Chapter 13:

Factor Analysis

- §Introduction
- §Orthogonal factor model
- §Estimation of loadings and communalities
- §Choosing the number of factors
- §Rotation

Introduction

- In *factor analysis*, we represent each $y_1, \ldots, y_p$ as linear combinations of fewer random variables $f_1, f_2, \ldots, f_m$ $(m < p)$ called *factors*.
- The factors are *latent* variables that "generate" the y's, but cannot be measured or observed.
- The goal of factor analysis is to reduce the redundancy among the variables by using a smaller number of factors.
- Factor analysis groups variables by their correlation patterns.
- Different from PCA:
  - Principal components are defined as linear combinations of original variables. In factor analysis, the original variables are expressed as linear combinations of the factors.
  - In PCA, we explain a large part of the total variance of the variables. In factor analysis, we seek to account for the covariances/correlations among the variables.

Orthogonal Factor Model

- Consider $\mathbf{y} = (y_1, \ldots, y_p)'$ with $E(\mathbf{y}) = \boldsymbol{\mu}$ and $\text{Cov}(\mathbf{y}) = \boldsymbol{\Sigma}$.
- For y_i $(i = 1, \ldots, p)$ in $\mathbf{y}$, the model is

$$y_i = \mu_i + \lambda_{i1} f_1 + \cdots + \lambda_{im} f_m + \varepsilon_i,$$

  where $f_1, \ldots, f_m$ are *factors* and λ_{ij} $(j = 1, \ldots, m)$ are *loadings*.
- Ideally, m should be much smaller than p.
- **Assumptions:**
  - $E(f_j) = 0$; $\text{Var}(f_j) = 1$; $\text{Cov}(f_j, f_k) = 0, j \neq k$.
  - $E(\varepsilon_i) = 0$; $\text{Var}(\varepsilon_i) = \psi_i$; $\text{Cov}(\varepsilon_i, \varepsilon_k) = 0, i \neq k$.
  - $\text{Cov}(\varepsilon_i, f_j) = 0$ for all i and j.
- The variance of y_i is

$$\text{Var}(y_i) = \sum_{j=1}^{m} \lambda_{ij}^2 + \psi_i = h_i^2 + \psi_i,$$

  where h_i^2 is *communality* and ψ_i is *specific variance*.

Orthogonal Factor Model

- Model in matrix notation

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{\Lambda f} + \varepsilon,$$

where $\mathbf{y} = (y_1, \ldots, y_p)'$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$, $\mathbf{f} = (f_1, \ldots, f_m)'$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_p)'$ and *loading matrix*

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pm} \end{pmatrix}.$$

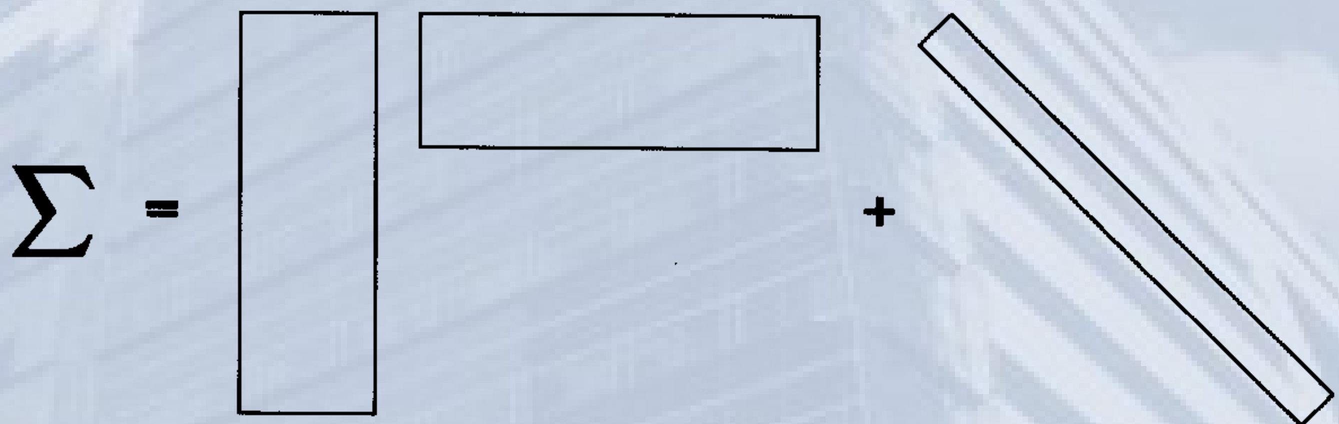
- **Assumptions:**
  - $E(\mathbf{f}) = \mathbf{0}$; $\quad \mathrm{Cov}(\mathbf{f}) = \mathbf{I}$.
  - $E(\varepsilon) = \mathbf{0}$; $\quad \mathrm{Cov}(\varepsilon) = \mathbf{\Psi} = \mathrm{diag}(\psi_1, \ldots, \psi_p)$.
  - $\mathrm{Cov}(\mathbf{f}, \varepsilon) = \mathbf{0}$.

