Software Development for Data Analysis

Factor Analysis or Exploratory Factor Analysis (EFA)

Exploratory Factor Analysis

- Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors.
- Factor analysis searches for such joint variations in response to unobserved latent variables.
- Factor analysis looks for independent dimensions, which limits its applicability in biological sciences.
- Users of factor analysis believe that it helps to deal with data sets where there is a large number of observed variables that are thought to reflect a smaller number of underlying/latent variables.
- Factor analysis is related to principal component analysis (PCA), but the two are not identical.

Factor Analysis or Exploratory Factor Analysis (EFA)

Exploratory Factor Analysis

- There has been significant controversy in the field over differences between the two techniques.
- In factor analysis, the researcher makes the assumption that an underlying causal model exists, whereas PCA is simply a variable reduction technique.
- Principal component analysis employs a mathematical transformation to the original data with no assumptions about the form of the covariance matrix.
- The aim of PCA is to determine a few linear combinations of the original variables that can be used to summarize the data set without losing much information.
- In EFA, the observed variables are modelled as linear combinations of the potential factors, plus "error" terms.

Exploratory Factor Analysis (EFA)

Factor Analysis

$$\begin{cases} X_1 = u_{11}F_1 + u_{12}F_2 + ... + u_{1q}F_q + e_1 \\ X_2 = u_{21}F_1 + u_{22}F_2 + ... + u_{2q}F_q + e_2 \\ \qquad\qquad ... \\ X_m = u_{m1}F_1 + u_{m2}F_2 + ... + u_{mq}F_q + e_m \end{cases} \qquad (1)$$

where:

- q is the number of factors ($q < m$),
- $X_j,\ j = \overline{1,m}$, are the centered or standardized causal variables (observed variables),
- $F_k,\ k = \overline{1,q}$, are the common factors,
- $e_j,\ j = \overline{1,m}$, are the specific factors, and
- $u_{jk},\ k = \overline{1,q}$, are the factorial coefficients (*factor loadings*).

Exploratory Factor Analysis (EFA)

Factor Analysis

- In terms of matrixes, the relation between the observed variables and factors can be written as:

  $\boldsymbol{X = Fu^t + e}$,

where:

- $\boldsymbol{X}$ is the observation tables,
- $\boldsymbol{u^t}$ the factorial coefficients matrix transposed,
- $\boldsymbol{e}$ is the specific factors matrix, and
- $\boldsymbol{F}$ is the common factors matrix, arranged on columns.

$$e = \begin{bmatrix} e_1 & 0 & \dots & 0 \\ 0 & e_2 & \dots & 0 \\ \dots & & \dots & \\ 0 & 0 & \dots & e_m \end{bmatrix}, \quad F = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1q} \\ f_{21} & f_{22} & \cdots & f_{2q} \\ \dots & & \dots & \\ f_{m1} & f_{m2} & \cdots & f_{mq} \end{bmatrix}$$

Exploratory Factor Analysis (EFA)

Model hypotheses

- The common factors ***F*** are standardized, and therefore have the variance 1 and the mean 0.
- They are build based on the information segregation principle, and consequently are not correlated to each other:

  $R(F_i, F_j)=0$, $Cov(F_i, F_j)=0$, $i = \overline{1,q}$, $j = \overline{1,q}$, $i \neq j$

- The specific factors have the mean 0.
- Specific factors variance is named specific variance (*psi*):

  $$\psi_j = \text{var}(e_j), \quad j = \overline{1,m}$$

- Among the common and the specific factors there is no correlation what so ever:

  $R(F_i, e_j) = 0$, $Cov(F_i, e_j) = 0$, $i = \overline{1,q}$, $j = \overline{1,m}$.

Exploratory Factor Analysis (EFA)

Model hypotheses

- These hypotheses are necessary in order to estimate uniquely the model's parameters.

