Unit 4.- Factorial analysis (FA)

Course: MULTIVARIATE STATISTICS

©Prof. Dr. José Luis Romero Béjar - Carlos Francisco Salto Díaz

November, 2023

Objective

The main **goal** of Factorial Analysis (FA)) is **to capture reality in the simplest possible way**, by identifying a 'few' **latent variables** that define this reality.

Latent variables

A **latent variable** is a **not observable** variable that is **inferred** based on a set of observable variables using a mathematical model.

Examples can be found in different areas of science:

- **Economy**. Quality of life is a latent variable that is inferred based on others through a mathematical model (FA, probit, logit, etc.).
- **Psychology**. The five variables that define personality: neuroticism, extraversion, openness to experiences, friendliness and responsibility, are latent variables.

PCA vs. FA

FA comprises a **set of techniques** that aim to identify hidden factors (latent variables) preferably **highly correlated with a group of observable variables but not with others**, with the objective mentioned above, to explain reality with the smallest number of variables possible, that is, **reduce the dimension**.

In this sense:

- PCA and FA **have in common** that both methods seek **to reduce the dimension of the problem**.
- They start from the **common hypothesis** that the variables are relatively **correlated**.
- PCA and FA **differ**, in that while the first searches for linear combinations of the original random variables, therefore **observables**, that **maximize the variance in each direction**, the second searches for latent factors, therefore **unobservable**, which **correlate in the maximum sense with certain groups of the observed variables**.

Problem statement

Let $X_1, X_2, \ldots, X_p$ be a set of p correlated random variables. The random vector that forms is denoted by $X = (X_1, X_2, \ldots, X_p)^t$. X is assumed to be centered, $E[X] = 0$ and its covariance matrix is denoted by $\Sigma = E[XX^t]$. Finally, we assume that the random

vector can be sampled in the form:

$$X = AF + L \tag{1}$$

where,

- F_{kx1} is a random vector of $k \leq p$ **common factors** (not observables that correlate with a set of observed variables).
- L_{px1} is a random vector of p **specific factors** (they only correlate with the corresponding observed variable).
- A_{pxk} is a matrix of constants, the **factorial weight matrix**, which will allow determining the influence of the factors on the observed variables and vice versa.

Prior assumptions

To solve the problem (2) the following assumptions are made, without loss of generality:

i. $E[F] = 0_{kx1}$

ii. $E[L] = 0_{px1}$

iii. $E[FL^t] = 0_{kxp}$

iv. $E[FF^t] = I_k$

v. $E[LL^t] = D_p = \begin{pmatrix} d_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & d_p \end{pmatrix}$, diagonal matrix.

Remark

Given the randomness of the factors, if in addition to assuming that they are centered, unitary variances are assumed, then **the factor matrix really represents the correlations of the factors with the observed variables**.

Important

- The common factors F influence X through the coefficients of the factor matrix A.
- Specific factors in L only influence the homologous variable (L_1 over X_1, L_2 over X_2, ...).
- A model like (1) is indicated when working with a large number of variables that may actually be caused by a few common factors.
- The FA model is **similar to a linear regression model** with the exception that here the response variable X is multivariate and that the regressors F are unobservable variables.
- The assumptions considered allow, in general, **to obtain a solution**, although **not unique**.

Bellow it will be justified that **the objective of FA is to estimate the matrices A and D**.

Fundamental equality

$$\Sigma = E[XX^t] = AA^t + D \tag{2}$$

The proof is very simple **(voluntary proposed exercise)** and can be found in the bibliographic reference of *Tussell, 2016, p.66*.

Communalities

If you write the equation (2) element by element, it is easy to see that:

- $\sigma_i^2 = \sigma_{ii} = \sum_{j=1}^k a_{ij}^2 + d_i$, $i = 1, \ldots, p$.
- $\sigma_{ij} = \sum_{l=1}^k a_{il} a_{lj}$, $i, j = 1, \ldots, p$, $i \neq j$.

The part of the variance of the random variables X_i identified by the common factors is called **communality** and is denoted by:

$$h_i^2 = \sum_{j=1}^{k} a_{ij}^2 \;, \forall i = 1, \ldots, p \tag{3}$$

Objetive in practice

Taking into account the fundamental equality (2) above, the **objective of FA**, from a computational point of view, will be **find matrices** A **and** D **for a given covariance matrix,** Σ, satisfying equality, so that A **has the smallest number of columns (factors) possible**.

Remark

In practice the matrix Σ is not known so **an estimate** S is considered, from which Σ is reconstructed as a product AA^t plus a diagonal matrix.

Example: statement

The following easy example illustrates the fit of a factor model with a single factor.

It is considered a random vector that stores the grades of three different subjects $X = (X_1, X_2, X_3)$. Starting from the following estimation of the correlation matrix between the ratings,

$$S = \begin{pmatrix} 1 & 0.83 & 0.78 \\ 0.83 & 1 & 0.67 \\ 0.78 & 0.67 & 1 \end{pmatrix},$$

it is intended to fit **a factorial model with a single factor**.

Example: solution

According to equation (1), the model for one factor will have the following expression:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} F_1 + \begin{pmatrix} L_1 \\ L_2 \\ L_3 \end{pmatrix}$$

This implies, through fundamental equality (2), that:

$$S = \begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} & a_{31} \end{pmatrix} + \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix}$$

Replacing and operating in the previous expression:

$$S = \begin{pmatrix} 1 & 0.83 & 0.78 \\ 0.83 & 1 & 0.67 \\ 0.78 & 0.67 & 1 \end{pmatrix} = \begin{pmatrix} a_{11}^2 & a_{11}a_{21} & a_{11}a_{31} \\ a_{21}a_{11} & a_{21}^2 & a_{21}a_{31} \\ a_{31}a_{11} & a_{31}a_{21} & a_{31}^2 \end{pmatrix} + \begin{pmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{pmatrix}$$

Example: solution

From the previous expression, the following system of 6 equations with 6 unknowns is obtained.

$$\begin{aligned} a_{11}^2 + d_1 &= 1 \\ a_{21}^2 + d_2 &= 1 \\ a_{31}^2 + d_3 &= 1 \\ a_{11}a_{21} &= 0.83 \\ a_{11}a_{31} &= 0.78 \\ a_{21}a_{31} &= 0.67 \end{aligned}$$

Therefore, the model adjusted with a single factor looks like this:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 0.983 \\ 0.844 \\ 0.793 \end{pmatrix} F_1 + \begin{pmatrix} L_1 \\ L_2 \\ L_3 \end{pmatrix}$$

Example: conclusions

- In this case it is seen as **the first rating is the one that most influences** (is most influenced) on the latent factor, although the other two also have a high rating with the factor.
- In a model with so few variables it does not seem that adjusting two factors greatly improves the interpretation of the results, although **as a practical voluntary exercise** it is proposed to **repeat this process to adjust a model with two factors** to this example.

Desirable situation

A **desirable situation** would be one in which the **factorial matrix showed a high correlation of each of the factors with a group of specific observable variables** and practically zero with the rest.

For instance:

Assuming that $X = (X_1, \ldots, X_7)$ is a random vector used to fit a factorial model with three factors with the factorial matrix bellow,

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

such a situation would indicate that the first factor influences the variables X_1, X_2, X_3 but not the others. The second factor in the variables X_4, X_5 and not in the others. Finally, the third factor only influences the variables X_6, X_7.

Some comments

- As said before, problems (1) and (2) do not have a single solution. This implies that it is unlikely that a representation as simple as the one shown in the previous example will be found as the first solution. As far as possible, the **factorial matrix should approximate one like the previous one**.
- A matrix G_k **orthogonal** ($G^{-1} = G^t$) represents an isometry in the Euclidean vector space $\mathbb{R}^k$ (**rotations**, reflections or composition of both).
- If an orthogonal matrix, G of order k, is considered, the equation (2) does not change according to the following expression:

$$\Sigma = E[XX^t] = AA^t + D = AGG^tA + D$$

Denoting $B = AG$, an expression equivalent to (2) is obtained in the form:

$$\Sigma = E[XX^t] = BB^t + D$$

More comments

- Finding the matrices A and D that solve the equation (2) is equivalent to finding the matrices B and D that also solve it.
- This implies that one can **change to a simpler view of reality** simply by **introducing a suitable rotation**. And therefore the equation (1) can be written as:

$$X = AGG^tF + L = BG^tF + L = BF_G + L$$

- The previous expression introduces the concept of **factor rotation**.
- There are two possible types of rotations: **orthogonal and oblique**.
- At this time, orthogonal rotations are considered, **due to the simplicity of their interpretation, since the weights of the factorial matrix represent the correlations between the variables and the factors**. This is not true in the case of obliques.

Rotations

The **main goal when performing a rotation is to find a simple structure** representing reality.

In this sense, the factorial matrix should satisfy the following **properties**:

- **Each row** of the factorial matrix must **contain** at least **one zero**.
- **Each column** of the factorial matrix must **contain** at least **k zeros**.
- **Each pair of columns** of the factorial matrix must **contain several variables** whose **weights are null in one column**, but not in the other.
- If there are **more than four factors** each **pair of columns of the factor matrix must contain a large number of variables with null weights** in both columns.
- Reciprocally, if there are **more than four factors**, in each pair of columns of the factor matrix **only a small number of variables should contain non-zero weights**.

