Chapter 2: Factor Analysis[a]

Factor models (FMs) are statistical models that attempt to explain the correlation amongst the observed variables $x^0_{(1)}, \ldots, x^0_{(p)}$ via a small number $k < p$ of unobservable factors.

Like PCA, FMs can be used to obtain a low dimensional representation of the observations (where low means k-dimensional).

FMs are also useful to infer unobservable variables or abstract concepts from proxy variables, such as

- intelligence from test results,
- 'Big Five' personality traits from personality surveys,
- salient attributes determining perception of a product.

Throughout this chapter we assume that the observations $\{x^0_i\}_{i=1}^n$ are n realizations of an $\mathbb{R}^p$-valued random variable X^0.

[a]The main reference for this chapter is [2, Chapter 9].

The factor model

For $k < p$ the k-factor model assumes that the random variable $X := X^0 - \mathbb{E}[X^0]$ is such that $X = \mathbf{\Lambda} F + U$ where

- $\mathbf{\Lambda} = [\lambda_{jl}]$ is a $p \times k$ matrix of constants, called loadings matrix,
- F is an $\mathbb{R}^k$-valued random variable, called factor, such that $\mathbb{E}[F] = 0$ and such that $\mathrm{Var}(F) = \boldsymbol{I}_k$,
- U is an $\mathbb{R}^p$-valued random variable such that $\mathbb{E}[U] = 0$ and such that $\mathrm{Var}(U) = \mathbf{\Psi}$ for some matrix $\mathbf{\Psi} = \mathrm{diag}(\psi_{11}, \ldots, \psi_{pp})$,
- $\mathrm{Cov}(F, U) = 0$.

Remark: If $\mathrm{Var}(X)$ is full rank then the FM always holds true if $k = p$, in which case $\mathbf{\Lambda}$ and F can be obtained using population PCA[a]. This is why in this chapter we focus on the case where $k < p$.

Under the k-FM, for all $j \in \{1, \ldots, p\}$ we have for the jth variable

$$X_j = \sum_{l=1}^{k} \lambda_{jl} F_l + U_j \Rightarrow \mathrm{Var}(X_j) = \sum_{l=1}^{k} \lambda_{jl}^2 + \psi_{jj}$$

where

- $h_j^2 := \sum_{l=1}^{k} \lambda_{jl}^2$ is called communality; this is the part of the variance of X_j that is shared with the other variables through the common factor F.
- $\psi_{jj} \geq 0$ is called the specific variance; this is the part of the variance of X_j which is not shared with the other variables.

[a] To see this recall that $X = \mathbf{\Gamma} Y$ where $\mathbf{\Gamma} \in O(p)$ and where Y is such that $\mathrm{Var}(Y) = \boldsymbol{L}$, with $\boldsymbol{L}$ a diagonal matrix having the eigenvalues of $\mathrm{Var}(X)$ as non-zero entries (see Chapter 1, page 24). If $\mathrm{Var}(X)$ is full rank the matrix $\boldsymbol{L}$ is invertible and $X = (\mathbf{\Gamma} \boldsymbol{L}^{1/2})(\boldsymbol{L}^{-1/2} Y)$, showing that for $k = p$ the FM holds with $\mathbf{\Lambda} = \mathbf{\Gamma} \boldsymbol{L}^{1/2}$, $F = \boldsymbol{L}^{-1/2} Y$ and $\mathbf{\Psi} = 0 \boldsymbol{I}_p$.

Variance decomposition and the fundamental theorem of factor analysis

Let $\mathbf{\Sigma} = \text{Var}(X)$ and note that the k-factor model implies that $\mathbf{\Sigma}$ can be decomposed as follows:

$$\mathbf{\Sigma} = \mathbf{\Lambda}\text{Var}(F)\mathbf{\Lambda}^\top + \text{Var}(U) = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}. \tag{2.1}$$

Interestingly, as shown in Theorem 2.1 below, the k-factor model holds for X if and only if its variance $\mathbf{\Sigma}$ can be decomposed as in (2.1) (under a weak assumption on $\mathbf{\Sigma}$).

