

# Latent Variable Model I - Factor Analysis

Factor Analysis (FA) shares with principal component analysis (PCA) similar objectives of dimension reduction and easy interpretation. While principal component analysis generally used as a mathematical technique, factor analysis is a statistical model, which has been the origin of the development of many popular statistical models with latent variables, including Structured Equation Models, Independent Component Analysis, and Probability Principal Component Analysis.

Data:  $n$  observations of  $p$ -variate vectors.

In standard factor analysis, the data are considered as  $n$  measurements of a common  $p$ -random vector.

We consider the common case that the  $n$  measurements are independent.

The probability distribution of the  $p$ -random vector is formulated as statistical model, called a factor model.

Goal of factor analysis: Factor analysis model seeks to explain the **covariance structure** of  $p$ -variate data with a few underlying unobservable random variables — called “common factors”.

Factor models construct the  $p$  random, observable variables as linear combinations of very few ( $\ll p$ ) underlying variables commonly called **factors** — unobserved, hidden, latent random variables — without altering the correlation structure of the data too much. It is desirable for the factors to have reasonable interpretations in terms of the subject matter.

Why one seeks “hidden” (a.k.a. latent) variables that are unobservable: It may not always be possible to measure the quantity of interest directly, such as “intelligence”.

Focus of factor model: The covariance structure of the  $p$  variables.

## 1 The Orthogonal Factor Model

(the population model)

The most basic factor model is the orthogonal factor model, where the latent factors are assumed to be uncorrelated.

The orthogonal factor model formulates a  $p$ -variate random vector  $\mathbf{X}$  of mean  $\boldsymbol{\mu}$  as a linear model of  $m \leq p$  underlying factors.

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon}$$

With dimensions indicated:

$$\mathbf{X}_{p \times 1} = \boldsymbol{\mu}_{p \times 1} + \mathbf{L}_{p \times m} \mathbf{F}_{m \times 1} + \boldsymbol{\varepsilon}_{p \times 1}$$

In detailed vector matrix form,

$$\begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_i \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \ell_{11} & \cdots & \ell_{1m} \\ \vdots & & \vdots \\ \ell_{i1} & \cdots & \ell_{im} \\ \vdots & & \vdots \\ \ell_{p1} & \cdots & \ell_{pm} \end{bmatrix} \begin{bmatrix} F_1 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_p \end{bmatrix} \quad (1)$$

In the above factor model:

$\mathbf{X}$  is the original, observable vector of  $p$  random variables.

$\mathbf{F}$  is the model-assumed, unobservable vector of  $m$  random components called factors.

Usually  $m \ll p$  to achieve dimension reduction.

The orthogonal factor model comes with a fair amount of assumptions and naming conventions.

- $\mathbf{X} = (X_1, \dots, X_p)'$  is the  $p$ -variate random vector that generated the observed data.  $E(\mathbf{X}) = \boldsymbol{\mu}$ .

- $\mathbf{F} = (F_1, \dots, F_m)'$  is a random vector with  $m \leq p$  **uncorrelated** components called **common factors**.

These common factors are assumed to be underlying **unobservable** random variables, usually normalized so that their mean vector is centered,

$$\mathbb{E}(\mathbf{F}) = \mathbf{0}_m, \quad m \leq p.$$

For the orthogonal factor model considered here, the common factors are uncorrelated (orthogonal), the variance-covariance matrix is

$$\text{Cov}(\mathbf{F}) = \mathbb{E}(\mathbf{F}\mathbf{F}') = \mathbf{I}_m, \quad m \leq p.$$

Consequently, common factors are independent under normality assumption.

- $\mathbf{L} = [\ell_{ij}]_{p \times m}$  is the matrix of coefficients called factor loadings.  $\ell_{ij}$  is the loading of the  $i$ th variable  $X_i$  on the  $j$ th common factor  $F_j$ . The loadings  $\ell_{ij}$  are model parameters to be estimated.

- $h_i^2 = \ell_{i1}^2 + \dots + \ell_{im}^2$  is called the  $i$ th communality, which is the portion in  $\text{Var}(X_i)$  contributed by the common factors.

- $\boldsymbol{\varepsilon} = (\epsilon_1, \dots, \epsilon_p)'$  is  $p$ -variate random vector of errors with independent components. Contrast to the common factors,  $\epsilon_i$  is called the  $i$ th specific factor.

