Photo: Vladimir Batagelj, ***UNI-LJ***

Factor Analysis

Anuška Ferligoj

UL, Ljubljana, Slovenia

NRU HSE, Moscow, Russia

Content

Introduction

Factor analysis studies the relationships among the measured variables in an effort to find new variables - factors, fewer in number than the original set of variables, which express that which is in **common** among the original variables.

Factor analysis attempts to simplify complex and diverse relationships that exist among a set of observed variables by uncovering common dimensions or factors that link together the measured variables, and consequently provides insight into the underlying structure of the data.

Applications

In the social sciences there are many crucial **concepts (constructs) that are not directly measurable** (e.g., social class, socio-economic development, satisfaction with work). Usually, such complex **concepts are measured indirectly by several well chosen indicators** (directly measured variables). Then we study if the interrelationships among these indicators can be explained by the assumed common variable - factor, which is in this case indirectly measured variable.

Sometimes the concept or construct is multidimensional. In this case more than one common variable or factor may account for the interrelationships among the measured variables or indicators.

Factor analysis can be used for such studies.

Goal

The goal of factor analysis is to find if the relationships among the measured variables (covariances or correlations) can be explained by a smaller number of variables - factors.

Some history

The early development of factor analysis was due to Spearman (1904). He studied correlations between test scores of various types and noted that many of the observed correlations could be accounted for by a simple model of scores.

For example for boys the correlations between their scores on tests on: Classics (X_1), French (X_2), and English (X_3) are

	X_1	X_2	X_3
X_1	1.		
X_2	0.83	1.	
X_3	0.78	0.67	1.

Spearman proposed the idea that each score has the form:

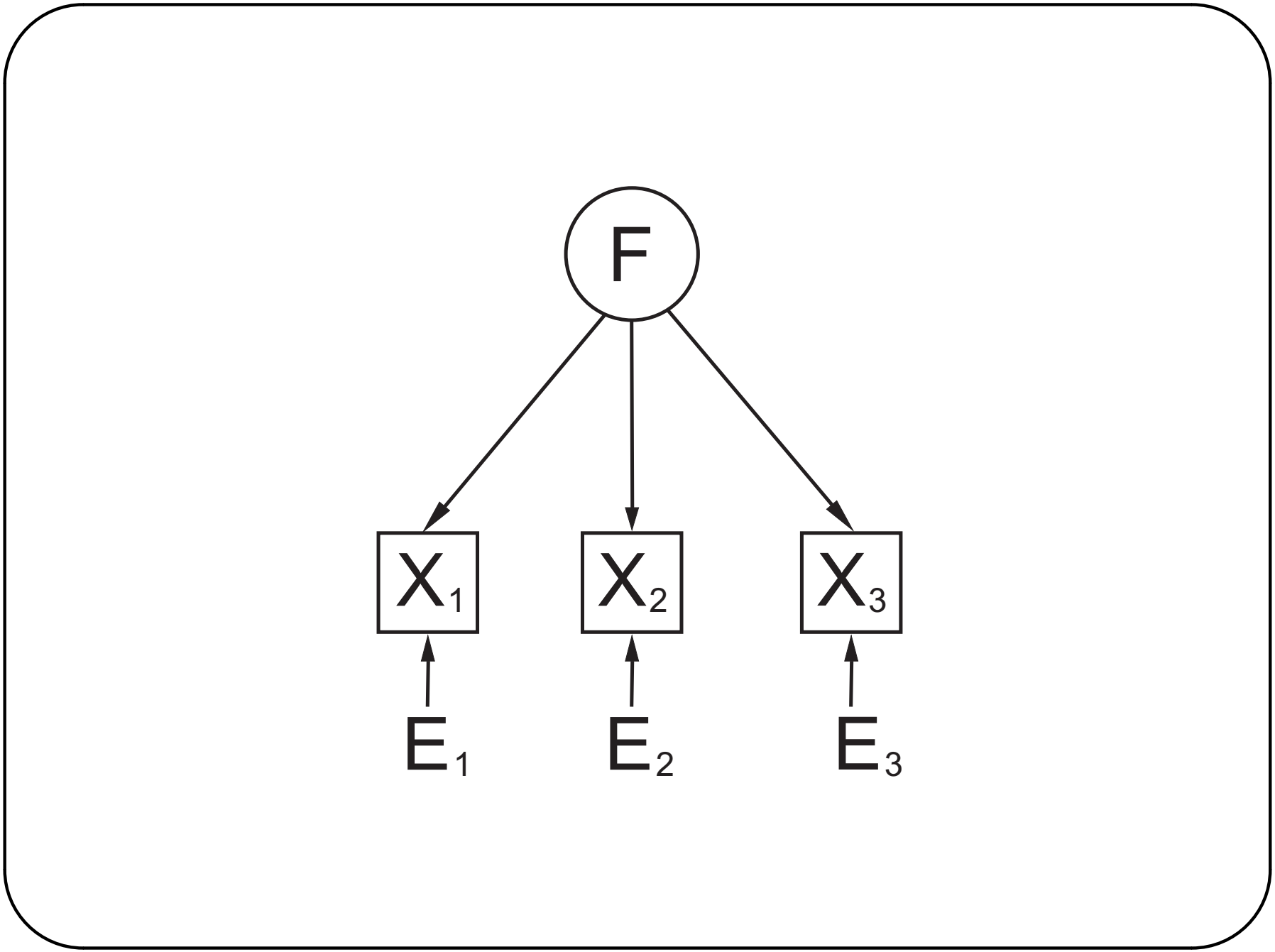
$$X_1 = \lambda_1 F + E_1$$

$$X_2 = \lambda_2 F + E_2$$

$$X_3 = \lambda_3 F + E_3$$

where

- X_i is the ith standardized score,
- λ_i is a constant (factor loading), and
- E_i is the part of X_i that is specific to the ith test only.
- Factor F means the common part of the test scores and measures the pupils' school performance.


F
X1
X2
X3
E1
E2
E3

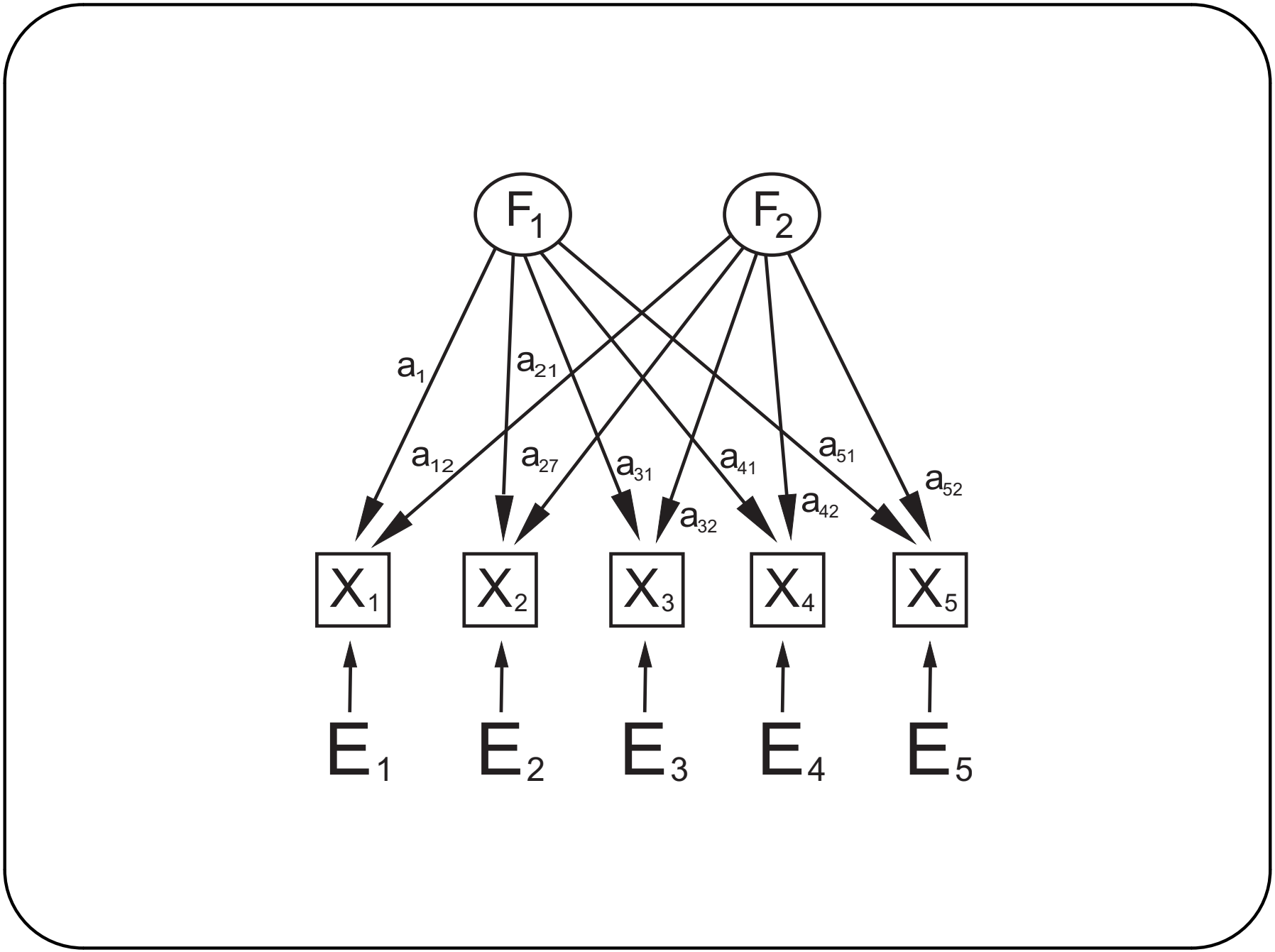
General factor analysis model

Let us have the following variables: X_i (i=1,...,m), F_r (r=1,...,k), and E_i (i=1,...,m). The general factor analysis model assumes that the relationship among the variables X_i, F_r in E_i is the following one:

$$X_i = \sum_{r=1}^{k} a_{ir} F_r + E_i, \qquad i = 1, ..., m$$

where $k < m$.

