Factor Analysis

Qingyang Li

1 Introduction

Factor analysis has the same objectives as principal component analysis:

- Discover the true dimension of the data.
- Try to interpret "new" variables.

The way factor analysis achieves these objectives are different from PCA:

- Principal components are defined as linear combination of the original variables. In factor analysis, the original variables are expressed as linear combinations of the factors.
- PCA is focused on explaining the variance structure of the data. FA is concerned with explaining the variance and covariance structure of the data.

2 Factor Analysis Model

2.1 Single Factor Analysis Model

suppose we observe the variables $x_1, x_2, \ldots, x_p$ for each individual. The single factor model is as follows.

$$\begin{aligned} x_1 - \mu_1 &= \lambda_1 f + \eta_1 \\ x_2 - \mu_2 &= \lambda_2 f + \eta_2 \\ &\vdots \\ x_p - \mu_p &= \lambda_p f + \eta_p \end{aligned}$$

The main components of this models are:

- x_j's are observed variables , called the manifest variables.
- f is the unobserved variable, called the common factor. The common factor is random component common to all original variables.
- λ_j's are called factor loadings. The loadings determine the strength of the relationship between the common factor and the observed variables.
- η_j's are called specific factors. The specific factors are random component specific for the j^{th} original variable.

2.2 K Factors Analysis Model

Let $\mathbf{x} \sim (\mu, \Sigma)$, where $\mathbf{x}$ is $p \times 1$. Notice that we do not use a multivariate normal distribution assumption.

$$\begin{aligned} x_1 - \mu_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1m} f_m + \eta_1 \\ x_2 - \mu_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2m} f_m + \eta_2 \\ &\vdots \\ x_p - \mu_p &= \lambda_{p1} f_1 + \lambda_{p2} f_2 + \cdots + \lambda_{pm} f_m + \eta_p \end{aligned}$$

The main components of this models are:

- x_j is the j^{th} random variable, where $j = 1, 2, \ldots, p$.
- f_k is the common factors, where $k = 1, 2, \ldots, m$ and $m < p$. f_k is independently and identically distributed with mean of zero and variance of 1. These factors are uncorrelated with each other.
- η_j is the specific factors. η_j is independently distributed with mean of zero and variance of ψ_j, where ψ_j is specific variance of η_j.
- f_k and η_j are independent for all $k = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, p$.
- λ_{jk} measures the contribution of the k^{th} common factors to the j^{th} original variable. These are called factor loadings. They will help to interpret common factors.

In general, we can use $\tilde{x}_j$ as "mean adjusted" and can write the model as:

$$\tilde{x}_j = \lambda_{j1} f_1 + \lambda_{j2} f_2 + \cdots + \lambda_{jm} f_m + \eta_j, j = 1, 2, \cdots, p$$

2.3 Factor Analysis Model In Matrix Form

$$\tilde{\mathbf{x}} = \mathbf{\Lambda f} + \eta$$

$$\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_p - \mu_p \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pm} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_p \end{bmatrix}$$

where,

- $\mathbf{f} \sim (E(\mathbf{f}) = \mathbf{0}, Cov(\mathbf{f}) = \mathbf{I})$, where $\mathbf{I}$ is identity matrix.
- $\eta \sim (E(\eta) = \mathbf{0}, Cov(\eta) = \mathbf{\Psi})$, where $\mathbf{\Psi} = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$
- $\mathbf{f}$ and η are independent.

2.4 Factor Analysis Model On Standardized Data

Similar to PCA, we more often work with standardized data so that we also have a variance of 1 for the random variable on the left hand side of the equation.

$$z_j = \lambda_{j1} f_1 + \lambda_{j2} f_2 + \cdots + \lambda_{jm} f_m + \eta_j, j = 1, 2, \cdots, p$$

In matrix form:

$$\mathbf{z} = \mathbf{\Lambda f} + \eta$$

where $\mathbf{z} = \begin{bmatrix} z_1 & z_2 & \cdots & z_p \end{bmatrix}^\intercal$ and $Cov(\mathbf{z}) = \mathbf{P}$. Note that the λ_{jk} will not be the same between using $\tilde{\mathbf{x}}$ or $\mathbf{z}$. I simply use the same notation for the factor loadings because otherwise the notation will get messier later

3 Covariance and Correlation Matrices

3.1 Background

- Let $\mathbf{A}$ be a matrix of constants, and let $\mathbf{y}$ be a vector of random variables.$Cov(\mathbf{Ay}) = \mathbf{A}Cov(\mathbf{y})\mathbf{A}^\intercal$
- Suppose $\mathbf{x}$ and $\mathbf{y}$ are independent random vectors. Then $Cov(\mathbf{x} + \mathbf{y}) = Cov(\mathbf{x}) + Cov(\mathbf{y})$

3.2 Generate Covariance Matrix by Using Matrix Form of FA Model

$$\begin{aligned}
\mathbf{\Sigma} &= Cov(\tilde{\mathbf{x}}) \\
&= Cov(\mathbf{\Lambda f} + \eta) \\
&= Cov(\mathbf{\Lambda f}) + Cov(\eta) \\
&= \mathbf{\Lambda}Cov(\mathbf{f})\mathbf{\Lambda}^\intercal + \mathbf{\Psi} \\
&= \mathbf{\Lambda I \Lambda}^\intercal + \mathbf{\Psi} \\
&= \mathbf{\Lambda \Lambda}^\intercal + \mathbf{\Psi}
\end{aligned}$$

$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\intercal + \mathbf{\Psi}$ is often called factor analysis equations.

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pm} \end{bmatrix} \begin{bmatrix} \lambda_{11} & \lambda_{21} & \cdots & \lambda_{p1} \\ \lambda_{12} & \lambda_{22} & \cdots & \lambda_{p2} \\ \vdots & \vdots & & \vdots \\ \lambda_{1m} & \lambda_{2m} & \cdots & \lambda_{pm} \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{k=1}^m \lambda_{1k}^2 & \sum_{k=1}^m \lambda_{1k}\lambda_{2k} & \cdots & \sum_{k=1}^m \lambda_{1k}\lambda_{pk} \\ \sum_{k=1}^m \lambda_{2k}\lambda_{1k} & \sum_{k=1}^m \lambda_{2k}^2 & \cdots & \sum_{k=1}^m \lambda_{2k}\lambda_{pk} \\ \vdots & \vdots & & \vdots \\ \sum_{k=1}^m \lambda_{pk}\lambda_{1k} & \sum_{k=1}^m \lambda_{pk}\lambda_{2k} & \cdots & \sum_{k=1}^m \lambda_{pk}^2 \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{k=1}^m \lambda_{1k}^2 + \psi_1 & \sum_{k=1}^m \lambda_{1k}\lambda_{2k} & \cdots & \sum_{k=1}^m \lambda_{1k}\lambda_{pk} \\ \sum_{k=1}^m \lambda_{2k}\lambda_{1k} & \sum_{k=1}^m \lambda_{2k}^2 + \psi_2 & \cdots & \sum_{k=1}^m \lambda_{2k}\lambda_{pk} \\ \vdots & \vdots & & \vdots \\ \sum_{k=1}^m \lambda_{pk}\lambda_{1k} & \sum_{k=1}^m \lambda_{pk}\lambda_{2k} & \cdots & \sum_{k=1}^m \lambda_{pk}^2 + \psi_p \end{bmatrix}$$

The following are the findings based on the final covariance matrix.

- Based on the above, $Var(x_j) = \sigma_{jj} = \sum_{k=1}^m \lambda_{jk}^2 + \psi_j$, where $j = 1, 2, \cdots, p$ and $k = 1, 2, \cdots, m$.
- The proportion of variance is $\frac{\sum_{k=1}^m \lambda_{jk}^2}{\sigma_{jj}}$. The numerator in the proportion is called communality of j^{th} original variables.
- $Var(x_j) = \sigma_{jj} = \sum_{k=1}^m \lambda_{jk}^2 + \psi_j$, which is equal to communality + specific variance. The specific variance is sometimes called the uniqueness.
- $Cov(x_j, f_k) = \lambda_{jk}$. Assume that the value of k is equal to m. Then, based on $Cov(X+Y, Z) = Cov(X, Z) + Cov(Y, Z)$, the proof is derived as following.

$$\begin{aligned} & Cov(\lambda_{j1}f_1 + \lambda_{j2}f_2 + \cdots + \lambda_{jm}f_m + \eta_j, f_m) \\ = Cov(\lambda_{j1}f_1, f_m) + Cov(\lambda_{j2}f_2, f_m) + \ldots + Cov(\lambda_{jm}f_m, f_m) + Cov(\eta_j, f_m) \\ & = 0 + 0 + \ldots + \lambda_{jm}Cov(f_m, f_m) + 0 \\ & = \lambda_{jm} \end{aligned}$$

3.3 Correlation Matrices On Standardized Data

P is also the covariance matrix of the standardized data.This implies that $\mathbf{\Lambda}$ is a matrix of correlations between the z_j (standardized data) and the f_k.

$$\mathbf{P} = \mathbf{\Lambda}\mathbf{\Lambda}^\intercal + \mathbf{\Psi}$$

- $Corr(z_j, f_k) = \lambda_{jk}$ Note that this also means $-1 \leq \lambda_{jk} \leq 1$ due to numerical range of correlations.
- $\sum_{k=1}^m \lambda_{jk}^2 + \psi_j = 1$ Because the diagonal elements of a correlation matrix are 1.
- The communality of the j^{th} standardized variable is $\sum_{k=1}^m \lambda_{jk}^2$.

4 Estimate The Factor Analysis Model

4.1 Background

The most often used procedure is maximum likelihood estimation. Let $\mathbf{x_1}, \ldots, \mathbf{x_N}$ be a random sample from a multivariate normal distribution $N_p(\mu, \Sigma)$. Then the likelihood function is

$$L(\mu, \Sigma | \mathbf{x_1}, \ldots, \mathbf{x_N}) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}[(\mathbf{x_i}-\mu)^\intercal \Sigma^{-1}(\mathbf{x_i}-\mu)]}$$

- The maximum likelihood estimations are found through iterative numerical methods. When the estimates change very little at successive iterations, the estimates are said to "converge" to the maximum likelihood estimations. The corresponding estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ are denoted symbolically as $\hat{\mathbf{\Lambda}}$ and $\hat{\mathbf{\Psi}}$.
- In R, the function factanal() automatically uses standardized data and maximum likelihood estimation.
- A way to assess how good the common factors are in accounting for the information in the data is to examine the difference between the standard estimate of the correlation matrix and the estimate obtained from the model structure.

$$\mathbf{R} - (\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^\intercal + \hat{\mathbf{\Psi}})$$

4.2 How Do We Choose An Appropriate Number of Common Factors

Likelihood Ratio Test (LRT):

H_0 : m common factors are sufficient

H_a : more common factors are needed

Using a bartlett correction, the modified statistic is

$$T.S = (N - 1 - \frac{2p + 4m + 5}{6}) N log(\frac{|\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^\intercal + \hat{\mathbf{\Psi}}|}{|[(N-1)/N]\hat{\mathbf{\Sigma}}|})$$

This statistic can be approximated by a $\chi^2_{\frac{(p-m)^2-p-m}{2}}$ from a large sample. We can reject H_0 if $T.S$ is larger than $1-\alpha$ quantile from a $\chi^2_{\frac{(p-m)^2-p-m}{2}}$ distribution.

5 None Uniqueness of The Common Factors

If $m > 1$, the factor loading matrix is not unique. Let $\mathbf{T}$ be an $m \times m$ orthogonal matrix.

$$\begin{aligned}
\mathbf{P} =& \mathbf{\Lambda}\mathbf{\Lambda}^\intercal + \mathbf{\Psi} \\
=& \mathbf{\Lambda}\mathbf{T}\mathbf{T}^\intercal\mathbf{\Lambda}^\intercal + \mathbf{\Psi} \text{ since } \mathrm{TT}^\intercal = \mathbf{I} \text{ is identity matrix} \\
=& (\mathbf{\Lambda}\mathbf{T})(\mathbf{\Lambda}\mathbf{T})^\intercal + \mathbf{\Psi} \\
=& \mathbf{\Lambda}^*(\mathbf{\Lambda}^*)^\intercal + \mathbf{\Psi}
\end{aligned}$$

Therefore, if $\mathbf{\Lambda}$ is a loading matrix, $\mathbf{\Lambda T}$ is also a loading matrix. A different $\mathbf{T}$ will lead to a different $\mathbf{\Lambda}^*$. We use the loading matrix to explain what common factors represent. If we have a

different factor loading matrix, we will have different interpretation of the common factors. Also there are infinite number of possible orthogonal matrices, then there will be infinite number of possible interpretation we can have for our factor analysis model as long as $m > 1$.

$$\begin{aligned}\mathbf{z} =& \mathbf{\Lambda f} + \eta \\ =& \mathbf{\Lambda T T^\intercal f} + \eta \\ =& (\mathbf{\Lambda T})(\mathbf{fT})^\intercal + \eta \\ =& \mathbf{\Lambda^* f^{*\intercal}} + \eta\end{aligned}$$

Multiplying $\mathbf{\Lambda}$ by an orthogonal matrix is called a rotation. When rotating, we try to find a $\mathbf{\Lambda}^*$ that allows us to more easily interpret the common factors. This usually means making the loadings close to 0 or 1 or -1. The reason is that if a factor loading is 0, then the common factor does not play a large part in forming an original variable. Similarly, if a loading is close to -1 or 1, the common factor plays a large part in forming an original variable.

5.1 Orthogonal Rotation Method

There are many established ways to choose an orthogonal matrix $\mathbf{T}$. The most often used is called the varimax method.

Let assume that $\mathbf{B} = \mathbf{\Lambda T}$, where $\mathbf{T}$ is an orthogonal matrix.

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & & \vdots \\ b_{p1} & b_{p2} & \cdots & b_{pm} \end{bmatrix}$$

We want to find a $\mathbf{T}$ that maximizes the following equation

$$V^* = \sum_{q=1}^{m} \left(\frac{\sum_{j=1}^{p} b_{jq}^4 - \frac{\sum_{j=1}^{p} b_{jq}^2}{p}}{p} \right)$$

where b_{jq} is the j^{th} row and q^{th} column elemnt of $\mathbf{B}$. For each column of $\mathbf{B}$, the formula essentially finds the variance of the squared elements. The formula that derives above equation is using biased sample variance equation,

$$\begin{aligned}\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n} = \frac{\sum_{i=1}^{n}(x^2 - 2x\bar{x} + \bar{x}^2)}{n} &= \frac{\sum_{i=1}^{n} x^2 - 2\bar{x}\sum_{i=1}^{n} x + \sum_{i=1}^{n}\bar{x}^2}{n} \\ = \frac{\sum_{i=1}^{n} x^2 - 2n\bar{x}^2 + \sum_{i=1}^{n}\bar{x}^2}{n} &= \frac{\sum_{i=1}^{n} x^2 - 2n\bar{x}^2 + n\bar{x}^2}{n} \\ = \frac{\sum_{i=1}^{n} x^2 - n\bar{x}^2}{n} &= \frac{\sum_{i=1}^{n} x^2 - n\frac{\sum_{i=1}^{n} x^2}{n^2}}{n} \\ &= \frac{\sum_{i=1}^{n} x^2 - \frac{\sum_{i=1}^{n} x^2}{n}}{n}\end{aligned}$$

The value of the elements in the $\mathbf{B}$ is $-1 < b_{jq} < 1$.Because V^* gives equal weight to original variables with small and large communalities, the rotated factor loadings are divided by the variable's communality:

$$V = \frac{1}{p^2}\sum_{q=1}^{m}\left(p\sum_{j=1}^{p}\frac{b_{jq}^4}{h_j^4} - \left(\sum_{j=1}^{p}\frac{b_{jq}^2}{h_j^2}\right)^2\right)$$

where $h_j^2 = \sum_{k=1}^{m} \lambda_{jk}^2$ is the communality for the j^{th} original variable.

6 Factor Scores

6.1 Bartlett's method (a.k.a.,weighted least-squares method)

Using standardized, the FA Model is $\mathbf{z} = \mathbf{\Lambda f} + \eta$. For the r^{th} observation, find the $\mathbf{f}$ that minimizes

$$(\mathbf{z}_r - \hat{\mathbf{\Lambda}}\mathbf{f})^\intercal \hat{\mathbf{\Psi}}^{-1} (\mathbf{z}_r - \hat{\mathbf{\Lambda}}\mathbf{f})$$

where $\mathbf{z}_r$ is a column vector of the standardized values for the r^{th} observation. Notice that $\mathbf{z}_r - \hat{\mathbf{\Lambda}}\mathbf{f}$ is a multivariate residual. It can be shown that the $\mathbf{f}$ that minimizes the above expression is

$$\hat{\mathbf{f}}_r = (\hat{\mathbf{\Lambda}}^\intercal \hat{\mathbf{\Psi}}^{-1} \hat{\mathbf{\Lambda}})^{-1} \hat{\mathbf{\Lambda}}^\intercal \hat{\mathbf{\Psi}}^{-1} \mathbf{z}_r$$

6.2 Thompson's method (a.k.a., regression method)

$$\hat{\mathbf{f}}_r = \hat{\mathbf{\Lambda}}^\intercal (\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}^\intercal + \hat{\mathbf{\Psi}})^{-1} \mathbf{z}_r$$