University of Ferrara

DIPARTIMENTO DI ECONOMIA E MANAGEMENT

Stefano Bonnini & Valentina Mini

Factor Analysis

And

Principal Component Analysis

Summary

- Introduction to Factor Analysis (FA)
- Factor Model
- Parameter estimation and factor rotation in FA
- How to proceed with FA…
- Principal Component Analysis (PCA)

Introduction to Factor Analysis (FA)

- FA is a multivariate technique for the analysis of data structure
- Goal: reduce the number of informative variables through the definition of new variables called factors
- Method: transformation of the structure of observed data into a new structure such that the data variability is explained by the factors

Introduction to Factor Analysis (FA)

Example: A marketing survey on the demand of the wine «Passito» has been performed.

A sample of n=386 people has been interviewed. The questionnaire includes several questions about their preferences and behaviors related to drinking wine

- Age: ____ - Sex: M ○ F ○ - Province of Residence: ____________

- Do you like drinking wine?

not at all	1	2	3	4	5	6	7	very much
	○	○	○	○	○	○	○	

- How often do you drink wine…

	never	rarely	sometimes	often	regularly
…at home with meals?	○	○	○	○	○
…in bars or pubs?	○	○	○	○	○
…at restaurants with meals?	○	○	○	○	○

- Do you know the wine Passito?

not at all	1	2	3	4	5	6	7	very well
	○	○	○	○	○	○	○	

Introduction to Factor Analysis (FA)

The variables:

Label	Description		Coding
ID	Personal ID of the interviewed		Increasing integer number
AgeClass	Age of the person		Age (years)
AGE_CLASS	Age class of the person		1-6
SEX	Sex of the person		M or F
PROV	Province where the interviewed lives		Province code
LIKE_WINE	How much do you like drinking wine?		Integrer number from 1 to 7
FREQ_HOME	How often do you drink wine at home with meals?		Integrer number from 1 to 5
FREQ_BAR	How often do you drink wine in bars/pubs?		Integrer number from 1 to 5
FREQ_REST	How often do you drink wine at restaurants with meals?		Integrer number from 1 to 5
KNOW_PAS	Do you know the wine Passito?		Integrer number from 1 to 7
FREQ_PAS	How often do you drink Passito?		Integrer number from 1 to 5
FREQ_P_HOL	How often do you drink Passito on holidays and celebrations?		Integrer number from 1 to 5
FREQ_P_ALO	How often do you drink Passito when you are alone?		Integrer number from 1 to 5
FREQ_P_MEA	How often do you drink Passito at the end of meals?		Integrer number from 1 to 5
FREQ_P_OFF	How often do you drink Passito offered by someone?		Integrer number from 1 to 5
HOW_MUCH	How much wine do you drink in one year?		Integrer number from 1 to 4
LIKE_PAS	How much do you like drinking Passito?		Integrer number from 1 to 7
LIKE_AROMA	How much do you like aroma and smell of Passito?		Integrer number from 1 to 7
LIKE_SWEET	How much do you like the sweetness of Passito?		Integrer number from 1 to 7
LIKE_ALCOHOL	How much do you like the alcohol content of Passito?		Integrer number from 1 to 7
LIKE_TASTE	How much do you like the intensity of taste of Passito?		Integrer number from 1 to 7
PRICE	How much could you pay for one bottle of Passito? (0.5 litre)		Integrer number from 1 to 5

Introduction to Factor Analysis (FA)

The dataset:

ID	AGE	AGE_CLAS	SEX	PROV	LIKE_WINE	FREQ_HOME	FREQ_BAR	FREQ_REST	KNOW_PAS	…
1	26	1	M	PD	6	2	4	4	4	
2	43	3	M	PD	7	3	1	4	6	
3	32	2	M	VR	6	4	3	3	6	
4	53	4	F	PD	6	4	2	5	5	
5	30	2	M	PD	4	2	3	4	2	
6	23	1	F	VR	5	3	2	4	5	
7	46	3	M	VE	5	2	3	6		
8	26	1	M	PD	6	3	2	5	5	
9	25	1	M	BL	6	3	4	4	7	
10	22	1	M	VE	5	3	4	4	5	
11	24	1	M	VE	4	1	3	3	3	
12	22	1	M	VE	7	5	4	5	7	
13	23	1	M	VI	7	3	5	5	7	
14	23	1	M	VE	7	4	4	4	4	
…		…	…	…	…	…	…	…	…	