- X_i is a measured variable,
- F_r is unobserved variable or common factor,
- E_i is unobserved or unique factor, and
- a_{ir} is unknown constant called factor loading.


F1
F2
a1
a21
a12
a27
a31
a41
a51
a52
a32
a42
X1
X2
X3
X4
X5
E1
E2
E3
E4
E5

Let us express the general model in matrix form. The following matrices can be constructed:

Data matrix:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

Factor matrix:

$$F = \begin{bmatrix} f_{11} & f_{12} & \dots & f_{1k} \\ f_{21} & f_{22} & \dots & f_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \dots & f_{nk} \end{bmatrix}$$

Matrix with factor loadings:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mk} \end{bmatrix}$$

Unique factor matrix:

$$E = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1m} \\ e_{21} & e_{22} & \dots & e_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n1} & e_{n2} & \dots & e_{nm} \end{bmatrix}$$

Than **the factor model** can be expressed in matrix form:

$$X = FA' + E$$

Assumptions of factor model

As there are too many unknown parameters to be estimated the following assumptions have to be introduced:

1. Unique factors are uncorrelated with each other
   $\text{cov}(E_i, E_j) = 0$, if $i \neq j$;
2. Each unique factor E_i is uncorrelated with a common one F_j
   $\text{cov}(E_i, F_j) = 0$ for each i and j;
3. Common factors are uncorrelated with each other
   $\text{cov}(F_i, F_j) = 0$, if $i \neq j$;
4. Variables X_i, F_i and E_i are centered
   $E(X_i) = E(F_i) = E(E_i) = 0$.

Factor equation

The following **factor equation** can be deduced from the factor model considering the assumptions:

$$\Sigma = AA' + \Psi$$

In general Σ is a variance-covariance matrix. If the measured variables are standardized the matrix Σ is a correlation matrix. Ψ is a diagonal matrix with the variances of unique factors on the diagonal.

Communality

If we compare the diagonal elements on the left and right side of the factor equation $\Sigma = AA' + \Psi$ the following equality can be seen

$$\sigma_i^2 = \sum_{j=1}^{k} a_{ij}^2 + \psi_{ii}$$

which means that the variance of the measured variable X_i can be partitioned into the **common variance** and the **unique variance** of X_i.

The common variance of a variable is also called the **communality**. The communality of a variable is the proportion of a variable's total variance that is accounted for by the common factors. If we denote the communality of the i-th variable by h_i^2 , we can write

$$\sigma_i^2 = h_i^2 + \psi_{ii}$$

Note that

$$h_i^2 = \sum_{j=1}^{k} a_{ij}^2$$

is simply the sum of the squared elements in the i-th row of the matrix A.

The unique variance of a variable, ψ_{ii} is called the **uniqueness** of the variable and reflects the extent to which the common factors fail to account for the variance of a variable – it is the proportion left unexplained by the common factors.

Parameter estimaton

From the known elements of the variance-covariance (correlation) matrix Σ the unknown parameters in the factor equation, the factor loadings A and the unique variances Ψ, have to be estimated.

Before the parameter estimation we have to ask ourselves two questions. Does a solution of the factor equation exist, and if it does, is it unique? This means that we have to consider two issues:

- **identificability** of the factor model and
- **uniqueness** of parameter estimates.

Identificability

The total number of parameters to be estimated is the number of factor loadings, namely $m \times k$, and m unique variances. There are $\frac{m(m+1)}{2}$ different variances and covariances in Σ. Hence, we can determine $\frac{m(m+1)}{2}$ equations. Generally, the necessary requirement for indentification is that the number of parameters be less or equal than the number of equations, therefore

$$mk + m \leq m(m+1)/2$$

or

$$k \leq \frac{(m-1)}{2}$$

Unfortuately, this does not guarantee that a solution will exist.

Uniqueness

The second question is, can we uniquelly estimate the unknown parameters A and Ψ from the given Σ satisying the equation

$$\Sigma = AA' + \Psi$$

If $k > 1$ and if it exists a unique matrix Ψ, than there are infinite number of matrices A that satisfy the factor equation. Let us demonstrate why: Let $M_{k\times k}$ be an orthonormal matrix ($MM' = I$) and

$$A^* = AM$$

Than

$$A^*A^{*\prime} = (AM)(AM)' = AMM'A' = AA' = \Sigma - \Psi$$

This means that also A^* is a solution of the factor equation. There are infinite number of solutions for the factor equation.

Therefore, some new assumptions have to be defined for a unique solution.

Estimation strategy

The strategy to find a unique estimation of a factor model is the following one:

1. the estimation of the communalities (common space) by a **factor method**,
2. the estimation of a simple structure of factor loadings by **factor rotation**.

Factor analysis is not completed if factor rotation was not performed.

Factor methods

There are several factor extraction procedures or factor methods for estimating the communalities. Some of them are:

- principal axis factor (PAF)
- maximum likelihood (ML)
- image factor analysis
- alpha factor analysis, ...