Orthogonal Factor Model

- The covariance matrix of $\mathbf{y}$ is

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

- Graphical illustration:



- **Goal:** estimate the $p \times p$ $\boldsymbol{\Sigma}$ in terms of simpler structure, involving only $p \times m$ loadings and p variances ψ_i.
- If the variables are NOT commensurate, we use $\mathbf{R}$ in place of $\mathbf{S}$. In practice, R is more often used.

Orthogonal Factor Model

- When $p = 2$ and $m = 2$, then the model is

$$\begin{aligned} y_1 &= \mu_1 + \lambda_{11} f_1 + \lambda_{12} f_2 + \varepsilon_1, \\ y_2 &= \mu_2 + \lambda_{21} f_1 + \lambda_{22} f_2 + \varepsilon_2. \end{aligned}$$

  and the variances and covariances are

$$\begin{aligned} \text{Var}(y_1) &= \lambda_{11}^2 + \lambda_{12}^2 + \psi_1 = h_1^2 + \psi_1 \\ \text{Var}(y_2) &= \lambda_{21}^2 + \lambda_{22}^2 + \psi_2 = h_2^2 + \psi_2 \\ \text{Cov}(y_1, y_2) &= \lambda_{11}\lambda_{21} + \lambda_{12}\lambda_{22} \\ \text{Cov}(y_i, f_j) &= \lambda_{ij}, \quad (i = 1, 2, j = 1, 2) \end{aligned}$$

- Note that λ_{ij} are not unique. Let $\mathbf{T}\mathbf{T}' = \mathbf{I}$ and the model is

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\varepsilon} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{T}\mathbf{T}'\mathbf{f} + \boldsymbol{\varepsilon} = \boldsymbol{\mu} + \boldsymbol{\Lambda}^*\mathbf{f}^* + \boldsymbol{\varepsilon},$$

  where $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\mathbf{T}$ and $\mathbf{f}^* = \mathbf{T}'\mathbf{f}$. Then, it can be shown that $\boldsymbol{\Sigma}$ of the 'new' model remains the same.

Estimation of λ_{ij} and ψ_i

Principal component method

- **Idea:** find estimators $\hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{\Psi}}$ such that $\mathbf{R} \approx \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}$
- **First**, ignore $\hat{\mathbf{\Psi}}$ and factor $\mathbf{R}$ into $\mathbf{R} = \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}'$, where

$$\hat{\mathbf{\Lambda}} = \mathbf{C}_m \mathbf{D}_m^{1/2} = \left(\sqrt{\theta_1}\mathbf{c}_1, \ldots, \sqrt{\theta_m}\mathbf{c}_m\right),$$

  $\theta_1 > \theta_2 > \cdots > \theta_m$ are first m eigenvalues of $\mathbf{R}$, and $\mathbf{c}_1, \ldots, \mathbf{c}_m$ are corresponding eigenvectors.
- **Then**, $\hat{\psi}_i = 1 - \hat{h}_i^2 = 1 - \sum_{j=1}^{m} \hat{\lambda}_{ij}^2$ (1 is the diagonal of $\mathbf{R}$.)
- **Finally**, we have $\mathbf{R} \approx \hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}$, where $\hat{\mathbf{\Psi}} = \text{diag}(\hat{\psi}_1, \ldots, \hat{\psi}_p)$.
- Variance explained by jth factor $= \theta_j = \hat{\lambda}_{1j}^2 + \cdots + \hat{\lambda}_{pj}^2$.
- Variance proportion for jth factor $= \theta_j/\text{tr}(\mathbf{R}) = \theta_j/p$.
- Show Example 13.3.1 (p.419).