Introduction to Factor Analysis (FA)

Factor properties:

- Uncorrelated with each other

- Unobserved latent variables (unknown a priori) which reproduce the existing correlations between original variables

- Original variables are linear combinations of factors

Introduction to Factor Analysis (FA)

Assumptions:

- FA can be applied to a set of numeric standardizable variables

- Number of statistical units should be at least 5 times the number of original variables

Summary

- Introduction to Factor Analysis (FA)
- Factor Model
- Parameter estimation and factor rotation in FA
- How to proceed with FA…
- Principal Component Analysis (PCA)

Factor Model

- $X_1,\ldots,X_k$ response variables such that

  - $E(X_j) = \mu_j$, $Var(X_j) = \sigma_{jj} = \sigma_j^2$, $Cov(X_j X_r) = \sigma_{jr}$;
    $j,r = 1,..,k$

- $X_j = \lambda_{j1} F_1 + \lambda_{j2} F_2 + \ldots + \lambda_{js} F_s + \ldots + \lambda_{jq} F_q + U_j + \mu_j$,

  $= \Sigma_s \lambda_{js} F_s + U_j + \mu_j$, $j=1,\ldots,k$

  - $\lambda_{j1}, \lambda_{j2}, \ldots, \lambda_{jq}$ $(j=1,\ldots,k)$: parameters (constants) called factor loadings
  - $F_1, F_2,\ldots, F_q$ common factors (random variables)
  - U_j, unique or specific factor $(j=1,\ldots,k)$

Factor Model

Model assumptions:

- ✓ $E(F_s)=0,$ $\quad s=1,\dots,q$
- ✓ $Var(F_s)=1,$ $\quad s=1,\dots,q$
- ✓ $Cov(F_s, F_t)=0,$ $\quad s,t=1,\dots,q;\ s \neq t$

- ✓ $E(U_j)=0,$ $\quad j=1,\dots,k$
- ✓ $Var(U_j)= {}_u\sigma_{jj} = {}_u\sigma_j^2$ $\quad j=1,\dots,k$
- ✓ $Cov(U_j, U_r)=0$ $\quad j,r=1,\dots,k;\ j \neq r$

- ✓ $Cov(F_s, U_j)=0$ $\quad s=1,\dots,q;\ j=1,\dots,k$

Factor Model

Matrix representation:

- $\boldsymbol{X} = [X_1,\ldots,X_k]'$ random vector of response variables
- $\boldsymbol{F} = [F_1,\ldots,F_q]'$ random vector of unique factors
- $\boldsymbol{U} = [U_1,\ldots,U_k]'$ random vector of unique factors
- $\Lambda = [\lambda_{js}]$ *$k \times q$ matrix of constants (parameters)*
- $\boldsymbol{\mu} = [\mu_1,\ldots, \mu_k]'$ vector of means

