Multivariate Analysis
Factor Analysis

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through this chapter, you should be able to

1. Perform factor analysis,
2. Explain indeterminancy in factor analysis,
3. Obtain factor scores.

Motivation: Latent Variable Modelling

	--	-	-/+	+	++
1. I am a 'worrier'	○	○	○	○	○
2. I make friends easily	○	○	○	○	○
3. I have a vivid imagination	○	○	○	○	○
4. I trust others	○	○	○	○	○
5. I complete tasks successfully	○	○	○	○	○
6. I get angry easily	○	○	○	○	○
7. I really enjoy large parties and gatherings	○	○	○	○	○
8. I think art is important	○	○	○	○	○

The Model

The factor analysis model for a $p \times 1$ random vector $\boldsymbol{X}$ is

$$X_i = \mu_i + \ell_{i1}F_1 + \ell_{i2}F_2 + \cdots + \ell_{im}F_m + e_i, \quad i = 1, 2, ..., p,$$

where F's are the common factors, ℓ's are the (factor) loadings, e_i is the error (or specific factor). In the model, we should have $m < p$.

In matrix notation,

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{F} + \boldsymbol{e},$$

where $\boldsymbol{L}$ is the factor loading matrix of size $p \times m$.

Scale Indeterminacy

The factor analysis model is indeterminate, since we do not observe $\boldsymbol{F}$.

1. $\mathbb{E}(\boldsymbol{F})$ is not unique: for any $m \times 1$ $\boldsymbol{a}$,

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{F} + \boldsymbol{e} = \underbrace{\boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{a}}_{\boldsymbol{\mu}^*} + \boldsymbol{L}\underbrace{(\boldsymbol{F} - \boldsymbol{a})}_{\boldsymbol{F}^*} + \boldsymbol{e}.$$

2. $\operatorname{cov}(\boldsymbol{F})$ is not unique: for any $m \times m$ invertible matrix $\boldsymbol{T}$,

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{F} + \boldsymbol{e} = \boldsymbol{\mu} + \underbrace{(\boldsymbol{L}\boldsymbol{T})}_{\boldsymbol{L}^*}\underbrace{(\boldsymbol{T}^{-1}\boldsymbol{F})}_{\boldsymbol{F}^*} + \boldsymbol{e}.$$

Assumptions

Our assumptions are

$$\begin{aligned} \mathbb{E}(\boldsymbol{F}) &= \mathbf{0}, \\ \operatorname{cov}(\boldsymbol{F}) &= \boldsymbol{I}, \\ \mathbb{E}(\boldsymbol{e}) &= \mathbf{0}, \\ \operatorname{cov}(\boldsymbol{e}) &= \boldsymbol{\Psi} \text{ (diagonal)}, \\ \operatorname{cov}(\boldsymbol{F}, \boldsymbol{e}) &= \mathbf{0}. \end{aligned}$$

The model

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{F} + \boldsymbol{e}$$

satisfying the above assumptions is an orthogonal factor model.

Model Implied Covariance Matrix

The orthogonal factor model implies

1. an expectation for $\boldsymbol{X}$: $\mathbb{E}(\boldsymbol{X}) = \boldsymbol{\mu} + \boldsymbol{L}\mathbb{E}(\boldsymbol{F}) + \mathbb{E}(\boldsymbol{e}) = \boldsymbol{\mu}$.
2. a covariance matrix for $\boldsymbol{X}$:

$$\begin{aligned} \operatorname{cov}(\boldsymbol{X}) &= \operatorname{cov}(\boldsymbol{LF} + \boldsymbol{e}) \\ &= \operatorname{cov}(\boldsymbol{LF}) + \operatorname{cov}(\boldsymbol{e}) \\ &= \boldsymbol{LL}^T + \boldsymbol{\Psi}, \end{aligned}$$

which is called a model-implied covariance matrix.

This means that

$$\begin{array}{ccccc} \sigma_{ii} & = & \sum_{j=1}^m \ell_{ij}^2 & + & \psi_i \\ \text{variance} & & \text{communality} & & \text{uniqueness (or specific variance)} \end{array}.$$

We often denote $h_i = \sum_{j=1}^m \ell_{ij}^2$.

