many forms of rotated loadings, whether standardised or not.

However, one does not know how large the sample size need be before the approximations provided by the asymptotic results are adequate in practice. Results reported by de Menezes (1999), and discussed in Chapter 6 below for the latent class model, show that the asymptotic theory may provide a very poor approximation even if the sample size is of the order of 1000. Similar investigations for factor analysis seem to be lacking but, given the complexities of multi-parameter models, much more comparative information is needed before the asymptotic results can be used with confidence.

An alternative approach to studying sampling behaviour is through resampling or simulation methods of the bootstrap or jackknife variety. These too have a long history in factor analysis but have not been widely used because of the amount of computing needed. However, the situation is rapidly changing and it is now possible to contemplate using the approach with complex models and large sample sizes. Tucker *et al.* (1969) carried out a detailed simulation study and Seber (1984) reported a series of simulation studies on samples of size 50 by Francis (1974).

Pennell (1972) used the jackknife to find confidence intervals for factor loadings. In the first edition (Bartholomew 1987) we reported an application of the bootstrap by Chatterjee (1984). He used data taken from Johnson and Wichern (1982) concerning seven variables measured on 50 randomly chosen salesmen. Three variables were measures of sales performance and four were from tests of aptitude. The method of fitting was the principal factor method with the ψs set equal to zero (i.e. a principal components analysis). One would have preferred a method which allowed estimation of the ψs, but with only two or three factors turning out to be significant the results are not likely to be seriously affected and the computational ease of this method commended it for this particular study. Chatterjee settled the question of how many repeated samples to draw empirically, and it appeared that 300 gave reasonable stability. Some of the results are given in Table 3.1 for factor loadings (standardised λs) for the first three factors.

Table 3.1 Estimated loadings and their standard deviations for Johnson and Wichern's (1982) data.

	Original estimates			Bootstrap estimates on 300 samples				
Var.	Fac. 1	Fac. 2	Fac. 3	Fac. 1	s.d.	Fac. 2	s.d.	Fac. 3	s.d.
1	0.973	−0.110	0.054	0.972	0.006	−0.096	0.055	0.038	0.052
2	0.943	0.029	0.312	0.945	0.013	0.010	0.091	0.192	0.158
3	0.945	0.010	−0.144	0.943	0.020	0.014	0.079	0.076	0.168
4	0.660	0.646	−0.318	0.657	0.095	0.577	0.276	−0.179	0.254
5	0.783	0.286	−0.005	0.774	0.064	0.268	0.165	0.041	0.373
6	0.649	−0.620	−0.427	0.644	0.093	−0.516	0.358	−0.266	0.229
7	0.914	−0.193	0.306	0.916	0.017	−0.185	0.091	0.197	0.173

Table 3.1 shows that the first factor is well determined but that we should be wary of attributing much significance to the other two. None of the loadings for factor 3 differ from zero by much more than their standard deviation, and only two or three of those for factor 2 come anywhere near significance. An advantage of the bootstrap method is that we can also look at the frequency distributions of the estimators. Chatterjee (1984) gives a number of examples, of which that given in Table 3.2 is particularly instructive. We notice that the large standard deviation is due to the extreme skewness of the distribution. Six percent of the samples actually gave negative loadings (if *all* loadings on a factor were negative the signs should have been reversed) but even on the positive half of the scale the scatter is considerable.

Although this example is very limited in both scope and method, it provides a warning against taking estimated loadings at their face value for sample sizes as small as 50. The results reported by Seber (1984) and others support this. Much more work is needed to extend and consolidate our limited knowledge in this important area. Results reported in later chapters for categorical variables strongly suggest that much larger samples (say, 500 or more) are needed if parameters are to be estimated with precision. The lack of reproducibility of the results of factor analysis, which has somewhat tarnished its image among practitioners, doubtless owes much to the use of inadequate sample sizes.

A much simpler approach conceptually, which can be used if the sample is large enough, is known as cross-validation. By splitting the sample randomly into two (or more) equal parts and fitting the model to each part, some limited idea can be gained about the stability of the estimates. An illustration for the latent class model, due to Pickering and Forbes (1984), is reported in Chapter 6.

Table 3.2 Frequency distribution of factor 2 loading for variable 4 (300 samples).

Mid-point of interval	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0	−0.1	−0.2	−0.3	−0.4	−0.5	−0.6	−0.7
Frequency	7	43	111	52	40	22	4	2	1	0	0	3	1	6	2	5	1

3.7 Goodness of fit and choice of q

If q is specified *a priori*, the goodness of fit of the factor model can be judged using the likelihood ratio statistic for testing the hypothesis $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$ (H_0) against the alternative that $\mathbf{\Sigma}$ is unconstrained (H_1). The statistic is then

$$\begin{aligned} -2\{L(H_0) - L(H_1)\} &= n\{\log|\hat{\mathbf{\Sigma}}| + \operatorname{trace}\hat{\mathbf{\Sigma}}^{-1}\mathbf{S} - \log|\mathbf{S}| - p\} \\ &= n\{\operatorname{trace}\hat{\mathbf{\Sigma}}^{-1}\mathbf{S} - \log|\hat{\mathbf{\Sigma}}^{-1}\mathbf{S}| - p\} \end{aligned} \tag{3.35}$$

where $\hat{\mathbf{\Sigma}} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}$ is the estimated covariance matrix. If $\mathbf{\Psi} > \mathbf{0}$ this statistic is asymptotically distributed as χ^2 with degrees of freedom

$$\nu = \frac{1}{2}p(p+1) - \left\{pq + p - \frac{1}{2}q(q-1)\right\} = \frac{1}{2}\{(p-q)^2 - (p+q)\}. \tag{3.36}$$

This is the difference between the number of parameters in $\mathbf{\Sigma}$ and the number of free parameters on the null hypothesis. Bartlett (1950) showed that the approximation can be improved by replacing n in (3.35) by $n - 1 - \frac{1}{6}(2p+5) - \frac{2}{3}q$. The behaviour of the test when n is small was investigated by Geweke and Singleton (1980), whose results suggest that it is adequate for n as low as 30 with one or two factors. However, Schönemann (1981) argues that this is too optimistic because they used untypical examples with high communalities.

Since q is not usually specified in advance the test is often made the basis of a procedure for choosing the best value. Starting with $q = 1$, we then take successive values in turn until the fit of the model is judged to be adequate. Viewed as a testing procedure this is not strictly valid because it does not adjust the significance levels to allow for the sequential character of the test. It rather depends on regarding the p-value of the test as a measure of the adequacy of the model.

3.7.1 Model selection criteria

The trouble with a procedure of this kind is that the larger we make q the better the fit, but it provides us with no criterion for judging when to stop. An alternative approach is provided by Akaike's *information criterion* for model selection. The situation is that we have a set of linear models indexed by q and a selection has to be made. Akaike (1983) proposed that method for use in factor analysis and showed that it required q to be chosen to make

$$\text{AIC} = -2L + 2\nu$$

a minimum, where ν is the number of free parameters in the model, and L is the maximised value of the log-likelihood function for the estimated model. Note that L and ν are both functions of q. The criterion can be justified in a variety of ways but, in essence, it effects a trade-off between the bias introduced by fitting the wrong number of factors and the precision with which the parameters are estimated—as q

is increased the bias decreases but the error increases. An alternative version, due to Schwarz (1978), also known as the *Bayesian information criterion* (BIC), replaces the term 2ν by $\nu \ln n$. A further criterion based on the residuals of the fitted correlation matrix is proposed by Bozdogan and Ramirez (1986), who also report a comparison of all three criteria using Monte Carlo methods applied to models used in the study of Francis (1974). All methods perform reasonably well, but for the examples considered no method is uniformly best. These model selection ideas show considerable promise and their performance over a wide range of models merits further investigation.

Other criteria for selecting q which do not involve distributional assumptions will be mentioned in Section 3.11.

3.8 Fitting without normality assumptions: least squares methods

If nothing is assumed about the distributions of $\mathbf{y}$ and $\mathbf{e}$, it remains true that the covariance matrix predicted by the model is given by $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. We could then aim to estimate $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ in such a way that $\boldsymbol{\Sigma}$ was as close to $\mathbf{S}$, the sample covariance matrix, as is possible. For this we need some scalar measure of distance between $\boldsymbol{\Sigma}$ and $\mathbf{S}$ which must then be minimised with respect to the parameters. The distance function

$$\Delta(\boldsymbol{\Sigma}, \mathbf{S}) = -\text{trace}\, \boldsymbol{\Sigma}^{-1}\mathbf{S} + \log\, |\, \boldsymbol{\Sigma}^{-1}\mathbf{S}\,| \tag{3.37}$$

which arose in the course of maximum likelihood estimation is one possibility. It will only yield maximum likelihood estimators if the normal assumptions hold, but if (3.37) were regarded as a reasonable way of measuring distance the estimators would be justified on a much broader basis.