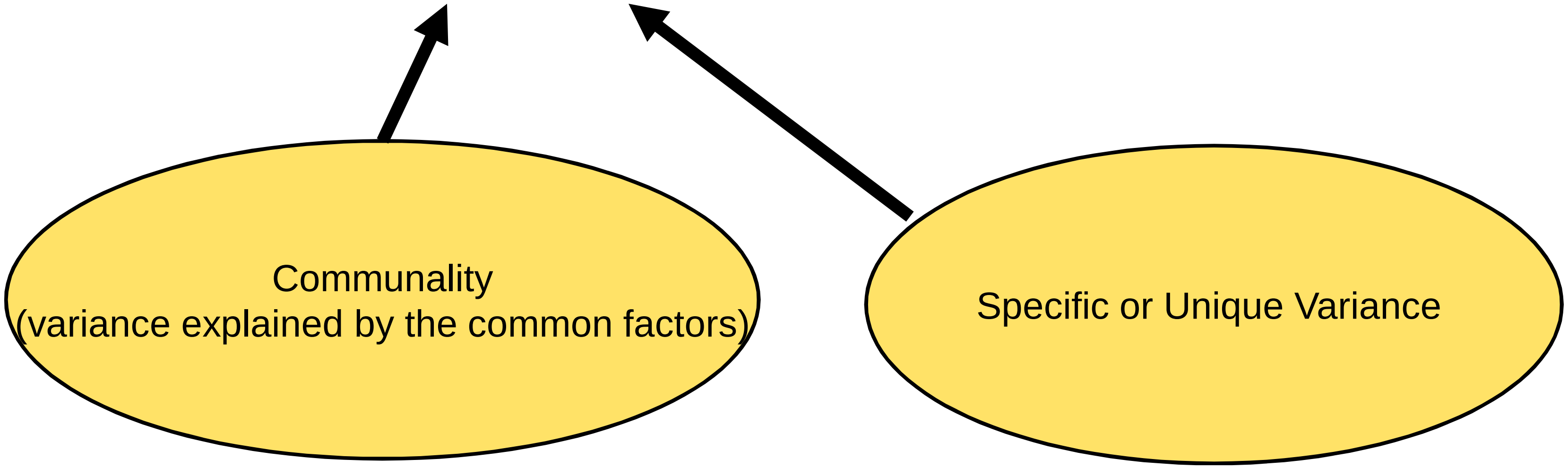
- $\boldsymbol{X} = \Lambda \boldsymbol{F} + \boldsymbol{U} + \boldsymbol{\mu}$

  - $E(X)= \mu$, $Var(X)= \Sigma = [\sigma_{jr}]$
  - $E(F)= 0$, $Var(F)= I = diag(1,1,\ldots,1)$
  - $E(U)= 0$, $Var(U)= diag({}_u\sigma_{11},\ldots, {}_u\sigma_{kk})= {}_u\Sigma$
  - $Cov(F,U)= 0$

Factor Model

Variance decompocition (VD):

$$Var(X_j) = \sigma_{jj} = \Sigma_s \lambda_{js}^2 + {}_u\sigma_{jj}$$
$$= h_j^2 + {}_u\sigma_{jj}, \quad j=1,\dots,k$$



$\lambda_{js} = E(X_j, F_s) = Cov(X_j, F_s) \rightarrow$ measure of the linear dependence between X_j and F_s

With matrix notation: $\Sigma = \Lambda\Lambda' + {}_u\Sigma$

Factor Model

- FA can be applied to <u>standardizable</u> numeric variables
- The number of units should be at least 5 times the number of original response variables: $n \geq 5 \times k$
- The common factors should explain at least 70% of the global variability of the original response variables
- The problem of detecting ***F*** and Λ has no unique solution

Summary

- Introduction to Factor Analysis (FA)
- Factor Model
- Parameter estimation and factor rotation in FA
- How to proceed with FA…
- Principal Component Analysis (PCA)

Parameter estimation and factor rotation in FA

If factor model assumptions are true for $\boldsymbol{F}$, then a rotation of **F** provides new factors **F*** for which the assumptions are still true and which correspond to different factor loadings Λ^* .

Formally:

given the orthogonal $q \times q$ matrix **G** (such that $\boldsymbol{G}\boldsymbol{G}' = \boldsymbol{I}$)

$$
\begin{aligned}
\boldsymbol{X} = \quad & \Lambda \boldsymbol{F} + \boldsymbol{U} + \boldsymbol{\mu} = \\
& = \Lambda \boldsymbol{G}\boldsymbol{G}'\boldsymbol{F} + \boldsymbol{U} + \boldsymbol{\mu} = \\
& = (\Lambda \boldsymbol{G})(\boldsymbol{G}'\boldsymbol{F}) + \boldsymbol{U} + \boldsymbol{\mu} = \\
& = \Lambda^* \boldsymbol{F}^* + \boldsymbol{U} + \boldsymbol{\mu}
\end{aligned}
$$

Parameter estimation and factor rotation in FA

To overcome the indeterminacy of factor loadings we can impose the constrain $\Lambda^{*\prime}{}_u\Sigma^{-1}\Lambda^* =$ diagonal

Parameter estimates:

- ✓ $\hat{\boldsymbol{\mu}}$ = sample mean of $\boldsymbol{X}$, i.e. $\hat{\mu}_j = \bar{x}_j = \sum_{i=1}^{n} x_{ij}/n$
- ✓ $\hat{\lambda}_{js} = l_{js}$
  - ○ l_{js} can be computed through the application of
    - Principal Factor Analysis (based on correlations)
    - Maximum likelihood Factor Analysis
- ✓ $\hat{\sigma}_{jj} = s_{jj} = \sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2/(n-1)$
- ✓ ${}_u\hat{\sigma}_{jj} = s_{jj} - \sum_{s=1}^{q} l_{js}^2$

Parameter estimation and factor rotation in FA

- The previous constrain on factor loadings, simplifies the computation of the estimates, thus it is mathematically convenient, but it can create some problems of interpretation in some cases

- Appropriate constrained transformation could be the one that allows to get …:
  - Few factor loadings distant from zero
  - Several factor loadings close to zero

- A suitable factor rotation can provide such result

Parameter estimation and factor rotation in FA

- Factor rotation methods:
  - Varimax: orthogonal rotation that provides few factor loadings far from zero and several factor loadings close to zero
  - Equimax
  - Quartimax
  - …

Summary

- Introduction to Factor Analysis (FA)
- Factor Model
- Parameter estimation and factor rotation in FA
- How to proceed with FA…
- Principal Component Analysis (PCA)

How to proceed with FA…

1. Standardization of variables
2. Compution of the matrix of covariances (correlations) of the original response variables
3. Factor extraction and factor loadings estimation
4. Factor rotation for a easier interpretation of the factors
5. Interpretation of the factors
6. Computation of the coefficients of factor scores (weights of the linear combinations which represent the factors as function of the original observed variables)

How to proceed with FA…

How many factors?

- No unique answer…it depends on the problem and on the observed data
- **Apriori method**: based on the experience of the researcher and on the theory
- **Method based on eigenvalues**: only factors with eigenvalues >1 must be considered; the eigenvalue represents the amount of variance explained by the factor
- **Method based on the least percentage of explained variance**: only factors which explain at least 10% of variance must be considered

How to proceed with FA…

Graphical solution of the Scree Test (Scree Plot):

1. Eigenvalues are represented in a graph where one axis correspond to the factors and the other axis correspond to the eigenvalues (factors are sorted according to the eigenvalue)

2. The first q factors are extracted according to one of the following rules:
   a. The $(q+1)$-th factor has eigenvalue less than a specific threshold (e.g. 1)
   b. The difference between the $(q+1)$-th and the q-th eigenvalue is not considerable

Summary

- Introduction to Factor Analysis (FA)
- Factor Model
- Parameter estimation and factor rotation in FA
- How to proceed with FA…
- Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

- Also known as Hotelling transformation and Karhunen-Loeve expansion
- Among the oldest and most common methods of multivariate analysis
- Proposed by Pearson in 1901 and then (independently) by Hotelling in 1933
- It provides an effective method for representing multivariate data in a space with a reduced dimensionality (*parsimonious summarization of data*), for simplifying the statistical analysis
- Useful method for explorative analyses or prediction models

Principal Component Analysis (PCA)

- Goals:
  - o Reduce the dimensionality of the dataset
  - o Detect new informative variables which can replace the observed original variables
  - o Use a graphical representation of data to get some preliminary information previous to a following analysis
  - o Reduce the number of explanatory variables in a multiple regression model in the presence of multicollinearity

Principal Component Analysis (PCA)

- Result:

  - The original variability of the observed response variables $X_1,\dots,X_k$ (which usually are correlated between each other) can be described by new uncorrelated variables $Y_1,\dots,Y_k$, which are linear combinations of the original observed variables $X_1,\dots,X_k$

  - The variables $Y_1,\dots,Y_k$ are sorted according to the degree of importance, i.e. Y_1 is the variable which «explains» the greatest proportion of variability; Y_2 is the variable (uncorrelated with Y_1) which «explains» the greatest proportion of the remaining variability; etc.