Existence of Decomposition

The orthogonal factor analysis model is basically:

Whether we can decompose $\boldsymbol{\Sigma}$ *into* $\boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi}$, *where* $\boldsymbol{\Psi} > 0$ *is diagonal?*

Valid Decomposition

$$\begin{bmatrix} 19 & 30 & 2 \\ 30 & 57 & 5 \\ 2 & 5 & 38 \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \\ -1 \end{bmatrix} \begin{bmatrix} 4 & 7 & -1 \end{bmatrix} + \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Invalid Decomposition

$$\begin{bmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{bmatrix} = \begin{bmatrix} 1.2549 \\ 0.7173 \\ 0.5577 \end{bmatrix} \begin{bmatrix} 1.2549 \\ 0.7173 \\ 0.5577 \end{bmatrix}^T + \begin{bmatrix} -0.5748 & 0 & 0 \\ 0 & 0.4852 & 0 \\ 0 & 0 & 0.6891 \end{bmatrix}.$$

Indeterminacy

Even though we let $\text{cov}(\boldsymbol{F}) = \boldsymbol{I}$, the model

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{F} + \boldsymbol{e}$$

is still indeterminate. Let $\boldsymbol{T}$ be any $m \times m$ orthogonal matrix. Then,

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{F} + \boldsymbol{e} = \boldsymbol{\mu} + \underbrace{(\boldsymbol{L}\boldsymbol{T})}_{\boldsymbol{L}^*}\underbrace{(\boldsymbol{T}^T\boldsymbol{F})}_{\boldsymbol{F}^*} + \boldsymbol{e}, \\ \text{cov}(\boldsymbol{X}) &= \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi} = (\boldsymbol{L}\boldsymbol{T})(\boldsymbol{L}\boldsymbol{T})^T + \boldsymbol{\Psi}, \\ \text{cov}(\boldsymbol{F}^*) &= \boldsymbol{T}^T\text{cov}(\boldsymbol{F})\boldsymbol{T} = \text{cov}(\boldsymbol{F}) = \boldsymbol{I}. \end{aligned}$$

Hence, $\boldsymbol{L}$ is determined only up to an orthogonal matrix $\boldsymbol{T}$.

Scale Invariant

Consider the model

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{F} + \boldsymbol{e}.$$

Suppose that we make a linear transformation $\boldsymbol{Z} = \boldsymbol{V}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$. Then the covariance matrix of $\boldsymbol{Z}$ satisfies

$$\begin{aligned} \operatorname{cov}(\boldsymbol{Z}) &= \boldsymbol{V}^{-1/2}\operatorname{cov}(\boldsymbol{X})\boldsymbol{V}^{-1/2} \\ &= \left(\boldsymbol{V}^{-1/2}\boldsymbol{L}\right)\left(\boldsymbol{V}^{-1/2}\boldsymbol{L}\right)^{T} + \boldsymbol{V}^{-1/2}\boldsymbol{\Psi}\boldsymbol{V}^{-1/2}, \end{aligned}$$

which correspond to the model

$$\boldsymbol{Z} = \boldsymbol{V}^{-1/2}\boldsymbol{L}\boldsymbol{F} + \boldsymbol{V}^{-1/2}\boldsymbol{e}.$$

Popular Estimation Methods

Suppose that we have observed $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_n$. We can compute the sample covariance matrix $\boldsymbol{S}$ or sample correlation matrix $\boldsymbol{R}$.

Popular methods of parameter estimation for a factor analysis model are

1. principal component (or principal factor),
2. maximum likelihood,
3. least squares.

Spectral Decomposition

The spectral decomposition implies that the covariance matrix $\mathbf{\Sigma}$ satisfies

$$\mathbf{\Sigma} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T = \underbrace{\begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e}_1 & \cdots & \sqrt{\lambda_p}\boldsymbol{e}_p \end{bmatrix}}_{\boldsymbol{L} \text{ is } p \times p} \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e}_1^T \\ \vdots \\ \sqrt{\lambda_p}\boldsymbol{e}_p^T \end{bmatrix} + \mathbf{0}.$$

If the last $p - m$ eigenvalues are small, we can ignore them and approximate $\mathbf{\Sigma}$ by

$$\mathbf{\Sigma} \approx \underbrace{\begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e}_1 & \cdots & \sqrt{\lambda_m}\boldsymbol{e}_m \end{bmatrix}}_{\boldsymbol{L} \text{ is } p \times m} \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{e}_1^T \\ \vdots \\ \sqrt{\lambda_m}\boldsymbol{e}_m^T \end{bmatrix} + \begin{bmatrix} \psi_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \psi_p \end{bmatrix},$$

where $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} \ell_{ij}^2$.