Estimation of λ_{ij} and ψ_i

Principal factor method

- If **R** is invertible, initial $\widehat{h}_i^2 = R_i^2$, where R_i^2 is the squared multiple correlation of y_i and other $p-1$ variables. (SAS: priors = SMC)
- If **R** is NOT invertible, initial $\widehat{h}_i^2$ = absolute value of the largest correlation in each row of **R**. (SAS: priors = MAX)
- Calculate initial $\widehat{\psi}_i = 1 - \widehat{h}_i^2$ and $\widehat{\mathbf{\Psi}} = \mathrm{diag}(\widehat{\psi}_1, \ldots, \widehat{\psi}_p)$.
- Given initial $\widehat{\mathbf{\Psi}}$, we factor $\mathbf{R} - \widehat{\mathbf{\Psi}} = \widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Lambda}}'$, where $\widehat{\mathbf{\Lambda}}$ is the $p \times m$ matrix obtained from the eigen-decomposition of $\mathbf{R} - \widehat{\mathbf{\Psi}}$.
- Then, we use $\widehat{\mathbf{\Lambda}}$ to obtain new

$$\widehat{h}_i^2 = \sum_{j=1}^{m} \widehat{\lambda}_{ij}^2 \quad \text{and} \quad \widehat{\psi}_i = 1 - \widehat{h}_i^2.$$

- Variance proportion for jth factor $= \theta_j / \mathrm{tr}(\mathbf{R} - \widehat{\mathbf{\Psi}}) = \theta_j / \sum_{i=1}^{p} \theta_i$.
- Show Example 13.3.2 (p.423).

Estimation of λ_{ij} and ψ_i

Iterated principal factor method

1. Start with initial $\hat{h}_i^2$ and $\hat{\psi}_i$.
2. Calculate $\hat{\mathbf{\Lambda}}$ from $\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' = \mathbf{R} - \hat{\mathbf{\Psi}}$.
3. Obtain new $\hat{h}_i^2 = \sum_{j=1}^{m} \hat{\lambda}_{ij}^2$.
4. Substitute the new $\hat{h}_i^2$ into the diagonal of $\mathbf{R} - \hat{\mathbf{\Psi}}$ and go back to Step (2) until $\hat{h}_i^2$ converge.

Remarks:

- The failure of convergence can happen ($\hat{h}_i^2 > 1$).
- If correlations are large or p is large, the three methods will produce similar results.

Show Example 13.3.3 (p.425).

Choosing m

1. Choose m equal to the number of factors necessary for the variance accounted for to achieve a predetermined percentage, say 80%.
2. Choose m equal to the number of eigenvalues greater than the average eigenvalue.
3. Use the scree test based on a plot of the eigenvalues of $\mathbf{S}$ or $\mathbf{R}$. If the graph drops sharply, followed by a straight line with much smaller slope, choose m equal to the number of eigenvalues before the straight line begins.
4. Test $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$

$$\chi^2 = \left(n - \frac{2p + 4m + 11}{6}\right) \ln \left(\frac{|\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}}|}{|\mathbf{S}|} \right)$$

Reject H_0 if $\chi^2 > \chi^2_\nu$ $(\nu = [(p-m)^2 - p - m]/2)$ and it implies that m is too small and more factors are needed. (NOT recommended!)

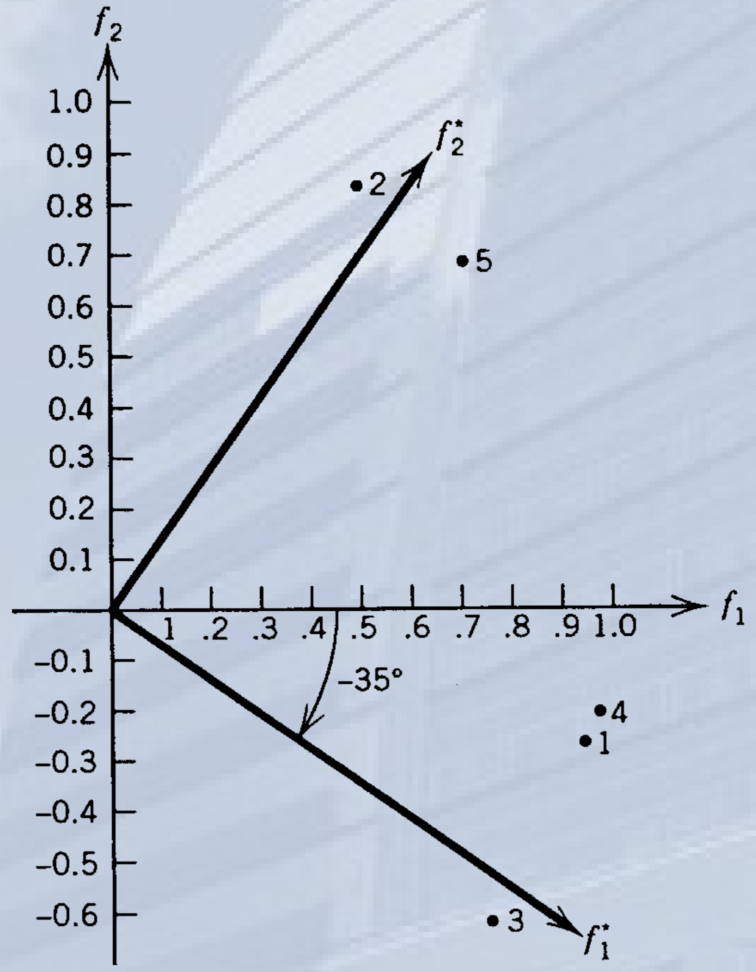
Rotation

- Recall that

$$\text{Cov}(\mathbf{y}) = \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi} = \mathbf{\Lambda}^*\mathbf{\Lambda}^{*'} + \mathbf{\Psi},$$

  where $\mathbf{\Lambda}^* = \mathbf{\Lambda}\mathbf{T}$ and $\mathbf{T}\mathbf{T}' = \mathbf{I}$.
- The original $\mathbf{\Lambda}$ can be rotated to obtain a new $\mathbf{\Lambda}^* = \mathbf{\Lambda}\mathbf{T}$, which provides the same estimate of the covariance matrix as before.
- The loadings in the ith row are the coordinates of variable y_i in the space defined by the factors. By rotating the loadings we hope to place new axes close to as many points as possible in order to make the factors more interpretable.
- The number of factors on which a variable has high loadings, is called "complexity" of the variable.
- In the ideal situation, the variables all have a complexity of 1.

Orthogonal Rotation



- Orthogonal rotations preserve communalities because the distance to the origin is unchanged.
- The variance accounted for by each factor θ_j will change.
- Graphical approach ($m = 2$): choose rotation angle by visual inspection. Show Example 13.5.2(a) (p.432).
- Varimax method ($m \geq 2$):maximize the variance of the squared loadings in each column. Show Example 13.5.2(b) (p.434).

Interpretation

- Interpret the factors by examination of rotated loadings $\boldsymbol{\Lambda}^*$.
- Identify the highest loading in each row for all p variables. Increase m if some variable has no significant loadings on any factor.
- The threshold to determine the significance of a loading is subjective. Many people use 0.3 to 0.6 to make complexity 1.
- After identifying potentially significant loadings, we then attempt to discover some meaning in the factors and, ideally, to label or name them. This can readily be done if the group of variables associated with each factor makes sense to the researcher.
- If the groupings are not so logical, a revision can be tried, such as adjusting the size of loading deemed to be important, changing m, using a different method of estimating the loadings, or employing another type of rotation.