The principal axis factor method is presented in more detail.

Principal axis factor method (PAF)

The factor equation

$$\Sigma = AA' + \Psi$$

can be written as follows

$$\Sigma - \Psi = AA'$$

Let us assume that all variables are standardized. The left side of the equation is the correlation matrix with the diagonal elements replaced by the respective variables' communality estimates.

In general the communalities can be determined if the matrix A is known. Matrix A can be determined from the matrix $\Sigma - \Psi$.

Principal axis factor method (PAF) solves the problem iteratively.

- First, the diagonal elements in the correlation matrix are replaced by the respective variables' communality estimates (e.g., by the largest correlation coefficient in the raw of the correlation matrix or by multiple correlation coefficient of a variable by the other variables).
- Repeat
  - We can obtain the matrix A by calculating the eigenvalues and eigenvectors of the corrected correlation matrix. The estimate of the matrix A can be obtained by ordering the eigenvalues from the largest to the smallest and place the eigenvectors on this order into the matrix A.
  - Then new communalities can be calculated from the estimated matrix A by the sum of the squared elements in each row of A. The diagonal elements in the correlation matrix are replaced by these new communalities.

The convergence of this procedure is not proved yet, but empiricaly works very well.

By estimating the matrix Ψ we obtain the common factor space.

The matrix A is obtained in such a way that the variance of the first common factor is maximal. Orthogonal to the first factor, the second factor with the maximal variance is obtained, etc. Such a matrix A is one among many and most of the times it does not find the common dimensions. Therefore, we do not consider this solution for the interpretation.

Only after factor rotation we can obtain an adequate matrix A.

Factor rotation

When solving the factor equation

$$\Sigma = AA' + \Psi$$

where Σ is known and the matrices A in Ψ are unknown, we realized that it is not possible to estimate the matrix A uniquely. We mentioned that by multiplying the matrix A by an orthonormal matrix M a new matrix A^* is obtaned which also satifies the factor equation. By matrix M we rotate the axes.

Using one of the factor extraction procedures (e.g., PAF) we estimate communalities and with them the common factor space. We also obtain an estimate of factor loadings. The rotated matrix represents an alternative interpretation of the data, which is in mathematical sense equally valid.

Example

Let us analyze 12 factors that influence the business success of small companies in Slovenia (J. Prašnikar, 1994).

The population consists of small companies employing at least 1 and up to 50 employees in all sectors of the economy except agriculture in Slovenia. A random sample was drown from the data-bank of the Chamber of Commerce of Slovenia and from the Crafts Chamber. Out of the 200 companies selected, 151 agreed to participate in the study. Each company was visited and a quationnaire was filed by personal interview (CAPI). The survey was conducted in 1993.

Variables

Survey question:

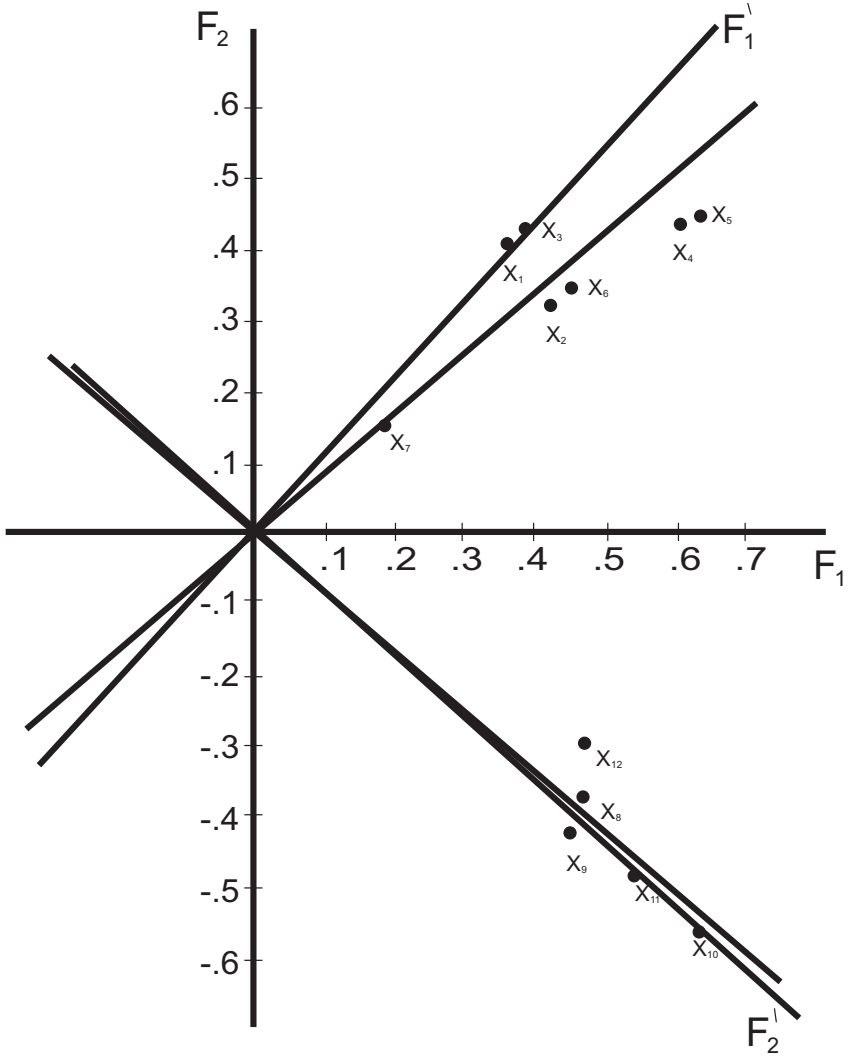
Please indicate your opinion about the influence of the following factors on business success of your company on 5 point scale (1 - not important at all, 5 - very important)

