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 is 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 dataset consists of n observations of a common p-random vector.

We consider the common case that the n observations or 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 ($<< p$) 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

1.1 Population factor analysis model

We consider a random vector $\boldsymbol{X}$ with p univariate components, from which the data are generated and sampled. The factor model imposes a statistic model structure in $\boldsymbol{X}$.

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 $\boldsymbol{X}$ of mean $\boldsymbol{\mu}$ as a linear model of $m \leq p$ underlying factors.

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

With dimensions indicated:

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

In detailed vector matrix form,

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

In the above factor model:

$\boldsymbol{X}$ is the original, observable vector of p random variables.
$\boldsymbol{F}$ is the model-assumed, unobservable vector of m random components called factors.
Usually $m << p$ to achieve dimension reduction.

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

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

- $\boldsymbol{F} = (F_1, \cdots, F_m)'$ 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,

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

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

  $$Cov(\boldsymbol{F}) = \mathbb{E}(\boldsymbol{F}\boldsymbol{F}') = \boldsymbol{I}_m, \qquad m \leq p.$$

  Consequently, common factors are independent under normality assumption.

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

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

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

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

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

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

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

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

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

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

Covariance structure for the orthogonal factor model

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

$$Cov(\boldsymbol{X}) = \boldsymbol{L}\boldsymbol{L}' + \Psi, \qquad Cov(\boldsymbol{X}, \boldsymbol{F}) = \boldsymbol{L} \tag{2}$$

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

To show $Cov(\boldsymbol{X}) = \boldsymbol{L}\boldsymbol{L}'$, write

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

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

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

To show $Cov(\boldsymbol{X}, \boldsymbol{F}) = \boldsymbol{L}$, note that

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

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

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

□

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 $\boldsymbol{F}^* = \boldsymbol{T}'\boldsymbol{F}$, an arbitrary orthogonal linear transformation of $\boldsymbol{F}$ with m common factors. The transformation is represented by an $m \times m$ orthogonal matrix $\boldsymbol{T}$, with the property

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

  Then,

  $$\boldsymbol{X} - \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{F} + \boldsymbol{\varepsilon} = \boldsymbol{L}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{F} + \boldsymbol{\varepsilon} = \boldsymbol{L}^*\boldsymbol{F}^* + \boldsymbol{\varepsilon} \tag{3}$$

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

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

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

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

  and the same covariance

  $$Cov(\boldsymbol{X}) = \boldsymbol{L}^*\boldsymbol{L}^{*\prime} + \Psi = \boldsymbol{L}TT'\boldsymbol{L}' + \Psi = \boldsymbol{L}\boldsymbol{L}' + \Psi \tag{4}$$

  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 Σ_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 Σ can have a proper factor model with a given number of factors.

1.2 Sample factor analysis

In the above, a factor model is constructed on random vector $\boldsymbol{X}$ 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.

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

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

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

where each row is a random sample point from the p-variate random vector $\boldsymbol{X}$ in the population model.

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.

$$[X_{jk}]'_{n\times p} = [X_{kj}]_{p\times n} = \begin{bmatrix}
X_{11} & X_{21} & \cdots & X_{j1} & \cdots & \cdots & X_{n1} \\
X_{12} & X_{22} & \cdots & X_{j2} & \cdots & \cdots & X_{n2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
X_{1k} & X_{2k} & \cdots & X_{jk} & \cdots & \cdots & X_{nk} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
X_{1p} & X_{2p} & \cdots & X_{jp} & \cdots & \cdots & X_{np}
\end{bmatrix} \tag{5}$$

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

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

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

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

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 (5).