Principal Component Method

Suppose that we have computed the sample covariance matrix $\boldsymbol{S}$ or the sample correlation matrix $\boldsymbol{R}$.

- The estimated factor loading matrix is

$$\hat{\boldsymbol{L}} = \left[\sqrt{\hat{\lambda}_1}\hat{\boldsymbol{e}}_1 \quad \sqrt{\hat{\lambda}_2}\hat{\boldsymbol{e}}_2 \quad \cdots \quad \sqrt{\hat{\lambda}_m}\hat{\boldsymbol{e}}_m\right].$$

- The estimated specific variances are

$$\hat{\psi}_i = s_{ii} - \sum_{j=1}^{m} \hat{\ell}_{ij}^2.$$

- The diagonal elements of $\boldsymbol{S}$ are reproduced by $\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}}$, but not the off-diagonal elements.

One advantage of the principal component method is that adding more factors does not change the loadings of early factors.

Determine Number of Factors

To determine the number of factors m,

1. subject knowledge,
2. using residual matrix: the entries in $\boldsymbol{S} - \hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T - \hat{\boldsymbol{\Psi}}$ are small.
3. Kaiser criterion: number of eigenvalues of $\boldsymbol{R}$ greater than 1. Easy but dirty.
4. cumulative eigenvalues: the proportion of total sample variance due to jth factor is

$$\frac{\hat{\lambda}_j}{\sum_{i=1}^{p} s_{ii}}.$$

Choose m such that a suitable proportion of the total sample variance has been explained.

Principal Factor/Axis Method

Consider the sample correlation matrix $\boldsymbol{R}$. Suppose that the specific variance ψ_i is ψ_i^*. Then, the communality satisfies

$$h_i^2 \quad = \quad 1 - \psi_i^*.$$

The reduced sample correlation matrix is

$$\boldsymbol{R}_r \quad = \quad \begin{bmatrix} h_1^{*2} & r_{12} & \cdots & r_{1p} \\ r_{12} & h_2^{*2} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & h_p^{*2} \end{bmatrix},$$

It should satisfy $\boldsymbol{R}_r \approx \boldsymbol{L}_r \boldsymbol{L}_r^T$, and our loading matrix is $\boldsymbol{L}_r$.

Principal Factor/Axis Method: Algorithm

Algorithm 1: Pseudo code of the principal axis method

Assign initial values ψ_i^* for all i, e.g., from eigen decomposition ;
Let the number of factors be m ;
while *Convergence not met* **do**
    Compute the reduced correlation matrix $\boldsymbol{R}_r$ using $\boldsymbol{R}$ and ψ_i^*'s ;
    Apply spectral decomposition to $\boldsymbol{R}_r$;
    The loading matrix is $\boldsymbol{L}_r = \left[\sqrt{\hat{\lambda}_1}\hat{\boldsymbol{e}}_1 \quad \sqrt{\hat{\lambda}_2}\hat{\boldsymbol{e}}_2 \quad \cdots \quad \sqrt{\hat{\lambda}_m}\hat{\boldsymbol{e}}_m\right]$;
    Update ψ_i^* using $\boldsymbol{L}_r\boldsymbol{L}_r^T$;
end

Maximum Likelihood

Now we also assume that

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{F} + \boldsymbol{e} \quad \sim \quad N_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right).$$

Then, the log-likelihood function is

$$\begin{aligned}\log L\left(\boldsymbol{\theta}\right) &= \text{constant} - \frac{n}{2}\log\left|\boldsymbol{\Sigma}\right| - \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}_i - \boldsymbol{\mu}\right) \\ &= \text{constant} - \frac{n}{2}\log\left|\boldsymbol{\Sigma}\right| - \frac{1}{2}\sum_{i=1}^{n}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}\right)^T\right],\end{aligned}$$

where $\boldsymbol{\theta}$ is the vector of unknown parameters.

