Machine Learning Methods
Factor Analysis

By Zhu Xuelin
Email: xuelin@u.nus.edu

September 28, 2024

1 From PCA to Factor Analysis

When it comes to the dimension reduction, the PCA is the most famous tool, and it provides the best linear transformation which attains the maximum explained proportion of total variance of the raw data. In the previous derivation, we know that if the goal dimension p is given, then the loading matrix is NOT unique, but up to a rotation matrix, which provides us with a way to get a better interpretation.

Example 1. Suppose the first 2 PC's are extracted by the raw data matrix $\mathbb{Y}_{N\times q}$, and the principal component transformation is

$$\mathbf{y}_{q\times 1} \mapsto \mathbf{x}_{2\times 1}: \quad \mathbf{x} = U^\top \mathbf{y} \quad \text{with} \quad U_{q\times 2} = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 \end{pmatrix}.$$

By doing this, we are approximating $\mathbf{y}$ by

$$\mathbf{y}_{q\times 1} \approx U_{q\times 2}\mathbf{x}_{2\times 1} = x_1\mathbf{u}_1 + x_2\mathbf{u}_2.$$

Mathematically, in this case, we are using the 2-dim subspace $\text{col}(U) \subset \mathbb{R}^q$ to approximate the raw data space. Recall that we can rotate the loading matrix U and the coordinates $\mathbf{x}$ simultaneously, without losing any explained proportion of variance. This means we can choose another base of $\text{col}(U)$ with another coordinates $\mathbf{x}^*$ such that

$$x_1^*\mathbf{u}_1^* + x_2^*\mathbf{u}_2^* = x_1\mathbf{u}_1 + x_2\mathbf{u}_2,$$

to represent this 2-dim subspace $\text{col}(U)$ as illustrated in the next example.

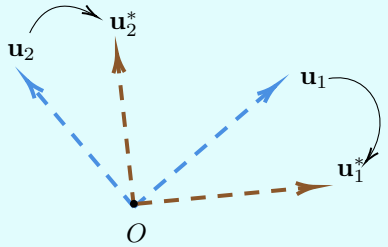


Figure 1: 2 dimension subspace: col(U) $\subset \mathbb{R}^q$

Therefore, we can choose the basis (do rotation) so that it could give some meaning such as

$$\mathbf{y}_3 = \begin{pmatrix} height \\ b.p. \\ weight \end{pmatrix} \approx x_1 \begin{pmatrix} 0.45 \\ 0.53 \\ 0.74 \end{pmatrix} + x_2 \begin{pmatrix} 0.62 \\ -0.34 \\ 0.54 \end{pmatrix}$$

$$\text{(By rotation)} \quad \Rightarrow \quad \approx x_1^* \begin{pmatrix} 0.05 \\ 0.03 \\ 0.98 \end{pmatrix} + x_2^* \begin{pmatrix} 0.02 \\ -0.94 \\ 0.14 \end{pmatrix}.$$

Using the second basis, we can see $b.p.$ is closely related to x_2^* and $weight$ is closely related to x_1^*. Though we might not know its true meaning, we can regard $\mathbf{x}^*$ as latent factors.

This goal of better interpretation is where factor analysis starts. Of course, getting factors by rotating the PC's is a good way. Another way relies on the following probabilistic model.

2 Exploratory Factor Analysis

We start with the basic FA model, exploratory factor analysis. Because we usually don't have any information about the data as well as the latent factors, for the simplicity in the estimation, we assume the components of the latent variable $\mathbf{X}$ (factors) are independent of each other. After using EFA to find the estimated subspace $\text{col}(U)$, we can then rotate the factors to make them match our needs like dependency or sparsity.

Assumption 2 (Exploratory factor analysis)**.** The observed data $\mathbf{Y}$ is q-dimensional. Assume there is a latent vector $\mathbf{X}$ affecting $\mathbf{Y}$, and the underlying model is

$$\mathbf{Y}_q = \boldsymbol{\mu}_q + \Lambda_{q\times p}\mathbf{X}_{p\times 1} + W_q \quad \text{where} \quad \begin{cases} W \perp\!\!\!\perp \mathbf{X}, \\ \mathbf{X} \sim \mathbf{N}(\mathbf{0}, I_p), \\ W \sim \mathbf{N}(\mathbf{0}, \Psi_q), \end{cases}$$

where $\Psi = \text{diag}(\psi_1, \ldots, \psi_q) > 0$.

Parameters are mean vector $\boldsymbol{\mu}$, error variance Ψ_q and the loading matrix Λ. Note that this assumption contains the identifiable issue, since inserting a rotation matrix $O_{p\times p}$ and letting $\Lambda^* = \Lambda O^\top$ and $\mathbf{X}^* = O\mathbf{X}$, leads to the same model.

To interpret this model, we can regard $\mathbf{Y}$ as the scores of each aspect from the questionnaire, i.e.

$$\mathbf{Y} = \begin{pmatrix} \text{calculation score} \\ \text{algebra score} \\ \vdots \\ \text{analysis score} \end{pmatrix},$$

$\boldsymbol{\mu}$ as the common mean score among people, $\mathbf{X}$ contains some factors that have impact on the score, Λ is the loading matrix with coefficients for how factors affect the questionnaire scores, and W is the error.

Since Assumption 2 contains distribution, simple derivation leads to

$$\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix} \sim \mathrm{N}\left(\begin{bmatrix}\mathbf{0}_p\\ \boldsymbol{\mu}_q\end{bmatrix}, \begin{bmatrix} I_p & \Lambda_{p\times q}^\top \\ \Lambda_{q\times p} & \Psi + \Lambda\Lambda^\top\end{bmatrix}\right) \quad \Rightarrow \quad \mathbf{Y} \sim \mathrm{N}_q(\boldsymbol{\mu}, \Psi + \Lambda\Lambda^\top).$$

how to make it identifiable? Intuitively, a model is not identifiable because it contains to many parameters. If we add some constraints on parameters, it may work (not rigorous).

Note that the identifiable issue comes from the rotation matrix O, which contains p^2 elements (free parameters). It is orthogonal, so

$$\begin{cases} O_j^\top O_j = 1 & 1 \le j \le p, \\ O_j^\top O_k = 0 & \forall j \ne k, \end{cases}$$

which gives $p + \binom{p}{2} = p(p+1)/2$ constraints. Therefore, we can add $p(p-1)/2$ more constraints, making the number of constraints equalling to then number of free parameters. One possible way is to restrict

$$\Lambda_{q\times p} = \begin{bmatrix} * & 0 & \cdots & 0 \\ * & * & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & * \\ * & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & * \end{bmatrix}_{q\times p},$$

whose restrictions are put on the right upper triangular area, and the number is $p(p-1)/2$. The fact is, by doing this way, we can make it identifiable, and it remains to estimate parameters. But we will leave the estimation to the next lecture note, and focus on other related issues in this one.

3 General Factor Analysis

Sometimes, we may want the factors to be NOT independent, but correlated. This leads to the generalization of the FA model.

Assumption 3 (General factor analysis)**.** The observed data $\mathbf{Y}$ is q-dimensional. Assume there is a latent vector $\mathbf{X}$ affecting $\mathbf{Y}$, and the underlying model is

$$\mathbf{Y}_q = \boldsymbol{\mu}_q + \Lambda_{q\times p}\mathbf{X}_{p\times 1} + W_q \quad \text{where} \quad \begin{cases} W \perp\!\!\!\perp \mathbf{X}, \\ \mathbf{X} \sim \mathbf{N}(\mathbf{0}, \Phi_p), \\ W \sim \mathbf{N}(\mathbf{0}, \Psi_q), \\ Y_i \perp\!\!\!\perp Y_j \mid \mathbf{X}, \ \ \forall i \ne j, \end{cases}$$

where Φ_p is not necessarily diagonal, and $\Psi = \mathrm{diag}(\psi_1, \ldots, \psi_q) > 0$.

Despite correlation between $\mathbf{X}$, the covariance structure of observable variable $\mathbf{Y}$ is all captured by $\mathbf{X}$, since once given $\mathbf{X}$, Y_i's are independent. And in general, we can relax the distribution assumption since we

only care about the covariance structure:

$$\mathrm{Cov}(\mathbf{Y}) = \Psi + \Lambda\Phi\Lambda^\top.$$

Similar to the EFA, this model is also not identifiable, since we can insert an orthogonal matrix $O_{p\times p}$:

$$\mathbf{Y}_q = \boldsymbol{\mu}_q + \Lambda_{q\times p} O^\top_{p\times p} O_{p\times p}\mathbf{X}_p + W_q = \boldsymbol{\mu}_q + \Lambda_{q\times p}\mathbf{X}_p + W_q.$$

And the number of free parameters is p^2. Expect for the restriction in EFA, we can also restrict

$$\Lambda_{q\times p} = \begin{pmatrix} I_{p\times p} \\ *_{(q-p)\times p} \end{pmatrix},$$

to solve it since I_p gives p^2 constraints. (More constrains on the Λ, we will have less constraints on the Φ.) For more constraints methods, refer *Bai & Li, 2012, Annals of Statistics*.

4 Practical Issues

Before we move on to the estimation, we now focus on three issues:

(a) after EFA, how to rotate the factor?

(b) given data matrix $\mathbb{Y}$ and after estimation $\hat{\Lambda}$, how to get the factors $\mathbb{X}$?

(c) how to choose the number of p?

4.1 Factor Rotation

Generally, if we do not have convincing information, it is suggested to do EFA first, assuming factors are mutually independent. After we get the estimated $\hat{\Lambda}$ in the EFA, now it is the time to consider rotation. Normally speaking, the pattern we expect is

$$\Lambda = \begin{pmatrix} 1.10 & -1.15 & 0.14 & \cdots & 0.35 \\ 0.68 & 1.35 & 0.61 & \cdots & -0.25 \\ -0.63 & 0.98 & -0.22 & \cdots & 0.25 \\ -0.64 & -2.95 & 0.24 & \cdots & -1.28 \\ -1.37 & -0.15 & -1.30 & \cdots & -1.02 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.41 & 0.37 & -0.99 & \cdots & 0.32 \\ 1.49 & -0.72 & 1.41 & \cdots & 0.16 \end{pmatrix} \quad \Rightarrow \quad \Lambda^* = \Lambda O^\top = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

It is obvious that we want sparsity structure in the loading matrix. Three common ways are briefly introduced:

(a) L_1 or L_2 loss function:

$$\min_{O: O^\top O = I_p} \mathrm{loss}(\hat{\Lambda} O^\top).$$

(b) varimax is another loss function.

(c) oblique rotation. (not orthogonal rotation, and it brings dependency of $\mathbf{X}$).

4.2 Estimation of Factors

Now we consider the EFA setting ($\text{Cov}\mathbf{X} = \Phi = I_p$) and assume the loading matrix is estimated already as $\hat{\Lambda}$, the error variance is estimated as $\hat{\Psi}$. Note that the assumption is

$$\mathbf{Y} = \boldsymbol{\mu} + \Lambda\mathbf{X} + W, \quad \text{where } W \sim \mathbf{N}(0, \Psi),$$

where $\hat{\boldsymbol{\mu}}, \hat{\Lambda}, \hat{\Psi}$ are already estimated, and $\mathbb{Y}$ is observed. So, we can treat $\mathbf{X}$ as regression coefficients, $\hat{\Lambda}$ as known covariates, and $\hat{\Psi}$ as the weights to get the WLS estimate of $\hat{\mathbf{x}}$:

$$\hat{\mathbf{x}}_i = \left(\hat{\Lambda}^\top \hat{\Psi} \hat{\Lambda}\right)^{-1} \hat{\Lambda}^\top \hat{\Psi} \mathbf{y}_i.$$