X_1 – PROD-MET	improvement of productive methods
X_2 – MARK-MET	improvement of marketing methods
X_3 – PRODUCT	improvement of products
X_4 – RELATION	good relations among employees
X_5 – SKIL-EMP	skilled employees
X_6 – SKIL-MAN	skilled managers
X_7 – FAMILY	support of the family
X_8 – ECON-ASO	support of economic associations
X_9 – POL-CON	political connections
X_{10} – LOC-AUT	support of local authorities
X_{11} – STATE	support of the state
X_{12} – COMPANY	support of other companies

Two factors were obtained by PAF:

	F_1	F_2	h^2
X_1 – PROD-MET	.38	.41	.31
X_2 – MARK-MET	.42	.32	.28
X_3 – PRODUCT	.39	.42	.33
X_4 – RELATION	.60	.43	.54
X_5 – SKIL-EMP	.63	.44	.58
X_6 – SKIL-MAN	.46	.33	.32
X_7 – FAMILY	.20	.15	.06
X_8 – ECON-ASO	.47	-.39	.37
X_9 – POL-CON	.46	-.44	.41
X_{10} – LOC-AUT	.63	-.59	.75
X_{11} – STATE	.54	-.50	.55
X_{12} – COMPANY	.48	-.30	.32
eigenvalue	2.82	2.00	
% p.v.	23.5	16.7	40.1

Let us present the obtained results in a two dimensional coordinate system, where the axes are the obtained factors and the points are variables.


F2
F1'
.6
.5
.4
.3
.2
.1
-.1
-.2
-.3
-.4
-.5
-.6
.1
.2
.3
.4
.5
.6
.7
F1
X3
X1
X5
X4
X6
X2
X7
X12
X8
X9
X11
X10
F2'

From the graphical presentation it can be seen that there are two distinctive groups of variables in the two dimensional space:

- variables from X_1 to X_6 and
- variables from X_8 to X_{12}.

Only the variable X_7 (the support from the family) is not a member of any of these two groups.

If the coordinate system is rotated in such a way that the axes are as close as possible to each of the mentioned two groups of variables new factor loadings can be obtained by projections of variables on the new coordinates.

The rotated factor loading matrix is the following one:

	F_1	F_2
X_1 – PROD-MET	**.56**	-.03
X_2 – MARK-MET	**.52**	.06
X_3 – PRODUCT	**.57**	-.03
X_4 – RELATION	**.73**	.11
X_5 – SKIL-EMP	**.75**	.13
X_6 – SKIL-MAN	**.56**	.09
X_7 – FAMILY	.25	.03
X_8 – ECON-ASO	.06	**.60**
X_9 – POL-CON	.02	**.64**
X_{10} – LOC-AUT	.04	**.86**
X_{11} – STATE	.03	**.74**
X_{12} – COMPANY	.13	**.55**

The obtained factor structure is much simpler for the interpretation:

1. **The first factor** has very large and positive loadings on all variables which measure the factors which have to be done **inside** the company to obtain its business success.
2. **The second factor** has large loadings on the rest of variables measuring all kind of **outside** supports to obtain the company business success.

Thurston's criteria

Thurston (1947) developed the criteria of 'simple structure' as a guide for rotation. The three major criteria are the following ones:

1. Any column of the factor loadings matrix should have several small values, as close to zero as possible.
2. Any row of the matrix should have only a few entries far from zero.
3. Any two columns of the matrix should exhibit a different pattern of high and low loadings.

Most of the rotation procedures use Thurston's criteria to construct appropriate criterion functions which are optimized to obtain simple structures of factor loadings.

Rotation methods

There are two methods in which the factor axes can be rotated:

- **orthogonal rotation** (the rotated factors are perpendicular)
- **oblique rotation** (the rotated factors are not perpendicular)

Orthogonal rotation

There are at least three orthogonal rotation methods:

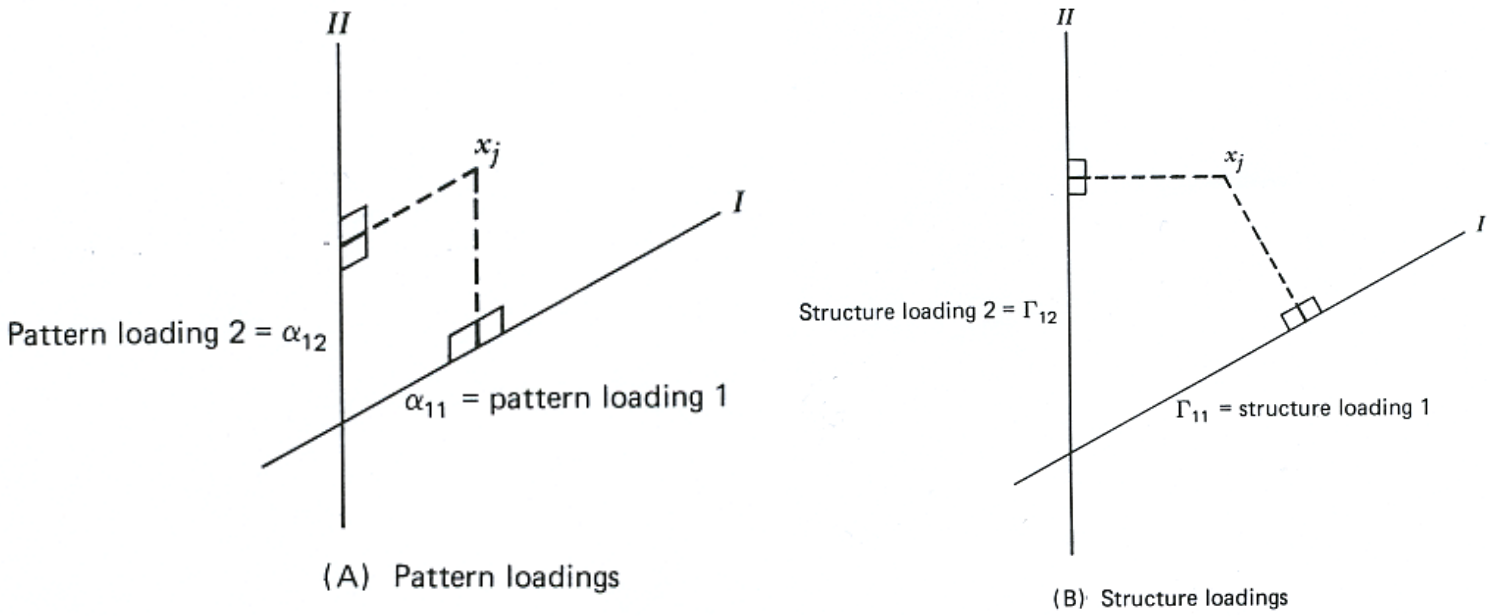
- **QUARTIMAX** simplifies the structure by rows in factor loading matrix. The consequence is that a general factor is usually generated with all or most of the variables having high loadings.
- **VARIMAX** simplifies the structure by columns.
- **EQUIMAX** simplifies the structure by rows and columns.

Oblique rotation

There are several oblique rotation methods: OBLIMIN, OBLIMAX, QUARTIMIN, COVARMIN in BIQUARTIMIN.

In the case of oblique rotation the projection of each variable on an axis which represents the loading of a variable on that factor can be done at least on two ways:

- parallel projection, with which **'pattern' loadings** are obtained. These loadings are the regression coefficients between variables and factors.
- orthogonal projection, with which **structure loadings** are obtained. These loadings are the correlations coefficients between variables and factors.



In the case of orthogonal rotation 'pattern' in 'structure' loadings are the same.

Factor scores

- Estimated are: A and $\Psi(h_i^2)$
- Not yet estimated factor scores: $F_{n\times k} = [f_{ij}]$
  f_{ij} is the value of the jth factor on ith unit

F **is not** a linear combination X_i

Regression estimates of factor scores

$$\hat{F} = XB$$

B has to be obtained. Let us assume that the variables are standardized. Than the elements of the vector B are standardized regression coefficients. The following regression estimates of the factor scores can be obtained by some linear algebra:

$$\hat{F} = X\Sigma^{-1}A$$

These are only regression estimates and not the real values of factor scores. This, for example, means that the correlation between orthogonal factors can be slightly different from 0.

Example

Population and sample

The data used in the example were gathered as a part of the research **Quality of measurement of egocentered social networks** by Ferligoj et al.

The target population of the research were the inhabitants of Ljubljana that were at least 18 years old at the time of the research. The sample consisted of 1033 randomly selected individuals. The data analyzed consist of 631 individuals and were gathered using computed aided personal interviewing (CAPI) between March and June 2000 (the others were interviewed by telepnone - CATI).

Variables

Factor analysis was performed on variables measuring emotional stability and extraversion (personal(lity) characteristics from the Big Five, International Personality Item Pool:

EMOCC – Seldom feel blue.
EMOCDR – Get upset easily. (*)
EMOCF – Am relaxed most of the time.
EMOCGR – Get irritated easily. (*)
EMOCJR – Am easily disturbed. (*)
EMOCKR – Worry about things. (*)
EMOCMR – Have frequent mood swings. (*)
EMOCQR – Change my mood a lot. (*)
EMOCSR – Often feel blue. (*)
EMOCTR – Get stressed out easily. (*)

EXTA – Am the life of the party.
EXTB – Don't mind being the center of attention.
EXTE – Talk to a lot of different people at parties.
EXTH – Start conversations.
EXTIR – Don't like to draw attention to myself. (*)
EXTLR – Don't talk a lot. (*)
EXTNR – Am quiet around strangers. (*)
EXTOR – Keep in the background. (*)
EXTP – Feel comfortable around people.
EXTRR – Keep in the background. (*)

The respondents expressed how accurate description for them each statements is on a 5 item ordinal scale: from 1 (very inaccurate) to 5 (very accurate).

