# Chapter 11
# Factor Analysis

A frequently applied paradigm in analyzing data from multivariate observations is to model the relevant information (represented in a multivariate variable $X$) as coming from a limited number of latent factors. In a survey on household consumption, for example, the consumption levels, $X$, of $p$ different goods during one month could be observed. The variations and covariations of the $p$ components of $X$ throughout the survey might in fact be explained by two or three main social behavior factors of the household. For instance, a basic desire of comfort or the willingness to achieve a certain social level or other social latent concepts might explain most of the consumption behavior. These unobserved factors are much more interesting to the social scientist than the observed quantitative measures ($X$) themselves, because they give a better understanding of the behavior of households. As shown in the examples below, the same kind of factor analysis is of interest in many fields such as psychology, marketing, economics, politic sciences, etc.

How can we provide a statistical model addressing these issues and how can we interpret the obtained model? This is the aim of factor analysis. As in Chapter 9 and Chapter 10, the driving statistical theme of this chapter is to reduce the dimension of the observed data. The perspective used, however, is different: we assume that there is a model (it will be called the "Factor Model") stating that most of the covariances between the $p$ elements of $X$ can be explained by a limited number of latent factors. Section 11.1 defines the basic concepts and notations of the orthogonal factor model, stressing the non-uniqueness of the solutions. We show how to take advantage of this non-uniqueness to derive techniques which lead to easier interpretations. This will involve (geometric) rotations of the factors. Section 11.2 presents an empirical approach to factor analysis. Various estimation procedures are proposed and an optimal rotation procedure is defined. Many examples are used to illustrate the method.

## 11.1 The Orthogonal Factor Model

The aim of factor analysis is to explain the outcome of $p$ variables in the data matrix $\mathcal{X}$ using fewer variables, the so-called *factors*. Ideally all the information in $\mathcal{X}$ can be reproduced by a smaller number of factors. These factors are interpreted as latent (unobserved) common characteristics of the observed $x \in \mathbb{R}^p$. The case just described occurs when every observed $x = (x_1, \ldots, x_p)^\top$ can be written as

$$x_j = \sum_{\ell=1}^{k} q_{j\ell} f_\ell + \mu_j, \quad j = 1, \ldots, p. \tag{11.1}$$

Here $f_\ell$, for $\ell = 1, \ldots, k$ denotes the factors. The number of factors, $k$, should always be much smaller than $p$. For instance, in psychology $x$ may represent $p$ results of a test measuring intelligence scores. One common latent factor explaining $x \in \mathbb{R}^p$ could be the overall level of "intelligence". In marketing studies, $x$ may consist of $p$ answers to a survey on the levels of satisfaction of the customers. These $p$ measures could be explained by common latent factors like the attraction level of the product or the image of the brand, and so on. Indeed it is possible to create a representation of the observations that is similar to the one in (11.1) by means of principal components, but only if the last $p - k$ eigenvalues corresponding to the covariance matrix are equal to zero. Consider a $p$-dimensional random vector X with mean $\mu$ and covariance matrix $\mathsf{Var}(X) = \Sigma$. A model similar to (11.1) can be written for $X$ in matrix notation, namely

$$X = \mathcal{Q}F + \mu, \tag{11.2}$$

where $F$ is the $k$-dimensional vector of the $k$ factors. When using the factor model (11.2) it is often assumed that the factors $F$ are centered, uncorrelated and standardized: $\mathsf{E}(F) = 0$ and $\mathsf{Var}(F) = \mathcal{I}_k$. We will now show that if the last $p - k$ eigenvalues of $\Sigma$ are equal to zero, we can easily express $X$ by the factor model (11.2).

The spectral decomposition of $\Sigma$ is given by $\Gamma \Lambda \Gamma^\top$. Suppose that only the first $k$ eigenvalues are positive, i.e., $\lambda_{k+1} = \cdots = \lambda_p = 0$. Then the (singular) covariance matrix can be written as

$$\Sigma = \sum_{\ell=1}^{k} \lambda_\ell \gamma_\ell \gamma_\ell^\top = (\Gamma_1 \Gamma_2) \begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \Gamma_1^\top \\ \Gamma_2^\top \end{pmatrix}.$$

In order to show the connection to the factor model (11.2), recall that the PCs are given by $Y = \Gamma^\top (X - \mu)$. Rearranging we have $X - \mu = \Gamma Y = \Gamma_1 Y_1 + \Gamma_2 Y_2$, where the components of $Y$ are partitioned according to the partition of $\Gamma$ above, namely

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \Gamma_1^\top \\ \Gamma_2^\top \end{pmatrix} (X - \mu), \quad \text{where} \quad \begin{pmatrix} \Gamma_1^\top \\ \Gamma_2^\top \end{pmatrix} (X - \mu) \sim \left(0, \begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix}\right).$$

In other words, $Y_2$ has a singular distribution with mean and covariance matrix equal to zero. Therefore, $X - \mu = \Gamma_1 Y_1 + \Gamma_2 Y_2$ implies that $X - \mu$ is equivalent to $\Gamma_1 Y_1$,

which can be written as

$$X = \Gamma_1 \Lambda_1^{1/2} \Lambda_1^{-1/2} Y_1 + \mu.$$

Defining $Q = \Gamma_1 \Lambda_1^{1/2}$ and $F = \Lambda_1^{-1/2} Y_1$, we obtain the factor model (11.2).

Note that the covariance matrix of model (11.2) can be written as

$$\Sigma = \mathsf{E}(X - \mu)(X - \mu)^\top = Q\mathsf{E}(FF^\top)Q^\top = QQ^\top = \sum_{j=1}^{k} \lambda_j \gamma_j \gamma_j^\top. \qquad (11.3)$$

We have just shown how the variable $X$ can be completely determined by a weighted sum of $k$ (where $k < p$) uncorrelated factors. The situation used in the derivation, however, is too idealistic. In practice the covariance matrix is rarely singular.

It is common praxis in factor analysis to split the influences of the factors into common and specific ones. There are, for example, highly informative factors that are common to all of the components of $X$ and factors that are specific to certain components. The factor analysis model used in praxis is a generalization of (11.2):

$$X = QF + U + \mu, \qquad (11.4)$$

where $Q$ is a $(p \times k)$ matrix of the (non-random) *loadings* of the *common factors* $F (k \times 1)$ and $U$ is a $(p \times 1)$ matrix of the (random) *specific factors*. It is assumed that the factor variables $F$ are uncorrelated random vectors and that the specific factors are uncorrelated and have zero covariance with the common factors. More precisely, it is assumed that:

$$\begin{aligned}
\mathsf{E}\,F &= 0, \\
\mathsf{Var}(F) &= \mathcal{I}_k, \\
\mathsf{E}\,U &= 0, \qquad\qquad\qquad\qquad (11.5) \\
\mathsf{Cov}(U_i, U_j) &= 0, \quad i \neq j \\
\mathsf{Cov}(F, U) &= 0.
\end{aligned}$$

Define

$$\mathsf{Var}(U) = \Psi = \mathrm{diag}(\psi_{11}, \ldots, \psi_{pp}).$$

The generalized factor model (11.4) together with the assumptions given in (11.5) constitute the *orthogonal factor model*.

---

**Orthogonal Factor Model**

| $X$ | $=$ | $Q$ | $F$ | $+$ | $U$ | $+$ | $\mu$ |
|---|---|---|---|---|---|---|---|
| $(p \times 1)$ | | $(p \times k)$ | $(k \times 1)$ | | $(p \times 1)$ | | $(p \times 1)$ |

| | | |
|---|---|---|
| $\mu_j$ | $=$ | mean of variable $j$ |
| $U_j$ | $=$ | $j$-th specific factor |
| $F_\ell$ | $=$ | $\ell$-th common factor |
| $q_{j\ell}$ | $=$ | loading of the $j$-th variable on the $\ell$-th factor |

The random vectors $F$ and $U$ are unobservable and uncorrelated.

Note that (11.4) implies for the components of $X = (X_1, \ldots, X_p)^\top$ that

$$X_j = \sum_{\ell=1}^{k} q_{j\ell} F_\ell + U_j + \mu_j, \quad j = 1, \ldots, p. \tag{11.6}$$

Using (11.5) we obtain $\sigma_{X_j X_j} = \mathsf{Var}(X_j) = \sum_{\ell=1}^{k} q_{j\ell}^2 + \psi_{jj}$. The quantity $h_j^2 = \sum_{\ell=1}^{k} q_{j\ell}^2$ is called the *communality* and $\psi_{jj}$ the *specific variance*. Thus the covariance of $X$ can be rewritten as

$$\begin{aligned}
\Sigma = \mathsf{E}(X - \mu)(X - \mu)^\top &= \mathsf{E}(QF + U)(QF + U)^\top \\
&= Q\,\mathsf{E}(FF^\top)Q^\top + \mathsf{E}(UU^\top) = Q\,\mathsf{Var}(F)Q^\top + \mathsf{Var}(U) \\
&= QQ^\top + \Psi.
\end{aligned} \tag{11.7}$$

In a sense, the factor model explains the variations of $X$ for the most part by a small number of latent factors $F$ common to its $p$ components and entirely explains all the correlation structure between its components, plus some "noise" $U$ which allows specific variations of each component to enter. The specific factors adjust to capture the individual variance of each component. Factor analysis relies on the assumptions presented above. If the assumptions are not met, the analysis could be spurious. Although principal components analysis and factor analysis might be related (this was hinted at in the derivation of the factor model), they are quite different in nature. PCs are linear transformations of $X$ arranged in decreasing order of variance and used to reduce the dimension of the data set, whereas in factor analysis, we try to model the variations of $X$ using a linear transformation of a fixed, limited number of latent factors. The objective of factor analysis is to find the loadings $Q$ and the specific variance $\Psi$. Estimates of $Q$ and $\Psi$ are deduced from the covariance structure (11.7).