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}_p, \quad \text{Cov}(\boldsymbol{\varepsilon}) = \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \text{diag}\{\psi_1, \dots, \psi_p\} = \boldsymbol{\Psi}.$$

- $\psi_i = \text{Var}(\epsilon_i)$  is the portion of  $\text{Var}(X_i)$  due to the specific factor  $\epsilon_i$ .  $\psi_i$  is called the specific variance or uniqueness of variable  $X_i$ . As shown below,

$$\psi_i = \text{Var}(X_i) - (\text{communality of } X_i) = \sigma_{ii} - h_i^2$$

- The errors and the common factors, both are unobservables, are assumed mutually independent:  $\boldsymbol{\varepsilon} \perp \mathbf{F}$ . Then  $\text{cov}(\epsilon_i, F_j) = E(\epsilon_i F_j) = 0$  for any  $i = 1, \dots, p, j = 1, \dots, m$ . Therefore,

$$\text{Cov}(\boldsymbol{\varepsilon}, \mathbf{F}) = \mathbb{E}(\boldsymbol{\varepsilon}\mathbf{F}') - \mathbb{E}(\boldsymbol{\varepsilon})\mathbb{E}(\mathbf{F})' = \mathbb{E}(\boldsymbol{\varepsilon}\mathbf{F}') = \mathbf{0}_{p \times m},$$

$$\text{Cov}(\mathbf{F}, \boldsymbol{\varepsilon}) = \mathbb{E}(\mathbf{F}\boldsymbol{\varepsilon}') - \mathbb{E}(\mathbf{F})\mathbb{E}(\boldsymbol{\varepsilon})' = \mathbb{E}(\mathbf{F}\boldsymbol{\varepsilon}') = \mathbf{0}_{m \times p}.$$

The above assumptions and the model relation (1) constitute the orthogonal factor model.

### Covariance structure for the orthogonal factor model

The orthogonal factor model  $\mathbf{X} = \boldsymbol{\mu} + \mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon}$  implies a specific covariance structure of the original variable  $\mathbf{X}$  and a covariance relation between  $\mathbf{X}$  and the unobservable common factors  $\mathbf{F}$ .

$$\text{Cov}(\mathbf{X}) = \mathbf{L}\mathbf{L}' + \boldsymbol{\Psi}, \quad \text{Cov}(\mathbf{X}, \mathbf{F}) = \mathbf{L} \quad (2)$$

*Proof.* Use the notation  $Cov(\mathbf{X}) = \Sigma = [\sigma_{ij}]_{p \times p}$ .

$$\begin{aligned}\sigma_{ij} &= cov(X_i, X_j) = cov(X_i - \mu_i, X_j - \mu_j) \\ &= \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] - \mathbb{E}(X_i - \mu_i)\mathbb{E}(X_j - \mu_j) \\ &= \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] \quad \text{since } \mathbb{E}(F_k) = \mathbb{E}(\epsilon_j) = 0, \text{ any } k, j \\ &= \mathbb{E}[(\ell_{i1}F_1 + \dots + \ell_{im}F_m + \epsilon_i)(\ell_{j1}F_1 + \dots + \ell_{jm}F_m + \epsilon_j)] \\ &= \ell_{i1}\ell_{j1} + \dots + \ell_{im}\ell_{jm} + E(\epsilon_i\epsilon_j) \quad \text{since } cov(F_k, \epsilon_j) = 0, \text{ any } k, j\end{aligned}$$

The first term is the  $(i, j)$ th element of matrix  $\mathbf{LL}'$ . The second term is the  $(i, j)$ th element of matrix  $\Psi$ , with entry 0 for  $i \neq j$ , of diagonal entry  $\psi_i$  for  $i = j$ . Hence  $Cov(\mathbf{X}) = \mathbf{LL}' + \Psi$ , with entries

$$\begin{aligned}\sigma_{ij} &= cov(X_i, X_j) = \ell_{i1}\ell_{j1} + \dots + \ell_{im}\ell_{jm}, \quad i \neq j \\ \sigma_{ii} &= var(X_i) = \ell_{i1}^2 + \dots + \ell_{im}^2 + \psi_i\end{aligned}$$