  - $Y_1,\dots,Y_k$, are called PRINCIPAL COMPONENTS

Principal Component Analysis (PCA)

- Assumptions:
    - $X_1,\ldots,X_k$ follow a (multivariate) distribution with mean vector μ and covariance matrix Σ;
    - The values in μ and Σ are finite;
    - The rank of Σ is $q < k$;
    - The dataset is given by the $n \times k$ matrix $[x_{ij}]$, $i=1,..,n;$ $j=1,\ldots,q$

Principal Component Analysis (PCA)

- S and R matrices:
  - A suitable estimate of Σ is provided by the sampling covariance matrix $S=[s_{ij}]$ which includes the necessary information for PCA
  - As a matter of fact the information for PCA is usually provided by the matrix of sampling correlations $R=[r_{ij}]$, especially when the magnitudes, the units of measurement or the variabilities of the original variables are very much different
  - Principal Components (PC) extraction from R is equivalent to PC extraction from S after standardization of the original variables

Principal Component Analysis (PCA)

- First Principal Component:

1. $Y_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{k1}X_k$

2. Detect the values $a_{11}, \cdots, a_{k1}$ which maximize the variance of Y_1, formally:
   - find $a_{11}^*, a_{21}^*, \cdots, a_{k1}^*$ such that
     1. $\max[Var(Y_1)] = Var(a_{11}^*X_1 + a_{21}^*X_2 + \cdots + a_{k1}^*X_k) = \\ = \sum_{j,r} a_{j1}^* a_{r1}^* s_{jr}$
     2. $\sum_j \left(a_{j1}^*\right)^2 = 1$
   - $\lambda_1 = \max[Var(Y_1)] = \sum_{j,r} a_{j1}^* a_{r1}^* s_{jr}$ max eigenvalue of S
   - $(a_{11}^*, \cdots, a_{k1}^*)'$ eigenvector of S which corresponds to λ_1

Principal Component Analysis (PCA)

- Second Principal Component:

  1. $Y_2 = a_{12}X_1 + a_{22}X_2 + \cdots + a_{k2}X_k$

  2. Detect the values $a_{12}, \cdots, a_{k2}$ which maximize the variance of Y_2, formally:
     - find $a_{12}^*, a_{22}^*, \cdots, a_{k2}^*$ such that
       1. $\max[Var(Y_2)] = Var(a_{12}^*X_1 + a_{22}^*X_2 + \cdots + a_{k2}^*X_k) = \\ = \sum_{j,r} a_{j2}^* a_{r2}^* s_{jr}$
       2. $\sum_j \left(a_{j2}^*\right)^2 = 1$
       3. $\sum_j a_{j1}^* a_{j2}^* = 0$
     - $\lambda_2 = \max[Var(Y_2)] = \sum_{j,r} a_{j2}^* a_{r2}^* s_{jr}$ *2nd* max eigenvalue of S
     - $(a_{12}^*, \cdots, a_{k2}^*)'$ eigenvector of S which corresponds to λ_2

- Following Principal Components: same iterative procedure…

Principal Component Analysis (PCA)

- Main differences between FA and PCA:

1. In FA we distinguish between common factors and unique factors while in PCA we have only common factors

2. In FA the communality is unknown and must be estimated while in PCA it is equal to 1

3. In FA the number of common factors is less than the number of observed original variables ($q<k$) while in PCA the number of components is equal to the number of observed orginale variables ($q=k$)

4. In FA the estimation of the communality follows an iterative method while PCA does not include iterations

R exercises

Problem 1 - Passito

- Perform a FA on the 17 response variables of the questionnaire which represent habits, behaviors and preferences of wine drinkers (from variable LIKE_WINE to variable PRICE) to detect new $q<17$ variables which «explain» data

- Perform a PCA on the 17 response variables of the questionnaire with the same goal

R exercises

Problem 2 - Mall

- Perform a FA on the 5 observed response variables to detect new $q<5$ variables which «explain» data
- Perform a PCA on the 5 response variables with the same goal

R exercises

Problem 3 – Eating Habits

- Perform a FA on the 12 observed response variables (from *Alcoholic.Beverages* to *Milk*) to detect new *q<12* variables which «explain» data
- Perform a PCA on the 12 response variables with the same goal