- Hence, the variance of an observed variable is:

$$\mathrm{Var}(X_j) = \mathrm{Var}(u_{j1}\cdot F_1 + u_{j2}\cdot F_2 + ... + u_{jq}\cdot F_{\mathrm{q}} + e_j)$$

$$= u_{j1}^2\cdot\mathrm{Var}(F_1) + u_{j2}^2\cdot\mathrm{Var}(F_2) + ... + u_{jq}^2\cdot\mathrm{Var}(F_q) + \mathrm{Var}(e_j)$$

$$= \sum_{k=1}^{q} u_{jk}^2 + \psi_j \,. \qquad j = \overline{1,m}$$

Exploratory Factor Analysis (EFA)

Model hypotheses

- The sum $h_j = \sum_{k=1}^{q} u_{jk}^2$ represents the **commonality** of the variable X_j .
- It has to be observed that if variables X_j were standardized, then the commonalities can be no greater than 1. The commonality represents the common variability, due to common factors.

- And covariance between 2 observed variables X_i și X_j is:

  $\text{Cov}(X_i,X_j) = \text{Cov}(u_{i1}F_1+...+u_{iq}F_q+e_i, u_{j1}F_1+...+u_{jq}F_q+e_j)$

  $= \sum_{k=1}^{q} u_{ik}u_{jk}$.

- The covariance between an observed variable and a common factor is:

  $\text{Cov}(X_i,F_j) = \text{Cov}(u_{i1}F_1+...+u_{iq}F_q+e_i, F_j) = u_{ij}$

Exploratory Factor Analysis (EFA)

Model hypotheses

- Taking into account the previous relations, the covariance matrix of the observed dataset can be written as:

$$\boldsymbol{V = u\, u^t + \psi}\ ,$$

where: $\psi = \begin{bmatrix} \psi_1 & 0 & \dots & 0 \\ 0 & \psi_2 & \dots & 0 \\ \dots & & \dots & \\ 0 & 0 & \dots & \psi_m \end{bmatrix}$

is the diagonal matrix of specific factor variances.

Exploratory Factor Analysis (EFA)

Factor existence estimation

- The hypothesis of latent factors existence is induced by a significant correlation among the causal variables. The common variability of the observed, causal variables leads to the idea of explaining it through the existence of hidden factors. Consequently, based on correlation matrices of the causal variables, there could be estimated the common factors existence.
- There are various tests designed to verify the latent factors existence hypothesis, based on the matrix of correlations.

Exploratory Factor Analysis (EFA)

Factor existence estimation - Bartlett sphericity test

- It is a χ^2 test which compares the matrix of correlations with the identity (unit) matrix. If a multivariate cloud was perfectly spherical, then its covariance matrix would be proportional to the identity matrix. Test hypotheses:
  - H0: There are no factors
  - H1: There is at least one common factor
- In the case of absolutely not correlated causal variables, the determinant of the correlation matrix is 1.

Exploratory Factor Analysis (EFA)

Factor existence estimation - Bartlett sphericity test

- The statistics of the test is:

  $\chi_c^2 = -\left(n - 1 - \frac{2 \cdot m + 5}{6}\right) \ln|R|$, where:

  - R is the correlation matrix of the causal variables,
  - n is the number of instances or observations,
  - and m is the number of variables.

- These values follow a χ^2 (chi-square) distribution with m·(m-1)/2 degrees of freedom.
- If $\chi_c^2 > \chi^2(m \cdot (m-1)/2), \alpha)$, then the null hypothesis is rejected with a $1 - \alpha$ level of confidence.

Exploratory Factor Analysis (EFA)

Factor existence estimation - KMO (Kaiser-Meyer-Olkin) index

- It is based on correlation matrices as well.
- The global KMO index is computed as follows:

$$KMO = \frac{\sum_i \sum_{j \neq i} r^2_{ij}}{\sum_i \sum_{j \neq i} r^2_{ij} + \sum_i \sum_{j \neq i} a^2_{ij}}$$, where:

  - r_{ij} is the linear correlation coefficient between the variables X_i and X_j;
  - a_{ij} represents the partial correlation coefficient between X_i and X_j (correlation between residuals).