Maximum Likelihood Estimator

The MLE maximizes the log-likelihood function

$$\log L(\boldsymbol{\theta}) = \text{constant} - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}\text{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T\right].$$

1. The MLE of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$.
2. The MLEs of $\boldsymbol{L}$ and $\boldsymbol{\Psi}$ must be obtained by numerical methods from

$$\begin{aligned}\log L(\boldsymbol{\theta}) &= \text{constant} - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}\text{tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T\right] \\ &= \text{constant} - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_n\right).\end{aligned}$$

1. Since the model is not identified, we often use the restriction $\boldsymbol{L}^T\boldsymbol{\Psi}\boldsymbol{L}$ is diagonal and the diagonal entries are written in decreasing order.

Large Sample Test for the Number of Common Factors

An advantage of maximum likelihood is that we can test

$$H_0: \ m \text{ factors are sufficient to describe the data}$$
$$H_1: \ \boldsymbol{\Sigma} \text{ has no constraints}$$

The likelihood ratio statistic is given by

$$\begin{aligned} -2\log\Lambda &= -2\log\frac{\exp\left\{-\frac{n}{2}\log|\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T+\hat{\boldsymbol{\Psi}}|-\frac{n}{2}\mathrm{tr}\left[\left(\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T+\hat{\boldsymbol{\Psi}}\right)^{-1}\boldsymbol{S}_n\right]\right\}}{\exp\left\{-\frac{n}{2}\log|\boldsymbol{S}_n|-\frac{n}{2}\mathrm{tr}\left(\boldsymbol{S}_n^{-1}\boldsymbol{S}_n\right)\right\}} \\ &= n\left\{\mathrm{tr}\left[\left(\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T+\hat{\boldsymbol{\Psi}}\right)^{-1}\boldsymbol{S}_n\right]-\log\left|\left(\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T+\hat{\boldsymbol{\Psi}}\right)^{-1}\boldsymbol{S}_n\right|-p\right\}, \end{aligned}$$

which has an asymptotic χ^2_s distribution under H_0 with
$s = 2^{-1}p\,(p+1) - \left[pm + m - 2^{-1}m\,(m-1)\right]$.

Bartlett Correction

In the psychometrics literature,

$$F\left(\boldsymbol{\Sigma}, \boldsymbol{S}_n\right) = \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_n\right) - \log\left|\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_n\right| - p$$

is often called a normal-theory fit function.
The likelihood ratio statistic is then

$$\begin{aligned} -2\log\Lambda &= n\left\{\operatorname{tr}\left[\left(\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}}\right)^{-1}\boldsymbol{S}_n\right] - \log\left|\left(\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}}\right)^{-1}\boldsymbol{S}_n\right| - p\right\} \\ &= nF\left(\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}}, \boldsymbol{S}_n\right). \end{aligned}$$

The chi-square approximation is improved if we instead consider

$$\left[n - 1 - \frac{1}{6}\left(2p + 4m + 5\right)\right] F\left(\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}^T + \hat{\boldsymbol{\Psi}}, \boldsymbol{S}_n\right).$$

Other Estimation Methods

We can use fit functions other than the fit function derived from the log-likelihood.

- Normal-theory fit function:

$$F\left(\boldsymbol{\Sigma}, \boldsymbol{S}_n\right) = \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_n\right) - \log\left|\boldsymbol{\Sigma}^{-1}\boldsymbol{S}_n\right| - p$$

- Unweighted least squares:

$$F\left(\boldsymbol{\Sigma}, \boldsymbol{S}_n\right) = \frac{1}{2}\operatorname{tr}\left\{\left(\boldsymbol{S}_n - \boldsymbol{\Sigma}\right)^2\right\}.$$

- Generalized least squares:

$$F\left(\boldsymbol{\Sigma}, \boldsymbol{S}_n\right) = \frac{1}{2}\operatorname{tr}\left\{\left[\left(\boldsymbol{S}_n - \boldsymbol{\Sigma}\right)\boldsymbol{S}_n^{-1}\right]^2\right\}.$$

For most estimators from fit functions, $\hat{\boldsymbol{\Psi}}$ is not guaranteed to be positive definite. It is called a Heywood case.

Parallel Analysis

The number of factors can also be determined using a simulation-based parallel analysis.