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

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

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

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

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

In this transposed form, jth column of the data matrix contains the jth observation, jth column of the factor matrix $\boldsymbol{F}$ or jth row of the factor matrix $\tilde{\boldsymbol{F}}$ represents the (latent) factor values of the jth observation or jth object. The notation of the transformed $\boldsymbol{F}$ 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

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

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

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

- The population factor matrix has the properties

$$\mathbb{E}(\boldsymbol{F}) = \boldsymbol{0}_m, \qquad Cov(\boldsymbol{F}) = E(\boldsymbol{F}\boldsymbol{F}') = \boldsymbol{I}_m, \qquad (m \le p)$$

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

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

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

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

2 Estimation of factor models

A factor model contains many parameters to be estimated from data: μ_i, ℓ_{ij}, and ψ_i for $i = 1, \cdots, p,\ j = 1, \cdots, m$.

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 Σ can have p orthonormal eigenvectors:

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

with

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

Then the covariance matrix Σ can have a spectral decomposition by its eigenvalues λ_i and the orthonormal eigenvectors $\boldsymbol{e}_i$.

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

By the non-negativeness of λ_i, we may rewrite the covariance matrix as

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

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

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

with

$$\boldsymbol{L}_p = \left[\sqrt{\lambda_1}\boldsymbol{e}_1 \vdots \sqrt{\lambda_2}\boldsymbol{e}_2 \vdots \cdots \vdots \sqrt{\lambda_p}\boldsymbol{e}_p\right] = [\ell_{ij}]_{p\times p}$$

The p-variate vector $\boldsymbol{e}_i$ is the ith principal component (short for the ith principal component direction vector), for $i = 1, \cdots, p$. The factor loadings on the jth factor

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

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

The factorization $\Sigma = \boldsymbol{L}_p\boldsymbol{L}_p'$ 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 $\boldsymbol{L}_p$ with fewer columns:

$$\boldsymbol{L} = \boldsymbol{L}_m = \left[\sqrt{\lambda_1}\boldsymbol{e}_1 \vdots \ \cdots \ \vdots \sqrt{\lambda_m}\boldsymbol{e}_m\right] = [\ell_{ij}]_{p\times m}, \qquad m < p.$$

Denote

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

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

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

Other than being simple and useful, another justification is that L_mL_m' is the optional m-dimensional approximation of Σ by Frobenius norm of the residual matrix, as stated

Estimation steps in applications (the PC method)