- The KMO_j index for each variable is computed as follows:

$$KMO_j = \frac{\sum_{j \neq i} r^2_{ij}}{\sum_{j \neq i} r^2_{ij} + \sum_{j \neq i} a^2_{ij}}, \text{j} = \overline{1, m},$$

- The above indices reveal which variables are less correlated with the others, and hence offer less common variability.

Exploratory Factor Analysis (EFA)

Factor existence estimation - KMO (Kaiser-Meyer-Olkin) index

- The partial correlation coefficient between two variables X_i and X_j is computed as follows:

  $a_{ij} = \frac{t_{ij}}{\sqrt{t_{ii} - t_{jj}}}$, where t_{ij} is the general term of the matrix $T = R^{-1}$.
- The partial correlation represent the linear link between two variables, while the effect on them of the other model variables is neutralized.
- KMO index value interpretation:

  0.90 - 1.00: very good factorability

  0.80 - 0.89: good factorability

  0.70 - 0.79: average

  0.60 - 0.69: mediocre

  0.50 - 0.59: weak

  0.00 - 0.49: no factors

Exploratory Factor Analysis (EFA)

Model parameters estimation. Factors extraction.

- The algorithm consists in determining the factorial coefficients: the common factors and the specific factors.
- The factor extraction algorithm starts from the hypothesis of a model containing a single common factor, and then is tested the discrepancy between the covariance matrix of observed variables and the one produced by the model:

$$V \approx \hat{V} = u \cdot u^t + \psi$$

- If the test is rejected, the discrepancy is statistically too large, then is tested a model having 2 common factors and so on, until the discrepancy test is passed.

Exploratory Factor Analysis (EFA)

Model parameters estimation. Factors extraction.

There are various methods that can be employed to extract factors, depending on the criteria used for testing the discrepancy between the 2 covariance matrixes:

- The maximum likelihood method
- Principal component analysis
- The least squares method
- Principal axis factoring

Exploratory Factor Analysis (EFA)

The maximum likelihood method

- The basis is that the observations follows a normal multivariate distribution, and the observations have been taken independently. This requirements represent a disadvantage for this method.
- Factor extractions process is an iterative one, and consists in maximizing a probability function as follows:

$$l\left(\hat{V}\right) = -\frac{1}{2} n \cdot \log\left|2\pi \cdot \hat{V}\right| - \frac{1}{2} n \cdot trace\left(\hat{V}^{-1} V\right)$$

  where $\hat{V}$ is the covariance matrix estimated through the model.
- ***trace()*** is defined to be the sum of the elements on the main diagonal (the diagonal from the upper left to the lower right).

Exploratory Factor Analysis (EFA)

Principal Component Analysis method

- Factor analysis and PCA are sometimes mistakenly used in a interchangeable manner. Some specialized applications employ PCA as a algorithmic engine for Factor Analysis.
- Having given ***C***, the $n \times m$ principal component matrix arrange on the columns: $C^1, C^2, ..., C^m$
- Then let ***A*** be the a $m \times m$ matrix of a^k coefficients, arranged on columns as well: $a^1, a^2, ..., a^m$.
- Since $C^k = X \cdot a^k$, for $k = 1,2,...,m$, then in terms of matrixes: $C = X \cdot A$.
- Matrix ***A*** has any 2 columns orthogonal, and therefore $A^{-1} = A^t$.
- Then $X = C \cdot A^{-1} = C \cdot A^t$, or as columns:

$$X = \sum_{j=1}^{m} C_j (a_j)^t$$

Exploratory Factor Analysis (EFA)

Principal Component Analysis method

- The standard deviation of a principal component is $\sigma(C_j)$, and is the square root of component's variance $\sqrt{\alpha_j}$.
- Therefore: $X = \sum_{j=1}^{m} \sqrt{\alpha_j} \frac{C_j}{\sqrt{\alpha_j}} \cdot (a_j)^t = \sum_{j=1}^{m} C_j^S \cdot \left(R_j\right)^t$

where C_j^S are **the scores** or the standardized components, and R_j is the column vector of correlation coefficients between the observed variables and C_j component (**factor loadings** in PCA).