Algorithm 2: Pseudo code of the principal axis method

```
Fit a factor model using the data and obtain eigenvalues of LL^T ;
while Simulation do
    Generate random data of same dimension as raw data by either
     resampling the raw data or simulating from (independent)
     normal data ;
    Obtain eigenvalues using the random data ;
end
Plot the eigenvalues from the real data and the random data ;
```

Versus PCA

FA	PCA
Dimension reduction	Dimension reduction
Reflective, with causal interpretation	Formative
Model assumes latent causal variables exist.	Model does not assume latent variables exist.
Scale invariant	Not scale invariant
$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi}$	Spectral decomposition

Orthogonal Rotation

The model-implied covariance matrix of a factor model is

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi}.$$

For any $m \times m$ orthogonal matrix $\boldsymbol{T}$, the model-implied covariance matrix is unchanged:

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi} = (\boldsymbol{L}\boldsymbol{T})(\boldsymbol{L}\boldsymbol{T})^T + \boldsymbol{\Psi},$$

and the factor model is equivalent to

$$\boldsymbol{X} = \boldsymbol{\mu} + (\boldsymbol{L}\boldsymbol{T})\boldsymbol{T}^T\boldsymbol{F} + \boldsymbol{e}$$

with $\text{cov}\left(\boldsymbol{T}^T\boldsymbol{F}\right) = \boldsymbol{T}^T\text{cov}\left(\boldsymbol{F}\right)\boldsymbol{T} = \text{cov}\left(\boldsymbol{F}\right) = \boldsymbol{I}$.

Orthogonal Rotation

We can multiply $\boldsymbol{L}$ by an orthogonal matrix $\boldsymbol{T}$ (equivalently $\boldsymbol{F}$ by $\boldsymbol{T}^T$) to rotate the factor to ease interpretation. Let $\boldsymbol{B}$ be the rotated loading matrix.

- Varimax rotation maximizes

$$\sum_{k=1}^{m}\left\{\frac{1}{p}\sum_{i=1}^{p}b_{ik}^4-\left(\frac{1}{p}\sum_{i=1}^{p}b_{ik}^2\right)^2\right\},$$

  the columnwise variance of the squared factor loadings.

- Quartimax rotation minimizes

$$-\sum_{i=1}^{p}\sum_{k=1}^{m}b_{ik}^4.$$

Oblique rotation

Suppose that $\boldsymbol{T}$ is nonsingular, but not orthogonal. Then,

$$\boldsymbol{\Sigma} \quad = \quad \boldsymbol{L}\boldsymbol{L}^T + \boldsymbol{\Psi} = (\boldsymbol{L}\boldsymbol{T})\,\boldsymbol{T}^{-1}\left(\boldsymbol{T}^{-1}\right)^T(\boldsymbol{L}\boldsymbol{T})^T + \boldsymbol{\Psi},$$

and the factor model is equivalent to

$$\boldsymbol{X} \quad = \quad \boldsymbol{\mu} + (\boldsymbol{L}\boldsymbol{T})\,\boldsymbol{T}^{-1}\boldsymbol{F} + \boldsymbol{e}.$$

In this model,

$$\mathrm{cov}\left(\boldsymbol{T}^{-1}\boldsymbol{F}\right) \quad = \quad \boldsymbol{T}^{-1}\mathrm{cov}\left(\boldsymbol{F}\right)\left(\boldsymbol{T}^{-1}\right)^T = \boldsymbol{T}^{-1}\left(\boldsymbol{T}^{-1}\right)^T.$$

This is the oblique rotation with new factor $\boldsymbol{T}^{-1}\boldsymbol{F}$ and new loading matrix $\boldsymbol{L}\boldsymbol{T}$. The covariance matrix of the rotated factor is not $\boldsymbol{I}$.

Oblique rotation

Let $\boldsymbol{B}$ be the rotated loading matrix.

- Promax rotation first applies the variamx rotation, and then rotate the initial loading matrix towards the varimax rotated matrix.
- Quartimin rotation minimizes

$$\sum_{\ell=1}^{m}\sum_{k\neq\ell}\sum_{i=1}^{p} b_{ik}^2 b_{i\ell}^2.$$

- Geomin rotation, given a small $\epsilon > 0$ (e.g., $\epsilon = 0.01$), minimizes

$$\sum_{i=1}^{p}\left[\prod_{k=1}^{m}\left(b_{ik}^2+\epsilon\right)\right]^{1/m}.$$

Bartlett Score

Factor scores are the estimated values of the common factors.