- Let the sample covariance matrix
$$\boldsymbol{S} = [s_{ij}]_{p\times p} = \hat{\Sigma}$$
That is, use the sample covariance matrix as the estimator of the population covariance matrix.
- Denote the eigenvalues and eigenvectors of $\boldsymbol{S}$ as $\hat{\lambda}_i$ and $\hat{\boldsymbol{e}}_i$ for $i = 1, \cdots, p$, 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
$$\tilde{\boldsymbol{L}} = \left[\sqrt{\hat{\lambda}_1}\hat{\boldsymbol{e}}_1 \vdots \cdots \vdots \sqrt{\hat{\lambda}_m}\hat{\boldsymbol{e}}_m\right] = [\tilde{\ell}_{ij}]_{p\times m}$$
- Define the estimates of the specific variances as
$$\tilde{\Psi} = diag\{\tilde{\psi}_1, \cdots, \tilde{\psi}_p\}, \qquad \tilde{\psi}_i = s_{ii} - (\tilde{\ell}_{i1}^2 + \cdots + \tilde{\ell}_{im}^2)$$
- 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 $\boldsymbol{S}$, the estimated loadings $\tilde{\ell}_{ij}$ $(j = 1, \cdots, m)$ do not change with the choice of m, the number of factors.
- The residual matrix is defined as
$$\boldsymbol{S} - (\tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}' + \tilde{\Psi})$$
By the definitions of $\tilde{\psi}_i$, the diagonal elements of the residual matrix are zero, by the principal component method.
The squares of magnitude of the off-diagonal elements are bounded by the sum of squares of the last $p - m$ eigenvalues of $\hat{\Sigma}$ that are neglected by the factor model:
$$\text{Sum of squared entries of the residual matrix } \left(\boldsymbol{S} - (\tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}' + \tilde{\Psi})\right) \leq \hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_p^2.$$
The above upper bound can be derived from
$$\begin{aligned}\text{Sum of squared entries of } \left(\boldsymbol{S} - (\tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}' + \tilde{\Psi})\right) &\leq \text{Sum of squared entries of } \left(\boldsymbol{S} - \tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}'\right) \\ &= \text{Sum of squared entries of } \left(\hat{\lambda}_{m+1}\hat{\boldsymbol{e}}_{m+1}\hat{\boldsymbol{e}}_{m+1} + \cdots + \hat{\lambda}_p\hat{\boldsymbol{e}}_p\hat{\boldsymbol{e}}_p'\right) \\ &= \hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_p^2\end{aligned}$$
The last equality comes from the eigenvalue-eigenvector decomposition of matrix
$$\boldsymbol{S} = \sum_{i=1}^{p} \hat{\lambda}_p\hat{\boldsymbol{e}}_i\hat{\boldsymbol{e}}_i', \qquad \boldsymbol{S} - \tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}' = \sum_{i=m+1}^{p} \hat{\lambda}_p\hat{\boldsymbol{e}}_i\hat{\boldsymbol{e}}_i'$$
and the matrix trace property that $\sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2 = tr(AA')$ for $A = [a_{ij}]_{p\times p}$.
Therefore, one may choose the number of factors by selecting m such that
$$\hat{\lambda}_{m+1}^2 + \cdots + \hat{\lambda}_p^2 \approx 0 \qquad \text{(if good enough and feasible)}$$
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 jth common factor is
$$\frac{\hat{\ell}_{1j}^2 + \cdots + \hat{\ell}_{pj}^2}{s_{11} + \cdots + s_{pp}}$$
Using the PC methods, $\begin{bmatrix} \ell_{1j} \\ \vdots \\ \ell_{pj} \end{bmatrix} = \sqrt{\lambda_j}\boldsymbol{e}_j$. Hence
$$\frac{\hat{\ell}_{1j}^2 + \cdots + \hat{\ell}_{pj}^2}{s_{11} + \cdots + s_{pp}} = \frac{\hat{\lambda}_j\|\boldsymbol{e}_j\|^2}{tr(\boldsymbol{S})} = \frac{\hat{\lambda}_j}{\sum_{i=1}^{p}\hat{\lambda}_i}$$

- 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 jth observation of the kth 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 $\boldsymbol{R}$ of the original x_{jk}'s. Since

$$tr(\boldsymbol{R}) = trace(\boldsymbol{R}) = p$$

  the proportion of the total sample variance due to the jth common factor becomes

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

2.2 The Maximum Likelihood Estimation Method

Maximum likelihood method is usually preferred in parameter estimation for factor models.

To use the maximum likelihood method, we need to make distribution assumptions on the random variable $\boldsymbol{X}$ that the observations are sampled from.

Maximum likelihood approach in factor analysis assumes that the variables are of joint normal distribution.

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

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

is of p-variate normal distribution.

Let $\boldsymbol{X}_j \in \mathbb{R}^p$ denote the jth observation, $j = 1, \cdots, n$.
Then each $\boldsymbol{X}_j$ corresponds to an $\boldsymbol{F}_j \in \mathbb{R}^m$, an observation of the common factor $\boldsymbol{F} \in \mathbb{R}^m$.

To use the maximum likelihood method for factor analysis, we assume that the observations $\boldsymbol{X}_j, j = 1, \cdots, n$ are independently sampled from $\boldsymbol{X}$. (the assumption can be modified for more complex factor models).

Therefore $\boldsymbol{X}_j$ *i.i.d.* $\sim N_p(\boldsymbol{\mu}, \Sigma)$ for $j = 1, \cdots, n$.