Rotations: quartimax approach

- The **quartimax** approach chooses $A_G = AG$ for which the variance per **rows** of the squares of the rotated factorial loadings $\hat{a}_{ij}$ is maximum.

$$\max_{\hat{a}_{ij}} \left(\frac{1}{pk} \sum_{j=1}^{k} \sum_{i=1}^{p} (\hat{a}_{ij}^2 - \frac{1}{pk} \sum_{j=1}^{k} \sum_{i=1}^{p} \hat{a}_{ij}^2)^2 \right)$$

- This approach **ensures that a given variable is highly correlated with one factor** and **very little correlated with the rest** of the factors.

Rotations: varimax approach

- The **varimax** approach chooses $A_G = AG$ for which the variance per **columns** of the squares of the rotated factorial loadings $\hat{a}_{ij}$ is maximum.

$$\max_{\hat{a}_{ij}} \left(\frac{1}{p} \sum_{j=1}^{k} \sum_{i=1}^{p} (\hat{a}_{ij}^2 - \frac{1}{p} \sum_{i=1}^{p} \hat{a}_{ij}^2)^2 \right)$$

- This approach tries to ensure that there are **factors with high correlations with a small number of variables and null correlations with the rest**. In this way the variance of the factors is redistributed.
- This is the **most common approach** when working with orthogonal rotations of factors.

Goals

- To **determine** the appropriate **number of factors**.

  Choosing the appropriate number of factors to represent the observed covariances is very important, since **between a solution with k factors and another with $k+1$ factors, very different factor matrices can be obtained**. This **did not happen in the PCA**, since the principal components are always the same let's take k or $k+1$ from them.

- To **estimate a factorial matrix** A, by some of the methods introduced below, which will then be rotated according to whether it is interesting to simplify the interpretation of reality or not.

Principal factor method

- Technique that, like PCA, is based on the **calculation of eigenvalues and eigenvectors**, but in this case, not on the covariance matrix, but on a **reduced covariance matrix**

$$S^* = S - \hat{D},$$

where $\hat{D}$ is an estimate of the diagonal matrix of specific variances (variances of the specific factors) and S, as before, an estimate of the covariance matrix.

- By subtracting $\hat{D}$, **the diagonal of the matrix S^* contains the different communalities** (parts of the variances of each variable explained by the latent factors).
- Conversely to principal components analysis, factor analysis **does not attempt to collect all the observed variance of the data, but that shared by common factors**.
- Finally, **the principal factor method consists of applying a principal component analysis for the matrix S^***.
- **The eigenvectors** now represent the **columns of the factorial matrix**.

Maximum likelihood method

- The data must be distributed according to a **multivariate Gaussian** distribution.
- **A distance matrix is defined between the observed covariance matrix and its values predicted by the factor analysis model**. This distance is defined as follows:

$$F = \ln |AA^t + D| + \mathrm{trace}(S|AA^t + D|^{-1}) - \ln |S| - p$$

- The estimates of the factorial weight matrix, A, are obtained **minimizing this distance**.
- This minimization problem is **equivalent to maximizing the likelihood function** of the k factorial model under the assumption of normality.
- The maximum likelihood method has an associated **statistical test to estimate the appropriate number of factors**.

FA Practice 2

In this practice, from among 25 items of a personality test, the variables that correspond to **each of the five aspects of the personality** of an individual **will be identified**. The five characteristics that define the personality of an individual are: A - *Agreeableness or friendliness*; C - *Consciousness or responsibility*; E - *Extraversion*; N - *Neuroticism* and - *Openness to experiences*.

To carry it out, you must **download and execute** the file **Practice_2_AF.Rmd** available on the PRADO platform.

Topics covered:

- Perform a prior exploratory analysis of the data to identify possible **missing data** and **extreme values**.
- **Make decisions and deal with** missing data and extreme values.
- Check the assumptions and perform a **FA**.
- **Choosing the optimal number** of factors.
- Interpretation of different **graphic outputs of interest** for this method.
- **R language**: functions debugging.

[1] Anderson, T.W. (2003, 3ª ed.). An Introduction to Multivariate Statistical Analysis. John Wiley & Sons.

[2] Gutiérrez, R. y González, A. (1991). Estadística Multivariable. Introducción al Análisis Multivariante. Servicio de Reprografía de la Facultad de Ciencias. Universidad de Granada.

[3] Härdle, W.K. y Simar, L. (2015, 4ª ed.). Applied Multivariate Statistical Analysis. Springer.

[4] Johnson, R.A. y Wichern, D.W. (1988). Applied Multivariate Analysis. Prentice Hall International, Inc.

[5] Rencher, A.C. y Christensen, W.F. (2012, 3ª ed.). Methods of Multivariate Analysis. John Wiley & Sons.

[6] Salvador Figueras, M. y Gargallo, P. (2003). Análisis Exploratorio de Datos. Online en http://www.5campus.com/leccion/aed.

[7] Timm, N.H. (2002). Applied Multivariate Analysis. Springer.

[8] Vera, J.F. (2004). Análisis Exploratorio de Datos. ISBN: 84-688-8173-2.