For every subject, the Bartlett score minimizes

$$L(\boldsymbol{f}) = (\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{L}\boldsymbol{f})^T \boldsymbol{\Psi}^{-1} (\boldsymbol{x} - \boldsymbol{\mu} - \boldsymbol{L}\boldsymbol{f}).$$

The minimizer is

$$\boldsymbol{f}_B = \left(\boldsymbol{L}^T \boldsymbol{\Psi}^{-1} \boldsymbol{L}\right)^{-1} \boldsymbol{L}^T \boldsymbol{\Psi}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}).$$

Plugging in the estimators,

$$\hat{\boldsymbol{f}}_B = \left(\hat{\boldsymbol{L}}^T \hat{\boldsymbol{\Psi}}^{-1} \hat{\boldsymbol{L}}\right)^{-1} \hat{\boldsymbol{L}}^T \hat{\boldsymbol{\Psi}}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}).$$

Thompson Score

Suppose that

$$\begin{bmatrix} \boldsymbol{F} \\ \boldsymbol{X} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Phi} & \boldsymbol{\Phi} \boldsymbol{L}^T \\ \boldsymbol{L}\boldsymbol{\Phi} & \boldsymbol{\Sigma} \end{bmatrix} \right).$$

Then,

$$\boldsymbol{F} \mid \boldsymbol{X} = \boldsymbol{x} \sim N\left(\boldsymbol{\Phi} \boldsymbol{L}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}), \boldsymbol{\Phi} - \boldsymbol{\Phi} \boldsymbol{L}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{L} \boldsymbol{\Phi} \right).$$

The Thompson score (or regression score) is

$$\begin{aligned} \boldsymbol{f}_R &= \boldsymbol{\Phi} \boldsymbol{L}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \\ &= \left(\boldsymbol{L}^T \boldsymbol{\Psi}^{-1} \boldsymbol{L} + \boldsymbol{\Phi}^{-1} \right)^{-1} \boldsymbol{L}^T \boldsymbol{\Psi}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}). \end{aligned}$$

Plugging in the estimators,

$$\hat{\boldsymbol{f}}_R = \left(\hat{\boldsymbol{L}}^T \hat{\boldsymbol{\Psi}}^{-1} \hat{\boldsymbol{L}} + \hat{\boldsymbol{\Phi}}^{-1} \right)^{-1} \hat{\boldsymbol{L}}^T \hat{\boldsymbol{\Psi}}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}).$$

Issues of Factor Scores

However,

$$\begin{aligned} \operatorname{cov}(\boldsymbol{f}_B) &= \left(\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{-1}\boldsymbol{L}\left(\boldsymbol{L}^T\boldsymbol{\Psi}^{-1}\boldsymbol{L}\right)^{-1}, \\ \operatorname{cov}(\boldsymbol{f}_R) &= \boldsymbol{\Phi}\boldsymbol{L}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{L}\boldsymbol{\Phi}, \end{aligned}$$

which are not necessary equal to $\boldsymbol{\Phi}$.

We cannot use the factor scores as if they were observed. If they are used in a regression model, they should be treated as errors-in-variables.

Anderson-Rubin Score

The Anderson-Rubin method minimizes

$$\sum_{j=1}^{n} (\boldsymbol{x}_j - \boldsymbol{\mu} - \boldsymbol{L}\boldsymbol{f}_j)^T \boldsymbol{\Psi}^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu} - \boldsymbol{L}\boldsymbol{f}_j)$$

$$\text{s.t. } \frac{1}{n-1} \sum_{j=1}^{n} \boldsymbol{f}_j \boldsymbol{f}_j^T = \text{cov}(\boldsymbol{F}).$$

Hence, the sample covariance matrix of the resulting factor score is equal to the (estimated) covariance matrix of the factors from the model.

Regression with Factor Score

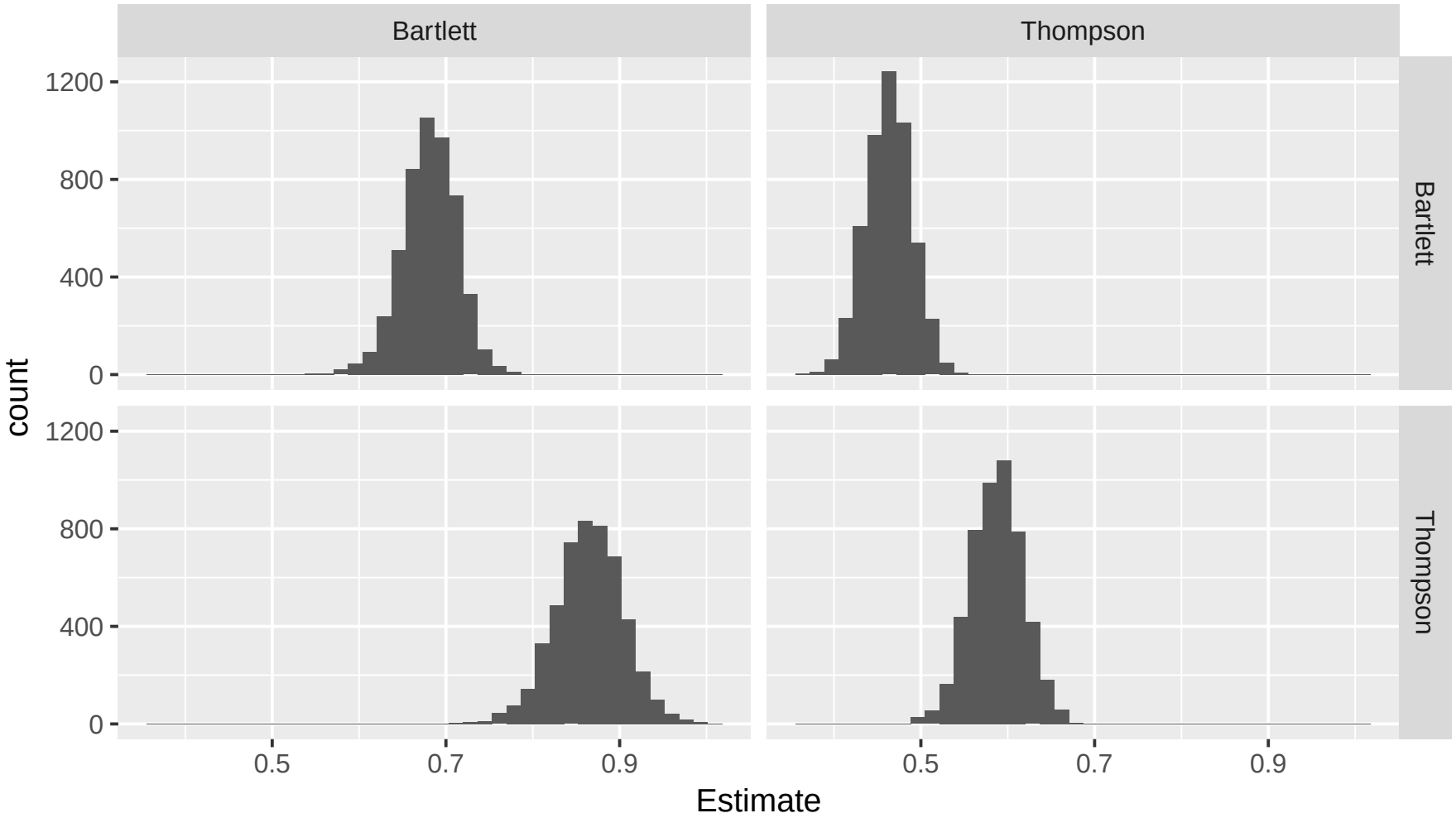
I generate 5000 data sets each with 400 observations from the model

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0.8 \\ 0.7 \\ 0.6 \\ 0.5 \end{bmatrix} F_X + \begin{bmatrix} e_1^X \\ e_2^X \\ e_3^X \\ e_4^X \end{bmatrix}, \quad \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0.8 \\ 0.7 \\ 0.6 \\ 0.5 \end{bmatrix} F_Y + \begin{bmatrix} e_1^Y \\ e_2^Y \\ e_3^Y \\ e_4^Y \end{bmatrix}$$

and

$$F_Y = 0 + 0.75 F_X + e.$$

Biased Results



Final Remarks

- The estimation methods that we introduced are mostly designed for continuous $\boldsymbol{X}$. If you have a discrete $\boldsymbol{X}$, other methods need to be used.
- The factor analysis that we introduced is often called exploratory factor analysis, as compared to confirmatory factor analysis.
  - Confirmatory factor analysis knows *a priori* which factor loadings should be zero, and no factor rotation is needed.