## Interpretation of the Factors

Assume that a factor model with $k$ factors was found to be reasonable, i.e., most of the (co)variations of the $p$ measures in $X$ were explained by the $k$ fixed latent factors. The next natural step is to try to understand what these factors represent. To interpret $F_\ell$, it makes sense to compute its correlations with the original variables $X_j$ first. This is done for $\ell = 1, \ldots, k$ and for $j = 1, \ldots, p$ to obtain the matrix $P_{XF}$. The sequence of calculations used here are in fact the same that were used to interprete the PCs in the principal components analysis.

The following covariance between $X$ and $F$ is obtained via (11.5),

$$\Sigma_{XF} = \mathsf{E}\{(QF + U)F^\top\} = Q.$$

The correlation is

$$P_{XF} = D^{-1/2}Q, \tag{11.8}$$

where $D = \text{diag}(\sigma_{X_1 X_1}, \ldots, \sigma_{X_p X_p})$. Using (11.8) it is possible to construct a figure analogous to Figure 10.6 and thus to consider which of the original variables $X_1, \ldots, X_p$ play a role in the unobserved common factors $F_1, \ldots, F_k$.

Returning to the psychology example where $X$ are the observed scores to $p$ different intelligence tests (the WAIS data set in Table B.12 provides an example), we would expect a model with one factor to produce a factor that is positively correlated with all of the components in $X$. For this example the factor represents the overall level of intelligence of an individual. A model with two factors could produce a refinement in explaining the variations of the $p$ scores. For example, the first factor could be the same as before (overall level of intelligence), whereas the second factor could be positively correlated with some of the tests, $X_j$, that are related to the individual's ability to think abstractly and negatively correlated with other tests, $X_i$, that are related to the individual's practical ability. The second factor would then concern a particular dimension of the intelligence stressing the distinctions between the "theoretical" and "practical" abilities of the individual. If the model is true, most of the information coming from the $p$ scores can be summarized by these two latent factors. Other practical examples are given below.

## Invariance of Scale

What happens if we change the scale of $X$ to $Y = \mathcal{C}X$ with $\mathcal{C} = \text{diag}(c_1, \ldots, c_p)$? If the $k$-factor model (11.6) is true for $X$ with $\mathcal{Q} = \mathcal{Q}_X$, $\Psi = \Psi_X$, then, since

$$\text{Var}(Y) = \mathcal{C}\Sigma\mathcal{C}^\top = \mathcal{C}\mathcal{Q}_X \mathcal{Q}_X^\top \mathcal{C}^\top + \mathcal{C}\Psi_X\mathcal{C}^\top,$$

the same $k$-factor model is also true for $Y$ with $\mathcal{Q}_Y = \mathcal{C}\mathcal{Q}_X$ and $\Psi_Y = \mathcal{C}\Psi_X\mathcal{C}^\top$. In many applications, the search for the loadings $\mathcal{Q}$ and for the specific variance $\Psi$ will be done by the decomposition of the correlation matrix of $X$ rather than the covariance matrix $\Sigma$. This corresponds to a factor analysis of a linear transformation of $X$ (i.e., $Y = D^{-1/2}(X - \mu)$). The goal is to try to find the loadings $\mathcal{Q}_Y$ and the specific variance $\Psi_Y$ such that

$$P = \mathcal{Q}_Y \mathcal{Q}_Y^\top + \Psi_Y. \tag{11.9}$$

In this case the interpretation of the factors $F$ immediately follows from (11.8) given the following correlation matrix:

$$P_{XF} = P_{YF} = \mathcal{Q}_Y. \tag{11.10}$$

Because of the scale invariance of the factors, the loadings and the specific variance of the model, where $X$ is expressed in its original units of measure, are given by

$$\mathcal{Q}_X = D^{1/2}\mathcal{Q}_Y$$
$$\Psi_X = D^{1/2}\Psi_Y D^{1/2}.$$

It should be noted that although the factor analysis model (11.4) enjoys the scale invariance property, the actual estimated factors could be scale dependent. We will come back to this point later when we discuss the method of principal factors.

## *Non-uniqueness of Factor Loadings*

The factor loadings are not unique! Suppose that $\mathcal{G}$ is an orthogonal matrix. Then $X$ in (11.4) can also be written as

$$X = (\mathcal{Q}\mathcal{G})(\mathcal{G}^\top F) + U + \mu.$$

This implies that, if a $k$-factor of $X$ with factors $F$ and loadings $\mathcal{Q}$ is true, then the $k$-factor model with factors $\mathcal{G}^\top F$ and loadings $\mathcal{Q}\mathcal{G}$ is also true. In practice, we will take advantage of this non-uniqueness. Indeed, referring back to Section 2.6 we can conclude that premultiplying a vector $F$ by an orthogonal matrix corresponds to a rotation of the system of axis, the direction of the first new axis being given by the first row of the orthogonal matrix. It will be shown that choosing an appropriate rotation will result in a matrix of loadings $\mathcal{Q}\mathcal{G}$ that will be easier to interpret. We have seen that the loadings provide the correlations between the factors and the original variables, therefore, it makes sense to search for rotations that give factors that are maximally correlated with various groups of variables.

From a numerical point of view, the non-uniqueness is a drawback. We have to find loadings $\mathcal{Q}$ and specific variances $\Psi$ satisfying the decomposition $\Sigma = \mathcal{Q}\mathcal{Q}^\top + \Psi$, but no straightforward numerical algorithm can solve this problem due to the multiplicity of the solutions. An acceptable technique is to impose some chosen constraints in order to get—in the best case—an unique solution to the decomposition. Then, as suggested above, once we have a solution we will take advantage of the rotations in order to obtain a solution that is easier to interprete.

An obvious question is: what kind of constraints should we impose in order to eliminate the non-uniqueness problem? Usually, we impose additional constraints where

$$\mathcal{Q}^\top \Psi^{-1} \mathcal{Q} \quad \text{is diagonal} \tag{11.11}$$

or

$$\mathcal{Q}^\top \mathcal{D}^{-1} \mathcal{Q} \quad \text{is diagonal.} \tag{11.12}$$

How many parameters does the model (11.7) have without constraints?

$$\mathcal{Q}(p \times k) \quad \text{has } p \cdot k \text{ parameters, and}$$

$$\Psi(p \times p) \quad \text{has } p \text{ parameters.}$$

Hence we have to determine $pk + p$ parameters! Conditions (11.11) respectively (11.12) introduce $\frac{1}{2}\{k(k-1)\}$ constraints, since we require the matrices to be diagonal. Therefore, the degrees of freedom of a model with $k$ factors is:

$$d = (\text{\# parameters for } \Sigma \text{ unconstrained}) - (\text{\# parameters for } \Sigma \text{ constrained})$$

$$= \frac{1}{2}p(p+1) - \left(pk + p - \frac{1}{2}k(k-1)\right)$$

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

If $d < 0$, then the model is undetermined: there are infinitely many solutions to (11.7). This means that the number of parameters of the factorial model is larger than the number of parameters of the original model, or that the number of factors $k$ is "too large" relative to $p$. In some cases $d = 0$: there is a unique solution to the problem (except for rotation). In practice we usually have that $d > 0$: there are more equations than parameters, thus an exact solution does not exist. In this case approximate solutions are used. An approximation of $\Sigma$, for example, is $\mathcal{Q}\mathcal{Q}^{\top} + \Psi$. The last case is the most interesting since the factorial model has less parameters than the original one. Estimation methods are introduced in the next section.

Evaluating the degrees of freedom, $d$, is particularly important, because it already gives an idea of the upper bound on the number of factors we can hope to identify in a factor model. For instance, if $p = 4$, we could not identify a factor model with 2 factors (this results in $d = -1$ which has infinitly many solutions). With $p = 4$, only a one factor model gives an approximate solution ($d = 2$). When $p = 6$, models with 1 and 2 factors provide approximate solutions and a model with 3 factors results in an unique solution (up to the rotations) since $d = 0$. A model with 4 or more factors would not be allowed, but of course, the aim of factor analysis is to find suitable models with a small number of factors, i.e., smaller than $p$. The next two examples give more insights into the notion of degrees of freedom.

*Example 11.1* Let $p = 3$ and $k = 1$, then $d = 0$ and

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix} = \begin{pmatrix} q_1^2 + \psi_{11} & q_1 q_2 & q_1 q_3 \\ q_1 q_2 & q_2^2 + \psi_{22} & q_2 q_3 \\ q_1 q_3 & q_2 q_3 & q_3^2 + \psi_{33} \end{pmatrix}$$

with $\mathcal{Q} = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix}$ and $\Psi = \begin{pmatrix} \psi_{11} & 0 & 0 \\ 0 & \psi_{22} & 0 \\ 0 & 0 & \psi_{33} \end{pmatrix}$. Note that here the constraint (11.8) is automatically verified since $k = 1$. We have

$$q_1^2 = \frac{\sigma_{12}\sigma_{13}}{\sigma_{23}}; \qquad q_2^2 = \frac{\sigma_{12}\sigma_{23}}{\sigma_{13}}; \qquad q_3^2 = \frac{\sigma_{13}\sigma_{23}}{\sigma_{12}}$$

and

$$\psi_{11} = \sigma_{11} - q_1^2; \qquad \psi_{22} = \sigma_{22} - q_2^2; \qquad \psi_{33} = \sigma_{33} - q_3^2.$$

In this particular case ($k = 1$), the only rotation is defined by $\mathcal{G} = -1$, so the other solution for the loadings is provided by $-\mathcal{Q}$.

*Example 11.2* Suppose now $p = 2$ and $k = 1$, then $d < 0$ and

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = \begin{pmatrix} q_1^2 + \psi_{11} & q_1 q_2 \\ q_1 q_2 & q_2^2 + \psi_{22} \end{pmatrix}.$$

We have infinitely many solutions: for any $\alpha$ ($\rho < \alpha < 1$), a solution is provided by

$$q_1 = \alpha; \qquad q_2 = \rho/\alpha; \qquad \psi_{11} = 1 - \alpha^2; \qquad \psi_{22} = 1 - (\rho/\alpha)^2.$$

The solution in Example 11.1 may be unique (up to a rotation), but it is not proper in the sense that it cannot be interpreted statistically. Exercise 11.5 gives an example where the specific variance $\psi_{11}$ is negative.

⚠   Even in the case of a unique solution ($d = 0$), the solution may be inconsistent with statistical interpretations.



## **Summary**

↪   The factor analysis model aims to describe how the original $p$ variables in a data set depend on a small number of latent factors $k < p$, i.e., it assumes that $X = \mathcal{Q}F + U + \mu$. The ($k$-dimensional) random vector $F$ contains the common factors, the ($p$-dimensional) $U$ contains the specific factors and $\mathcal{Q}(p \times k)$ contains the factor loadings.

↪   It is assumed that $F$ and $U$ are uncorrelated and have zero means, i.e., $F \sim (0, \mathcal{I})$, $U \sim (0, \Psi)$ where $\Psi$ is diagonal matrix and $\mathsf{Cov}(F, U) = 0$.
This leads to the covariance structure $\Sigma = \mathcal{Q}\mathcal{Q}^\top + \Psi$.

↪   The interpretation of the factor $F$ is obtained through the correlation $P_{XF} = D^{-1/2}\mathcal{Q}$.

↪   A normalized analysis is obtained by the model $P = \mathcal{Q}\mathcal{Q}^\top + \Psi$. The interpretation of the factors is given directly by the loadings $\mathcal{Q}$: $P_{XF} = \mathcal{Q}$.

↪   The factor analysis model is scale invariant. The loadings are not unique (only up to multiplication by an orthogonal matrix).

↪   Whether a model has an unique solution or not is determined by the degrees of freedom $d = 1/2(p - k)^2 - 1/2(p + k)$.

## 11.2  Estimation of the Factor Model

In practice, we have to find estimates $\widehat{\mathcal{Q}}$ of the loadings $\mathcal{Q}$ and estimates $\widehat{\Psi}$ of the specific variances $\Psi$ such that analogously to (11.7)

$$\mathcal{S} = \widehat{\mathcal{Q}}\widehat{\mathcal{Q}}^\top + \widehat{\Psi},$$

where $\mathcal{S}$ denotes the empirical covariance of $\mathcal{X}$. Given an estimate $\widehat{\mathcal{Q}}$ of $\mathcal{Q}$, it is natural to set

$$\widehat{\psi}_{jj} = s_{X_j X_j} - \sum_{\ell=1}^{k} \widehat{q}_{j\ell}^2.$$

We have that $\widehat{h}_j^2 = \sum_{\ell=1}^{k} \widehat{q}_{j\ell}^2$ is an estimate for the communality $h_j^2$.

In the ideal case $d = 0$, there is an exact solution. However, $d$ is usually greater than zero, therefore we have to find $\widehat{Q}$ and $\widehat{\Psi}$ such that $S$ is approximated by $\widehat{Q}\widehat{Q}^\top + \widehat{\Psi}$. As mentioned above, it is often easier to compute the loadings and the specific variances of the standardized model.

Define $\mathcal{Y} = \mathcal{H}\mathcal{X}\mathcal{D}^{-1/2}$, the standardization of the data matrix $\mathcal{X}$, where, as usual, $\mathcal{D} = \text{diag}(s_{X_1 X_1}, \ldots, s_{X_p X_p})$ and the centering matrix $\mathcal{H} = \mathcal{I} - n^{-1} 1_n 1_n^\top$ (recall from Chapter 2 that $\mathcal{S} = \frac{1}{n}\mathcal{X}^\top \mathcal{H}\mathcal{X}$). The estimated factor loading matrix $\widehat{Q}_Y$ and the estimated specific variance $\widehat{\Psi}_Y$ of $\mathcal{Y}$ are

$$\widehat{Q}_Y = \mathcal{D}^{-1/2}\widehat{Q}_X \quad \text{and} \quad \widehat{\Psi}_Y = \mathcal{D}^{-1}\widehat{\Psi}_X.$$

For the correlation matrix $\mathcal{R}$ of $\mathcal{X}$, we have that

$$\mathcal{R} = \widehat{Q}_Y \widehat{Q}_Y^\top + \widehat{\Psi}_Y.$$

The interpretations of the factors are formulated from the analysis of the loadings $\widehat{Q}_Y$.

*Example 11.3* Let us calculate the matrices just defined for the car data given in Table B.7. This data set consists of the averaged marks (from $1 = $ low to $6 = $ high) for 24 car types. Considering the three variables price, security and easy handling, we get the following correlation matrix:

$$\mathcal{R} = \begin{pmatrix} 1 & 0.975 & 0.613 \\ 0.975 & 1 & 0.620 \\ 0.613 & 0.620 & 1 \end{pmatrix}.$$

We will first look for one factor, i.e., $k = 1$. Note that (# number of parameters of $\Sigma$ unconstrained $-$ # parameters of $\Sigma$ constrained) is equal to $\frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k) = \frac{1}{2}(3-1)^2 - \frac{1}{2}(3+1) = 0$. This implies that there is an exact solution! The equation

$$\begin{pmatrix} 1 & r_{X_1 X_2} & r_{X_1 X_3} \\ r_{X_1 X_2} & 1 & r_{X_2 X_3} \\ r_{X_1 X_3} & r_{X_2 X_3} & 1 \end{pmatrix} = \mathcal{R} = \begin{pmatrix} \widehat{q}_1^2 + \widehat{\psi}_{11} & \widehat{q}_1 \widehat{q}_2 & \widehat{q}_1 \widehat{q}_3 \\ \widehat{q}_1 \widehat{q}_2 & \widehat{q}_2^2 + \widehat{\psi}_{22} & \widehat{q}_2 \widehat{q}_3 \\ \widehat{q}_1 \widehat{q}_3 & \widehat{q}_2 \widehat{q}_3 & \widehat{q}_3^2 + \widehat{\psi}_{33} \end{pmatrix}$$

yields the communalities $\widehat{h}_i^2 = \widehat{q}_i^2$, where

$$\widehat{q}_1^2 = \frac{r_{X_1 X_2} r_{X_1 X_3}}{r_{X_2 X_3}}, \qquad \widehat{q}_2^2 = \frac{r_{X_1 X_2} r_{X_2 X_3}}{r_{X_1 X_3}} \quad \text{and} \quad \widehat{q}_3^2 = \frac{r_{X_1 X_3} r_{X_2 X_3}}{r_{X_1 X_2}}.$$

Combining this with the specific variances $\widehat{\psi}_{11} = 1 - \widehat{q}_1^2$ , $\widehat{\psi}_{22} = 1 - \widehat{q}_2^2$ and $\widehat{\psi}_{33} = 1 - \widehat{q}_3^2$, we obtain the following solution

$$\widehat{q}_1 = 0.982 \qquad \widehat{q}_2 = 0.993 \qquad \widehat{q}_3 = 0.624$$
$$\widehat{\psi}_{11} = 0.035 \qquad \widehat{\psi}_{22} = 0.014 \qquad \widehat{\psi}_{33} = 0.610.$$

Since the first two communalities $(\widehat{h}_i^2 = \widehat{q}_i^2)$ are close to one, we can conclude that the first two variables, namely price and security, are explained by the single factor quite well. This factor can be interpreted as a "price+security" factor.

### *The Maximum Likelihood Method*

Recall from Chapter 6 the log-likelihood function $\ell$ for a data matrix $\mathcal{X}$ of observations of $X \sim N_p(\mu, \Sigma)$:

$$\ell(\mathcal{X}; \mu, \Sigma) = -\frac{n}{2} \log |2\pi \Sigma| - \frac{1}{2} \sum_{i=1}^{n}(x_i - \mu)\Sigma^{-1}(x_i - \mu)^{\top}$$
$$= -\frac{n}{2} \log |2\pi \Sigma| - \frac{n}{2} \operatorname{tr}(\Sigma^{-1}\mathcal{S}) - \frac{n}{2}(\overline{x} - \mu)\Sigma^{-1}(\overline{x} - \mu)^{\top}.$$

This can be rewritten as

$$\ell(\mathcal{X}; \widehat{\mu}, \Sigma) = -\frac{n}{2}\{\log |2\pi \Sigma| + \operatorname{tr}(\Sigma^{-1}\mathcal{S})\}.$$

Replacing $\mu$ by $\widehat{\mu} = \overline{x}$ and substituting $\Sigma = \mathcal{Q}\mathcal{Q}^{\top} + \Psi$ this becomes

$$\ell(\mathcal{X}; \widehat{\mu}, \mathcal{Q}, \Psi) = -\frac{n}{2}[\log\{|2\pi(\mathcal{Q}\mathcal{Q}^{\top} + \Psi)|\} + \operatorname{tr}\{(\mathcal{Q}\mathcal{Q}^{\top} + \Psi)^{-1}\mathcal{S}\}]. \quad (11.13)$$

Even in the case of a single factor ($k = 1$), these equations are rather complicated and iterative numerical algorithms have to be used (for more details see Mardia et al. (1979, p. 263ff)). A practical computation scheme is also given in Supplement 9A of Johnson and Wichern (1998).

### **Likelihood Ratio Test for the Number of Common Factors**

Using the methodology of Chapter 7, it is easy to test the adequacy of the factor analysis model by comparing the likelihood under the null (factor analysis) and alternative (no constraints on covariance matrix) hypotheses.

Assuming that $\widehat{\mathcal{Q}}$ and $\widehat{\Psi}$ are the maximum likelihood estimates corresponding to (11.13), we obtain the following LR test statistic:

$$-2\log\left(\frac{\text{maximized likelihood under } H_0}{\text{maximized likelihood}}\right) = n\log\left(\frac{|\widehat{\mathcal{Q}}\widehat{\mathcal{Q}}^{\top} + \widehat{\Psi}|}{|\mathcal{S}|}\right), \quad (11.14)$$

which asymptotically has the $\chi^2_{\frac{1}{2}\{(p-k)^2 - p - k\}}$ distribution.

The $\chi^2$ approximation can be improved if we replace $n$ by $n - 1 - (2p + 4k + 5)/6$ in (11.14) (Bartlett, 1954). Using Bartlett's correction, we reject the factor analysis model at the $\alpha$ level if

$$\{n - 1 - (2p + 4k + 5)/6\} \log \left( \frac{|\widehat{\mathcal{Q}}\widehat{\mathcal{Q}}^\top + \widehat{\Psi}|}{|\mathcal{S}|} \right) > \chi^2_{1-\alpha;\{(p-k)^2-p-k\}/2}, \quad (11.15)$$

and if the number of observations $n$ is large and the number of common factors $k$ is such that the $\chi^2$ statistic has a positive number of degrees of freedom.

## *The Method of Principal Factors*

The *method of principal factors* concentrates on the decomposition of the correlation matrix $\mathcal{R}$ or the covariance matrix $\mathcal{S}$. For simplicity, only the method for the correlation matrix $\mathcal{R}$ will be discussed. As pointed out in Chapter 10, the spectral decompositions of $\mathcal{R}$ and $\mathcal{S}$ yield different results and therefore, the method of principal factors may result in different estimators. The method can be motivated as follows: Suppose we know the exact $\Psi$, then the constraint (11.12) implies that the columns of $\mathcal{Q}$ are orthogonal since $\mathcal{D} = \mathcal{I}$ and it implies that they are eigenvectors of $\mathcal{Q}\mathcal{Q}^\top = \mathcal{R} - \Psi$. Furthermore, assume that the first $k$ eigenvalues are positive. In this case we could calculate $\mathcal{Q}$ by means of a spectral decomposition of $\mathcal{Q}\mathcal{Q}^\top$ and $k$ would be the number of factors.

The principal factors algorithm is based on good preliminary estimators $\widetilde{h}_j^2$ of the communalities $h_j^2$, for $j = 1, \ldots, p$. There are two traditional proposals:

- $\widetilde{h}_j^2$, defined as the square of the multiple correlation coefficient of $X_j$ with $(X_l)$, for $l \neq j$, i.e., $\rho^2(V, W\widehat{\beta})$ with $V = X_j$, $W = (X_\ell)_{\ell \neq j}$ and where $\widehat{\beta}$ is the least squares regression parameter of a regression of $V$ on $W$.
- $\widetilde{h}_j^2 = \max_{\ell \neq j} |r_{X_j X_\ell}|$, where $\mathcal{R} = (r_{X_j X_\ell})$ is the correlation matrix of $\mathcal{X}$.

Given $\widetilde{\psi}_{jj} = 1 - \widetilde{h}_j^2$ we can construct the *reduced correlation matrix*, $\mathcal{R} - \widetilde{\Psi}$. The Spectral Decomposition Theorem says that

$$\mathcal{R} - \widetilde{\Psi} = \sum_{\ell=1}^{p} \lambda_\ell \gamma_\ell \gamma_\ell^\top,$$

with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$. Assume that the first $k$ eigenvalues $\lambda_1, \ldots, \lambda_k$ are positive and large compared to the others. Then we can set

$$\widehat{q}_\ell = \sqrt{\lambda_\ell} \, \gamma_\ell, \quad \ell = 1, \ldots, k$$

or

$$\widehat{\mathcal{Q}} = \Gamma_1 \Lambda_1^{1/2}$$

with

$$\Gamma_1 = (\gamma_1, \ldots, \gamma_k) \quad \text{and} \quad \Lambda_1 = \text{diag}(\lambda_1, \ldots, \lambda_k).$$

In the next step set

$$\widehat{\psi}_{jj} = 1 - \sum_{\ell=1}^{k} \widehat{q}_{j\ell}^2, \quad j = 1, \ldots, p.$$

Note that the procedure can be iterated: from $\widehat{\psi}_{jj}$ we can compute a new reduced correlation matrix $\mathcal{R} - \widehat{\Psi}$ following the same procedure. The iteration usually stops when the $\widehat{\psi}_{jj}$ have converged to a stable value.

*Example 11.4* Consider once again the car data given in Table B.7. From Exercise 10.4 we know that the first PC is mainly influenced by $X_2$–$X_7$. Moreover, we know that most of the variance is already captured by the first PC. Thus we can conclude that the data are mainly determined by one factor ($k = 1$).

The eigenvalues of $\mathcal{R} - \widehat{\Psi}$ for $\widehat{\Psi} = (\max_{j \neq i} |r_{X_i X_j}|)$ are

$$(5.448, 0.003, -.246, -0.646, -0.901, -0.911, -0.948, -0.964)^\top.$$

It would suffice to choose only one factor. Nevertheless, we have computed two factors. The result (the factor loadings for two factors) is shown in Figure 11.1.

We can clearly see a cluster of points to the right, which contain the factor loadings for the variables $X_2$–$X_7$. This shows, as did the PCA, that these variables are highly dependent and are thus more or less equivalent. The factor loadings for $X_1$ (economy) and $X_8$ (easy handling) are separate, but note the different scales on the horizontal and vertical axes! Although there are two or three sets of variables in the plot, the variance is already explained by the first factor, the "price+security" factor.



**Fig. 11.1** Loadings of the evaluated car qualities, factor analysis with $k = 2$  MVAfactcarm
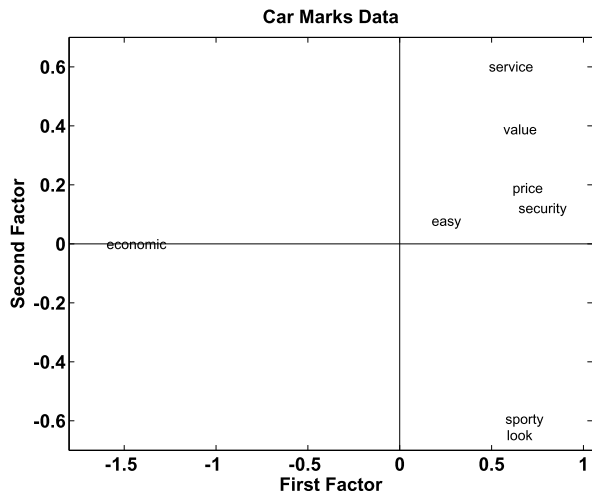
## *The Principal Component Method*

The *principal factor method* involves finding an approximation $\tilde{\Psi}$ of $\Psi$, the matrix of specific variances, and then correcting $\mathcal{R}$, the correlation matrix of $X$, by $\tilde{\Psi}$. The *principal component method* starts with an approximation $\hat{\mathcal{Q}}$ of $\mathcal{Q}$, the factor loadings matrix. The sample covariance matrix is diagonalized, $\mathcal{S} = \Gamma \Lambda \Gamma^{\top}$. Then the first $k$ eigenvectors are retained to build

$$\hat{\mathcal{Q}} = [\sqrt{\lambda_1} \gamma_1, \ldots, \sqrt{\lambda_k} \gamma_k]. \tag{11.16}$$

The estimated specific variances are provided by the diagonal elements of the matrix $\mathcal{S} - \hat{\mathcal{Q}} \hat{\mathcal{Q}}^{\top}$,

$$\hat{\Psi} = \begin{pmatrix} \hat{\psi}_{11} & & & 0 \\ & \hat{\psi}_{22} & & \\ & & \ddots & \\ 0 & & & \hat{\psi}_{pp} \end{pmatrix} \quad \text{with } \hat{\psi}_{jj} = s_{X_j X_j} - \sum_{\ell=1}^{k} \hat{q}_{j\ell}^2. \tag{11.17}$$

By definition, the diagonal elements of $\mathcal{S}$ are equal to the diagonal elements of $\hat{\mathcal{Q}} \hat{\mathcal{Q}}^{\top} + \hat{\Psi}$. The off-diagonal elements are not necessarily estimated. How good then is this approximation? Consider the residual matrix

$$\mathcal{S} - (\hat{\mathcal{Q}} \hat{\mathcal{Q}}^{\top} + \hat{\Psi})$$

resulting from the principal component solution. Analytically we have that

$$\sum_{i,j} (\mathcal{S} - \hat{\mathcal{Q}} \hat{\mathcal{Q}}^{\top} - \hat{\Psi})_{ij}^2 \leq \lambda_{k+1}^2 + \cdots + \lambda_p^2.$$

This implies that a small value of the neglected eigenvalues can result in a small approximation error. A heuristic device for selecting the number of factors is to consider the proportion of the total sample variance due to the $j$-th factor. This quantity is in general equal to

(A)  $\lambda_j / \sum_{j=1}^{p} s_{jj}$ for a factor analysis of $\mathcal{S}$,
(B)  $\lambda_j / p$ for a factor analysis of $\mathcal{R}$.

*Example 11.5* This example uses a consumer-preference study from Johnson and Wichern (1998). Customers were asked to rate several attributes of a new product. The responses were tabulated and the following correlation matrix $\mathcal{R}$ was constructed:

| Attribute (*Variable*) | | | | | | |
|---|---|---|---|---|---|---|
| Taste | 1 | 1.00 | 0.02 | **0.96** | 0.42 | 0.01 |
| Good buy for money | 2 | 0.02 | 1.00 | 0.13 | 0.71 | **0.85** |
| Flavor | 3 | 0.96 | 0.13 | 1.00 | 0.50 | 0.11 |
| Suitable for snack | 4 | 0.42 | 0.71 | 0.50 | 1.00 | **0.79** |
| Provides lots of energy | 5 | 0.01 | 0.85 | 0.11 | **0.79** | 1.00 |

**Table 11.1** Estimated factor loadings, communalities, and specific variances

| Variable | Estimated factor loadings | | Communalities | Specific variances |
|---|---|---|---|---|
| | $\hat{q}_1$ | $\hat{q}_2$ | $\hat{h}_j^2$ | $\hat{\psi}_{jj} = 1 - \hat{h}_j^2$ |
| 1. Taste | 0.56 | 0.82 | 0.98 | 0.02 |
| 2. Good buy for money | 0.78 | −0.53 | 0.88 | 0.12 |
| 3. Flavor | 0.65 | 0.75 | 0.98 | 0.02 |
| 4. Suitable for snack | 0.94 | −0.11 | 0.89 | 0.11 |
| 5. Provides lots of energy | 0.80 | −0.54 | 0.93 | 0.07 |
| Eigenvalues | 2.85 | 1.81 | | |
| Cumulative proportion of total (standardized) sample variance | 0.571 | 0.932 | | |

The bold entries of $\mathcal{R}$ show that variables 1 and 3 and variables 2 and 5 are highly correlated. Variable 4 is more correlated with variables 2 and 5 than with variables 1 and 3. Hence, a model with 2 (or 3) factors seems to be reasonable.

The first two eigenvalues $\lambda_1 = 2.85$ and $\lambda_2 = 1.81$ of $\mathcal{R}$ are the only eigenvalues greater than one. Moreover, $k = 2$ common factors account for a cumulative proportion

$$\frac{\lambda_1 + \lambda_2}{p} = \frac{2.85 + 1.81}{5} = 0.93$$

of the total (standardized) sample variance. Using the principal component method, the estimated factor loadings, communalities, and specific variances, are calculated from formulas (11.16) and (11.17), and the results are given in Table 11.1.

Take a look at:

$$\hat{Q}\hat{Q}^\top + \hat{\Psi} = \begin{pmatrix} 0.56 & 0.82 \\ 0.78 & -0.53 \\ 0.65 & 0.75 \\ 0.94 & -0.11 \\ 0.80 & -0.54 \end{pmatrix} \begin{pmatrix} 0.56 & 0.78 & 0.65 & 0.94 & 0.80 \\ 0.82 & -0.53 & 0.75 & -0.11 & -0.54 \end{pmatrix}$$

$$+ \begin{pmatrix} 0.02 & 0 & 0 & 0 & 0 \\ 0 & 0.12 & 0 & 0 & 0 \\ 0 & 0 & 0.02 & 0 & 0 \\ 0 & 0 & 0 & 0.11 & 0 \\ 0 & 0 & 0 & 0 & 0.07 \end{pmatrix}$$

$$= \begin{pmatrix} 1.00 & 0.01 & 0.97 & 0.44 & 0.00 \\ 0.01 & 1.00 & 0.11 & 0.79 & 0.91 \\ 0.97 & 0.11 & 1.00 & 0.53 & 0.11 \\ 0.44 & 0.79 & 0.53 & 1.00 & 0.81 \\ 0.00 & 0.91 & 0.11 & 0.81 & 1.00 \end{pmatrix}.$$

This nearly reproduces the correlation matrix $\mathcal{R}$. We conclude that the two-factor model provides a good fit of the data. The communalities $(0.98, 0.88, 0.98, 0.89, 0.93)$ indicate that the two factors account for a large percentage of the sample variance of each variable. Due to the nonuniqueness of factor loadings, the interpretation might be enhanced by rotation. This is the topic of the next subsection.

### *Rotation*

The constraints (11.11) and (11.12) are given as a matter of mathematical convenience (to create unique solutions) and can therefore complicate the problem of interpretation. The interpretation of the loadings would be very simple if the variables could be split into disjoint sets, each being associated with one factor. A well known analytical algorithm to rotate the loadings is given by the *varimax rotation method* proposed by Kaiser (1985). In the simplest case of $k = 2$ factors, a rotation matrix $\mathcal{G}$ is given by

$$\mathcal{G}(\theta) = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix},$$

representing a clockwise rotation of the coordinate axes by the angle $\theta$. The corresponding rotation of loadings is calculated via $\hat{\mathcal{Q}}^* = \hat{\mathcal{Q}}\mathcal{G}(\theta)$. The idea of the *varimax method* is to find the angle $\theta$ that maximizes the sum of the variances of the squared loadings $\hat{q}_{ij}^*$ within each column of $\hat{\mathcal{Q}}^*$. More precisely, defining $\tilde{q}_{jl} = \hat{q}_{jl}^*/\hat{h}_j^*$, the *varimax criterion* chooses $\theta$ so that

$$V = \frac{1}{p} \sum_{\ell=1}^{k} \left[ \sum_{j=1}^{p} (\tilde{q}_{jl}^*)^4 - \left\{ \frac{1}{p} \sum_{j=1}^{p} (\tilde{q}_{jl}^*)^2 \right\}^2 \right]$$

is maximized.

*Example 11.6* Let us return to the marketing example of Johnson and Wichern (1998) (Example 11.5). The basic factor loadings given in Table 11.1 of the first factor and a second factor are almost identical making it difficult to interpret the factors. Applying the varimax rotation we obtain the loadings $\tilde{q}_1 = (0.02, \mathbf{0.94}, 0.13, \mathbf{0.84}, \mathbf{0.97})^\top$ and $\tilde{q}_2 = (\mathbf{0.99}, -0.01, \mathbf{0.98}, 0.43, -0.02)^\top$. The high loadings, indicated as bold entries, show that variables 2, 4, 5 define factor 1, a nutricional factor. Variables 1 and 3 define factor 2 which might be referred to as a taste factor.
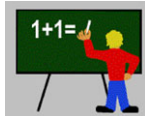
**Summary**

$\hookrightarrow$    In practice, $\mathcal{Q}$ and $\Psi$ have to be estimated from $\mathcal{S} = \hat{\mathcal{Q}}\hat{\mathcal{Q}}^\top + \hat{\Psi}$. The number of parameters is $d = \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k)$.

| Summary (continued) |
|---|
| ↪ If $d = 0$, then there exists an exact solution. In practice, $d$ is usually greater than 0, thus approximations must be considered. |
| ↪ The maximum-likelihood method assumes a normal distribution for the data. A solution can be found using numerical algorithms. |
| ↪ The method of principal factors is a two-stage method which calculates $\widehat{\mathcal{Q}}$ from the reduced correlation matrix $\mathcal{R} - \widetilde{\Psi}$, where $\widetilde{\Psi}$ is a pre-estimate of $\Psi$. The final estimate of $\Psi$ is found by $\widehat{\psi}_{ii} = 1 - \sum_{j=1}^{k} \widehat{q}_{ij}^2$. |
| ↪ The principal component method is based on an approximation, $\widehat{\mathcal{Q}}$, of $\mathcal{Q}$. |
| ↪ Often a more informative interpretation of the factors can be found by rotating the factors. |
| ↪ The varimax rotation chooses a rotation $\theta$ that maximizes $\mathcal{V} = \frac{1}{p} \sum_{\ell=1}^{k} \left[ \sum_{j=1}^{p} (\tilde{q}_{jl}^*)^4 - \left\{ \frac{1}{p} \sum_{j=1}^{p} (\tilde{q}_{jl}^*)^2 \right\}^2 \right]$. |

## 11.3  Factor Scores and Strategies

Up to now strategies have been presented for factor analysis that have concentrated on the estimation of loadings and communalities and on their interpretations. This was a logical step since the factors $F$ were considered to be normalized random sources of information and were explicitly addressed as nonspecific (common factors). The estimated values of the factors, called the *factor scores*, may also be useful in the interpretation as well as in the diagnostic analysis. To be more precise, the factor scores are estimates of the unobserved random vectors $F_l$, $l = 1, \ldots, k$, for each individual $x_i$, $i = 1, \ldots, n$. Johnson and Wichern (1998) describe three methods which in practice yield very similar results. Here, we present the regression method which has the advantage of being the simplest technique and is easy to implement.

The idea is to consider the joint distribution of $(X - \mu)$ and $F$, and then to proceed with the regression analysis presented in Chapter 5. Under the factor model (11.4), the joint covariance matrix of $(X - \mu)$ and $F$ is:

$$\mathsf{Var} \begin{pmatrix} X - \mu \\ F \end{pmatrix} = \begin{pmatrix} \mathcal{Q}\mathcal{Q}^\top + \Psi & \mathcal{Q} \\ \mathcal{Q}^\top & \mathcal{I}_k \end{pmatrix}. \tag{11.18}$$

Note that the upper left entry of this matrix equals $\Sigma$ and that the matrix has size $(p + k) \times (p + k)$.

Assuming joint normality, the conditional distribution of $F|X$ is multinormal, see Theorem 5.1, with

$$\mathsf{E}(F|X = x) = \mathcal{Q}^\top \Sigma^{-1}(X - \mu) \tag{11.19}$$

and using (5.7) the covariance matrix can be calculated:

$$\mathsf{Var}(F|X = x) = \mathcal{I}_k - \mathcal{Q}^\top \Sigma^{-1} \mathcal{Q}. \tag{11.20}$$

In practice, we replace the unknown $\mathcal{Q}$, $\Sigma$ and $\mu$ by corresponding estimators, leading to the estimated individual factor scores:

$$\widehat{f}_i = \widehat{\mathcal{Q}}^{\top} \mathcal{S}^{-1}(x_i - \overline{x}). \tag{11.21}$$

We prefer to use the original sample covariance matrix $\mathcal{S}$ as an estimator of $\Sigma$, instead of the factor analysis approximation $\widehat{\mathcal{Q}}\widehat{\mathcal{Q}}^{\top} + \widehat{\Psi}$, in order to be more robust against incorrect determination of the number of factors.

The same rule can be followed when using $\mathcal{R}$ instead of $\mathcal{S}$. Then (11.18) remains valid when standardized variables, i.e., $Z = \mathcal{D}_\Sigma^{-1/2}(X - \mu)$, are considered if $\mathcal{D}_\Sigma = \text{diag}(\sigma_{11}, \ldots, \sigma_{pp})$. In this case the factors are given by

$$\widehat{f}_i = \widehat{\mathcal{Q}}^{\top} \mathcal{R}^{-1}(z_i), \tag{11.22}$$

where $z_i = \mathcal{D}_S^{-1/2}(x_i - \overline{x})$, $\widehat{\mathcal{Q}}$ is the loading obtained with the matrix $\mathcal{R}$, and $\mathcal{D}_S = \text{diag}(s_{11}, \ldots, s_{pp})$.

If the factors are rotated by the orthogonal matrix $\mathcal{G}$, the factor scores have to be rotated accordingly, that is

$$\widehat{f}_i^* = \mathcal{G}^{\top} \widehat{f}_i. \tag{11.23}$$

A practical example is presented in Section 11.4 using the Boston Housing data.

## *Practical Suggestions*

No one method outperforms another in the practical implementation of factor analysis. However, by applying a *tâtonnement* process, the factor analysis view of the data can be stabilized. This motivates the following procedure.

1. Fix a reasonable number of factors, say $k = 2$ or $3$, based on the correlation structure of the data and/or screeplot of eigenvalues.
2. Perform several of the presented methods, including rotation. Compare the loadings, communalities, and factor scores from the respective results.
3. If the results show significant deviations, check for outliers (based on factor scores), and consider changing the number of factors $k$.

For larger data sets, cross-validation methods are recommended. Such methods involve splitting the sample into a training set and a validation data set. On the training sample one estimates the factor model with the desired methodology and uses the obtained parameters to predict the factor scores for the validation data set. The predicted factor scores should be comparable to the factor scores obtained using only the validation data set. This stability criterion may also involve the loadings and communalities.

## *Factor Analysis versus PCA*

Factor analysis and principal component analysis use the same set of mathematical tools (spectral decomposition, projections, …). One could conclude, on first sight,

that they share the same view and strategy and therefore yield very similar results. This is not true. There are substantial differences between these two data analysis techniques that we would like to describe here.

The biggest difference between PCA and factor analysis comes from the model philosophy. Factor analysis imposes a strict structure of a fixed number of common (latent) factors whereas the PCA determines $p$ factors in decreasing order of importance. The most important factor in PCA is the one that maximizes the projected variance. The most important factor in factor analysis is the one that (after rotation) gives the maximal interpretation. Often this is different from the direction of the first principal component.

From an implementation point of view, the PCA is based on a well-defined, unique algorithm (spectral decomposition), whereas fitting a factor analysis model involves a variety of numerical procedures. The non-uniqueness of the factor analysis procedure opens the door for subjective interpretation and yields therefore a spectrum of results. This data analysis philosophy makes factor analysis difficult especially if the model specification involves cross-validation and a data-driven selection of the number of factors.

## 11.4  Boston Housing

To illustrate how to implement factor analysis we will use the Boston housing data set and the by now well known set of transformations. Once again, the variable $X_4$ (Charles River indicator) will be excluded. As before, standardized variables are used and the analysis is based on the correlation matrix.

In Section 11.3, we described a practical implementation of factor analysis. Based on principal components, three factors were chosen and factor analysis was applied using the maximum likelihood method (MLM), the principal factor method (PFM), and the principal component method (PCM). For illustration, the MLM will be presented with and without varimax rotation.

Table 11.2 gives the MLM factor loadings without rotation and Table 11.3 gives the varimax version of this analysis. The corresponding graphical representations of the loadings are displayed in Figures 11.2 and 11.3. We can see that the varimax does not significantly change the interpretation of the factors obtained by the MLM. Factor 1 can be roughly interpreted as a "quality of life factor" because it is positively correlated with variables like $X_{11}$ and negatively correlated with $X_8$, both having low specific variances. The second factor may be interpreted as a "residential factor", since it is highly correlated with variables $X_6$, and $X_{13}$. The most striking difference between the results with and without varimax rotation can be seen by comparing the lower left corners of Figures 11.2 and 11.3. There is a clear separation of the variables in the varimax version of the MLM. Given this arrangement of the variables in Figure 11.3, we can interpret factor 3 as an employment factor, since we observe high correlations with $X_8$ and $X_5$.

We now turn to the PCM and PFM analyses. The results are presented in Tables 11.4 and 11.5 and in Figures 11.4 and 11.5. We would like to focus on the

**Table 11.2** Estimated factor loadings, communalities, and specific variances, MLM Q
MVAfacthous

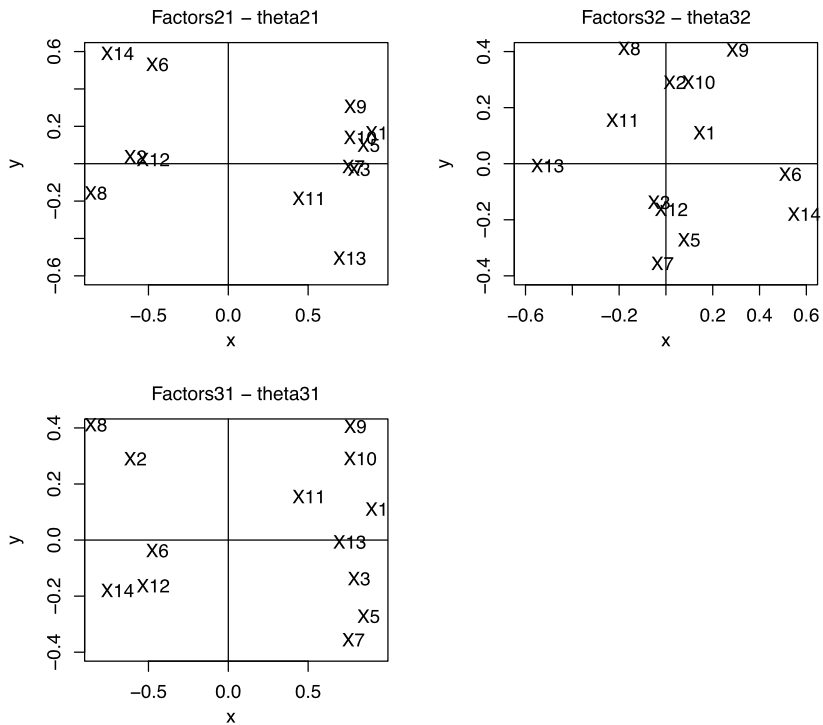| | Estimated factor loadings | | | Communalities | Specific variances |
|---|---|---|---|---|---|
| | $\hat{q}_1$ | $\hat{q}_2$ | $\hat{q}_3$ | $\hat{h}_j^2$ | $\hat{\psi}_{jj} = 1 - \hat{h}_j^2$ |
| 1. crime | 0.9295 | 0.1653 | 0.1107 | 0.9036 | 0.0964 |
| 2. large lots | −0.5823 | 0.0379 | 0.2902 | 0.4248 | 0.5752 |
| 3. nonretail acres | 0.8192 | −0.0296 | −0.1378 | 0.6909 | 0.3091 |
| 5. nitric oxides | 0.8789 | 0.0987 | −0.2719 | 0.8561 | 0.1439 |
| 6. rooms | −0.4447 | 0.5311 | −0.0380 | 0.4812 | 0.5188 |
| 7. prior 1940 | 0.7837 | −0.0149 | −0.3554 | 0.7406 | 0.2594 |
| 8. empl. centers | −0.8294 | −0.1570 | 0.4110 | 0.8816 | 0.1184 |
| 9. accessibility | 0.7955 | 0.3062 | 0.4053 | 0.8908 | 0.1092 |
| 10. tax-rate | 0.8262 | 0.1401 | 0.2906 | 0.7867 | 0.2133 |
| 11. pupil/teacher | 0.5051 | −0.1850 | 0.1553 | 0.3135 | 0.6865 |
| 12. African American | 0.4701 | −0.0227 | −0.1627 | 0.2480 | 0.7520 |
| 13. lower status | 0.7601 | −0.5059 | −0.0070 | 0.8337 | 0.1663 |
| 14. value | −0.6942 | 0.5904 | −0.1798 | 0.8628 | 0.1371 |



**Fig. 11.2** Factor analysis for Boston housing data, MLM Q  MVAfacthous

**Table 11.3** Estimated factor loadings, communalities, and specific variances, MLM, varimax rotation ![Q] MVAfacthous

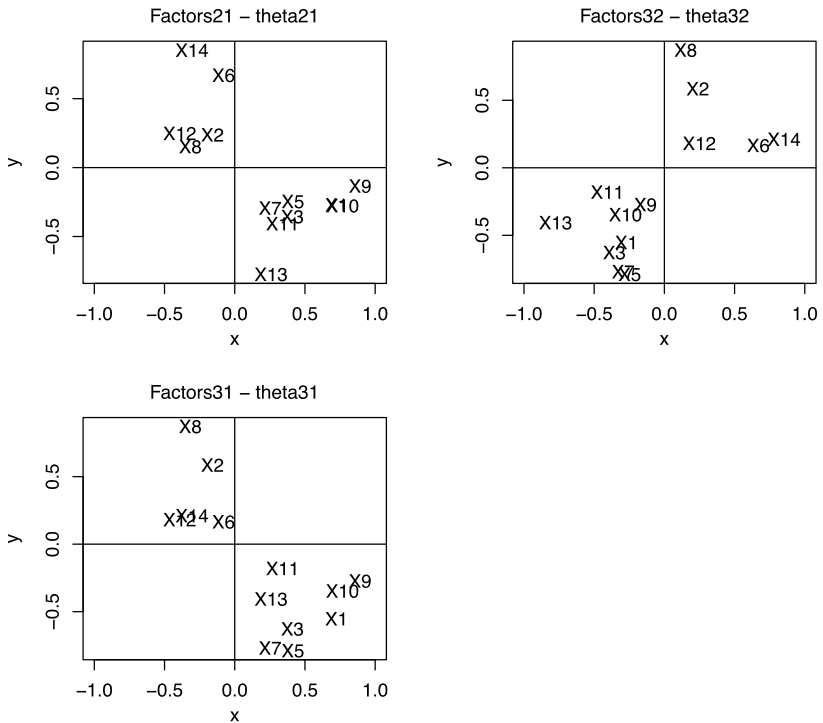| | Estimated factor loadings | | | Communalities | Specific variances |
|---|---|---|---|---|---|
| | $\hat{q}_1$ | $\hat{q}_2$ | $\hat{q}_3$ | $\hat{h}_j^2$ | $\hat{\psi}_{jj} = 1 - \hat{h}_j^2$ |
| 1. crime | 0.7247 | −0.2705 | −0.5525 | 0.9036 | 0.0964 |
| 2. large lots | −0.1570 | 0.2377 | 0.5858 | 0.4248 | 0.5752 |
| 3. nonretail acres | 0.4195 | −0.3566 | −0.6287 | 0.6909 | 0.3091 |
| 5. nitric oxides | 0.4141 | −0.2468 | −0.7896 | 0.8561 | 0.1439 |
| 6. rooms | −0.0799 | 0.6691 | 0.1644 | 0.4812 | 0.5188 |
| 7. prior 1940 | 0.2518 | −0.2934 | −0.7688 | 0.7406 | 0.2594 |
| 8. empl. centers | −0.3164 | 0.1515 | 0.8709 | 0.8816 | 0.1184 |
| 9. accessibility | 0.8932 | −0.1347 | −0.2736 | 0.8908 | 0.1092 |
| 10. tax-rate | 0.7673 | −0.2772 | −0.3480 | 0.7867 | 0.2133 |
| 11. pupil/teacher | 0.3405 | −0.4065 | −0.1800 | 0.3135 | 0.6865 |
| 12. African American | −0.3917 | 0.2483 | 0.1813 | 0.2480 | 0.7520 |
| 13. lower status | 0.2586 | −0.7752 | −0.4072 | 0.8337 | 0.1663 |
| 14. value | −0.3043 | 0.8520 | 0.2111 | 0.8630 | 0.1370 |



**Fig. 11.3** Factor analysis for Boston housing data, MLM after varimax rotation ![Q] MVAfacthous

**Table 11.4** Estimated factor loadings, communalities, and specific variances, PCM, varimax rotation 🔍 MVAfacthous

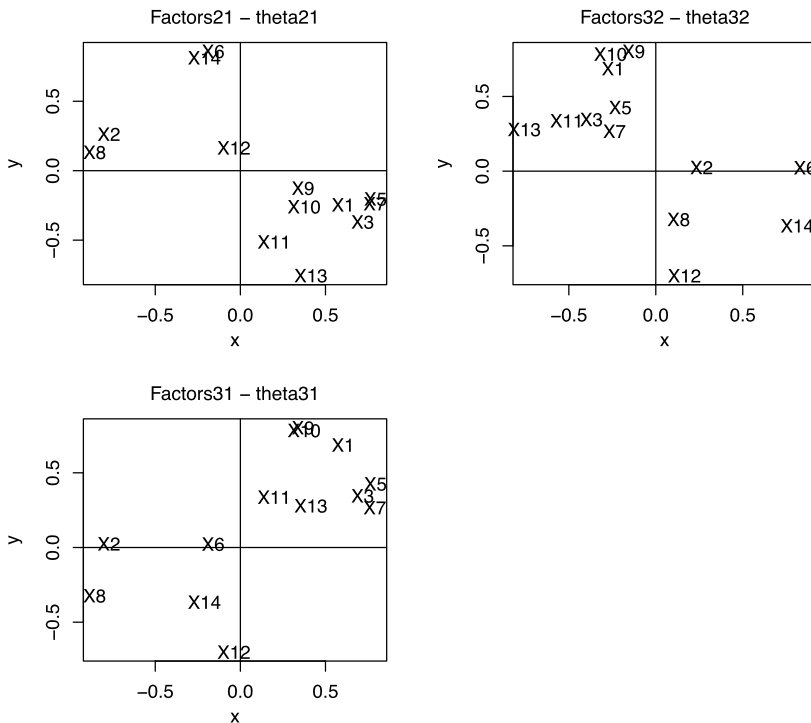|  | Estimated factor loadings | | | Communalities | Specific variances |
|---|---|---|---|---|---|
|  | $\hat{q}_1$ | $\hat{q}_2$ | $\hat{q}_3$ | $\hat{h}_j^2$ | $\hat{\psi}_{jj} = 1 - \hat{h}_j^2$ |
| 1. crime | 0.6034 | −0.2456 | 0.6864 | 0.8955 | 0.1045 |
| 2. large lots | −0.7722 | 0.2631 | 0.0270 | 0.6661 | 0.3339 |
| 3. nonretail acres | 0.7183 | −0.3701 | 0.3449 | 0.7719 | 0.2281 |
| 5. nitric oxides | 0.7936 | −0.2043 | 0.4250 | 0.8521 | 0.1479 |
| 6. rooms | −0.1601 | 0.8585 | 0.0218 | 0.7632 | 0.2368 |
| 7. prior 1940 | 0.7895 | −0.2375 | 0.2670 | 0.7510 | 0.2490 |
| 8. empl. centers | −0.8562 | 0.1318 | −0.3240 | 0.8554 | 0.1446 |
| 9. accessibility | 0.3681 | −0.1268 | 0.8012 | 0.7935 | 0.2065 |
| 10. tax-rate | 0.3744 | −0.2604 | 0.7825 | 0.8203 | 0.1797 |
| 11. pupil/teacher | 0.1982 | −0.5124 | 0.3372 | 0.4155 | 0.5845 |
| 12. African American | 0.1647 | 0.0368 | −0.7002 | 0.5188 | 0.4812 |
| 13. lower status | 0.4141 | −0.7564 | 0.2781 | 0.8209 | 0.1791 |
| 14. value | −0.2111 | 0.8131 | −0.3671 | 0.8394 | 0.1606 |



**Fig. 11.4** Factor analysis for Boston housing data, PCM after varimax rotation 🔍 MVAfacthous

**Table 11.5** Estimated factor loadings, communalities, and specific variances, PFM, varimax rotation Q MVAfacthous

| | Estimated factor loadings | | | Communalities | Specific variances |
|---|---|---|---|---|---|
| | $\hat{q}_1$ | $\hat{q}_2$ | $\hat{q}_3$ | $\hat{h}_j^2$ | $\hat{\psi}_{jj} = 1 - \hat{h}_j^2$ |
| 1. crime | 0.5477 | −0.2558 | −0.7387 | 0.9111 | 0.0889 |
| 2. large lots | −0.6148 | 0.2668 | 0.1281 | 0.4655 | 0.5345 |
| 3. nonretail acres | 0.6523 | −0.3761 | −0.3996 | 0.7266 | 0.2734 |
| 5. nitric oxides | 0.7723 | −0.2291 | −0.4412 | 0.8439 | 0.1561 |
| 6. rooms | −0.1732 | 0.6783 | 0.1296 | 0.0699 | 0.5046 |
| 7. prior 1940 | 0.7390 | −0.2723 | −0.2909 | 0.7049 | 0.2951 |
| 8. empl. centers | −0.8565 | 0.1485 | 0.3395 | 0.8708 | 0.1292 |
| 9. accessibility | 0.2855 | −0.1359 | −0.8460 | 0.8156 | 0.1844 |
| 10. tax-rate | 0.3062 | −0.2656 | −0.8174 | 0.8325 | 0.1675 |
| 11. pupil/teacher | 0.2116 | −0.3943 | −0.3297 | 0.3090 | 0.6910 |
| 12. African American | 0.1994 | 0.0666 | 0.4217 | 0.2433 | 0.7567 |
| 13. lower status | 0.4005 | −0.7743 | −0.2706 | 0.8333 | 0.1667 |
| 14. value | −0.1885 | 0.8400 | 0.3473 | 0.8611 | 0.1389 |
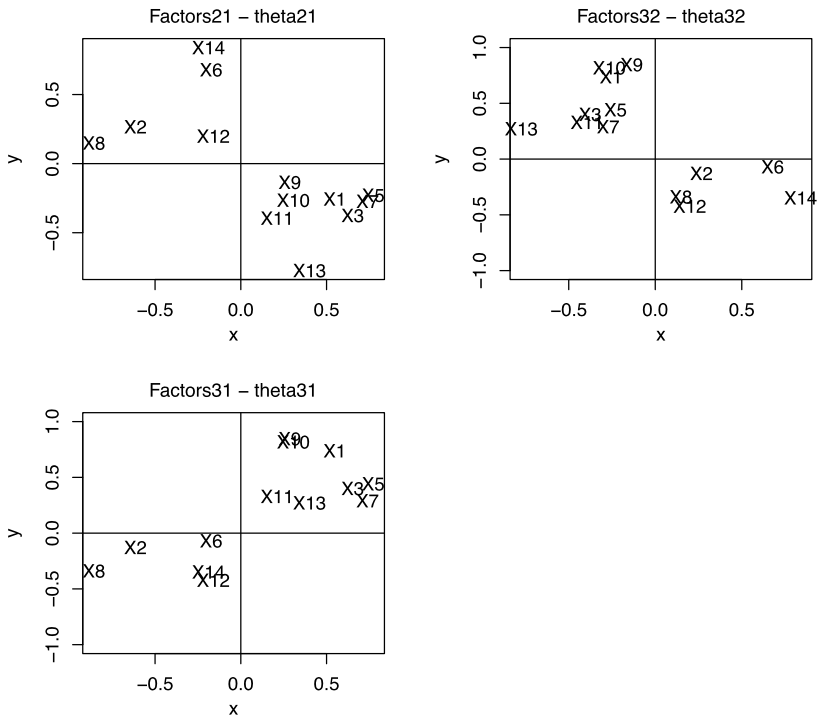


**Fig. 11.5** Factor analysis for Boston housing data, PFM after varimax rotation Q MVAfacthous

PCM, because this 3-factor model yields only one specific variance (unexplained variation) above 0.5. Looking at Figure 11.4, it turns out that factor 1 remains a "quality of life factor" which is clearly visible from the clustering of $X_5$, $X_3$, $X_{10}$ and $X_1$ on the right-hand side of the graph, while the variables $X_8$, $X_2$, $X_{14}$, $X_{12}$ and $X_6$ are on the left-hand side. Again, the second factor is a "residential factor", clearly demonstrated by the location of variables $X_6$, $X_{14}$, $X_{11}$, and $X_{13}$. The interpretation of the third factor is more difficult because all of the loadings (except for $X_{12}$) are very small.

## 11.5 Exercises

**Exercise 11.1** In Example 11.4 we have computed $\widehat{\mathcal{Q}}$ and $\widehat{\Psi}$ using the method of principal factors. We used a two-step iteration for $\widehat{\Psi}$. Perform the third iteration step and compare the results (i.e., use the given $\widehat{\mathcal{Q}}$ as a pre-estimate to find the final $\Psi$).

**Exercise 11.2** Using the bank data set, how many factors can you find with the Method of Principal Factors?

**Exercise 11.3** Repeat Exercise 11.2 with the U.S. company data set!

**Exercise 11.4** Generalize the two-dimensional rotation matrix in Section 11.2 to $n$-dimensional space.

**Exercise 11.5** Compute the orthogonal factor model for

$$\Sigma = \begin{pmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{pmatrix}.$$

[Solution: $\psi_{11} = -0.575$, $q_{11} = 1.255$]

**Exercise 11.6** Perform a factor analysis on the type of families in the French food data set. Rotate the resulting factors in a way which provides the most reasonable interpretation. Compare your result with the varimax method.

**Exercise 11.7** Perform a factor analysis on the variables $X_3$ to $X_9$ in the U.S. crime data set (Table B.10). Would it make sense to use all of the variables for the analysis?

**Exercise 11.8** Analyze the athletic records data set (Table B.18). Can you recognize any patterns if you sort the countries according to the estimates of the factor scores?

**Exercise 11.9** Perform a factor analysis on the U.S. health data set (Table B.16) and estimate the factor scores.

**Exercise 11.10** Redo Exercise 11.9 using the U.S. crime data in Table B.10. Compare the estimated factor scores of the two data sets.

**Exercise 11.11** Analyze the vocabulary data given in Table B.17.