We can even further use the Bayesian idea, treat $\mathbf{X} \sim \mathbf{N}(0, I_p)$ as the prior, and $\hat{\mathbf{x}}_i$ as the data. By Ridge regression, it follows that

$$\hat{\mathbf{x}}_i = \left(I + \hat{\Lambda}^\top \hat{\Psi} \hat{\Lambda}\right)^{-1} \hat{\Lambda}^\top \hat{\Psi} \mathbf{y}_i.$$

4.3 Selection of p

This is a serious issue, and it could be applied not only in FA, but also PCA. Simple methods are known as

(a) scree plot;

(b) explained proportion of total variance.

But none of them are supported by statistical theory. Now we propose two theoretical methods.

4.3.1 Parallel Analysis

Horn's parallel analysis (1965) uses the idea of permutation test. Suppose the observed data matrix is

$$\mathbb{Y}_{N\times q} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{Nq} \end{bmatrix} = \begin{pmatrix} \mathbf{y}_1^\top \\ \mathbf{y}_2^\top \\ \vdots \\ \mathbf{y}_N^\top \end{pmatrix},$$

and the sample covariance matrix of $\mathbb{Y}$ is S with eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_q$. Under EFA assumptions,

$$S \xrightarrow{a.s.} \Psi + \Lambda\Lambda^\top.$$

If we assume the covariance structure is captured by the p factors, then $\Lambda\Lambda^\top$ part should accounts for a large proportion in the diagonal of S, and the noise variance, Ψ accounts for a small proportion, i.e. the first p terms in the diagonal of

$$\Psi + \Lambda\Lambda^\top = \begin{pmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_p & & & \\ & & & \lambda_{p+1} & & \\ & & & & \ddots & \\ & & & & & \lambda_q \end{pmatrix}$$

should be large. Therefore, we can use the following algorithm to test

$$H_0 : \text{there is no signal} \quad \text{vs} \quad H_1 : \text{there is at least one signal.}$$

But note that this is not a strict hypothesis testing as we shall see. Under H_0, there is no significant factor, so λ's, the eigenvalues of S, come from the noise Ψ, but not $\Lambda\Lambda^\top$. Now, we utilize this idea to regenerate data.

Algorithm 4 (Parallel analysis)**.** Set a threshold $0 < \alpha < 1$ for when to stop. Do random permutation for each column of $\mathbb{Y}$:

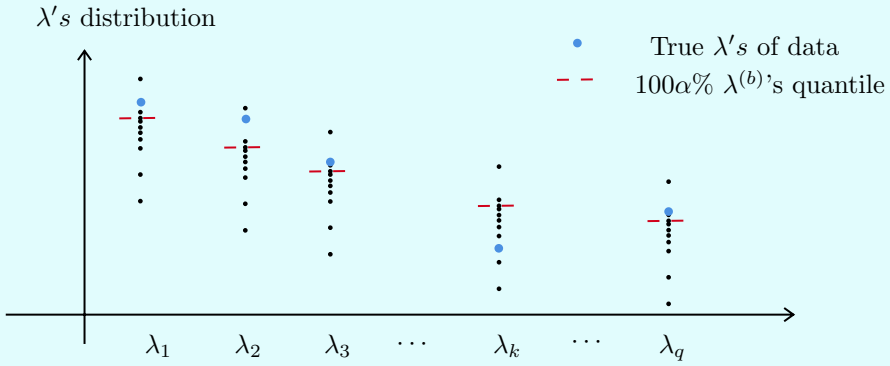
$$\mathbb{Y}_{N\times q} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{Nq} \end{bmatrix}.$$

(a) After permuting all columns, we get $\mathbb{Y}^{(1)}$. Calculate the eigenvalues $\lambda_1^{(1)} > \cdots > \lambda_q^{(1)}$ of $S^{(1)}$.