There are many other possible measures of distance which could be used in place of (3.37). It is natural to turn to least squares ideas. A simple unweighted least squares criterion would be

$$\Delta_1 = \sum_{i=1}^{p}\sum_{u=1}^{p}(s_{iu} - \sigma_{iu})^2 = \text{trace}(\mathbf{S} - \boldsymbol{\Sigma})^2. \tag{3.38}$$

Another, suggested by the role played by the matrix $\boldsymbol{\Psi}^{-1/2}\mathbf{S}\boldsymbol{\Psi}^{-1/2}$ in the Lawley and Maxwell approach to maximum likelihood estimation, is

$$\Delta_2 = \text{trace}\{[\boldsymbol{\Psi}^{-1/2}(\mathbf{S} - \boldsymbol{\Sigma})\boldsymbol{\Psi}^{-1/2}]^2\} = \text{trace}\{(\mathbf{S} - \boldsymbol{\Sigma})\boldsymbol{\Psi}^{-1}\}^2. \tag{3.39}$$

These are both special cases of a general class of measures stemming from fundamental work by Browne (1982, 1984) (see also Tanaka and Huba (1985)) which may be written

$$\Delta = \text{trace}\{(\mathbf{S} - \boldsymbol{\Sigma})\mathbf{V}\}^2. \tag{3.40}$$

These arise from a generalised least squares approach which allows the deviations $\{s_{ij} - \sigma_{ij}\}$ to be weighted in various ways. We shall not develop these ideas here, but reference to the authors mentioned above will show that these methods lead to robust methods of estimation under a range of distributional assumptions. The case $\mathbf{V} = \mathbf{S}^{-1}$ was investigated by Jöreskog and Goldberger (1972); see also Anderson (1984, Section 14.3.4). The attraction of Δ_1 and Δ_2 is that, like maximum likelihood, the optimisation requires the solution of an eigenvalue problem. In the case of Δ_1 the function to be minimised may be written

$$\Delta_1 = \sum_{i=1}^{p}\sum_{u=1}^{p}\left(s_{iu} - \delta_{iu}\psi_i - \sum_{j=1}^{q}\lambda_{ij}\lambda_{uj}\right)^2.$$

Differentiating with respect to λ_{rs},

$$\frac{\partial \Delta_1}{\partial \lambda_{rs}} = 4\left\{-\sum_{i=1}^{p}(s_{ri} - \delta_{ir}\psi_i)\lambda_{is} + \sum_{i=1}^{p}\lambda_{is}\sum_{j=1}^{q}\lambda_{ij}\lambda_{rj}\right\}$$
$$(r = 1, 2, \ldots, p;\ s = 1, 2, \ldots, q) \tag{3.41}$$

or

$$\frac{\partial \Delta_1}{\partial \mathbf{\Lambda}} = 4\{\mathbf{\Lambda}(\mathbf{\Lambda}'\mathbf{\Lambda}) - (\mathbf{S} - \mathbf{\Psi})\mathbf{\Lambda}\},$$

which gives the estimating equations

$$(\mathbf{S} - \mathbf{\Psi})\mathbf{\Lambda} = \mathbf{\Lambda}(\mathbf{\Lambda}'\mathbf{\Lambda}). \tag{3.42}$$

Differentiating with respect to ψ_r,

$$\frac{\partial \Delta_1}{\partial \psi_r} = -2\left(s_{rr} - \psi_r - \sum_{j=1}^{q}\lambda_{rj}^2\right)$$

or

$$\operatorname{diag}\frac{\partial \Delta_1}{\partial \mathbf{\Psi}} = -\operatorname{diag}\mathbf{S} + \mathbf{\Psi} + \operatorname{diag}\mathbf{\Lambda}\mathbf{\Lambda}',$$

leading to

$$\mathbf{\Psi} = \operatorname{diag}(\mathbf{S} - \mathbf{\Lambda}\mathbf{\Lambda}'). \tag{3.43}$$

These estimating equations are similar to those obtained for maximum likelihood estimation in (3.14) and (3.17) and are solved in a similar manner. Suppose first that $\boldsymbol{\Psi}$ is known and that $\mathbf{S} - \boldsymbol{\Psi}$ is positive definite. Then (3.42) will be satisfied if:

1. the columns of $\boldsymbol{\Lambda}$ consist of any q eigenvectors of $\mathbf{S} - \boldsymbol{\Psi}$;
2. $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}$ is a diagonal matrix with elements equal to the eigenvalues of $\mathbf{S} - \boldsymbol{\Psi}$ associated with the vectors in $\boldsymbol{\Lambda}$.

Thus if we have a starting value for $\boldsymbol{\Psi}$, the solution of (3.42) will yield a first approximation to $\boldsymbol{\Lambda}$ which can then be inserted in (3.43) to give a second estimate of $\boldsymbol{\Psi}$, and then the cycle can be continued until convergence occurs. The question of which eigenvectors of $\mathbf{S} - \boldsymbol{\Psi}$ are to be included in $\boldsymbol{\Lambda}$ can be answered as follows:

$$\begin{aligned}\Delta_1 &= \text{trace}(\mathbf{S} - \boldsymbol{\Psi} - \boldsymbol{\Lambda}\boldsymbol{\Lambda}')^2 \\ &= \text{trace}(\mathbf{S} - \boldsymbol{\Psi})^2 - 2\text{trace}(\mathbf{S} - \boldsymbol{\Psi})\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \text{trace}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}')^2 \\ &= \text{trace}(\mathbf{S} - \boldsymbol{\Psi})^2 - \text{trace}(\boldsymbol{\Lambda}\boldsymbol{\Lambda}')^2,\end{aligned}$$

using (3.42).

Now $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ has $(p - q)$ zero eigenvalues because it is of rank q. The others are also eigenvalues of $\mathbf{S} - \boldsymbol{\Psi}$ since if we replace $\mathbf{S} - \boldsymbol{\Psi}$ in (3.42) by $\boldsymbol{\Lambda}\boldsymbol{\Lambda}'$ the equation is obviously satisfied. Let the eigenvalues which the two matrices have in common be $\theta_1, \theta_2, \ldots, \theta_q$ and the remaining eigenvalues of $\mathbf{S} - \boldsymbol{\Psi}$ be $\theta_{q+1}, \ldots, \theta_p$; then

$$\Delta_1 = \sum_{i=1}^{p} \theta_i^2 - \sum_{i=1}^{q} \theta_i^2 = \sum_{i=q+1}^{p} \theta_i^2. \tag{3.44}$$

For this to be a minimum, $\theta_{q+1}, \ldots, \theta_p$ must be the smallest eigenvalues and hence $\boldsymbol{\Lambda}$ must consist of the eigenvectors associated with the q largest eigenvalues.

In this method $\boldsymbol{\Psi}$ is chosen so that $\mathbf{S} - \boldsymbol{\Psi}$ is positive definite and so that the sum of squares of the $p - q$ smallest eigenvalues of $\mathbf{S} - \boldsymbol{\Psi}$ is as small as possible.

3.9 Other methods of fitting

The foregoing method is known as the *principal factor* (or *principal axis*) method because of its similarity to principal components analysis to which it is equivalent if $\boldsymbol{\Psi} = \mathbf{0}$. If we use the correlation matrix $\mathbf{R}$ instead of $\mathbf{S}$ the estimates obtained for the scale-invariant parameters $\boldsymbol{\Lambda}^*$ and $\boldsymbol{\Psi}^*$ defined in Section 3.12.2 will not be identical to those arrived at by first using $\mathbf{S}$ and then transforming as they were with the maximum likelihood method.

Estimation for Δ_2 using (3.39) proceeds in an exactly similar manner. In fact, since

$$\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{-1/2} = \boldsymbol{\Lambda}^*\boldsymbol{\Lambda}^{*\prime} + \mathbf{I} = (\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Lambda})(\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Lambda})' + \mathbf{I},$$

all we have to do is to replace $\mathbf{\Lambda}$ by $\mathbf{\Lambda}^* = \mathbf{\Psi}^{-1/2}\mathbf{\Lambda}$ in (3.42) and $\mathbf{S} - \mathbf{\Psi}$ by $\mathbf{S}^* - \mathbf{I}$. The estimating equation for $\mathbf{\Lambda}^*$ is then identical to that for maximum likelihood. Rather surprisingly, it turns out that the differing distance functions both lead to the eigenvalues and vectors of $\mathbf{S}^*$. However, this is true only for fixed $\mathbf{\Psi}$. If we bring $\mathbf{\Psi}$ into the picture its partial derivatives will be different from those in the maximum likelihood case and, in fact, a good deal more complicated. With this method the choice of $\mathbf{\Psi}$ is made so that $\mathbf{\Psi}^{-1/2}\mathbf{S}\mathbf{\Psi}^{-1/2}$ has eigenvalues all no less than 1 and so that the sum of the squared differences from 1 of the $p - q$ smallest eigenvalues is as small as possible.