The statements marked with (*) were negative statements and were recoded: (1=5) (2=4) (3=3) (4=2) (5=1).

The variables are measuring two dimensions of the personality characteristics. The first 10 variables are measuring emotional stability and the other 10 variables extraversion.

Therefore, we expect the following factor solution:

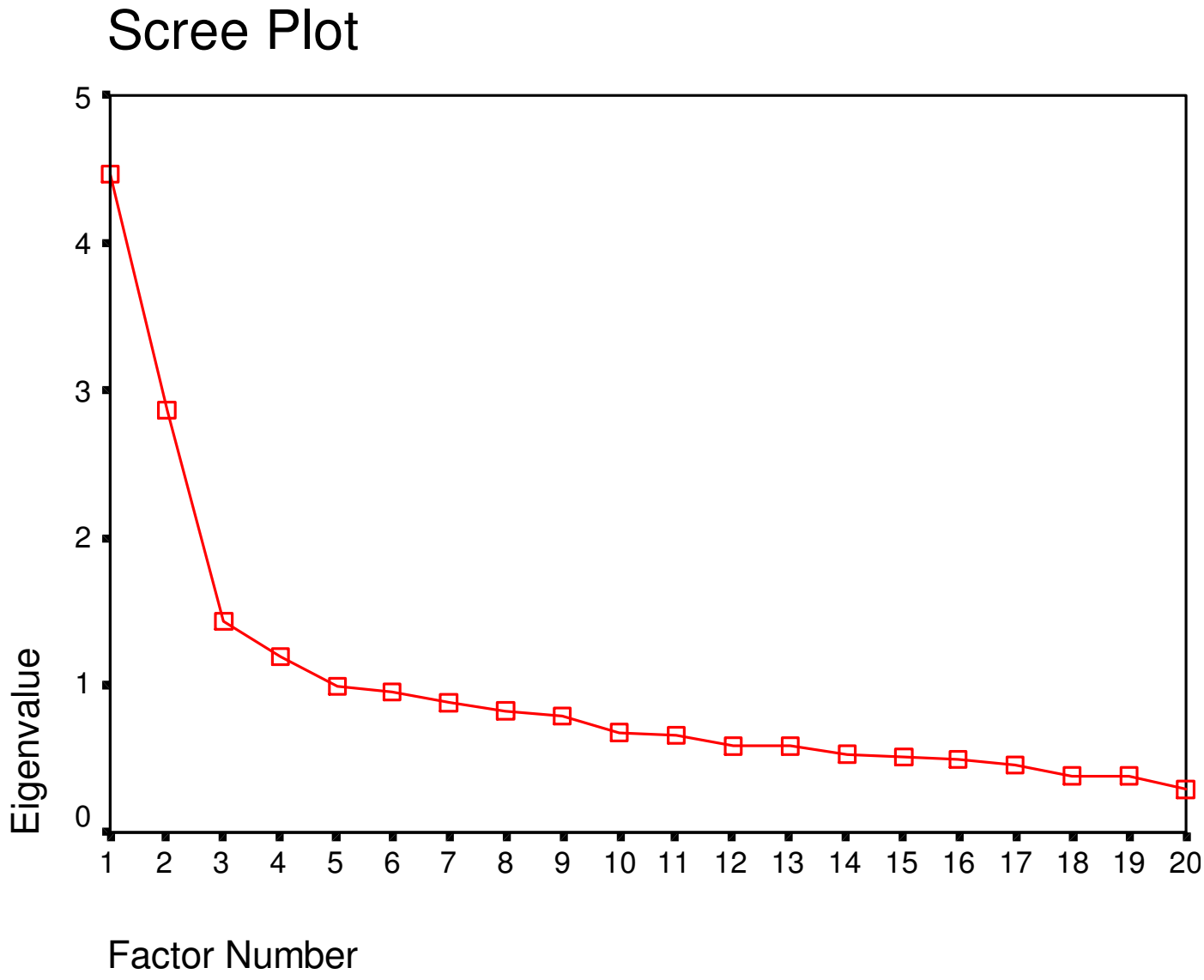
- one factor should have large factor loadings on the variables of emotional stability and very low ones on extraversion (**the factor of emotional stability**)
- the other factor should have have large factor loadings on the variables of extraversion and very low ones on emotional stability (**the factor of extraversion**)

We test the validity of the measurement instrument for two dimensions of the personality characteristics.

First we perform the first step of factor analysis: the estimation of the common space. We do this by principal axis factor method.

We first look to the scree diagram to see if the data really show 2 dimensions.

Then we look to the percentage of the common variance obtained by the two factors and the estimated communalities.