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

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

By our earlier derivation (as in the lecture notes on "The multivariate normal distribution"), the exponent can be re-expressed using matrix trace,

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

The joint likelihood function of a ransom sample $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ can be written as

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

The same derivation shows that the maximum of the likelihood function is achieved at the MLE $(\hat{\boldsymbol{\mu}}, \hat{\Sigma}) = (\bar{\boldsymbol{x}}, \boldsymbol{S}_n)$.

First, we have shown that, given any Σ, the likelihood function is maximized when the estimator of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$.

$$L(\hat{\boldsymbol{\mu}}, \hat{\Sigma}) = \max_{\Sigma} L(\hat{\boldsymbol{\mu}}, \Sigma)$$

where

$$L(\hat{\boldsymbol{\mu}}, \Sigma) = L(\bar{\boldsymbol{x}}, \Sigma) = \max_{\boldsymbol{\mu}} L(\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} e^{-\frac{1}{2}tr\{\Sigma^{-1}(n\boldsymbol{S}_n)\}} \tag{6}$$

Then, given the MLE $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}$, we obtain the MLE $\hat{\Sigma} = \boldsymbol{S}_n$,

$$\max_{\boldsymbol{\mu},\Sigma} L(\boldsymbol{\mu}, \Sigma) = \max_{\Sigma} L(\hat{\boldsymbol{\mu}}, \Sigma) = L(\hat{\boldsymbol{\mu}}, \hat{\Sigma}) = L(\bar{\boldsymbol{x}}, \boldsymbol{S}_n) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{S}_n|^{n/2}} e^{-\frac{np}{2}} \tag{7}$$

Remarks on ML for FA

- The above maximum likelihood in (7) is with respect to unstructured Σ.

- In factor model, the model assumption is $\Sigma = \boldsymbol{L}\boldsymbol{L}' + \Psi$.

  With observed data, $\hat{\Sigma}$ is used (in the place of Σ) to further estimations of the parameters of the factor model under the imposed factor structure (see below).

- Recall that the factor loading matrix $\boldsymbol{L}$ is not unique: its entries are identifiable only up to orthogonal rotations, therefore further constraints are needed.

  A common used constraints is to requre $\boldsymbol{L}'\Psi^{-1}\boldsymbol{L}$ to be diagonal.

- With $\hat{\Sigma}$ and the additional constraints, $\boldsymbol{L}$ matrix and Ψ matrices are estimated iteratively (more details in the remarks on the constraints below).

- The resulting Maximum Likelihood estimates are

$$\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}, \qquad \hat{\boldsymbol{L}} = [\hat{\ell}_{ij}]_{p\times s}, \qquad \hat{\Psi} = diag\{\hat{\psi}_1, \cdots, \hat{\psi}_p\},$$

  subject to additional constraints.

- The ML estimates of the commonalities are

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

  The proportion of total sample variance contributed by the jth common factor is

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

- The goodness of fit of the model can be check by the magnitude of the entries of the residual matrix $\boldsymbol{S} - (\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}' + \hat{\Psi})$.

  For example, we may check the Frobenius norm of the residual matrix,

$$\|\boldsymbol{S} - (\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}' + \hat{\Psi})\|_F = \sqrt{\text{Sum of squared entries of the residual matrix } \left(\boldsymbol{S} - (\hat{\boldsymbol{L}}\hat{\boldsymbol{L}}' + \hat{\Psi})\right)}$$

- The goodness of fit can also be tested by likelihood ratio test, mostly the large sample likelihood ratio test such as the Bartlett test.

- Remarks on the constraints and computation in ML estimation* (Optional, not required)

  - Due to the non-uniqueness nature of factor loading matrix $\boldsymbol{L}$ under orthogonal transformation (as illustrated in (3) and (4)), the parameters in $\boldsymbol{L}$ are not uniquely identifiable. Therefore further parameter constraints are needed.

  - There are various ways to add the constraints on the model parameters.

  - A common constraint used in practice is

$$\boldsymbol{L}'\Psi^{-1}\boldsymbol{L} = \Delta \quad \text{(a diagonal matrix)} \tag{8}$$

  - Under the FA conditions

$$\Sigma = LL' + \Psi$$

    The constraint in (8) yields the Ψ

$$\left(\Psi^{-1/2}\Sigma\Psi^{-1/2}\right)\left(\Psi^{-1/2}L\right) = \left(\Psi^{-1/2}L\right)(I+\Delta) \tag{9}$$

    Proof.

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

    □

  - Note that (9) implies that the columns of $\Psi^{-1/2}L$ are (nonnormalized) eigenvectors of matrix $\Psi^{-1/2}\Sigma\Psi^{-1/2}$ corresponding to the eigenvalues on the diagonal of the diagonal matrix $\Delta + I_m$.

  - Using the ML estimate $S_n = \hat{\Sigma}$, L matrix and Ψ are estimated:

    * First, initial values of the specific variance $\hat{\psi}_i, i = 1, \cdots, m$ on the diagonal of Ψ are proposed.
    * With the initial $\hat{\Psi}$, an initial $\hat{L}$ can be estimated via the eigenvalue-eigenvector equation

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

    * With the estimated $\hat{L}$, Σ is expressed as $\hat{L}\hat{L}'+\Psi$ back to (6) to obtain an estimate of Ψ by maximizing the likelihood over all Ψ.
    * With estimate $\hat{\Psi}$, the process repeats.

  - Thus L matrix and Ψ matrices are estimated iteratively until convergence (details see 9A in J&W).

A large sample test for the number of common factors

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

H_o: $\Sigma = LL' + \Psi$ with m factors.
H_a: Σ does not have the imposed structure in H_o.

The likelihood ratio statistic for testing H_o is

$$-2\ln\Lambda = n\ln\frac{|\hat{\Sigma}|}{|S_n|}, \qquad \hat{\Sigma} = \hat{L}\hat{L}' + \hat{\Psi}$$

By the properties of maximum likelihood,

$$-2\ln\Lambda \sim \chi^2_d \qquad \text{approximately for large } n,$$

where $d = d_1 - d_0$ is the degree of freedom of the χ^2 distribution, with

d_0 = the dimension or the degree of freedom of the parameter space under H_o
d_1 = the dimension of the parameter space under $H_o \cup H_a$

For the non-structured S_n, by the symmetry of covariance matrix, the number of distinct elements are

$$d_1 = \frac{1}{2}p(p+1)$$

Under the factor model, $\Sigma = LL' + \Psi$. In addition, the constraints in (8) requires $L'\Psi^{-1}L$ to be diagonal.

- L has $p \times m$ parameters
- Ψ has p diagonal parameters.
- The constraints in (8) requires off-diagonal elements $= 0$, thus there are $m(m-1)/2$ constraints.

Combined,

$$d_0 = pm + p - \frac{1}{2}m(m-1)$$

Therefore,

$$d = d_1 - d_0 = \frac{1}{2}p(p+1) - [p(m+1) - \frac{1}{2}m(m-1)] = \frac{1}{2}[(p-m)^2 - p - m]$$

A common approximation method is **Bartlett's approximation**, which uses the Bartlett's correction, replacing n in the likelihood ratio statistic by $n - 1 - (2p + 4m + 5)/6$. Then

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

The asymptotic distribution is used for the hypothesis test to evaluate the adequacy of using m factors in the model. At test level α, we reject H_o when

$$(n-1-(2p+4m+5)/6)\ln\frac{|\hat{L}\hat{L}'+\hat{\Psi}|}{|S_n|} > \chi^2_{[(p-m)^2-p-m]/2,\,\alpha}$$

where $\chi^2_{*,\alpha}$ denotes the right α quantile, defined by

$$\mathbb{P}(W \geq \chi^2_{k,\alpha}) = \alpha, \qquad where \quad W \sim \chi^2_k.$$

3 Factor Rotation* (Optional)

The factors F_i and factor loadings ℓ_{ij} are unique only up to an orthonormal transformation of the common factors, as shown in the non-uniqueness of factors in (3).

This seemingly disadvantageous property is used intentionally in practice to rotate the axes in the m dimensional $\boldsymbol{F}$ space to get a set of common factors with better interpretations.

The interpretation of a factor is easier or more preferred when the factor loadings of the variables are either large or negligible, that is, some (original) variables are of large loadings on a factor thus closely related to the factor, and some other variables are of negligible loadings on the factor thus almost not related to that factor.

Therefore, it is preferable that the magnitudes of variable loadings on a factor to be as different as possible.
In other words, large variations in the magnitude of factor loadings for each and all factors are preferred.

This desired property can be quantified as requesting loading magnitudes with large variance.

For example, it is reasonable to maximize

$$\sum_{j=1}^{m} (\text{variance of squares of loadings for the } j\text{th factor})$$

This is the idea of the Kaiser varimax criterion.

The (raw) varimax criterion

$$VC(\boldsymbol{L}) = \sum_{j=1}^{m} \left[\frac{1}{p} \sum_{i=1}^{p} (\ell_{ij}^2)^2 - \left(\frac{1}{p} \sum_{i=1}^{p} \ell_{ij}^2 \right)^2 \right] = \frac{1}{p} \sum_{j=1}^{m} \left[\sum_{i=1}^{p} \ell_{ij}^4 - \left(\sum_{i=1}^{p} \ell_{ij}^2 \right)^2 \Big/ p \right]$$

The standardized varimax criterion

$$VC(\boldsymbol{L}) = \frac{1}{p} \sum_{j=1}^{m} \left[\sum_{i=1}^{p} \ell_{ij}^{*\,4} - \left(\sum_{i=1}^{p} \ell_{ij}^{*\,2} \right)^2 \Big/ p \right], \qquad \ell_{ij}^* = \frac{\ell_{ij}}{h_i}.$$

The factor rotation process has the following outline.

- Start with a factor matrix $\boldsymbol{F}$ with loading matrix $\boldsymbol{L}$.
- Consider $VC(\boldsymbol{LT})$ for all $m \times m$ orthogonal matrix $\boldsymbol{T}$.
- Choose the optimal $\boldsymbol{T}^*$ such that
$$VC(\boldsymbol{LT}^*) = \max_{\boldsymbol{T}} VC(\boldsymbol{LT})$$
- The optimal common factor vector is $\boldsymbol{F}^* = \boldsymbol{T}'\boldsymbol{F}$ by the varimax criterion.

4 FA and PCA

Factor Analysis and Principal Component Analysis share many common properties.

- Both are commonly used at the beginning stage of exploratory data analysis.
- Both achieve dimension reduction by using a smaller number of variables (consisting of linear combinations of original variables) to explain a data set of many more variables.
- The scaled principal components from PCA may simply serve as factors in FA.
- Both methods are not effective or useful if the original variables are (almost) uncorrelated to start with.

They also have many differences, especially when maximum likelihood method is used in the factor model.

- Population PCA is commonly used as a mathematical procedure, which has no assumptions on data structure.
  Factor analysis assumes that data are from a specific statistical model, with many assumptions.
- PCA aims to achieve the transformation of $observed\ variables \to principle\ component\ variables$.
  FA focus on recovering the structure of the transformation of $hidden\ factors \to observed\ variables$.
- PCA measures data variations by variable variances.
  FA models data variations by focusing on the covariance structure.
- PCA is not scale invariant (different results for original and scaled data with variance 1).
  FA using ML method is scale invariant.
- In PCA, considering k + 1 instead of k components does not change variable loadings in the first k components.
  In FA, considering k + 1 instead of k factors may change the first k factors when using ML method.
- Calculation of PCA scores is relatively computational straightforward.
  Calculation of factor scores is computationally more complex.

Note Relevant chapter in the book by Johnson and Wichern: Chapter 8.