Notice that  $\sigma_{ii} = h_i^2 + \psi_i$ . Furthermore, since

$$\mathbb{E}[(X_i - \mu_i)F_j] = \mathbb{E}[(\ell_{i1}F_1 + \dots + \ell_{im}F_m + \epsilon_i)F_j] = \mathbb{E}(\ell_{ij}F_j^2) + \mathbb{E}(\epsilon_i F_j) = \ell_{ij}var(F_j) + 0 = \ell_{ij},$$

the  $(i, j)$ th entry of  $\mathbf{L}$ . We have obtained the covariance relation of the original variables and common factors

$$Cov(\mathbf{X}, \mathbf{F}) = Cov[\mathbf{X} - \boldsymbol{\mu}, \mathbf{F}] = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})\mathbf{F}'] = \mathbf{L}$$

□

#### Remarks (on orthogonal factor models)

- Latent variables  
FA — the (orthogonal) factor model, attempts to describe the observed data by latent, unobservable variables.
- Dimension reduction  
FA achieves dimension reduction by expressing the original, higher dimensional data in fewer underlying latent variables.
- Non-uniqueness  
There is a lack of identifiability of the factors.  
Consider an  $m$ -vector  $\mathbf{F}^* = \mathbf{T}'\mathbf{F}$ , an arbitrary orthogonal linear transformation of  $\mathbf{F}$  with  $m$  common factors. The transformation is represented by an  $m \times m$  orthogonal matrix  $\mathbf{T}$ , with the property

$$\mathbf{T}\mathbf{T}' = \mathbf{T}'\mathbf{T} = \mathbf{I}_m$$

Then,

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{F} + \boldsymbol{\varepsilon} = \mathbf{L}^*\mathbf{F}^* + \boldsymbol{\varepsilon} \quad (3)$$

The model has the new loading  $\mathbf{L}^* = \mathbf{L}\mathbf{T}$  and new common factor  $\mathbf{F}^* = \mathbf{T}'\mathbf{F}$ . Now

$$\mathbf{X} - \boldsymbol{\mu} = \mathbf{L}^*\mathbf{F}^* + \boldsymbol{\varepsilon}$$

is also an orthogonal  $m$ -factor model, with the desired properties (exercises)

$$E(\mathbf{F}^*) = \mathbf{0}_m, \quad Cov(\mathbf{F}^*) = \mathbf{I}_m, \quad Cov(\mathbf{F}^*, \boldsymbol{\varepsilon}) = \mathbf{0}_{m \times p}$$

and the same covariance

$$Cov(\mathbf{X}) = \mathbf{L}^*\mathbf{L}^{*'} + \Psi = \mathbf{L}\mathbf{T}\mathbf{T}'\mathbf{L}' + \Psi = \mathbf{L}\mathbf{L}' + \Psi$$

but different factors and loadings.

Therefore,

- The common factors and factor loadings are unique only up to an orthogonal transformation.
- Given only the variance-covariance matrix of the original variable  $\Sigma_x$  (without other constraints), the common factors and factor loadings cannot be recovered uniquely.

- Possible non-existence of factor models

Because factor models impose demanding variance-covariance structure requirements, not all covariance matrices  $\Sigma$  can have a proper factor model with a given number of factors.

#### Sample data factor model structure

In the above, a factor model is constructed on random vector  $\mathbf{X}$  representing the population model. In applications, sample data are used to fit the factor model and to estimate the parameters.

Assume that there are  $n$  observations. Recall the convention of the data format.

$$[x_{jk}]_{n \times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix} \begin{array}{l} \leftarrow \text{1st } (p\text{-variate}) \text{ observation} \\ \leftarrow \text{2nd observation} \\ \vdots \\ \leftarrow j\text{th observation} \\ \vdots \\ \leftarrow n\text{th observation} \end{array}$$

The data are viewed as observed values of a random sample of  $n$  random vectors, which are denoted as

$$[X_{jk}]_{n \times p} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1k} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2k} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{j1} & X_{j2} & \cdots & X_{jk} & \cdots & X_{jp} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nk} & \cdots & X_{np} \end{bmatrix}$$

where each row is a random sample point from a  $p$ -variate random vector.