Exploratory Factor Analysis (EFA)

Principal Component Analysis method

- The previous equality shows how the initial matrix X can be retrieved starting from the scores and the factorial correlations.
- The factorial correlations are decreasing with each principal component taken into account.
- If we consider only the first q phases of PCA, then the table X can be approximated with:

$$\sum_{j=1}^{q} C_j^S \cdot \left(R_j\right)^t$$

- While the quantity:

$$\sum_{j=q+1}^{m} C_j^S \cdot \left(R_j\right)^t$$

may be considered as residual, negligible quantity, with a non-significant contribution at table ***X*** reconstitution.

Exploratory Factor Analysis (EFA)

Principal Component Analysis method

- Consequently:

$$X = \sum_{j=1}^{q} C_j^S \cdot \left(R_j\right)^t + \sum_{j=q+1}^{m} C_j^S \cdot \left(R_j\right)^t = C^S \cdot R^t + e \,,$$

where:

- $\boldsymbol{C}^S$ is the matrix of **scores**,
- $\boldsymbol{R}$ is the matrix of correlations between the principal components and the observed variables (**factor loadings**), and
- $\boldsymbol{e}$ is the residual contribution to $\boldsymbol{X}$ reconstitution.
- Unfolding the above sums and comparing them with (1) equations, there can be observed that the factors F_k are actually PCA **scores**, and u coefficients from EFA are **factor loadings** from PCA.

Exploratory Factor Analysis (EFA)

Number of factors estimation. Bartlett's test

- Exploratory Factor Analysis starts from the assumption of a number of factors, q. This presumed number may be insufficient.
- A good model is one in which the estimated covariance matrix, derived from the model, nears as much as possible the covariance matrix of the observed, causal variables.
- In order to test the model, statistical tests are employed, such *goodness-of-fit* test is the one proposed by Bartlett, a chi-square test.
- A perfect description of the observed date through a number of factors q implies that $\hat{V} = V$

  Multiplying with $\hat{V}^{-1}$ on the left we obtain $\hat{V}^{-1}\hat{V} = \hat{V}^{-1}V$, and hence:

  $$\hat{V}^{-1}\hat{V} = (u \cdot u^t + \psi)^{-1}V = I,$$

  where I is the identity (unit) matrix of rank m.

Exploratory Factor Analysis (EFA)

Number of factors estimation. Bartlett's test

- The statistics of the test is computed as follows:

$$\chi_c^2 = \left(n - 1 - \frac{2m + 4q\ -\ 5}{6}\right)\left(\left(trace\left((u \cdot u^t +\ \psi)^{-1}V\right)\right) - ln(|(u \cdot u^t +\ \psi)^{-1}V|)\ - m\right)$$

It has to be noticed that the statistics of the test is 0 for a perfect description of the observed data through $\boldsymbol{q}$ factors. The trace of the matrix $\left((u \cdot u^t +\ \psi)^{-1}V\right)$ is $\boldsymbol{m}$, the trace of identity matrix, and the logarithm from determinant is 0 (the determinant of the identity matrix is 1).

Therefore, the lower the value of the statistics, the greater the chance that the factors describe adequately the observed data. And the vice versa, the greater the value of the statistics, the chances that the factors to describe the data adequately are diminishing.

Exploratory Factor Analysis (EFA)

Number of factors estimation. Bartlett's test

Hypotheses:

- H0 - the factors of the model describe the observed data accurately;
- H1 – the model factors do not describe the observed data adequately.
- If $\chi_C^2 > \chi^2(r, \alpha)$, then the null hypothesis, H0, is rejected with a $1 - \alpha$ level of confidence, where $r = \frac{(m-q)^2 - m - q}{2}$ is the number of degrees of freedom.