For Δ given by (3.40) with $\mathbf{V} = \mathbf{S}^{-1}$ the same equation for $\mathbf{\Lambda}$ for fixed $\mathbf{\Psi}$ is obtained, but the (implicit) estimating equation for $\mathbf{\Psi}$ is

$$\operatorname{diag} \mathbf{S}^{-1}\{(\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}) - \mathbf{S}\}\mathbf{S}^{-1} = \operatorname{diag} \mathbf{0};$$

see Anderson (1984). Here the sum of the squared differences from 1 of the reciprocals of the $p - q$ smallest eigenvalues of $\mathbf{\Psi}^{-1/2}\mathbf{S}\mathbf{\Psi}^{-1/2}$ is minimised.

We have noted that if $\mathbf{\Psi}$ were known, these methods would be much simpler, being straightforward eigenproblems. Since $\mathbf{\Psi}$ only enters into the diagonal elements of $\mathbf{\Sigma}$, there is a prospect of avoiding the difficulties by eliminating these terms from Δ_1. We would then minimise

$$\Delta_1' = \sum_{i=1,\ i\neq u}^{p} \sum_{u=1}^{p} \left(s_{iu} - \sum_{j=1}^{q} \lambda_{ij}\lambda_{uj} \right)^2 . \tag{3.45}$$

This approach is described in Harman and Jones (1966) and is usually known as the 'minres' method. Various methods of obtaining estimates have been given by Comrey (1962), Comrey and Ahumada (1964), Okamoto and Ihara (1983) and Zegers and ten Berge (1983). Unfortunately the omission of the diagonal terms destroys the structure which led to the easily solved eigenequations. Given that the minres method is unweighted and that the iterative methods for the Δ-family are well within the scope of modern computers, the method offers few advantages and is not included in modern computer packages.

There are yet other methods of estimating the normal linear factor model. Knott (2005) has a contribution to, and references for, the *minimum trace* method.

3.10 Approximate methods for estimating $\mathbf{\Psi}$

The iterative methods require a starting value for $\mathbf{\Psi}$, and although one can use an arbitrary value such as $\mathbf{\Psi} = \operatorname{diag} \mathbf{S}$ there are approximations which will reduce the number of iterations required. Apart from their practical value they give some insight into the interpretation of the analysis. We have a non-negative definite matrix

$$\mathbf{\Sigma} - \mathbf{\Psi} \tag{3.46}$$

and so, assuming now that $\boldsymbol{\Psi}$ is invertible,

$$\boldsymbol{\Psi}^{-1} - \boldsymbol{\Sigma}^{-1} \tag{3.47}$$

is non-negative definite, (see, for instance, Rao (1973, p. 70, Exercise 9)), from which it follows that

$$\sigma^{ii} \leq \psi_r^{-1}. \tag{3.48}$$

Since σ^{ii} can be estimated from the inverse of $\mathbf{S}$, we can estimate an upper bound for ψ_i and this may be used as a starting value for the iteration.

This result may be given another interpretation: s^{ii}, which we would use as an approximation, may be expressed using standard regression results as

$$s^{ii} = s_{ii}(1 - R_i^2), \tag{3.49}$$

where R_i^2 is the multiple correlation coefficient of x_i regressed on the remaining xs. Now

$$s_{ii}(1 - R_i^2) = \text{var}(x_i \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$$

and

$$\psi_i = \text{var}(x_i \mid \mathbf{y}).$$

We would expect the latter to be smaller than the former because $\mathbf{y}$ is not precisely determined by any finite set of xs (see (3.6)). A simpler but less precise bound for ψ follows from the fact that $R_i^2 \geq \max_{j \neq i} r_{ij}^2$, where r_{ij} is the correlation between x_i and x_j. Hence an estimated upper bound for ψ_i is $s_{ii}(1 - \max r_{ij}^2)$.

The approximation $\hat{\boldsymbol{\Psi}} \approx \left(\text{diag}\, \mathbf{S}^{-1}\right)^{-1}$ derived from (3.48) is the basis of *image factor analysis*. This was proposed by Guttman (1953); see also Jöreskog (1969). Like other methods of fitting the NLFM which seem to have fallen into disuse, it offered a relatively simple method of fitting at a time when computing facilities were very modest. Treating $\boldsymbol{\Psi}$ as if it were known and proportional to $(\text{diag}\, \mathbf{S}^{-1})^{-1}$ means that the fitting can be carried out by the principal axis method without iteration. An iterative version can be used in which $\boldsymbol{\Psi}$ is updated using the latest values of $\hat{\boldsymbol{\Psi}}$, but Basilevsky (1994, Section 5.33) has shown that this leads to inconsistent estimators. With modern computing facilities there is no longer any need for such methods.

3.11 Goodness of fit and choice of q for least squares methods

Little appears to be known about the sampling behaviour of the methods of fitting discussed in Sections 3.8–3.10, but there is an important result due to Amemiya and Anderson (1985) which shows that the limiting χ^2 distribution of the goodness-of-fit

statistic of (3.40) is valid under very general circumstances. They show that if the elements of **e** are independent and if **y** and **e** have finite second moments then (3.35) calculated using the maximum likelihood estimators has the same distribution as in the normal case. This result also holds for another goodness-of-fit statistic which is sometimes used, namely

$$\frac{1}{2}n \text{ trace}\{(\mathbf{S} - \hat{\boldsymbol{\Sigma}})\hat{\boldsymbol{\Sigma}}^{-1}\}^2$$

which, in the normal case, is asymptotically equivalent to (3.35). This is a further reason for using the maximum likelihood method of fitting, whether or not one makes the normality assumptions.

One consequence of these results is that we can use the same methods for choosing q as were proposed for the maximum likelihood method in Section 3.7. There are also two other methods which do not depend on distributional assumptions. Both are based on the eigenvalues of the sample correlation matrix and the role which they have in principal components analysis. The Kaiser–Guttman criterion chooses q equal to the number of eigenvalues greater than unity. The rationale is that the average contribution of a manifest variable to the total variation is 1, and that a principal component which did not contribute at least as much variation as a single variable represents no advantage. The carry-over of this argument from principal components to factor analysis rests on the similarity between the two techniques noted in Chapters 1 and 9. Simulation results obtained by Fachel (1986) suggest that if p is large this method is likely to overestimate q.

The second method, due to Cattell, is known as the 'scree test'. If the eigenvalues are plotted against their rank order they will lie on a decreasing curve. One then looks for an 'elbow' in the curve, as this would indicate the point at which the further addition of factors shows diminishing returns in terms of variation explained. A simulation study by Hakistian *et al.* (1982) comparing these two methods with the use of the likelihood ratio statistic does not lead to clear-cut conclusions.

3.12 Further estimation issues

3.12.1 Consistency

In the discussion of estimation methods we have tacitly assumed that the parameters could be consistently estimated. A necessary condition for this to be possible is that there shall be at least as many sample statistics as there are parameters to be estimated. The number of parameters in a q-factor model is $pq + p$, but in order to obtain a unique solution we have to impose $\frac{1}{2}q(q-1)$ constraints (usually by requiring the off-diagonal elements of $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ to be zero). The number of free parameters is then

$$pq + p - \frac{1}{2}q(q-1).$$

The sample covariance matrix $\mathbf{S}$ has $\frac{1}{2}p(p+1)$ distinct elements, so for consistent estimation to be possible we must have

$$\frac{1}{2}p(p+1) - pq - p + \frac{1}{2}q(q-1) = \frac{1}{2}[(p-q)^2 - (p+q)] \geq 0. \tag{3.50}$$

Equation (3.50) implies that there is an upper bound to the number of factors which can be fitted which is given by

$$q \leq \frac{1}{2}\{2p + 1 - (8p+1)^{1/2}\}. \tag{3.51}$$

The condition in (3.50) is not sufficient because it does not guarantee that the estimates of the ψ_i will be non-negative. Building on earlier work by Anderson and Rubin (1956), Kano (1983, 1986a,b) has provided general conditions under which maximum likelihood and generalised least squares estimators are consistent. However, in Kano (1986a) he provides an example of a model which does not admit any consistent estimator. The question of consistency when p is not fixed is investigated in Kano (1986b).

3.12.2 Scale-invariant estimation

In practice the factor model is almost always fitted using the sample correlation matrix rather than the covariance matrix. The reason usually given for this is that the scaling adopted for the manifest variables is often arbitrary, especially in social science applications. Using correlations rather than covariances amounts to standardising the xs using their sample standard deviations, and this ensures that changing the scale of the xs has no effect on the analysis. However, the joint distribution of the correlation coefficients is not the same as that of the covariances and so it is not obvious that the estimators obtained in the way described above will be true maximum likelihood estimators or that the asymptotic goodness-of-fit test will be valid. A good deal of confusion has surrounded this topic so we shall approach the matter from first principles.

We start from the proposition that if the units of measurement of the xs are arbitrary then there can be no scientific interest in any aspect of the model which depends on the choice of units. The effect of a scale change on the model of (3.3) can be seen by transforming to $\mathbf{x}^* = \mathbf{C}\mathbf{x}$, where $\mathbf{C}$ is a diagonal matrix with positive elements. Expressed in terms of $\mathbf{x}^*$, the model becomes

$$\mathbf{x}^* = \mathbf{C}\boldsymbol{\mu} + \mathbf{C}\boldsymbol{\Lambda}\mathbf{y} + \mathbf{C}\mathbf{e}, \tag{3.52}$$

with

$$\text{var}(\mathbf{x}^*) = \mathbf{C}\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\mathbf{C} + \mathbf{C}\boldsymbol{\Psi}\mathbf{C}.$$

In general, the parameters $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ do not therefore meet our requirements since their estimated values and the interpretation to be put upon them will depend on the scales of the xs. Suppose, however, that we choose $\mathbf{C} = (\text{diag}\,\mathbf{\Sigma})^{-1/2}$ so that each x is divided by its theoretical standard deviation; then we may write (3.52) as

$$\mathbf{x}^* = \boldsymbol{\mu}^* + \mathbf{\Lambda}^*\mathbf{y} + \mathbf{e}^*, \tag{3.53}$$

with $\text{var}(\mathbf{x}^*) = \mathbf{\Lambda}^*\mathbf{\Lambda}^{*\prime} + \mathbf{\Psi}^*$, where $\mathbf{\Lambda}^* = \mathbf{C}\mathbf{\Lambda}$ and $\mathbf{\Psi}^* = \mathbf{C}\mathbf{\Psi}\mathbf{C}$.

Clearly, any change in the scale of $\mathbf{x}$ now has no effect on the value of $\mathbf{x}^*$ and hence the parameter estimates obtained for $\mathbf{\Lambda}^*$ and $\mathbf{\Psi}^*$ will be independent of the scaling of $\mathbf{x}$. These parameters will be said to be *scale-invariant* and it is on them that our interest will be centred. We shall also refer to $\mathbf{\Lambda}^*$ and $\mathbf{\Psi}^*$ as the parameters of the model in *standard form*.

The important difference between this approach and that starting with the sample correlations is that in the latter case the xs would be standardised using the *sample* standard deviations. Their distribution would not then be normal, and that would be inconsistent with the assumptions we have made about the form and distribution of the right-hand side of (3.3).

To estimate $\mathbf{\Lambda}^*$ and $\mathbf{\Psi}^*$ one could first estimate $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ for any arbitrary scaling of the xs using the sample covariance matrix and then transform them by

$$\hat{\mathbf{\Lambda}}^* = (\text{diag}\,\hat{\mathbf{\Sigma}})^{-1/2}\hat{\mathbf{\Lambda}} \quad \text{and} \quad \hat{\mathbf{\Psi}}^* = (\text{diag}\,\hat{\mathbf{\Sigma}})^{-1/2}\,\hat{\Psi}(\text{diag}\,\hat{\mathbf{\Sigma}})^{-1/2}.$$

However, it turns out that the usual practice of treating the correlation matrix as if it were the covariance matrix does yield the maximum likelihood estimators of the scale-invariant parameters. This was shown by Krane and McDonald (1978) and an outline of the justification is as follows.

We reparameterise the likelihood by writing $\mathbf{T} = \mathbf{C}^{-1/2}\mathbf{\Sigma}\mathbf{C}^{-1/2}$, where $\mathbf{C} = \text{diag}\,\mathbf{\Sigma}$. From (3.11) we then have

$$\begin{aligned} -2L/n &= \text{constant} + \log|\,\mathbf{C}^{1/2}\mathbf{T}\mathbf{C}^{1/2}\,| + \text{trace}(\mathbf{C}^{1/2}\mathbf{T}\mathbf{C}^{1/2})^{-1}\mathbf{S} \\ &= \text{constant} + \log|\,\mathbf{C}\,| + \log|\,\mathbf{T}\,| + \text{trace}\,\mathbf{C}^{-1/2}\mathbf{T}^{-1}\mathbf{C}^{-1/2}\mathbf{S} \\ &= \text{constant} + \log|\,\mathbf{C}\,| + \log|\,\mathbf{T}\,| + \text{trace}\,\mathbf{T}^{-1}\mathbf{C}^{-1/2}\mathbf{S}\mathbf{C}^{-1/2}. \end{aligned} \tag{3.54}$$

The matrix $\mathbf{T}$ involves only the parameters $\mathbf{\Lambda}^*$, since its diagonal elements are now unity. The expression in (3.54) may be maximised in two stages, first with respect to $\mathbf{C}$ and then with respect to $\mathbf{T}$. Krane and McDonald showed, as one might have anticipated, that $\mathbf{C} = \text{diag}\,\mathbf{S}$, so at the second stage the quantity to be maximised is

$$\log|\,\mathbf{T}\,| + \text{trace}\,\mathbf{T}^{-1}\mathbf{R},$$

where $\mathbf{R} = (\text{diag}\,\mathbf{S})^{-1/2}\mathbf{S}(\text{diag}\,\mathbf{S})^{-1/2}$ which is exactly what we would have done if we had treated the sample correlation matrix as a covariance matrix. It is easily verified

that the log-likelihood ratio statistic is the same whichever parameterisation is used, so the procedure for estimation and testing goodness of fit is fully justified.

3.12.3 Heywood cases

All the methods of estimating the parameters of the factor model involve minimising a measure of the distance between $\mathbf{S}$ and $\boldsymbol{\Sigma}$. However, the parameter space is restricted by the condition $\boldsymbol{\Psi} \geq \mathbf{0}$ so the usual procedure of setting all the partial derivatives equal to zero may yield a solution with a negative ψ. In such cases the minimum we seek will lie on a boundary of the admissible region at a point where one or more of the ψs is zero. When this happens we have what is called a Heywood case, after Heywood (1931). In practice this will be recognised either by the appearance of a negative estimate or by the convergence of an estimate to zero, depending on the algorithm used.

There is no inconsistency in the occurrence of a zero residual variance and, taken at its face value, it would simply mean that the variation of the manifest variable in question was wholly explained by the latent variables. In practice this rarely seems plausible and the rather frequent occurrence of Heywood cases has caused a good deal of unease among practitioners. Taken with the fact that it is known empirically that zero estimates can easily arise when the true parameter values are not zero, there is good reason for not taking such estimates at their face value.

There is a good deal of evidence, mainly of an empirical kind, about the circumstances under which Heywood cases are likely to occur. Much of this is based on simulation studies where there is no question of the model itself being invalid. The results are widely scattered in the literature but we have drawn heavily in what follows on van Driel (1978), Anderson and Gerbing (1984), Boomsma (1985) and Fachel (1986).

If a Heywood case arises when the data conform to the linear factor model it will probably be the result of sampling error. A key factor is therefore sample size. For a model with positive ψs the probability of a Heywood case tends to zero as n tends to infinity. We are therefore dealing with a small-sample phenomenon. The risk of a Heywood case depends on other factors mentioned below, but as a rough guide it appears that it is high with sample sizes of 100 or less and low with samples of 500 or more.

For a given sample size the risk decreases as p, the number of variables, increases. Obviously the risk will also be greater if one or more ψs is very small, but this is not something which would be known in advance. Nevertheless, one can obtain some clues from the data. For example, in Section 3.10 and the remarks that followed, we saw that an estimated upper bound for ψ could be obtained and, in particular, that ψ_i would be small if x_i was highly correlated with any other variable. The presence of one or more high correlations is therefore indicative of a potential Heywood case.

One of the commonest cause of Heywood cases is the attempt to extract more factors than are present. This is readily demonstrated by simulation but might have been anticipated on the grounds that artificially inflating the communality forces the residuals towards zero.

Once the causes of Heywood cases are understood, the ways of dealing with them become clearer. At the stage of designing an enquiry one should aim for a large sample with a good number of variables. But in selecting variables it is important to avoid introducing new variables which add little to those already there. This will merely create high correlations without contributing significantly to the information about the latent variables.

At the analysis stage the options are more limited. Over-factoring can be avoided by paying careful attention to the various criteria suggested in Section 3.7. If a highly correlated pair of variables appears to be the cause, one of them can be dropped with little loss. However, it does not follow that dropping a variable which is implicated in a Heywood case is always advisable. If it arises because of a single small residual variance we should then be omitting one of the more valuable variables which was a relatively pure indicator of the latent variables. In any case, experience shows that such a course often leads to a Heywood case involving another variable, which defeats the object.

If the foregoing precautions fail there are at least three courses still open:

1. We can use a Bayesian approach like that suggested by Martin and McDonald (1975). Here we maximise not the likelihood but the posterior density, using a prior distribution for the ψs which assigns zero probability to negative values. Martin and McDonald propose a form which is both tractable and plausible in that it implies a distribution which is almost uniform except that it decreases to zero at the point $\psi_i = 0$. Lee (1981) also investigated the form of the posterior density under different informative prior distributions, some of which have been designed to deal with Heywood cases. Furthermore, a simulation-based Bayesian estimation approach discussed in Sections 2.10 and 4.11 has been also used by Shi and Lee (1998) for the factor analysis model for continuous variables that can also handle categorical variables.

2. We may stop the iteration at some arbitrarily small value of ψ_i such as 0.05 or 0.01. In effect, this is a special case of course of action 1 with a uniform prior on the interval $(\epsilon,\ 1)$ where ϵ is the chosen cut-off point. It may be justified on the grounds that the likelihood (or other distance measure) will be very close to its optimum and all that matters for purposes of interpretation is to know that ψ_i is 'small'.

3. A third method which shows promise but needs further investigation rests on the following simple idea. If a variable x_i has a small ψ_i we could increase the latter by adding to x_i an independent random variable with known variance σ^2. Denoting the new variable by x_i', its variance would then be

$$\text{var}(x_i') = \sum_{j=1}^{q} \lambda_{ij}^2 + \psi_i + \sigma^2,$$

but the covariances would be unchanged.

If x_i' were used instead of x_i in estimating the parameters, we would obtain an estimate of $\psi_i + \sigma^2$. Knowing σ^2, we could then obtain an estimate of ψ_i

by subtraction. If this still led to a negative ψ_i the procedure would need to be repeated with a larger value of σ^2. It is not, in fact, necessary to add the artificial variable to each value of x_i; we could simply add σ^2 to the appropriate diagonal element of $\mathbf{S}$.

If the analysis is carried out on the correlation matrix the effect of replacing x_i by x_i' is to multiply the off-diagonal elements in the ith row of the sample correlation matrix by $(1+\sigma^2)^{-1/2}$. The relationship between the parameters of the original model and the modified one is then given by

$$\lambda_{ij}^* = \lambda_{ij}'(1+\sigma^2)^{1/2}, \quad \psi_i^* = 1-(1+\sigma^2)\sum_{j=1}^{q}\lambda_{ij}'^2, \tag{3.55}$$

where the asterisk denotes the modified loading. The result holds for any value of σ^2 but it seems reasonable to choose a value just large enough to avoid the occurrence of a Heywood case. In practice $\sigma^2 = 1$ seems to be the right order of magnitude. It should be noted that the results on the standard errors of the estimates and on goodness of fit will be invalidated by this device.

3.13 Rotation and related matters

In Section 2.11 we introduced the idea of orthogonal and oblique rotations in the factor space. These transformations have the property that they leave the joint distribution of the manifest variables unchanged and so the fit of the model is unaffected. A choice of rotation thus has to be made on non-statistical grounds. We saw that patterns of loadings exhibiting simple structure were relatively easy to interpret. The rotations available in the major software packages for the NLFM all embody some algorithm for getting as close as possible to simple structure.

If only two factors have been fitted the position will usually be clear from a plot of the loadings. In the space of the factor loadings one is looking to position the axes so that the points lie close to one or other axis. This is often all that is necessary, and we shall illustrate the method on the examples in Section 3.17.

3.13.1 Orthogonal rotation

One characteristic of simple structure is that the loadings on any factor are either zero or large. Since the squares of standardised loadings lie between 0 and 1, an algorithm which seeks to maximise their variance will move them towards opposite ends of the range. This is the varimax approach which was proposed by Kaiser (1958). In an alternative version λ_{ij}^2 is replaced by $\lambda_{ij}^2/(1-\Psi_i)$ $(i = 1, 2, \ldots, p)$ in the criterion for maximisation. The maximisation is achieved iteratively, and details are given in Magnus and Neudecker (1988).

Another way of describing simple structure is in terms of the variance of the squares of the factor loadings across the rows of $\mathbf{\Lambda}$ rather than down the columns. This idea is implemented in the quartimax method. It would, of course, be possible to combine the characteristics of varimax and quartimax in a single criterion.

3.13.2 Oblique rotation

An oblique rotation offers a better chance of finding simple structure, but at the price of complicating the interpretation. In practical terms there is no obvious reason why factors of substantive interest should be uncorrelated as orthogonality implies. For example, if one is seeking to identify two factors called verbal ability and numerical ability in an educational test, there is every reason to expect them to be correlated and, if so, they will not be uncovered by a search restricted to orthogonal factors. Algorithms similar to those in the orthogonal case, but without the constraint of orthogonality, are available. Details of the options available may be found in the specialist literature, including Tabachnick and Fidell (1996) and McDonald (1985). We shall use the OBLIMIN routine provided by Mplus (Muthén and Muthén 2010).

When the factors are uncorrelated we saw that the standardised factor loadings could be interpreted as the correlation coefficients of the manifest variables and the factors. Under oblique rotation this is no longer true. The calculation in (a) of Section 3.2 now gives

$$E(\mathbf{x} - \boldsymbol{\mu})\mathbf{y}' = \boldsymbol{\Lambda}\boldsymbol{\Phi}, \tag{3.56}$$

where $\boldsymbol{\Phi}$ is the correlation matrix of the ys. The matrix of correlations is sometimes called the *structure loading* matrix to distinguish it from the *pattern loading* matrix $\boldsymbol{\Lambda}$.

3.13.3 Related matters

There are other methods of fitting the factor model which are best regarded as ways of selecting a particular rotation. One of these is given by Rao (1955). This approach views the problem in the context of canonical correlation, where the object is to find linear combinations of one group of variables which are maximally correlated with linear combinations of a second group. In factor analysis the xs form one group and the ys the other. Using the methods of canonical correlation, we find that linear combination of the xs and of the ys which are most highly correlated. Then we find two further linear combinations, uncorrelated with the corresponding member of the first pair, and so on. This procedure leads to estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ which satisfy the maximum likelihood equations and also the constraint that $\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ be diagonal. In effect, therefore, it is selecting the rotation with this property and, in the process, providing another reason for choosing that particular solution.

Alpha factor analysis can be regarded in the same light. This was proposed by Kaiser and Caffrey (1965) and based on the psychometric notion of generalisability which we considered in Section 2.17. It is based on the eigendecomposition of the correlation matrix of the common parts of the manifest variable. Thus for the model

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{y} + \mathbf{e}, \tag{3.57}$$

$\mathbf{c} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{y}$ is the vector of the common parts of $\mathbf{x}$. Its correlation matrix is

$$\mathbf{R}^* = (\mathbf{I} - \boldsymbol{\Psi})^{-1/2}(\mathbf{R} - \boldsymbol{\Psi})(\mathbf{I} - \boldsymbol{\Psi})^{-1/2}, \tag{3.58}$$

where $\mathbf{R}$ is the correlation matrix of the xs.

The standard least squares fit of principal axis factoring then yields

$$\hat{\mathbf{\Lambda}} = (\mathbf{I} - \mathbf{\Psi})^{1/2}\mathbf{M}\mathbf{\Theta}, \tag{3.59}$$

where $\mathbf{\Theta}^2$ is the diagonal matrix containing the q largest eigenvalues of $\hat{\mathbf{R}}^*$ and $\mathbf{M}$ the matrix of associated eigenvectors. The unknown $\mathbf{\Psi}$ can be estimated by an iterative procedure.

The solution is thus the rotation which maximises the generalisability as measured by the correlations of the common parts. Given the questionable nature of the psychometric measure of generalisability, other methods of selecting rotations are to be preferred. An alternative version of alpha factor analysis was proposed by Bentler (1968) which turned out to be equivalent to Rao's method.

The software packages CEFA (Browne *et al.* 1998) and Mplus (Muthén and Muthén 2010) allow a choice between very many different sorts of rotations and standardisations.

3.14 Posterior analysis: the normal case

Having determined the dimension of the factor space and established that the model is a reasonable fit, we may wish to assign scale values to individuals in the latent dimensions. We may need these for selection purposes as, for example, when we wish to choose those with the highest ability. Or we may wish to use scale values as an input to some further analysis to see how ability is related to some other variable. It is sometimes argued that this latter need can best be met within the framework of a structural equation model (see Chapter 8) where the role of the latent variables is implicit and does not need to be made explicit. In Chapter 8 we shall advance reasons for treating this recommendation with great caution. For the present we simply note that this does not eliminate the need for factor scores.

Within the framework we have adopted and set out in Chapter 2, the so-called *factor scores problem* is solved by the posterior distribution of $\mathbf{y}$ given $\mathbf{x}$. Scale values for the dimensions of $\mathbf{y}$ can be obtained by using some measure of location of the posterior distribution such as $E(\mathbf{y} \mid \mathbf{x})$. The precision of any predicted variable can be measured by the posterior variance. This procedure was set out in general terms in Chapter 2, where it was noted that the posterior expectation was close to a linear function of the sufficient statistic. We have already given the posterior distribution for the normal case in (3.6) (and also in (1.13)) which shows that the relationship for the normal linear latent variable model is precisely linear. Whether we use the expectation or the component is therefore of no significance since one is a linear function of the other.

The posterior covariance matrix of $\mathbf{y} \mid \mathbf{x}$ was given in (3.6) in the form $(\mathbf{I} + \mathbf{\Gamma})^{-1}$. It is much easier to interpret the posterior analysis when $\mathbf{\Gamma}$ is diagonal because this makes the ys independent.

The traditional way of approaching the factor scores problem has been to start with (3.3). This is an equation in *random* variables, but realised values of these

random variables will satisfy the same equation. However, the only realised values we are able to observe are the *x*s. This leaves the *y*s and the *e*s undetermined. The set of equations cannot, therefore, be solved for the *y*s alone. For this reason the factor scores are said to be indeterminate and various devices have been adopted to get best-fitting values, in some sense, for **y**.

The case for adopting the posterior approach was argued in Bartholomew (1981) but had been anticipated by Dolby (1976). The 'indeterminacy' is reflected in the fact that **y** remains a random variable after **x** has been observed though its variance will have been reduced by the knowledge of **x**. The choice between the posterior and the traditional approach remains a matter of lively debate, as reference to Maraun (1996) and the ensuing discussion shows. In purely practical terms there is little at stake and we shall confine our treatment to the posterior method. (It may be noted that it is only in relation to the linear factor model that this controversy arises. With other latent variable models, in particular the latent class model, the posterior approach has been used without question.)

The greatest apparent drawback of the posterior method is that the scale values depend on the choice of prior distribution. However, this is an inevitable consequence of the essential indeterminacy of the prior. Theorem 2.15.1 shows that, using the component scores, no higher level of scaling than ordinal is possible. Any assumption about the form of the prior distribution is therefore a matter of convention. By choosing a standard normal prior we are implicitly choosing to measure the latent variable on a scale which renders its distribution normal. The predicted value of *y* for a given individual is therefore with reference to that prior scaling.

3.15 Posterior analysis: least squares

If we are unwilling to make assumptions about the forms of the distributions in the linear model we can still aim to find functions of the *x*s which are, in some sense, as near as possible to the *y*s. In the case of the linear model it is sufficient to consider only linear functions, as we now show. Let us consider the q linear functions $\mathbf{X}$ obtained by pre-multiplying $\mathbf{x} - \boldsymbol{\mu}$ by a $q \times p$ matrix of the form $(\mathbf{B}\boldsymbol{\Lambda})^{-1}\mathbf{B}$. Then from (3.3),

$$\mathbf{X} = (\mathbf{B}\boldsymbol{\Lambda})^{-1}\mathbf{B}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{y} + (\mathbf{B}\boldsymbol{\Lambda})^{-1}\mathbf{B}\mathbf{e}, \tag{3.60}$$

and hence we may write

$$X_j = y_j + u_j \quad (j = 1, 2, \ldots, q). \tag{3.61}$$

We could then regard X_j as a factor score for y_j since u_j is a random error independent of y_j. Since the transformation in (3.60) involves the arbitrary matrix $\mathbf{B}$ we shall seek to choose the transformation so that $\text{var}(u_j)$ is as small as possible for each j. In this way X_j is made as close as possible to y_j in a mean square sense.

It can easily be shown that choosing $(\mathbf{B}\boldsymbol{\Lambda})^{-1}\mathbf{B} = \mathbf{C} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}$ leads to the smallest possible values for the variances of the u_j. These smallest values are the diagonal elements of $\boldsymbol{\Gamma}^{-1}$. One way to see this is to show that for fixed $\mathbf{a}$,

$$\operatorname{var} \mathbf{a}'\mathbf{u} = \mathbf{a}'\mathbf{C}\boldsymbol{\Psi}\mathbf{C}'\mathbf{a} \geq \mathbf{a}'\boldsymbol{\Gamma}^{-1}\mathbf{a}.$$

Since $\mathbf{C}\boldsymbol{\Lambda} = \mathbf{I}$,

$$\mathbf{a}'\boldsymbol{\Gamma}^{-1}\mathbf{a} = \mathbf{a}'\mathbf{C}\boldsymbol{\Psi}^{1/2}\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^{-1}\mathbf{a}.$$

The Cauchy inequality gives

$$\begin{aligned}(\mathbf{a}'\boldsymbol{\Gamma}^{-1}\mathbf{a})^2 &\leq (\mathbf{a}'\mathbf{C}\boldsymbol{\Psi}^{1/2}\boldsymbol{\Psi}^{1/2}\mathbf{C}'\mathbf{a})(\mathbf{a}'\boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^{-1}\mathbf{a}) \\ &= (\mathbf{a}'\mathbf{C}\boldsymbol{\Psi}\mathbf{C}'\mathbf{a})(\mathbf{a}'\boldsymbol{\Gamma}^{-1}\mathbf{a}).\end{aligned}$$

So,

$$\mathbf{a}'\boldsymbol{\Gamma}^{-1}\mathbf{a} \leq \mathbf{a}'\mathbf{C}\boldsymbol{\Psi}\mathbf{C}'\mathbf{a} = \operatorname{var}(\mathbf{a}'\mathbf{u}).$$

We have

$$\mathbf{X} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \quad \text{and} \quad \operatorname{var}(\mathbf{u}) = \boldsymbol{\Gamma}^{-1}. \tag{3.62}$$

These are known as Bartlett's scores (Bartlett 1937), and the advantage often claimed for them is that $E(\mathbf{X} \mid \mathbf{y}) = \mathbf{y}$, as (3.61) shows. However, an expectation of $\mathbf{x}$ conditional on $\mathbf{y}$ is hardly relevant when it is $\mathbf{x}$ that is known and $\mathbf{y}$ that is to be predicted. Of more interest is the comparison with (3.9). The only difference is that $(\mathbf{I} + \boldsymbol{\Gamma})^{-1}$ in (3.9) is replaced by $\boldsymbol{\Gamma}^{-1}$ in (3.62). If $\boldsymbol{\Gamma}$ is diagonal the effect of this difference is merely to introduce different scaling of the Xs which is of little practical significance. The distribution-free argument can thus be viewed as establishing the result of (3.9) on a broader basis. However, a slightly different approach leads to scores which are identical to those of (3.9), as we now show.

In this method, due to Thomson (1951), we choose $\mathbf{X} = \mathbf{C}(\mathbf{x} - \boldsymbol{\mu})$, where $\mathbf{C}$ is a $q \times p$ matrix, and try to make X_j to as close to y_j as possible in the sense of minimising

$$\phi_j = E(X_j - y_j)^2 \tag{3.63}$$

where the expectation is with respect to both $\mathbf{y}$ and $\mathbf{x}$.

The choice $\mathbf{C} = \boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}$ leads to the smallest ϕ_j because for a fixed vector $\mathbf{a}$, $\mathbf{a}'(\mathbf{X} - \mathbf{y})$ has variance

$$\mathbf{a}'(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}' - \mathbf{C}\boldsymbol{\Lambda} - \boldsymbol{\Lambda}'\mathbf{C}' + \mathbf{I})$$

which may be written

$$\mathbf{a}'[(\mathbf{C}-\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1})\boldsymbol{\Sigma}(\mathbf{C}-\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1})'+\mathbf{I}-\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]\mathbf{a}.$$

This is clearly no less than

$$\mathbf{a}'[\mathbf{I}-\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}]\mathbf{a}$$

and that lower bound is achieved for

$$\mathbf{C}=\boldsymbol{\Lambda}'\boldsymbol{\Sigma}^{-1}=(\mathbf{I}+\boldsymbol{\Gamma})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}. \tag{3.64}$$

When defined in this way the scores are known as 'regression' scores. In this case it is *not* true that $E(\mathbf{X}\,|\,\mathbf{y})=\mathbf{y}$, but for the reason given above this is not a relevant objection.

3.16 Posterior analysis: a reliability approach

An entirely different approach to determining factor scores was proposed by Knott and Bartholomew (1993) (see also Bartholomew and Knott (1993)). This starts from the *reliability* of the score and aims to choose a score for which the reliability is a maximum. Let $\phi(\mathbf{x})$ denote any function of the xs which is proposed as a score for the latent variable y. Suppose we were able to make two independent determinations of $\phi(\mathbf{x})$, $\phi(\mathbf{x}_1)$ and $\phi(\mathbf{x}_2)$ say, for all members of a population. Then the better the function ϕ, the closer we would expect $\phi(\mathbf{x}_1)$ and $\phi(\mathbf{x}_2)$ to be. Closeness can be measured by the correlation of $\phi(\mathbf{x}_1)$ and $\phi(\mathbf{x}_2)$ – known as the *test–retest* correlation. Knott and Bartholomew showed that the function h which maximised this correlation satisfies

$$E\{\phi(\mathbf{x}_1)\,|\,\mathbf{x}_2\}=\lambda\phi(\mathbf{x}_2). \tag{3.65}$$

In the case of the NLFM it turns out that the function ϕ is identical to the components and hence proportional to the conditional expectation $E(y\,|\,\mathbf{x})$. For other models this is not necessarily so though it appears, on the limited evidence available, that the components are close to optimal. This approach therefore provides an interesting further justification for the existing methods rather than a different type of score. For a fuller account of the use of reliability in this context, see Bartholomew (1996, Section 1.2).

3.17 Examples

We illustrate the application of the foregoing theory on two examples. The first is small-scale and is designed to show some of the options available in a typical computer package. The second is on a larger scale and is more typical of examples

encountered in the applied literature. Both examples exhibit features which are quite common but seldom remarked upon in textbook treatments.

Example 3.17.1 In the first case the data are taken from a study by Smith and Stanley (1983) on the relationship between reaction times and intelligence test scores. Here we consider only the factor analysis of the ability variables. Scores were available for 112 individuals on the following six variables:

1. a non-verbal measure of general intelligence (Spearman's *g*) using Cattell's culture-fair test;
2. picture completion test;
3. block design;
4. mazes
5. reading comprehension;
6. vocabulary.

Full details may be found in the original paper. The correlation coefficients, covariances and variances, supplied by the authors, are set out in Table 3.3. The data were first analysed using the maximum likelihood routine with the correlation matrix input.

Table 3.3 Correlation coefficients (right upper) and variances and covariances (left lower) for Smith and Stanley's (1983) data.

	1	2	3	4	5	6
1	26.641	0.466	0.552	0.340	0.576	0.510
2	5.991	6.700	0.572	0.193	0.263	0.239
3	33.520	18.137	149.831	0.445	0.354	0.356
4	6.023	1.782	19.424	12.711	0.184	0.219
5	20.755	4.936	31.430	4.757	52.604	0.794
6	29.701	7.204	50.753	9.075	66.762	135.292

Since the *x*s have no natural common scale it is sensible to base the interpretation on the standardised loadings and communalities, which are given in Table 3.4 for a two-factor fit. We discuss the reasons for choosing two factors later.

The interpretation of these results is not immediately clear. The communalities vary a good deal. In the case of reading comprehension, block design and vocabulary the high figure means that a large part of their variation is accounted for by the two factors. The other three variables, especially mazes, have smaller communalities and so are much poorer indicators of the factors.

The particularly high communality of 0.96 for reading comprehension indicates that we are close to a Heywood case. This output was obtained from Mplus (Muthén and Muthén 2010).

Table 3.4 Standardised loadings, with standard errors in brackets, and communalities of a two-factor model estimated by maximum likelihood for Smith and Stanley's (1983) data.

Variable (i)	$\hat{\lambda}_{i1}$	$\hat{\lambda}_{i2}$	$1 - \hat{\psi}_i$
1	0.53 (0.09)	0.52 (0.09)	0.54
2	0.19 (0.09)	0.61 (0.08)	0.41
3	0.25 (0.08)	0.85 (0.07)	0.78
4	0.13 (0.09)	0.46 (0.09)	0.23
5	0.97 (0.07)	0.13 (0.05)	0.96
6	0.79 (0.07)	0.18 (0.07)	0.66

One has to deduce from the factor loadings what these two factors might be. The loadings in the first column are all positive, indicating that the first factor contributes to all variables, and this might tentatively be identified with some general ability. In the second column the loadings indicate that the second factor is largely concerned with the first four items. These are all non-verbal items, and so this seems to point to a non-verbal dimension to ability.

It might be that the interpretation would become clearer if we look at rotations. Some idea of what to try can be gained from a plot of the loadings, and this is given in Figure 3.1.

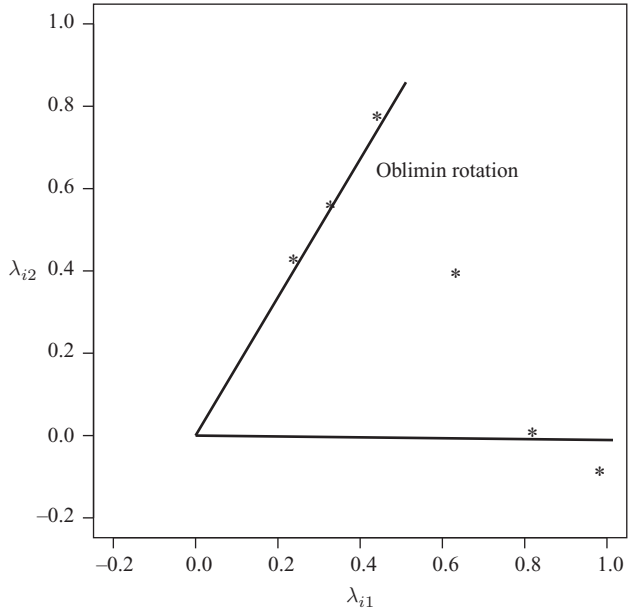

Figure 3.1 Factor loadings given in Table 3.4. The axes for OBLIMIN rotation are also shown.

The aim of rotation, we recall, is to achieve 'simple structure'. In geometrical terms this means seeking a rotation which makes the points lie as nearly as possible along the two axes. It is clear from the figure that this cannot be achieved with an orthogonal rotation. We could certainly rotate to make items 1, 2 and 3 lie very close to one axis, but then the remaining points would be nowhere near the other. Equally an axis passing close to 6 and 5 would leave the remaining points farther away from the other axis than before rotation.

The figure shows that something close to simple structure could be obtained by an oblique rotation, and one resulting from the OBLIMIN option of Mplus is illustrated. This still leaves item 1 with a substantial loading on both dimensions, but this is not surprising given that it is itself a composite measure. The rotated loadings are given in Table 3.5.

Table 3.5 Factor loadings for the OBLIMIN rotation, with standard errors in brackets, for Smith and Stanley's (1983) data (the communalities are unchanged by rotation and so are omitted).

Variable (i)	$\hat{\lambda}_{i1}$	$\hat{\lambda}_{i2}$
1	0.38 (0.10)	0.48 (0.10)
2	−0.01 (0.08)	0.65 (0.09)
3	−0.03 (0.03)	0.90 (0.07)
4	−0.02 (0.10)	0.49 (0.10)
5	1.00 (0.07)	−0.04 (0.02)
6	0.78 (0.09)	0.06 (0.08)

Leaving aside item 1 for the moment, we have factor 1 loading almost entirely on the verbal items, 5 and 6, and factor 2 loading on the remaining non-verbal items. This strongly suggests two dimensions of ability – one verbal and one non-verbal. We would expect the measure of general intelligence to contribute to both factors but more strongly, perhaps, to the non-verbal dimension because of the non-verbal character of that test.