Theorem 2.1 (Fundamental theorem of factor analysis) *Assume that there exists a $k < p$ such that*

$$\mathbf{\Sigma} = \tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}^\top + \tilde{\mathbf{\Psi}} \tag{2.2}$$

for some $p \times k$ matrix $\tilde{\mathbf{\Lambda}}$ and some $p \times p$ diagonal matrix $\tilde{\mathbf{\Psi}}$ having strictly positive diagonal entries. Then, the k-factor model holds for X with $\mathbf{\Lambda} = \tilde{\mathbf{\Lambda}}$ and with $\mathbf{\Psi} = \tilde{\mathbf{\Psi}}$.

Remark: Theorem 2.1 notably assumes that $\mathbf{\Sigma}$ is full rank.

Remark: In Theorem 2.1, for a fix matrix $\tilde{\mathbf{\Psi}}$ the decomposition (2.2) is unique up to a rotation of $\tilde{\mathbf{\Lambda}}$; that is,

$$\mathbf{\Sigma} = \tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}^\top + \tilde{\mathbf{\Psi}} = \boldsymbol{M}\boldsymbol{M}^\top + \tilde{\mathbf{\Psi}}$$

if and only if $\boldsymbol{M} = \tilde{\mathbf{\Lambda}}\boldsymbol{G}$ for some $\boldsymbol{G} \in O(k)$ [see 2, Section 9.2.3].

Proof of Theorem 2.1

To simplify the notation let $\boldsymbol{\Lambda} = \tilde{\boldsymbol{\Lambda}}$ and $\boldsymbol{\Psi} = \tilde{\boldsymbol{\Psi}}$, so that $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$. Remark that under the assumptions of the theorem the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ are invertible.

Let

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{I}_p & \boldsymbol{\Lambda} \\ -\boldsymbol{\Lambda}^\top \boldsymbol{\Psi}^{-1} & \boldsymbol{I}_k \end{pmatrix}$$

and note that $\boldsymbol{M}$ is invertible since

$$\begin{aligned} \det(\boldsymbol{M}) &= \det(\boldsymbol{I}_k) \det\left(\boldsymbol{I}_p - \boldsymbol{\Lambda}\boldsymbol{I}_k^{-1}(-\boldsymbol{\Lambda}^\top \boldsymbol{\Psi}^{-1})\right) \\ &= \det\left((\boldsymbol{\Psi} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)\boldsymbol{\Psi}^{-1}\right) = \det(\boldsymbol{\Sigma}) \det(\boldsymbol{\Psi}^{-1}) > 0. \end{aligned}$$

Let $Y \sim \mathcal{N}_k(0, \boldsymbol{I}_k + \boldsymbol{\Lambda}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda})$ be independent of X and let

$$\begin{pmatrix} U \\ F \end{pmatrix} = \boldsymbol{M}^{-1} \begin{pmatrix} X \\ Y \end{pmatrix}.$$

Then,

$$\boldsymbol{M} \begin{pmatrix} U \\ F \end{pmatrix} = \begin{pmatrix} U + \boldsymbol{\Lambda} F \\ -\boldsymbol{\Lambda}^\top \boldsymbol{\Psi}^{-1} U + F \end{pmatrix} = \begin{pmatrix} X \\ Y \end{pmatrix}$$

from which we obtain that $X = \boldsymbol{\Lambda} F + U$.

To complete the proof it remains to verify that $\mathbb{E}[F] = 0$, $\mathbb{E}[U] = 0$, $\mathrm{Var}(U) = \boldsymbol{\Psi}$ and that $\mathrm{Cov}(F, U) = 0$.

First, we have $\mathbb{E}\big[(U, F)\big] = \boldsymbol{M}^{-1} \mathbb{E}[(X, Y)] = (0, 0)$ as required.

To proceed further recall that if $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$, are four matrices such that $\boldsymbol{D}$ and $\boldsymbol{E} := \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}$ are invertible matrices then

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{E}^{-1} & -\boldsymbol{E}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{E}^{-1} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{E}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \end{pmatrix}. \tag{2.3}$$

Then, applying (2.3) to invert $\boldsymbol{M}^{-1}$ we obtain

$$\boldsymbol{M}^{-1} = \left[\begin{pmatrix} \boldsymbol{\Psi} & \boldsymbol{\Lambda} \\ -\boldsymbol{\Lambda}^\top & \boldsymbol{I}_k \end{pmatrix} \begin{pmatrix} \boldsymbol{\Psi}^{-1} & 0 \\ 0 & \boldsymbol{I}_k \end{pmatrix} \right]^{-1} = \begin{pmatrix} \boldsymbol{\Psi} & 0 \\ 0 & \boldsymbol{I}_k \end{pmatrix} \begin{pmatrix} \boldsymbol{\Psi} & \boldsymbol{\Lambda} \\ -\boldsymbol{\Lambda}^\top & \boldsymbol{I}_k \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Psi} & 0 \\ 0 & \boldsymbol{I}_k \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}^{-1} & -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1} & \boldsymbol{I}_k - \boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \end{pmatrix}.$$

Proof of Theorem 2.1 (end)

Using the above expression for $\boldsymbol{M}^{-1}$ we have

$$\begin{aligned}
\mathrm{Var}\big((U,F)\big) &= \boldsymbol{M}^{-1}\mathrm{Var}\big((X,Y)\big)\big(\boldsymbol{M}^{-1}\big)^\top \\
&= \begin{pmatrix} \boldsymbol{\Psi} & 0 \\ 0 & \boldsymbol{I}_k \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}^{-1} & -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1} & \boldsymbol{I}_k - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma} & 0 \\ 0 & \boldsymbol{I}_k + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}^{-1} & -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1} & \boldsymbol{I}_k - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} \end{pmatrix}^\top\begin{pmatrix} \boldsymbol{\Psi} & 0 \\ 0 & \boldsymbol{I}_k \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{\Psi} & 0 \\ 0 & \boldsymbol{I}_k \end{pmatrix}\begin{pmatrix} \boldsymbol{I}_p & -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}(\boldsymbol{I}_k + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \\ \boldsymbol{\Lambda}^\top & (\boldsymbol{I}_k - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})(\boldsymbol{I}_k + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}) \end{pmatrix}\begin{pmatrix} \boldsymbol{\Sigma}^{-1} & \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} \\ -\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1} & \boldsymbol{I}_k - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} \end{pmatrix}\begin{pmatrix} \boldsymbol{\Psi} & 0 \\ 0 & \boldsymbol{I}_k \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{\Psi} & 0 \\ 0 & \boldsymbol{I}_k \end{pmatrix}\begin{pmatrix} \text{(i)} & \text{(ii)} \\ \text{(ii)}^\top & \text{(iv)} \end{pmatrix}\begin{pmatrix} \boldsymbol{\Psi} & 0 \\ 0 & \boldsymbol{I}_k \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Psi} & 0 \\ 0 & \boldsymbol{I}_k \end{pmatrix},
\end{aligned}$$

where

$$\begin{aligned}
\text{(i)} &= \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}(\boldsymbol{I}_k + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1} \\
&= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)\boldsymbol{\Sigma}^{-1} \\
&= \boldsymbol{\Sigma}^{-1}\big(\boldsymbol{\Sigma} + \boldsymbol{\Sigma} - \boldsymbol{\Psi} + (\boldsymbol{\Sigma} - \boldsymbol{\Psi})\boldsymbol{\Psi}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{\Psi})\big)\boldsymbol{\Sigma}^{-1} \\
&= 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi})\boldsymbol{\Psi}^{-1}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1}) \\
&= 2\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Psi}^{-1} - \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Psi}^{-1} \\
\text{(ii)} &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}(\boldsymbol{I}_k + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})(\boldsymbol{I}_k - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}) \\
&= \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}(\boldsymbol{I}_k + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}) \\
&= -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\big(\boldsymbol{\Psi}^{-1} - \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\big)\boldsymbol{\Lambda} \\
&= -\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{\Psi})\big(\boldsymbol{\Psi}^{-1} - \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Psi}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{\Psi})\boldsymbol{\Sigma}^{-1}\big)\boldsymbol{\Lambda} \\
&= -(\boldsymbol{I}_p - \boldsymbol{\Sigma}^{-1}\boldsymbol{\Psi})\big(\boldsymbol{\Psi}^{-1} - \boldsymbol{\Sigma}^{-1} - (\boldsymbol{\Psi}^{-1} - \boldsymbol{\Sigma}^{-1})\big)\boldsymbol{\Lambda} = 0 \\
\text{(iv)} &= \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} + (\boldsymbol{I}_k - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})(\boldsymbol{I}_k + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})(\boldsymbol{I}_k - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}) \\
&= \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} + (\boldsymbol{I}_k - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})(\boldsymbol{I}_k - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}) \\
&= \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} + \boldsymbol{I}_k - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} \\
&\quad - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} \\
&= \boldsymbol{I}_k + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} \\
&\quad + \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Psi}^{-1} + \boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1})\boldsymbol{\Lambda} \\
&= \boldsymbol{I}_k + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{\Psi})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} \\
&\quad + \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Psi}^{-1} + \boldsymbol{\Psi}^{-1}(\boldsymbol{\Sigma} - \boldsymbol{\Psi})\boldsymbol{\Sigma}^{-1})\boldsymbol{\Lambda} \\
&= \boldsymbol{I}_k + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda} - \boldsymbol{\Lambda}^\top(\boldsymbol{\Psi}^{-1} - \boldsymbol{\Sigma}^{-1})\boldsymbol{\Lambda} \\
&\quad + \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Psi}^{-1} + \boldsymbol{\Psi}^{-1} - \boldsymbol{\Sigma}^{-1})\boldsymbol{\Lambda} = \boldsymbol{I}_k.
\end{aligned}$$

The proof is complete.

Two important properties of factor analysis

- **Invariance of the factors:** Let $X' = \boldsymbol{K}X$ for some diagonal matrix $\boldsymbol{K}$. Then, if $X = \boldsymbol{\Lambda}F + U$ we have

$$X' = \boldsymbol{K}\boldsymbol{\Lambda}F + \boldsymbol{K}U, \quad \text{Var}(\boldsymbol{K}U) = \boldsymbol{K}\boldsymbol{\Psi}\boldsymbol{K}$$

  so that the k-factor model holds for X' with factor loadings $\boldsymbol{K}\boldsymbol{\Lambda}$, specific variances $\boldsymbol{K}\boldsymbol{\Psi}\boldsymbol{K}$ and the same factor F.

  $\Rightarrow$ Unlike PCA, factor analysis (FA) is unaffected by the scaling of the variables (i.e. if the variables are rescaled then the factor F remains unchanged).

- **Non-Uniqueness of the factor loadings:** For $\boldsymbol{G} \in O(k)$ we have

$$X = \boldsymbol{\Lambda}F + U = (\boldsymbol{\Lambda}\boldsymbol{G})(\boldsymbol{G}^\top F) + U$$

  where $\text{Var}(\boldsymbol{G}^\top F) = \boldsymbol{G}^\top\boldsymbol{G} = \boldsymbol{I}_k$ and $\text{Cov}(\boldsymbol{G}^\top F, U) = 0$.

  $\Rightarrow$ If the FM holds for X then the loadings matrix $\boldsymbol{\Lambda}$ and the factor F are not unique.

Following the remark made just after Theorem 2.1, for a given $\boldsymbol{\Psi}$ the factor loadings are unique up to a rotation. This indeterminacy in the definition of $\boldsymbol{\Lambda}$ can be resolved by forcing this matrix to satisfy a constraint, such as [see 2, Section 9.2.3]

$$\text{(a) } \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda} \text{ is diagonal or (b) } \boldsymbol{\Lambda}^\top\boldsymbol{D}^{-1}\boldsymbol{\Lambda} \text{ is diagonal,}$$

where in either case the diagonal elements are written e.g. in non-increasing order, and where in (b) $\boldsymbol{D}$ is some diagonal matrix.

Remark: Conditions (a)-(b) are scale invariant and in (a) we must have $\psi_{jj} > 0$ for $j = 1, \ldots, p$.

Computing the model parameters

Because of the scale invariance of FA the model parameters are often estimated from the correlation matrix $\boldsymbol{R}$ of the data (and not from $\boldsymbol{S}$). Below we follow this common practice.

The goal is then to compute $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ from $\boldsymbol{R}$ in such a way that

$$\boldsymbol{R} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi} \quad \text{where } \boldsymbol{\Lambda} \text{ satisfies (a) or (b).} \tag{2.4}$$

To see if a solution to this problem exists remark that $\boldsymbol{\Lambda}$ contains kp parameters and $\boldsymbol{\Psi}$ contains p parameters. However, under either (a) or (b) there are $k(k+1)/2 - k = k(k-1)/2$ constraints[a] on $\boldsymbol{\Lambda}$, and thus under (a) or (b) the total number of free parameters in the matrix $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$ is equal to $n_{\text{fp}} := kp + p - k(k-1)/2$. On the other hand, the equality $\boldsymbol{R} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$ imposes $n_{\text{cons}} := p(p+1)/2$ constraints on the matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$.

Let

$$\Delta_{k,p} = n_{\text{cons}} - n_{\text{fp}} = \frac{(p-k)^2}{2} - \frac{(p+k)}{2} \tag{2.5}$$

and consider the following three cases:

- If $\Delta_{k,p} = 0$ there are as many free parameters as constraints and (2.4) usually has an exact and unique solution.
- If $\Delta_{k,p} < 0$ there are more free parameters than constraints and (2.4) has infinitely many solutions[b] $\Rightarrow$ the FM is not well-defined.
- If $\Delta_{k,p} > 0$ there are more constraints than free parameters and (2.4) has no solution. In this case we look for an approximate solution.

[a] This is because both (a) and (b) imposes that all the off diagonal elements of a $k \times k$ symmetric matrix are equal to zero.

[b] Note that neither (a) nor (b) ensures that $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ are unique. Instead, under (a) or (b) $\boldsymbol{\Lambda}$ is unique given that we know a matrix $\boldsymbol{\Psi}$ such that the FM holds (see below).

Approximate solution to (2.4) when $\Delta_{k,p} > 0$

The scenario $\Delta_{k,p} > 0$ is the most frequent one in practice, in which case the objective is to compute loadings $\hat{\mathbf{\Lambda}}$ and specific variances $\hat{\mathbf{\Psi}}$ in such a way that $\boldsymbol{R} \approx \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^\top + \hat{\mathbf{\Psi}}$, where $\hat{\mathbf{\Lambda}}$ satisfies (a) or (b).

To this aim assume for now that we know a diagonal matrix $\mathbf{\Psi}$ such that the matrix $\boldsymbol{R} - \mathbf{\Psi}$ is

(i) positive semi-definite,

(ii) has rank $k < p$.

Let $\gamma_1 \geq \cdots \geq \gamma_p \geq 0$ be the eigenvalues of the matrix $\boldsymbol{R} - \mathbf{\Psi}$, $\boldsymbol{B}$ be the corresponding matrix of orthonormal eigenvectors, $\mathbf{\Gamma} = \text{diag}(\gamma_1, \ldots, \gamma_p)$ and $\mathbf{\Gamma}_k = \text{diag}(\gamma_1, \ldots, \gamma_k)$.

Then, using the fact that $\gamma_{k+1} = \cdots = \gamma_p = 0$, we have

$$\boldsymbol{R} - \mathbf{\Psi} = \boldsymbol{B}\mathbf{\Gamma}\boldsymbol{B}^\top = \boldsymbol{B}_{1:k}\mathbf{\Gamma}_k\boldsymbol{B}_{1:k}^\top = \left(\boldsymbol{B}_{1:k}\mathbf{\Gamma}_k^{1/2}\right)\left(\boldsymbol{B}_{1:k}\mathbf{\Gamma}_k^{1/2}\right)^\top$$

showing that if we know $\mathbf{\Psi}$ such that (i) and (ii) hold then the decomposition $\boldsymbol{R} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}$ holds for $\mathbf{\Lambda} = \boldsymbol{B}_{1:k}\mathbf{\Gamma}_k^{1/2}$. Moreover, condition (b) holds for $\boldsymbol{D} = \boldsymbol{I}_p$ since $\mathbf{\Lambda}^\top\mathbf{\Lambda} = \mathbf{\Gamma}_k$ when $\mathbf{\Lambda} = \boldsymbol{B}_{1:k}\mathbf{\Gamma}_k^{1/2}$.

In practice a diagonal matrix $\mathbf{\Psi}$ such that (ii) holds is of course unknown (and may not exist when $\Delta_{k,p} > 0$), and the idea of the algorithm introduced below is to alternate between computing $\mathbf{\Lambda}$ as above for a given $\mathbf{\Psi}$ and computing $\mathbf{\Psi}$ for a given matrix $\mathbf{\Lambda}$, using $\mathbf{\Psi} = \boldsymbol{R} - \mathbf{\Lambda}\mathbf{\Lambda}^\top$.

The iterated principal factor analysis algorithm

Input: A preliminary estimate $\hat{\boldsymbol{\Psi}}^{(0)}$ of $\boldsymbol{\Psi}$

for $s \geq 1$ **do**

(i) Compute the eigenvalues $\gamma_1^{(s)} \geq \cdots \geq \gamma_p^{(p)}$ of the reduced correlation matrix $\tilde{\boldsymbol{R}}^{(s)} := \boldsymbol{R} - \hat{\boldsymbol{\Psi}}^{(s-1)}$ and a corresponding set $\{v_j^{(s)}\}_{j=1}^p$ of orthonormal eigenvectors.

(ii) Set $\hat{\boldsymbol{\Lambda}}^{(s)} = [\hat{\lambda}_{jl}^{(s)}] = \boldsymbol{B}^{(s)}\big(\boldsymbol{\Gamma}^{(s)}\big)^{1/2}$ where

$$\boldsymbol{B}^{(s)} = [v_1^{(s)} \dots v_k^{(s)}], \quad \boldsymbol{\Gamma}^{(s)} = \operatorname{diag}(\gamma_1^{(s)}, \dots, \gamma_k^{(s)}).$$

(iii) Let $\hat{\boldsymbol{\Psi}}^{(s)} = \operatorname{diag}(\psi_{11}^{(s)}, \dots, \psi_{pp}^{(s)})$ where

$$\hat{\psi}_{jj}^{(s)} = \max\Big\{1 - \sum_{l=1}^{k} (\hat{\lambda}_{jl}^{(s)})^2, 0\Big\}, \quad \forall j \in \{1, \dots, p\}.$$

if Convergence=TRUE **then**

(iv) **return** $\hat{\boldsymbol{\Lambda}}^{(s)}$ and $\hat{\boldsymbol{\Psi}}^{(s)}$.

(v) **break**

end if

end for

Remark: The above definition of $\hat{\boldsymbol{\Psi}}^{(s)}$ ensures that the equality $\boldsymbol{R} = \hat{\boldsymbol{\Lambda}}^{(s)}(\hat{\boldsymbol{\Lambda}}^{(s)})^\top + \hat{\boldsymbol{\Psi}}^{(s)}$ holds at least for the diagonal elements.

Remark: A standard choice for $\hat{\boldsymbol{\Psi}}^{(0)}$ is

$$\hat{\boldsymbol{\Psi}}^{(0)} = \operatorname{diag}\Big(1 - \max_{l \neq 1} |r_{1l}|, \dots, 1 - \max_{l \neq p} |r_{pl}|\Big)^{\text{a}}.$$

[a] Another possible choice is $\hat{\boldsymbol{\Psi}}^{(0)} = \operatorname{diag}\big(1 - \sum_{l \neq 1} |r_{1l}|, \dots, 1 - \sum_{l \neq p} |r_{1p}|\big)$. In this case, the initial reduced correlation matrix $\tilde{\boldsymbol{R}}^{(1)}$ is diagonal domiant and symmetric, and thus is positive semi-definite.

Varimax rotation to increase interpretability

As we saw above, the FM is defined up to a rotation of the loadings matrix $\boldsymbol{\Lambda}$, and the aim of the varimax rotation is to find a rotation that makes the estimated FM easy to interpret.

Let $(\delta_{jl}^{(\boldsymbol{G})}) = \boldsymbol{\Lambda}\boldsymbol{G}$ be the loadings obtained by rotating $\boldsymbol{\Lambda}$ using the matrix $\boldsymbol{G} \in O(p)$, and let $(\tilde{F}_1^{(\boldsymbol{G})}, \ldots, \tilde{F}_k^{(\boldsymbol{G})}) = \boldsymbol{G}^\top F$ be the vector containing the corresponding rotated factors.

Then, the varimax rotation of the loadings $\boldsymbol{\Lambda}$ is the matrix $\boldsymbol{G}_* \in O(p)$ that maximises the function

$$O(p) \ni \boldsymbol{G} \mapsto \sum_{l=1}^{k} \sum_{j=1}^{p} \Big(d_{jl}(\boldsymbol{G})^2 - \bar{d}_{\cdot l}(\boldsymbol{G}) \Big)^2,$$

where $d_{jl}(\boldsymbol{G}) = \delta_{jl}^{(\boldsymbol{G})} / h_j \in [-1, 1]$ and $\bar{d}_{\cdot l}(\boldsymbol{G}) = \frac{1}{p} \sum_{j=1}^{p} d_{jl}(\boldsymbol{G})^2$, recalling that h_j^2 is the jth diagonal element of $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top = (\boldsymbol{\Lambda}\boldsymbol{G})(\boldsymbol{\Lambda}\boldsymbol{G})^\top$.

Remark: $d_{jl}(\boldsymbol{G})$ is the correlation between the factor $\tilde{F}_l^{(\boldsymbol{G})}$ and the variable X_j when $\psi_{jj} = 0$.

For all l, the sample variance of $d_{1l}(\boldsymbol{G}_*)^2, \ldots, d_{pl}(\boldsymbol{G}_*)^2$ being large it typically follows that we have $d_{jl}(\boldsymbol{G}_*) \approx 0$ for many $j \in \{1, \ldots, p\}$ and $d_{jl}(\boldsymbol{G}_*) \approx \pm 1$ otherwise.

Therefore, the loadings matrix $\boldsymbol{\Lambda}\boldsymbol{G}_*$ typically contains a few large loadings (in absolute value) and many near-zero loadings. When this happens the varimax rotation allows to reduce the number of factors on which depends a given component of X, which facilitates their interpretation.

Estimation of the factors and dimension reduction

Given an estimate $(\hat{\mathbf{\Lambda}}, \hat{\mathbf{\Psi}})$ of $(\mathbf{\Lambda}, \mathbf{\Psi})$ we can estimate the factor f_i associated to observation x_i by estimating the coefficient $\beta \in \mathbb{R}^k$ in the linear regression model

$$X_{ij} = \sum_{l=1}^{k} \hat{\lambda}_{jl}\beta_l + \epsilon_{ij}, \quad j = 1, \ldots, p, \quad \text{Var}(\epsilon_i) = \hat{\mathbf{\Psi}}.$$

Using the generalized least squares estimate of β, we obtain the following estimate $\hat{f}_i$ of the factor f_i:

$$\hat{f}_i = \boldsymbol{A}_k^{(\text{FA})} x_i, \quad \boldsymbol{A}_k^{(\text{FA})} := \left(\hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Psi}}^{-1} \hat{\mathbf{\Lambda}}\right)^{-1} \hat{\mathbf{\Lambda}}^\top \hat{\mathbf{\Psi}}^{-1}.$$

PCA versus FA: Recall that the q-dimensional representation of x_i obtained with PCA is $x_i' = \boldsymbol{A}_{1:q}^\top x_i$.

- Main differences between PCA and FA:
  - The rows of $\boldsymbol{A}_k^{(\text{FA})}$ are not orthogonal, while those of $\boldsymbol{A}_{1:q}^\top$ are.
  - $\hat{f}_i$ is model based, while x_i' is model free.

- Which method to choose for dimension reduction?
  - There is no definitive answer to this question, and quite often both PCA and FA lead to similar results (see the example below).
  - It makes sense to use FA if the FM is a reasonable assumption.

Remark: Like PCA, factor analysis can be used as a dimension reduction technique in regression models. This is done by replacing in principal component regression the matrix $\boldsymbol{A}_{1:q}^\top$ by the matrix $\boldsymbol{A}_k^{(\text{FA})}$.

Illustration of FA: The marks dataset (continued)

We consider the marks dataset, already used in the previous chapter on PCA, to illustrate the use of FA. We recall that this dataset contains the marks of $n = 88$ students in $p = 5$ subjects.

For this dataset we obtain the following correlation matrix $\boldsymbol{R}$:

variable	Matrix $\boldsymbol{R}$				
mechanics	1.00	0.55	0.55	0.41	0.39
vectors		1.00	0.61	0.48	0.44
algebra			1.00	0.71	0.66
analysis				1.00	0.61

For $k > 2$ the value of $\Delta_{k,p}$ defined in (2.5) is negative, and thus the factor model is well-defined only for $k \in \{1, 2\}$.

In what follows we let $k = 2$ be the number of factors. In this case, $\Delta_{k,p} > 0$ and an approximate solution to (2.4) is computed using the iterated FA algorithm introduced in this chapter.

Illustration of FA: The marks dataset (continued)

The matrix $\hat{\mathbf{\Lambda}} = [\hat{\lambda}_{(1)}\ \hat{\lambda}_{(2)}]$ computed with the iterated FA algorithm is as follows:

variable	$\hat{\lambda}_{(1)}$	$\hat{\lambda}_{(2)}$
mechanics	-0.64	0.34
vectors	-0.71	0.29
algebra	-0.90	-0.09
analysis	-0.77	-0.23
statistics	-0.72	-0.23

while $\hat{\mathbf{\Psi}} = \text{diag}\big(0.47, 0.41, 0.19, 0.35, 0.43\big)$.

Letting $\hat{\boldsymbol{R}} = [\hat{r}_{ij}] = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^\top + \hat{\mathbf{\Psi}}$ and writing $\boldsymbol{R}$ as $\boldsymbol{R} = [r_{ij}]$, we observe that $|r_{ij} - \hat{r}_{ij}| < 0.006$ for all $i, j \in \{1, \ldots, p\}$. Therefore, the $k = 2$ factors model fits the data well (i.e. the approximation $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^\top + \hat{\mathbf{\Psi}} \approx \boldsymbol{R}$ is good).

To interpret the two factors of the estimated model recall that this latter leads to the approximation $x_i \approx \hat{\lambda}_{(1)}\hat{f}_{i1} + \hat{\lambda}_{(2)}\hat{f}_{i2}$.

Since all the components of $\hat{\lambda}_{(1)}$ are negative if follows that (minus) the first factor can be interpreted as "intelligence" (the larger $-\hat{f}_{i1}$ the higher the intelligence and thus the higher the marks in all the five subjects).

Recalling that the examination was open book for the first two subjects and closed book for the remaining three subjects, the second factor can be interpreted as the propensity to perform well in one type of examination and poorly in the other.

FA versus PCA: The marks dataset

PCA with $q = 2$ principal components leads to the approximation $x_i \approx a_{(1)}y_{i1} + a_{(2)}y_{i2}$, where the vectors $a_{(1)}$ and $a_{(2)}$ are given in Table 1.1.

From Table 1.1, we observe that $a_{(1)}$ is very similar to $\hat{\lambda}_{(1)}$ and that $a_{(2)}$ is very similar to $\hat{\lambda}_{(2)}$. Therefore, if the two approximations $x_i \approx \hat{\lambda}_{(1)}\hat{f}_{i1} + \hat{\lambda}_{(2)}\hat{f}_{i2}$ and $x_i \approx a_{(1)}y_{i1} + a_{(2)}y_{i2}$ are good then we must have $(\hat{f}_{i1}, \hat{f}_{i2}) \approx (y_{i1}, y_{i2})$, that is, the estimated two dimensional factor $\hat{f}_i$ must be similar to the $q = 2$ dimensional reduction x_i' of x_i obtained with PCA.

From Figure 2.1 below we remark that for the marks dataset PCA and FA indeed give a very similar two dimensional representation of the data points.

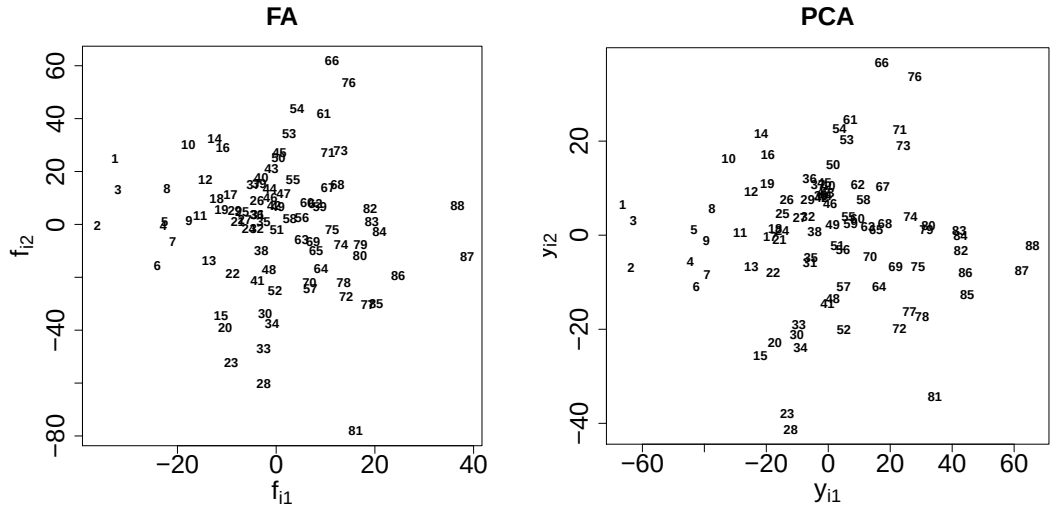

Figure 2.1: Two dimensional approximation of the marks dataset obtained with FA and with PCA.