(b) Repeat B times of step (a).

(c) Calculate the distribution and quantile of each re-sampled λ's.

(d) Choose the $k \in \{1, 2, \ldots, q\}$ when the true λ_k falls below the threshold in the re-sampled distribution.



Why does this algorithm work? When we permute the data matrix $\mathbb{Y}$ by column, the sample variance matrix S does not change, so the covariance structure of $\mathbf{Y}$'s components is preserved. However, since every observation is mutually independent, i.e. $\mathbf{y}_i \perp\!\!\!\perp \mathbf{y}_j$ and $\mathbf{x}_i \perp\!\!\!\perp \mathbf{x}_j$ for $i \neq j$, when we switch the order inside the column, the covariance structure of $\mathbf{X}$ gets killed. Therefore, the re-sampled data could be regarded from the structure

$$\mathrm{Cov}(\mathbf{y}_i^{(b)}) = \Psi.$$

And under H_0 : there is no signal, the true λ should not be significantly greater than the re-sampled $\lambda' s$, otherwise, we have the confidence to claim there is a signal. For more details, refer *Edgar Dobriban, 2020, Annals of Statistics.*

4.3.2 Hypothesis Testing

The last method relies on the probabilistic assumption. It could also be applied to probabilistic PCA. In essence, we are doing sequential hypothesis testing for

$$H_0 : \Sigma \text{ is diagonal} \quad \text{v.s.} \quad H_1 : \Sigma \text{ is not diagonal.}$$

For simplicity, we consider the setting of probabilistic PCA.

Assumption 5. For every observation, the raw data $\mathbf{Y} \in \mathbb{R}^q$ is q-dimensional, and is generated by

$$\mathbf{Y}_q = \boldsymbol{\mu}_q + \Lambda_{q\times p}\mathbf{X}_p + W_q \quad \text{where} \quad \begin{cases} \mathbf{X} \sim \mathrm{N}_p(\mathbf{0}, I_p), \\ W \sim \mathrm{N}_q(\mathbf{0}, \Psi_q), \\ W \perp\!\!\!\perp \mathbf{X} \text{ for every obs}, \\ \text{observations are independent.} \end{cases}$$

Equivalently, we are assuming the population

$$\mathbf{Y} \sim \mathrm{N}_q\left(\boldsymbol{\mu}, \Psi + \Lambda\Lambda^\top\right).$$

For the hypothesis

$$H_0 : \Sigma \text{ is diagonal} \quad \Longleftrightarrow \quad p = 0 \quad \Longleftrightarrow \quad \mathrm{Var}\mathbf{Y} = \Psi,$$

the likelihood ratio test is

$$\mathrm{LRT} = \frac{\sup_{\boldsymbol{\mu},\Psi} L(\boldsymbol{\mu}, \Psi \mid \mathbb{Y})}{\sup_{\boldsymbol{\mu},\Sigma} L(\boldsymbol{\mu}, \Sigma \mid \mathbb{Y})}, \quad \text{where } \Psi \text{ is diagonal and } \Sigma \text{ is symmetric.}$$

By the large sample theory, as N increase and p is any fixed number

$$2\log(\mathrm{LRT}) \xrightarrow{d} \chi^2\left(\frac{q(q+1)}{2} - \left(q(p+1) - \frac{p(p-1)}{2}\right)\right).$$

Setting $p = 0$, if H_0 is rejected, then we can set $p = 1, \ldots$ and so on. Sequentially, we will stop at some point k. But since this is a sequential hypothesis testing, the total type II error is the sum of k-th type II error and type I error of all previous hypothesis testings.