We might thus conclude that the analysis identifies two dimensions of ability, one verbal and one non-verbal, but that these are not uncorrelated (because the rotation is oblique). The package gives an estimate of the correlation between the two factors of 0.462 with an estimated standard error of 0.086.

In view of this analysis we might expect the two-factor structure to be clearer if we omitted item 1 from the analysis, thus leaving items which fall clearly into verbal and non-verbal categories.

A repeat of maximum likelihood estimation for the last five items reveals a Heywood case. Mplus produces an unsatisfactory solution with a very large negative variance for item 2. In effect, we have a Heywood case with a communality of 1 for item 2. An R routine has been specifically written to cope with Heywood cases based

on the maximum likelihood estimation presented in Section 3.4, the results of which are given in Table 3.6.

Table 3.6 Factor loadings and communalities of a two-factor model for Smith and Stanley's (1983) data (item 1 omitted).

Variable (i)	$\hat{\lambda}_{i1}$	$\hat{\lambda}_{i2}$	$1 - \hat{\psi}_i$
2	−0.57	0.05	0.33
3	−1.00	0.00	1.00
4	−0.45	0.06	0.20
5	−0.35	0.78	0.73
6	−0.36	0.86	0.87

Heywood cases are not uncommon with small sample sizes of a hundred or so as we have already remarked. However, taking the loadings at their face values, the interpretation is essentially the same as before. The verbal ability factor emerges clearly in the second column whereas the first factor loads more heavily on the non-verbal items, with the dominating item being block design.

Given the small sample size, one should not push the interpretation beyond claiming some evidence for verbal and non-verbal dimensions of ability. If we wished to scale individuals on any of the dimensions we have uncovered, we would need the coefficients of the factor scores (or components), and these are provided by the standard packages. However, those coefficients can be very unstable in small samples because they involve the factor $\psi^{-1/2}$. If $\hat{\psi}_i$ is small, and subject to a large sampling error, the weight of x_i will be large and uncertain. If a Heywood case occurs, as in Table 3.6, the estimate becomes infinite. Posterior analysis for this example is thus not worthwhile.

All of the foregoing analysis presupposes that a model with two factors provides a satisfactory fit and that two is the best number of factors. In this case the default value provided by Mplus was two and we set this question on one side in order to proceed directly to questions of interpretation. We now return to the question of goodness of fit.

For the example with all six items the overall global test of goodness of fit yields: for the one-factor model, $\chi^2 = 78.95$ with 9 degrees of freedom ($P = 0.000$); and for the two-factor model, $\chi^2 = 6.36$ with 4 degrees of freedom ($P = 0.174$). There is little point in going beyond two factors because when $q = 3$ the model is only just identifiable with the number of parameters equal to the number of statistics. In any case the fit with two factors is good.

If the fit is not good, a comparison of the observed and predicted correlation (or covariance) matrix may help to identify the source of the deviation. Table 3.7 gives for the present example the differences between the observed and fitted correlation matrices. It appears that, with the possible exception of items 2 and 4, the 2-factor model predicts the observed correlation matrix very closely.

Table 3.7 Differences between the observed and fitted correlation matrices for Smith and Stanley's (1983) data.

Variable	1	2	3	4	5	6
1	0	0.05	−0.02	0.03	0.00	0.00
2	—	0	0.01	−0.11	0.00	−0.02
3	—	—	0	0.02	0.00	0.00
4	—	—	—	0	0.00	0.03
5	—	—	—	—	0	0.00
6	—	—	—	—	—	0

Example 3.17.2 The second example is taken from Harman (1976) and is based on eight physical variables measured on 305 individuals. The eight variables are: height, arm span, length of forearm, length of lower leg, weight, bitrochanteric diameter, chest girth and chest width. The first four variables were assumed to measure 'lankiness' and the last four 'stockiness'. Table 3.8 gives the sample correlation matrix of the eight variables.

Table 3.8 Sample correlation matrix for the physical data.

Variable	1	2	3	4	5	6	7	8
1	1.000							
2	0.846	1.000						
3	0.805	0.881	1.000					
4	0.859	0.826	0.801	1.000				
5	0.473	0.376	0.380	0.436	1.000			
6	0.398	0.326	0.319	0.329	0.762	1.000		
7	0.301	0.277	0.237	0.327	0.730	0.583	1.000	
8	0.382	0.415	0.345	0.365	0.629	0.577	0.539	1.000

Inspection of the correlation matrix shows a quite clear grouping of the first four variables and the last four. We would therefore expect to find that a two-factor solution adequately explains the correlation matrix. We first address the question of the choice of q, the number of factors. Inspection of the scree plot of the eigenvalues in Figure 3.2 shows that there are two eigenvalues greater than 1. If we look at the goodness-of-fit test and the AIC in Table 3.9 the position is less clear. The χ^2 test is highly significant for $q \leq 3$ and the AIC shows no sign of reaching a minimum value by $q = 4$. However, the normality assumption on which the χ^2 test is based is suspect, and we note that, whether the factors lower in the order are significant or not, they individually contribute very little to the overall variation in scores. When $q = 5$ no solution could be obtained due to convergence problems, indicating that the factoring has been taken too far. We shall, therefore, concentrate on the two-factor model for purposes of interpretation.

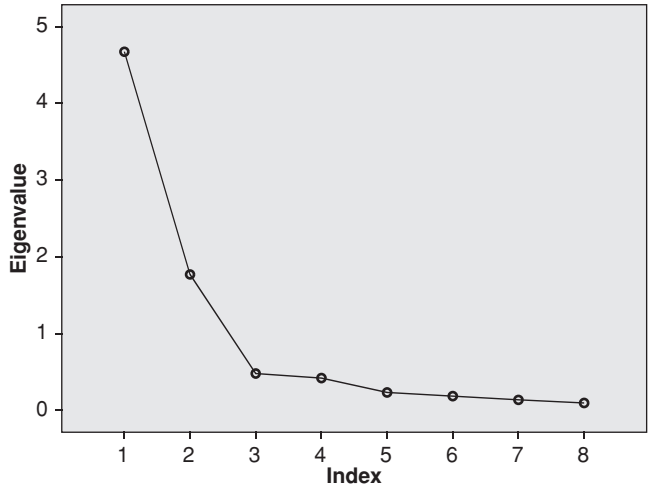

Figure 3.2 Scree plot for physical data.

Table 3.9 χ^2 and AIC for physical data.

Factors	χ^2	df	P	AIC
1	713.7	20	0.000	6254.5
2	88.6	13	0.000	5643.3
3	23.9	7	0.001	5590.6
4	0.84	2	0.658	5577.6

Table 3.10 Two-factor solution (unrotated and geomin rotated), with standard errors in brackets, for physical data.

	Unrotated				Rotated			
Variable	Factor 1		Factor 2		Factor 1		Factor 2	
1	0.885	(0.013)	0.219	(0.024)	**0.855**	(0.020)	0.115	(0.032)
2	0.939	(0.009)	0.108	(0.020)	**0.951**	(0.013)	−0.015	(0.022)
3	0.907	(0.011)	0.109	(0.022)	**0.917**	(0.012)	−0.010	(0.011)
4	0.876	(0.014)	0.184	(0.025)	**0.858**	(0.021)	0.078	(0.034)
5	0.304	(0.041)	0.905	(0.021)	0.003	(0.019)	**0.953**	(0.019)
6	0.254	(0.044)	0.756	(0.027)	0.002	(0.026)	**0.797**	(0.027)
7	0.192	(0.045)	0.739	(0.028)	−0.057	(0.043)	**0.787**	(0.033)
8	0.325	(0.046)	0.598	(0.037)	0.133	(0.047)	**0.612**	(0.040)

The parameter estimates obtained by maximum likelihood are given in Table 3.10, with standard errors in brackets, both for the unrotated and the rotated solution. The geomin oblique rotation, available in Mplus (Muthén and Muthén 2010), (Jennrich 2006) is used here. Both the unrotated and the oblique solution produced interpretable factors. However, the rotated solution produced a simple structure. In order to identify what the two factors are measuring, we have to ask what distinguishes the items with the higher loading from those with the lower. This, of course, requires some expertise in the field but the task might not be very difficult here because of the nature of the items. The loadings printed in bold type for factor 1 constitute a group which load heavily on that factor. On the second factor it is the complementary set of items which have the high loadings. The high loadings on the first four items are associated with length and height, and therefore factor 1 is related to aspects of physique that have to do with tallness and thinness. The high loadings on the last four items are related to weight, diameter and width, and so factor 2 measures the stockiness aspects of the physique. The two factors are significantly correlated ($r = 0.436$, (0.049)) where the standard error is given in brackets.