1 Models for Factor Analysis

- The basic idea underlying factor analysis is that p **observed random variables or manifest variables**, $x_1, \ldots, x_p$, can be expressed, except for an error term, as linear functions $k(< p)$ **common factors or latent random variables**, $f_1, \ldots, f_k$.

$$\begin{aligned} x_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1k} f_k + u_1, \\ x_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2k} f_k + u_2, \\ &\vdots \\ x_p &= \lambda_{p1} f_1 + \lambda_{p2} f_2 + \cdots + \lambda_{pk} f_k + u_p. \end{aligned}$$

- λ_{ij} are constants called the **factor loadings**.
- u_i are error terms, sometimes called **specific factors** (because u_i is specific to x_i, whereas f_j are common to several x_i.)
- In matrix form:

$$X = \Lambda f + U. \tag{1}$$

  where $X, U \in \mathbb{R}^{p\times 1}, \Lambda \in \mathbb{R}^{p\times k}$, and $f \in \mathbb{R}^{k\times 1}, k \leq p$.

- In general, there are a number of assumptions associated with the factor model:

$$\begin{aligned} &\mathbb{E}(U) = 0, \quad \mathbb{E}(f) = 0, \quad \mathbb{E}(X) = 0, \\ &\mathbb{E}(UU^T) = \Psi \text{ diagonal}, \\ &\mathbb{E}(fU^T) = 0 \text{ a matrix of zeros}, \\ &\mathbb{E}(ff^T) = I_k \text{ an identity matrix}. \end{aligned}$$

  - The first is a standard assumption for error terms in most statistical models. If $\mathbb{E}(U) \neq 0$, another factor f can be added such that $\mathbb{E}(U) = 0$.
  - The second is convenient and loses no generality.
  - The third one may not be true, but if it is not, (1) can be simply adapted to

$$X = \Lambda f + U + \mu, \tag{2}$$

    where $\mathbb{E}(X) = \mu$. Note that the form of the basic model for factor analysis given in (1) is fairly standard (in textbook 2), there are some different versions. For example, the model could be made non-linear and there could be another term on the right-hand side (in the textbook 1, An introduction to multivariate statistical analysis). In fact, the basic one (1) loses no real generality and is usually adopted.
  - The fourth one is stating that the error terms are uncorrelated. The most important and common thing is contained in Λf, and $u_i, u_j (i \neq j)$ are therefore uncorrelated.
  - The fifth one is stating that the common factors are uncorrelated with the specific factors.

– The last one can be relaxed so that the common factors may be correlated.

$$\begin{cases} \Sigma = \Lambda\Lambda^T + \Psi & f_i \text{ and } f_j \text{ are uncorrelated, which is related with PCA} \\ \Sigma = \Lambda\Phi\Lambda^T + \Psi & f_i \text{ and } f_j \text{ are correlated, where } \Phi = Cov\left((f_1, \ldots, f_k)^T\right) \end{cases}$$

2 Estimation for Factor Analysis

Estimation of the model is usually done initially in terms of the parameters in Λ and Ψ, while estimates of f are found at a later stage.

2.1 $\Sigma = \Lambda\Lambda^T + \Psi$

- Equation (2) in slides: $Var(x_i) = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_i$.

$$\begin{aligned} Var(x_i) &= Cov(x_i, x_i) \\ &= Cov(\lambda_{i1}f_1 + \ldots + \lambda_{ik}f_k + u_k, \lambda_{i1}f_1 + \ldots + \lambda_{ik}f_k + u_k) \\ &= \lambda_{i1}^2 Cov(f_1, f_1) + \ldots + \lambda_{ik}^2 Cov(f_k, f_k) + \psi_i \\ &= \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_i. \end{aligned}$$

- Equation (6) in slides: $Cov(x_i, x_j) = \sum_{l=1}^{k} \lambda_{il}\lambda_{jl}$.
- Combing Equations (2) and (6), we have Equation (7) in slides: $\Sigma = \Lambda\Lambda^T + \Psi$.
- Note that in practice, we have the sample covariance matrix S, rather than Σ.

2.2 The indeterminacy

The indeterminacy of the solution becomes obvious: if Λ, Ψ is a solution and T is an orthogonal matrix, then Λ^*, Ψ is also a solution, where $\Lambda^* = \Lambda T$. This follows since

$$\Lambda^*(\Lambda^*)^T = (\Lambda T)(\Lambda T)^T = \Lambda(TT^T)\Lambda^T = \Lambda\Lambda^T.$$

Because of the indeterminacy, estimation of Λ and Ψ typically proceeds in two stages.

- In the first, some restrictions are placed of Λ in order to find a unique initial solution.
- Having found an initial solution, other solutions which can be found by **rotation** of Λ, that is, multiplication by an orthogonal matrix T.
  - The 'best' of the rotated solutions is chosen according to some particular requirements.
  - There are several possible requirements, but all are designed to make the structure of Λ as simple as possible in some sense, with most elements of either 'close to zero' or 'far from zero', and with as few as possible of the elements taking intermediate values.

- How to estimate Λ and Ψ? And how to find an appropriate value of k?
- In general, Λ, Ψ, and k are not unique.
- Specially, if $k = 1$, Λ and Ψ are uniquely determined.

$$S = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{pmatrix} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} (\lambda_1, \lambda_2, \lambda_3) + \begin{pmatrix} \psi_1 & 0 & 0 \\ 0 & \psi_2 & 0 \\ 0 & 0 & \psi_3 \end{pmatrix}.$$

$1 - \lambda_i^2 = \psi_i, i = 1, 2, 3.$

$\lambda_1\lambda_2 = r_{12}.\ \lambda_1\lambda_3 = r_{13}.\ \lambda_2\lambda_3 = r_{23}.$

- First calculate $\lambda_1\lambda_2\lambda_3 = \sqrt{\lambda_1^2\lambda_2^2\lambda_3^2} = \sqrt{r_{12}r_{13}r_{23}}$.
- Then, we can derive λ_1, λ_2, and λ_3.
- Lastly, we can derive ψ_1, ψ_2, and ψ_3.

2.3 The iterated process of Factor Analysis

Suppose Ψ is known:

- Calculate $\Sigma - \Psi$. Consider the spectral decomposition of $\Sigma - \Psi$:
  - Calculate the eigenvalues λ_i and associated eigenvectors v_i (v_i are column vectors) of $\Sigma - \Psi$.
  - $A = diag\{\lambda_1, \ldots, \lambda_p\}$ and $\Gamma = (v_1, \ldots, v_p)$, where $\lambda_1 \geq \ldots \geq \lambda_p$ and $\Gamma\Gamma^T = I$.
  - $\Sigma - \Psi = \Gamma A^{\frac{1}{2}} A^{\frac{1}{2}} \Gamma^T = \Lambda\Lambda^T$, where $(A^{\frac{1}{2}})^T = A^{\frac{1}{2}}$, $\Lambda \in \mathbb{R}^{p\times k}$, and $\Gamma A^{\frac{1}{2}} \in \mathbb{R}^{p\times p}$.
  - $1 \leq k < p$.
  - Note that if $k = p$, f_i is the re-scaled PCs.
- Block $A^{\frac{1}{2}}$ and get $A_1^{\frac{1}{2}} = diag\{\lambda_1^{\frac{1}{2}}, \ldots, \lambda_k^{\frac{1}{2}}\}$ while correspondingly block Γ and get Γ_1. Let $\hat{\Lambda} = \Gamma_1 A_1^{\frac{1}{2}}$.

How to update Ψ:

- Suggest one Ψ.
- $\hat{\Lambda} = \Gamma_1 A_1^{\frac{1}{2}}$.
- Since Σ is known, $\hat{\Psi} = \Sigma - \hat{\Lambda}\hat{\Lambda}^T$.

3 The difference between PCA and FA

- The major distinction between PCA and FA is that there is a definite model underlying FA, but for most purposes no model is assumed in PCA.

- Both PCA and FA can be thought of as trying to present some aspect of the covariance matrix Σ as well as possible. PCA concentrates on the diagonal elements of Σ whereas FA concentrates on explaining only the off-diagonal elements of Σ by a small number of factors.
  - PCA: the objective is to maximize $\sum_{i=1}^{m} var(z_i)$. Note that $\sum_{i=1}^{p} var(z_i) = \sum_{i=1}^{p} var(x_i)$. The objective account for as much as possible of the sum of diagonal elements $(var(x_i))$ of Σ.
  - FA and consider the basic model $X = \Lambda f + U$: as Ψ is diagonal, the common factor term Λf accounts for the off-diagonal elements of Σ.
    (i) If all of the variables have considerable common variation, then $\psi_j, j = 1, \ldots, k$ of Ψ will all be low.
    (ii) If a variable x_j is almost independent of all other variables, than $\psi_j = Var(u_j)$ will be almost as large as $Var(x_j)$.
- The last difference also leads to another difference between these two techniques concerning the number of dimensions m which give a good representation of the p dimensional variable x.
  - In PCA, if any individual variables are almost independent of all other variables, then there will be a PC corresponding to each such variable, and that PC will be almost equivalent to the corresponding variable. Such 'single variable' PCs are generally included if a good representation of x is required.
  - In contrast, a common factor in FA must contribute to **at least two** of the variables. Thus, it is not possible to have a 'single variable' common factor. Such factors appear as specific factors (error terms) and do not contribute to the dimensionality of the model.
- Changing $m(k)$, the dimensionality of the model, can have much more drastic effects on FA than it does on PCA.
  - In PCA, if m is increased from m_1 to m_2, then an additional $m_2 - m_1$ PCs are included, but the original m_1 PCs are still present and unaffected.
  - In FA an increase from m_1 to m_2 produces m_2 factors, none of which need bear any resemblance to the original m_1 factors.
- The last difference between PCs and common factors is that PCs can be calculated exactly from X, whereas the latter typically cannot.