To relate sample data with the population model we derived earlier, consider the transpose of the random sample matrix. Now each column is a random sample point from a  $p$ -variate random vector.

$$[X_{jk}]'_{n \times p} = [X_{kj}]_{p \times n} = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{j1} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{j2} & \cdots & X_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{1k} & X_{2k} & \cdots & X_{jk} & \cdots & X_{nk} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{1p} & X_{2p} & \cdots & X_{jp} & \cdots & X_{np} \end{bmatrix}_{p \times n} \quad (4)$$

The  $j$ th sample item (a.k.a.  $j$ th measurement), a  $p$ -valued vector, is assume to have the structure

$$\begin{bmatrix} X_{j1} \\ \vdots \\ X_{ji} \\ \vdots \\ X_{jp} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_i \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \ell_{11} & \cdots & \ell_{1m} \\ \vdots & \vdots & \vdots \\ \ell_{i1} & \cdots & \ell_{im} \\ \vdots & \vdots & \vdots \\ \ell_{p1} & \cdots & \ell_{pm} \end{bmatrix} \begin{bmatrix} F_{1j} \\ \vdots \\ F_{mj} \end{bmatrix} + \begin{bmatrix} \epsilon_{1j} \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_{pj} \end{bmatrix}, \quad j = 1, \dots, n.$$

Where  $F_{ij}$  reflects the value of latent factor  $i$  for object  $j$  or item  $j$  ( $j = 1, \dots, n$ ).

Note that, the means and the entries in the loading matrix  $\mathbf{L}$  are the parameters of the FA model, common to every observation  $j, j = 1, \dots, n$ .

To consider the whole sample of  $n$  sample items, matrix form are needed to express each term. For example, the left hand side of the model equation becomes an  $p \times n$  matrix with  $n$  column vectors, which is the transposed sample matrix (4).

In matrix form, observed data matrix under factor model can be written in the (transposed) form

$$\mathbf{X}_{p \times n} = \boldsymbol{\mu}_{p \times n} + \mathbf{L}_{p \times m} \mathbf{F}_{m \times n} + \boldsymbol{\epsilon}_{p \times n}$$

where  $\boldsymbol{\mu}_{p \times n}$  denotes the  $p \times n$  matrix with column  $\equiv \boldsymbol{\mu}$ ,  $\mathbf{F}_{m \times n}$  is a non-observable  $p \times n$  matrix with  $j$ th column  $= [F_{1j} \cdots F_{mj}]'$ .

Sometimes the transform format of  $\tilde{\mathbf{F}} = \mathbf{F}^T$  is used in the model expression instead of  $\mathbf{F}$ , the data matrix becomes

$$\mathbf{X}_{p \times n} = \boldsymbol{\mu}_{p \times n} + \mathbf{L}_{p \times m} (\tilde{\mathbf{F}}_{n \times m})^T + \boldsymbol{\epsilon}_{p \times n}$$

In this transposed form,  $j$ th column of the data matrix contains the  $j$ th observation,  $j$ th column of the factor matrix  $\mathbf{F}$  or  $j$ th row of the factor matrix  $\tilde{\mathbf{F}}$  represents the (latent) factor values of the  $j$ th observation or  $j$ th object. The notation of the transformed  $\mathbf{F}$  is useful if the transposed-back expression for the data matrix is needed, where each row corresponds to an observation. This twice-transposed data and its FA structure has the dimensions

$$\mathbf{X}_{n \times p} = \boldsymbol{\mu}_{n \times p} + \tilde{\mathbf{F}}_{n \times m} \mathbf{L}'_{m \times p} + \boldsymbol{\epsilon}_{n \times p}$$

The observed  $n \times p$  data values are used to estimate the population factor model (1). The correlation structure of the factor model given data is the focus of estimation and inference.

- Loading matrix  $\mathbf{L}$  is the most wanted, especially the off-diagonal entries:

$$\text{Cov}(\mathbf{X}) = \mathbf{L}\mathbf{L}' + \Psi$$

- The population factor matrix has the properties

$$\mathbb{E}(\mathbf{F}) = \mathbf{0}_m, \quad \text{Cov}(\mathbf{F}) = E(\mathbf{F}\mathbf{F}') = \mathbf{I}_m, \quad (m \leq p)$$

The properties impose orthogonality or uncorrelated-ness between two common factors in the sample factor model:

$$\sum_{j=1}^n F_{ij} F_{kj} = \delta_{ik} = \begin{cases} 1, & i = k, \\ 0, & i \neq k, \end{cases} \quad i, k = 1, \dots, m.$$

- Analogously, independence of factor and error  $\mathbf{F} \perp\!\!\!\perp \boldsymbol{\epsilon}$  implies uncorrelated-ness in the population factor model, also imposes the constraint

$$\sum_{j=1}^n F_{ij} \epsilon_{kj} = 0, \quad \text{for } i = 1, \dots, m \text{ and } k = 1, \dots, p.$$

## 2 Estimation of factor models

A factor model contains many parameters to be estimated from data:  $\mu_i, \ell_{ij}$ , and  $\psi_i$  for  $i = 1, \dots, p, j = 1, \dots, m$ .

Two common methods are used to estimate the parameters in the factor model:

- The principal component method (PC)
- The maximum likelihood method (ML)

The PC method is intuitive and easy to carry out, however it is an approximation method.

The ML method needs constraints to ensure identifiability, and it may not exist for a given number of factors.

The ML approach has certain desirable properties and is often preferred.

### 2.1 The Principal Factor Estimation Method

The principal component method is easy to implement, thus commonly used in preliminary estimation of factor loadings.

Recall that the covariance matrix  $\Sigma$  can have  $p$  orthonormal eigenvectors:

$$\Sigma[\mathbf{e}_1 \cdots \mathbf{e}_p] = [\lambda_1 \mathbf{e}_1 \cdots \lambda_p \mathbf{e}_p]$$

with

$$\lambda_1 \geq \cdots \geq \lambda_p \geq 0, \quad \mathbf{e}_i' \mathbf{e}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

Then the covariance matrix  $\Sigma$  can have a spectral decomposition by its eigenvalues  $\lambda_i$  and the orthonormal eigenvectors  $\mathbf{e}_i$ .

$$\Sigma = [\lambda_1 \mathbf{e}_1 \cdots \lambda_p \mathbf{e}_p] \begin{bmatrix} \mathbf{e}_1' \\ \vdots \\ \mathbf{e}_p' \end{bmatrix} = \lambda_1 \mathbf{e}_1 \mathbf{e}_1' + \cdots + \lambda_p \mathbf{e}_p \mathbf{e}_p'$$

By the non-negativeness of  $\lambda_i$ , we may rewrite the covariance matrix as

$$\Sigma = [\lambda_1 \mathbf{e}_1 \cdots \lambda_p \mathbf{e}_p] \begin{bmatrix} \mathbf{e}_1' \\ \vdots \\ \mathbf{e}_p' \end{bmatrix} = [\sqrt{\lambda_1} \mathbf{e}_1 \cdots \sqrt{\lambda_p} \mathbf{e}_p] \begin{bmatrix} \sqrt{\lambda_1} \mathbf{e}_1' \\ \vdots \\ \sqrt{\lambda_p} \mathbf{e}_p' \end{bmatrix}$$

Thus the spectral decomposition can be used to factor the covariance matrix as

$$\Sigma = \mathbf{L}_p \mathbf{L}_p'$$

with

$$\mathbf{L}_p = \begin{bmatrix} \sqrt{\lambda_1} \mathbf{e}_1 & \sqrt{\lambda_2} \mathbf{e}_2 & \cdots & \sqrt{\lambda_p} \mathbf{e}_p \end{bmatrix} = [\ell_{ij}]_{p \times p}$$

The  $p$ -variate vector  $\mathbf{e}_i$  is the  $i$ th principal component (short for the  $i$ th principal component direction vector), for  $i = 1, \dots, p$ . The factor loadings on the  $j$ th factor

$$\begin{bmatrix} \ell_{1j} \\ \vdots \\ \ell_{pj} \end{bmatrix} = \sqrt{\lambda_j} \mathbf{e}_j$$

are the coefficients of the  $j$ th principal components scaled by multiplying  $\sqrt{\lambda_j}$ .

The factorization  $\Sigma = \mathbf{L}_p \mathbf{L}'_p$  is exact but not very useful, since the original  $p$  dimensional space is represented by  $p$  factors, no dimension reduction is achieved.

To reduce the dimension of the factors, consider a truncated version of  $\mathbf{L}_p$  with fewer columns:

$$\mathbf{L} = \mathbf{L}_m = \begin{bmatrix} \sqrt{\lambda_1} \mathbf{e}_1 & \cdots & \sqrt{\lambda_m} \mathbf{e}_m \end{bmatrix} = [\ell_{ij}]_{p \times m}, \quad m < p.$$

Denote

$$\Psi = \text{diag}\{\psi_1, \dots, \psi_p\}, \quad \text{where } \psi_i = \sigma_{ii} - (\ell_{i1}^2 + \cdots + \ell_{im}^2)$$

Then the covariance matrix may have a useful truncated version of the decomposition with  $m$  factors:

$$\Sigma \approx \mathbf{L} \mathbf{L}' + \Psi$$

#### Estimation steps in applications (the PC method)

- Let the sample covariance matrix

$$\mathbf{S} = [s_{ij}]_{p \times p} = \hat{\Sigma}$$

That is, use the sample covariance matrix as the estimator of the population covariance matrix.

- Denote the eigenvalues and eigenvectors of  $\mathbf{S}$  as  $\hat{\lambda}_i$  and  $\hat{\mathbf{e}}_i$  for  $i = 1, \dots, p$ , with  $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq 0$ .
- For  $m < p$ , define the factor matrix with  $m$  estimated common factors as

$$\tilde{\mathbf{L}} = \begin{bmatrix} \sqrt{\hat{\lambda}_1} \hat{\mathbf{e}}_1 & \cdots & \sqrt{\hat{\lambda}_m} \hat{\mathbf{e}}_m \end{bmatrix} = [\tilde{\ell}_{ij}]_{p \times m}$$

- Define the estimates of the specific variances as

$$\tilde{\Psi} = \text{diag}\{\tilde{\psi}_1, \dots, \tilde{\psi}_p\}, \quad \tilde{\psi}_i = s_{ii} - (\tilde{\ell}_{i1}^2 + \cdots + \tilde{\ell}_{im}^2)$$

- The communalities of the model are estimated as

$$\tilde{h}_i^2 = \tilde{\ell}_{i1}^2 + \cdots + \tilde{\ell}_{im}^2$$

#### Remarks

- Since the principal components method for factor model is based on truncating the full spectral decomposition of the sample covariance matrix  $\mathbf{S}$ , the estimated loadings  $\tilde{\ell}_{ij}$  ( $j = 1, \dots, m$ ) do not change with the choice of  $m$ , the number of factors.

- The residual matrix is defined as

$$\mathbf{S} - (\tilde{\mathbf{L}} \tilde{\mathbf{L}}' + \tilde{\Psi})$$

By the definitions of  $\tilde{\psi}_i$ , the diagonal elements of the residual matrix are zero, by the principal component method.

The magnitude of the off-diagonal elements are bounded by the sum of squares of the last  $p - m$  eigenvalues of  $\Sigma$  that are neglected by the factor model:

$$\text{Sum of squared entries of the residual matrix } \left( \mathbf{S} - (\tilde{\mathbf{L}} \tilde{\mathbf{L}}' + \tilde{\Psi}) \right) \leq \hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_p^2.$$

The above upper bound can be derived from

$$\begin{aligned} \text{Sum of squared entries of } \left( \mathbf{S} - (\tilde{\mathbf{L}} \tilde{\mathbf{L}}' + \tilde{\Psi}) \right) &\leq \text{Sum of squared entries of } \left( \mathbf{S} - \tilde{\mathbf{L}} \tilde{\mathbf{L}}' \right) \\ &= \text{Sum of squared entries of } \left( \hat{\lambda}_{m+1} \hat{\mathbf{e}}_{m+1} \hat{\mathbf{e}}_{m+1}' + \cdots + \hat{\lambda}_p \hat{\mathbf{e}}_p \hat{\mathbf{e}}_p' \right) \\ &= \hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_p^2 \end{aligned}$$

The last equality comes from the eigenvalue-eigenvector decomposition of matrix  $\sum_{i=m+1}^p \hat{\lambda}_i \hat{\mathbf{e}}_i \hat{\mathbf{e}}_i'$ , and the matrix trace property that  $\sum_{i=1}^p \sum_{j=1}^p a_{ij}^2 = \text{tr}(\mathbf{A} \mathbf{A}')$  for  $\mathbf{A} = [a_{ij}]_{p \times p}$ .

Therefore, one may choose the number of factors by selecting  $m$  such that

$$\hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_p^2 \approx 0 \quad (\text{if good enough and feasible})$$

That is, the sum of squares of the last  $p - m$  eigenvalues are acceptably small.

- The proportion of the total sample variance due to the  $j$ th common factor is

$$\frac{\hat{\ell}_{1j}^2 + \cdots + \hat{\ell}_{pj}^2}{s_{11} + \cdots + s_{pp}}$$

Using the PC methods,  $\begin{bmatrix} \ell_{1j} \\ \vdots \\ \ell_{pj} \end{bmatrix} = \sqrt{\lambda_j} \mathbf{e}_j$ . Hence

$$\frac{\hat{\ell}_{1j}^2 + \cdots + \hat{\ell}_{pj}^2}{s_{11} + \cdots + s_{pp}} = \frac{\hat{\lambda}_j \|\mathbf{e}_j\|^2}{\text{tr}(\mathbf{S})} = \frac{\hat{\lambda}_j}{\sum_{i=1}^p \hat{\lambda}_i}$$

- Similar to the case of principal component analysis, it is more common to use normalized variables of variance one and centered variables of mean zero.

For example, the  $j$ th observation of the  $k$ th variable  $x_{jk}$  can be scaled as  $x_{jk}/\sqrt{s_{kk}}$ , then mean adjusted as  $(x_{jk} - \bar{x}_k)/\sqrt{s_{kk}}$ . The sample covariance matrix for the normalized data becomes the correlation matrix  $\mathbf{R}$  of the original  $x_{jk}$ 's. Since  $\text{tr}(\mathbf{R}) = \text{trace}(\mathbf{R}) = p$ , the proportion of the total sample variance due to the  $j$ th common factor becomes

$$\frac{\hat{\lambda}_j}{\text{tr}(\mathbf{R})} = \frac{\hat{\lambda}_j}{p}$$

## 2.2 The Maximum Likelihood Estimation of Factors

Maximum likelihood method is usually preferred to estimate parameters in factor models. Maximum likelihood approach in factor analysis assumes that the variables are of joint normal distribution.

If the common factors  $\mathbf{F}$  and the specific factors  $\boldsymbol{\varepsilon}$  are assumed to be normally distributed, then

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{L}\mathbf{F} + \boldsymbol{\varepsilon} \sim N_p(\boldsymbol{\mu}, \Sigma)$$

is of  $p$ -variate normal distribution. Let  $\mathbf{X}_j$  denote the  $j$ th observation with common factors  $\mathbf{F}$  at values  $\mathbf{F}_j$ ,  $j = 1, \dots, n$ .

The joint likelihood function of the random sample  $\mathbf{X}_1, \dots, \mathbf{X}_n$  is

$$L(\boldsymbol{\mu}, \Sigma) = L(\boldsymbol{\mu}, \Sigma | \mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{j=1}^n f(\mathbf{x}_j) = \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} e^{-\frac{1}{2} \sum_{j=1}^n (\mathbf{x}_j - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x}_j - \boldsymbol{\mu})}$$

By the property (e.g., as derived in the chapter "The multivariate normal distribution")

$$-\frac{1}{2} \text{tr} \left\{ \Sigma^{-1} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})' \right\} = -\frac{1}{2} \text{tr} [\Sigma^{-1} (n\mathbf{S}_n)] - \frac{1}{2} \text{tr} \{ \Sigma^{-1} [n(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})'] \}$$

the joint likelihood function of a random sample  $\mathbf{X}_1, \dots, \mathbf{X}_n$  can also be written as

$$L(\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} e^{-\frac{1}{2} \{ \text{tr}[\Sigma^{-1} n\mathbf{S}_n] + n(\bar{\mathbf{x}} - \boldsymbol{\mu})' \Sigma^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \}}$$

Other than being a function of the components of  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)'$ , the likelihood function depends on the  $p \times m$  entries of  $\mathbf{L} = [\ell_{ij}]_{p \times m}$  and the  $p$  elements of  $\Psi = \text{diag}\{\psi_1, \dots, \psi_p\}$  through  $\Sigma$ , according to the model relation  $\Sigma = \mathbf{L}\mathbf{L}' + \Psi$ .

Constraints of ML estimation\* (\* means Optional)

Due to the non-uniqueness nature of  $\mathbf{L}$  as indicated in (3), the parameters in  $\mathbf{L}$  are not uniquely identifiable.

Parameter constraints are needed. There are various ways to add the constraints.

A common constraint used in practice is

$$\mathbf{L}'\Psi^{-1}\mathbf{L} = \Delta \quad (\text{a diagonal matrix}) \quad (5)$$

The above constraint yields the eigenvalue-eigenvector equation

$$\left( \Psi^{-1/2} \Sigma \Psi^{-1/2} \right) \left( \Psi^{-1/2} \mathbf{L} \right) = \left( \Psi^{-1/2} \mathbf{L} \right) (I + \Delta) \quad (6)$$

under the FA conditions

$$\Sigma = \mathbf{L}\mathbf{L}' + \Psi, \quad \mathbf{L}'\mathbf{L} = I,$$

because

$$\begin{aligned} \left( \Psi^{-1/2} \Sigma \Psi^{-1/2} \right) \left( \Psi^{-1/2} \mathbf{L} \right) &= \left( \Psi^{-1/2} (\mathbf{L}\mathbf{L}' + \Psi) \Psi^{-1/2} \right) \left( \Psi^{-1/2} \mathbf{L} \right) \\ &= \left( \Psi^{-1/2} \mathbf{L}\mathbf{L}' \Psi^{-1/2} \right) \left( \Psi^{-1/2} \mathbf{L} \right) + \left( \Psi^{-1/2} \Psi \Psi^{-1/2} \right) \left( \Psi^{-1/2} \mathbf{L} \right) \\ &= \left( \Psi^{-1/2} \mathbf{L} \right) \left( \mathbf{L}' \Psi^{-1} \mathbf{L} \right) + \Psi^{-1/2} \mathbf{L} \\ &= \left( \Psi^{-1/2} \mathbf{L} \right) \Delta + \Psi^{-1/2} \mathbf{L} \\ &= \left( \Psi^{-1/2} \mathbf{L} \right) (\Delta + I) \end{aligned}$$

Now the columns of  $\Psi^{-1/2}\mathbf{L}$  are (normalized) eigenvectors of matrix  $\Psi^{-1/2}\Sigma\Psi^{-1/2}$  with eigenvalues on the diagonal of the diagonal matrix  $\Delta + I$ .

Using the ML estimate  $S_n = \hat{\Sigma}$  and initial values of  $\Psi$ , an initial  $\mathbf{L}$  can be obtained.

$\mathbf{L}$  matrix is then estimated iteratively (details see 9A in J&W).

The resulting Maximum Likelihood estimates are

$$\hat{\mathbf{L}} = [\hat{\ell}_{ij}]_{p \times s}, \quad \hat{\Psi} = \text{diag}\{\hat{\psi}_1, \dots, \hat{\psi}_p\}, \quad \hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}, \quad \text{subject to } \hat{\mathbf{L}}' \hat{\Psi}^{-1} \hat{\mathbf{L}} \text{ being diagonal.}$$

The ML estimates of the communalities are

$$\hat{h}_i^2 = \hat{\ell}_{i1}^2 + \dots + \hat{\ell}_{im}^2, \quad i = 1, \dots, p.$$

The proportion of total sample variance contributed by the  $j$ th common factor is

$$\frac{\hat{\ell}_{1j}^2 + \dots + \hat{\ell}_{pj}^2}{s_{11} + \dots + s_{pp}}, \quad j = 1, \dots, m.$$

Notes: Number of parameters and constraints in ML method\*

- $\text{Cov}(\mathbf{X})$ :  $p(p+1)/2$  distinct elements.
- $\mathbf{L}$ :  $p \times m$  parameters,  $\text{Cov}(\mathbf{F}) = \mathbf{I}_m$ .
- $\Psi$ :  $p$  diagonal parameters.
- $m(m-1)/2$  constraints in (5).

Large sample LR test for number of common factors\* (Optional)

Under the normal distribution assumption, the adequacy of the factor model can be tested by

$H_o$ :  $\Sigma = \mathbf{L}\mathbf{L}' + \Psi$  with  $m$  factors.

$H_a$ :  $\Sigma$  does not have the restricted structure in  $H_o$ .

The likelihood ratio statistic for testing  $H_o$  is

$$-\ln \Lambda = n \ln \frac{|\hat{\Sigma}|}{|S_n|}, \quad \hat{\Sigma} = \hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Phi}$$

A common approximation method is Bartlett's approximation or Bartlett's correction

$$(n-1-(2p+4m+5)/6) \ln \frac{|\hat{\mathbf{L}}\hat{\mathbf{L}}' + \hat{\Phi}|}{|S_n|} \sim \chi_{[(p-m)^2-p-m]/2}^2 \quad \text{under } H_o$$

which is used for the hypothesis test.