Scree Plot
Eigenvalue
0
1
2
3
4
5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Factor Number

Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,460	22,301	22,301	3,820	19,100	19,100
2	2,867	14,337	36,638	2,219	11,097	30,197
3	1,436	7,180	43,818			
4	1,188	5,939	49,757			
5	,992	4,960	54,717			
6	,954	4,769	59,486			
7	,884	4,418	63,904			
8	,831	4,157	68,061			
9	,794	3,970	72,031			
10	,674	3,371	75,402			
11	,656	3,282	78,684			
12	,595	2,974	81,658			
13	,584	2,919	84,577			
14	,538	2,692	87,269			
15	,516	2,582	89,851			
16	,493	2,467	92,318			
17	,458	2,291	94,609			
18	,392	1,958	96,567			
19	,385	1,924	98,491			
20	,302	1,509	100,000			

Extraction Method: Principal Axis Factoring.

Communalities

	Initial	Extraction
EKSTA Dusa vsake družbe	,215	,186
EKSTB Ne moti - sredisce pozornosti	,164	,128
EMOCC Redkokdaj potrt	,178	,116
EMOCDR Zlahka vrze iz tira	,478	,491
EKSTE Na zabavah se pomenkujem z mnogo ljudmi	,251	,267
EMOCF Vecidel sproscen	,369	,264
EMOCGR Zlahka me kaj razdrazi	,519	,476
EKSTH Pogovore nacenjam jaz	,199	,198
EKSTIR Nerad pozornost nase	,187	6,478E-02
EMOCJR Zlahka me kaj vznemiri	,537	,543
EMOCKR Sem zaskrbljene narave	,283	,250
EKSTLR Sem redkobeseden	,361	,416
EMOCMR Velikokrat muhasto razpolozen	,217	,161
EKSTNR Neznane osebe - sem molcec	,354	,378
EKSTOR Imam malo povedati	,348	,348
EKSTP Med ljudmi pocutim sprosceno	,306	,192
EMOCQR Moje razpolozenje pogosto menja	,386	,363
EKSTRR Zadrzujem se v ozadju	,439	,433
EMOCSR Pogosto sem potrt	,381	,353
EMOCTR Zlahka se me poloti napetost	,379	,412

Extraction Method: Principal Axis Factoring.

The scree diagram confirms two dimensions. The results obtained by PAF show that the two factors explain 30.2 % of total variance (common variance).

It is recommended that the communalities are greater than 0.20.

13 variables satisfiy this recommentation. 6 variables have still satisfying communalities (greater than 0.10).

The communality of the variable EKSTIR (Don't like to draw attention to myself) has too low communality. This means that this indicator is not enough related to the other indicators measuring the extraversion.

First, let us look to the oblique rotation, especially to the correlation coefficient between two obtained factors:

Factor Correlation Matrix

Factor	1	2
1	1,000	,220
2	,220	1,000

Extraction Method: Principal Axis Factoring.
Rotation Method: Oblimin with Kaiser Normalization.

As the correlation coefficient is relatively small we perform orthogonal rotation.

	Factor	
	1	2
EKSTA Dusa vsake družbe	-,038	**,430**
EKSTB Ne moti - sredisce pozornosti	-,008	**,358**
EMOCC Redkokdaj potrt	**,326**	,097
EMOCDR Zlahka vrze iz tira	**,698**	,054
EKSTE Na zabavah se pomenkujem z mnogo ljudmi	,058	**,514**
EMOCF Vecidel sproscen	,274	**,435**
EMOCGR Zlahka me kaj razdrazi	**,688**	-,045
EKSTH Pogovore nacenjam jaz	-,043	**,443**
EKSTIR Nerad pozornost nase	,052	,249
EMOCJR Zlahka me kaj vznemiri	**,737**	-,016
EMOCKR Sem zaskrbljene narave	**,444**	,228
EKSTLR Sem redkobeseden	-,004	**,645**
EMOCMR Velikokrat muhasto razpolozen	**,401**	-,011
EKSTNR Neznane osebe - sem molcec	,169	**,591**
EKSTOR Imam malo povedati	,096	**,582**
EKSTP Med ljudmi pocutim sprosceno	,122	**,421**
EMOCQR Moje razpolozenje pogosto menja	**,601**	,043
EKSTRR Zadrzujem se v ozadju	,195	**,628**
EMOCSR Pogosto sem potrt	**,561**	,195
EMOCTR Zlahka se me poloti napetost	**,634**	,104

Extraction Method: Principal Axis Factoring.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Two measured variables (indicators) are problematic:

- EMOCF (Am relaxed most of the time): the indicator should measure emotional stability but has a large factor loading on the factor of extraversion. This is a major error of the measurement instrument used.
- EKSTIR (Don't like to draw attention to myself): there is no factor loading large enough. This is expected as its communality was very low.

We have shown that the measuremet instrument is not good. The reason might be in the translation from English to Slovenian.

Let us save factor scores obtained by the regression estimation (PAF and VARIMAX rotation). Let us calculate averages for both factors for each gender (males and females) and draw the obtained centroids in the coordinate system defined by both factors: