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

Principal Component Analysis

Anuška Ferligoj

UL, Ljubljana, Slovenia

NRU HSE, Moscow, Russia

Content

Introduction

Principal component analysis (PCA) is the most popular multivariate method.

- PCA was first described by **Karl Pearson** (1901).
- **Hotelling** (1933) further developed its mathematical formulation.

The main object of PCA is to describe the variability of n units in m–dimensional space (determined by m measured variables) by a set of noncorrelated variables - components that are linear combinations of the measured variables.

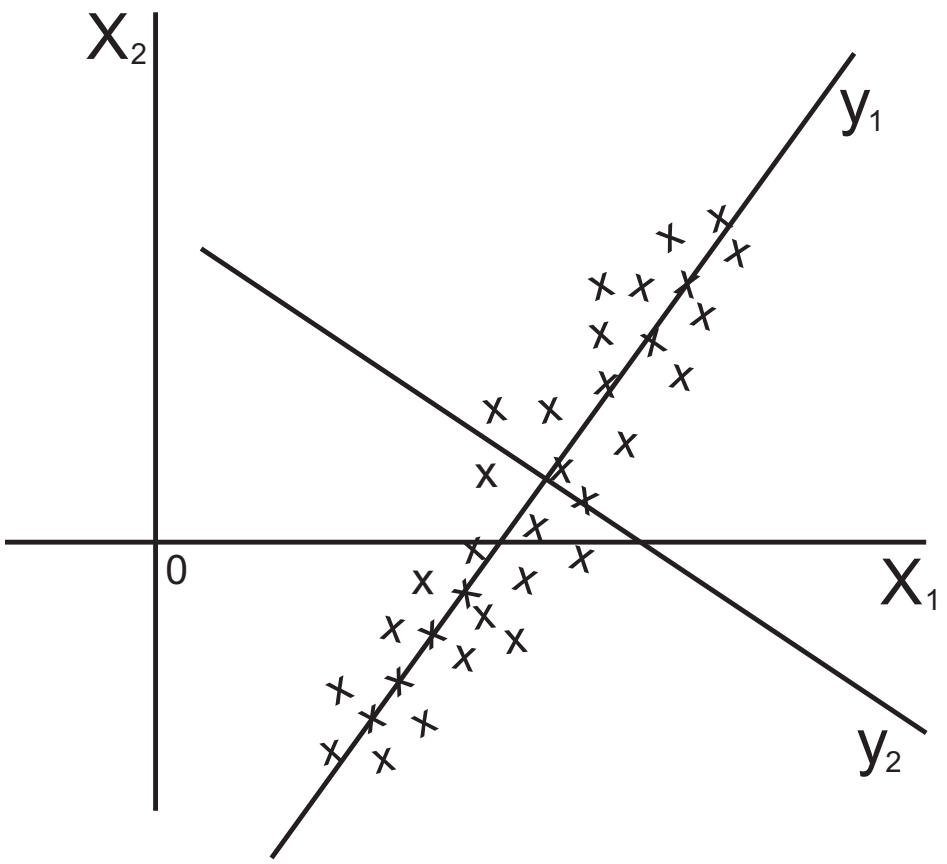
The components are ordered from the most important one to the least important one. Here, the importance means that the componet explains as large as possible amount of the variation of the measured variables.

Goal

The usual goal of the PCA is to determine a few first components that explain the largest possible amount of the variation of the measured variables. PCA enables to describe the data in few dimensions with the lowest possible loss of information.

An example

Usualy we measure socio–economic development of communes or countries by several socio—economic variables (indicators). From these variables we can determine an index of socio–economic development by first component obtained by PCA.


X2
y1
0
X1
y2

Problem

The goal of PCA is to determine such a linear combination of the measured variables that it explaines the largest possible amount of the variation of the measured variables. Therefore, the coefficients (loadings) of the linear combination are determined by the maximal variance of this linear combination – principal component. If the first linear combination of the measured variables $X_1, X_2, \ldots, X_m$ is

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1m}X_m$$

we are searching for such loadings a_{1j} that satisfy

$$\text{var}(Y_1) = max$$

Matrix formulation

$$Y_1 = Xa_1$$

X is the data matrix

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

and a_1 is the vector of the loadings

$$a_1 = \begin{bmatrix} a_{11} \\ \vdots \\ a_{1m} \end{bmatrix}$$

We are searching for the loadings a_1 for which the variance of Y_1 is as large as possible:

$$\text{var}(Y_1) = \text{var}(Xa_1) = max$$

This linear combination Y_1 is the first principal component. The condition for a unique solution is:

$$\sum_{i=1}^{m} a_{1i}^2 = a_1' a_1 = 1$$

After we obtain the first princpal component we search for the second one which is not correlated with the first one and has again maximal variance:

$$Y_2 = a_{21}X_1 + a_{22}X_2 + ... + a_{2m}X_m = Xa_2$$

$$a_2'a_2 = 1$$

and

$$a_2'a_1 = 0$$

The procedure is repeated and the jth component is

$$Y_j = Xa_j$$

with the condition:

$$a_j'a_j = 1$$

and

$$a_j'a_i = 0, \qquad i < j$$

PCA procedure

The problem that has to be solved is

$$\text{var}(Y_1) = \text{var}(Xa_1) = a_1'\Sigma a_1 = max$$

with the condition $a_1'a_1 = 1$. This is maximization with constraint problem that is usually solved by the Lagrange multiplier method. First we determine the Lagrange function

$$t = a'\Sigma a - \lambda(a'a - 1)$$

– Σ is variance–covariance matrix or correlation matrix if the variables are standardized;
– $a'\Sigma a$ is the variance of the linear combination that has to be maximized;
– λ is an unknown constant, known as Langrange multiplier, and
– $(a'a - 1)$ represents the constraint.

λ and a, for which the function t has the largest value are the solutions of the given principle component problem.

Eigenvalues and eigenvectors of the matrix Σ

Solving the optimization problem brings us to the calculation of the eigenvalues and eigenvector of the matrix Σ.

It can be shown that the **eigenvalues λ_i are the variances of the principle components** and the corresponding **eigenvectors are the component loadings**.

The eigenvector a_1 which corresponds to the largest eigenvalue λ_1 determines the first principal component. Next eigenvector that correspond to the next largest eigenvalue gives loadings of the second principle component, etc.

The obtained PCA solutions are meaningfull if the variances λ_i of the principle components are positive.

This is the case when the matrix Σ is positive definite.

This is true if the rank of the data matrix X is equal to the number of the measured variables (m).

This means that no one measured variable is a linear combination of the other measured variables. Also at least m rows have to be linearly independent.

Therefore, the number of units has to be much larger than the number of variables.

It can be easily proved that

$$|\Sigma| = \prod_{i=1}^{m} \lambda_i$$

and

$$\sum_{i=1}^{m} \sigma_{ii} = \sum_{i=1}^{m} \lambda_i$$

This means that the sum of the diagonal element of the matrix Σ (the sum of the variances of the measured variables) is equal to the sum of eigenvalues of the matrix Σ (the sum of the variances of the principal components).

The variability is preserved.

The proportion of the total variance due to jth principal component is

$$\frac{\lambda_j}{\sum_{i=1}^{m} \sigma_{ii}}$$

In most of the cases the measured variables are measured by different measurement scales. Therefore, the variables are standardized before calculating the principle components. In this case the variance–covariance matrix is the correlation matrix and the proportion of the total variance due to jth principal component is

$$\lambda_j / m$$

Usually the obtained principal components are rescaled and the rescaled loadings are

$$a_{ij}^* = a_{ij}\sqrt{\lambda_i}$$

The length of the ith component a_i is equal 1, the length of the rescaled component a_i^* is equal to λ_i.

The loading of the ith rescaled component a_{ij}^* is the **correlation** between ith component and jth variable.

When PCA?

PCA is meaningful if the measured variables correlate with each other. If they do not, the obtained principal components are measured variables. Therefore, it is wise to test the following hypothesis before using PCA:

$$H_0 : \Sigma = I$$

which means that the measured variables do not correlate.

Bartlett (1947) developed test statistics (with the assumption that the measured variables are multivariate normal distributed) for the above hypothesis:

$$\chi^2 = -(n - 1 - \frac{(2m + 5)}{6}) \ln |R|$$

which distribution is approximately χ^2 with $\frac{m(m-1)}{2}$ dgrees of freedom. R is the sample correlation matrix.

Bartlett's procedure was futher developed to test the hypothesis that the $(m - k)$ "smaller" eigenvalues of Σ are equal to 0.

How many principal components to retain?

In the literature there are many heuristical rules to determine the number of the important principal components. Here are some:

1. the selected principal component have to explain at least 80 % of the total variance;
2. the eigenvalues should be greater than the average value of all eigenvalues;
3. the percentage of the explained variance of the last selected component should be greater than 5;
4. the number of the components can be determined by graphical representation of the eigenvalues – scree diagram (This will be explained later)

Example 1

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. (2000).

The target population 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 that were gathered using computed aided personal interviewing (CAPI) between March and June 2000 (the others were interviewed by telephnone - CATI).

Variables

PCA was performed on variables measuring emotional stability (personal(lity) characteristics from the Big Five):

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. (*)

The respondents expressed how accurate is the description for them on a 5 point 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).

Here is the correlation matrix:

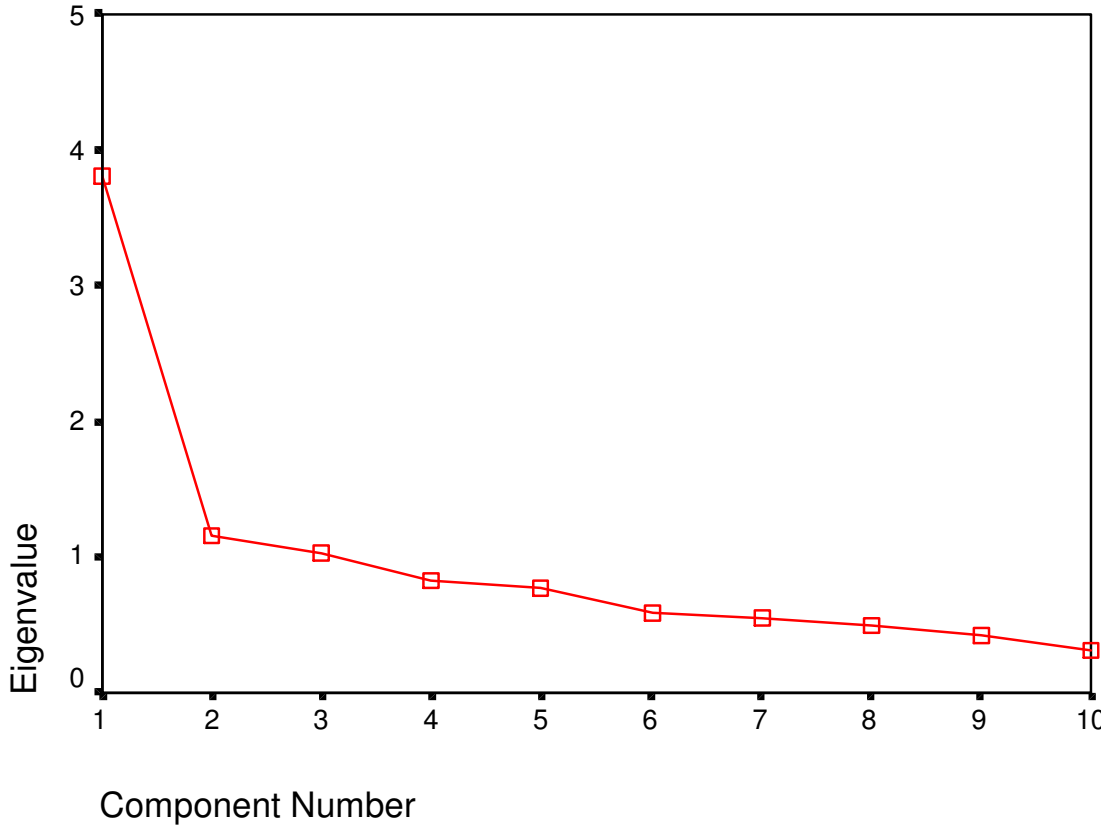
	EMOCC Redkokdaj potrt	EMOCDR Zlahka vrze iz tira	EMOCF Vecidel sproscen	EMOCGR Zlahka me kaj razdrazi	EMOCJR Zlahka me kaj vznemiri	EMOCKR Sem zaskrbljene narave	EMOCMR Velikokrat muhasto razpolozen	EMOCQR Moje razpolozenje pogosto menja	EMOCSR Pogosto sem potrt	EMOCTR Zlahka se me poloti napetost
EMOCC Redkokdaj p	1,000	,221	,209	,171	,166	,201	,163	,239	,348	,169
EMOCDR Zlahka vrze tira	,221	1,000	,212	,582	,585	,295	,214	,366	,334	,429
EMOCF Vecidel sproscen	,209	,212	1,000	,154	,188	,234	,088	,190	,275	,217
EMOCGR Zlahka me razdrazi	,171	,582	,154	1,000	,650	,266	,273	,359	,263	,379
EMOCJR Zlahka me k vznemiri	,166	,585	,188	,650	1,000	,309	,223	,338	,377	,457
EMOCKR Sem zaskrbljene narave	,201	,295	,234	,266	,309	1,000	,095	,248	,361	,401
EMOCMR Velikokrat muhasto razpolozen	,163	,214	,088	,273	,223	,095	1,000	,406	,281	,267
EMOCQR Moje razpolozenje pogosto menja	,239	,366	,190	,359	,338	,248	,406	1,000	,432	,440
EMOCSR Pogosto se potrt	,348	,334	,275	,263	,377	,361	,281	,432	1,000	,385
EMOCTR Zlahka se m poloti napetost	,169	,429	,217	,379	,457	,401	,267	,440	,385	1,000

The obtained eigenvalues:

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3,812	38,121	38,121
2	1,166	11,664	49,785
3	1,031	10,309	60,094
4	,831	8,313	68,407
5	,773	7,733	76,140
6	,591	5,915	82,055
7	,556	5,560	87,615
8	,497	4,973	92,588
9	,425	4,252	96,840
10	,316	3,160	100,000

The first component explaines 38 % of the total variance.

”Scree” diagram:



Scree diagram shows that there is one important dimension in the 10–dimensional space. This means that the 10 indicators of emotional stability really measure only one dimension of ”emotional stability”

The obtained component loadings are:

	Component									
	1	2	3	4	5	6	7	8	9	10
EMOCC Redkoko	,421	,528	,003	,584	-,361	,089	,236	-,068	,066	,032
EMOCDR Zlahka tira	,730	-,325	-,140	,205	,015	-,062	,067	,052	-,537	,016
EMOCF Vecidel sproscen	,398	,440	-,382	,085	,700	,050	,032	,008	,032	,003
EMOCGR Zlahka razdrazi	,709	-,466	-,037	,227	,041	,139	-,024	,136	,242	-,359
EMOCJR Zlahka vznemiri	,747	-,399	-,144	,123	,000	-,032	-,177	-,122	,231	,383
EMOCKR Sem zaskrbljene narav	,548	,201	-,406	-,436	-,301	,401	-,035	,224	-,024	,040
EMOCMR Velikol muhasto razpoloz	,467	,078	,719	-,051	,189	,432	-,092	-,131	-,084	,042
EMOCQR Moje razpolozenje pogo menja	,664	,135	,398	-,159	,049	-,322	,203	,441	,084	,086
EMOCSR Pogost potrt	,659	,384	,045	-,059	-,144	-,292	-,523	-,112	-,045	-,139
EMOCTR Zlahka poloti napetost	,701	-,030	-,030	-,392	-,041	-,145	,371	-,427	,047	-,093

Example 2

Data

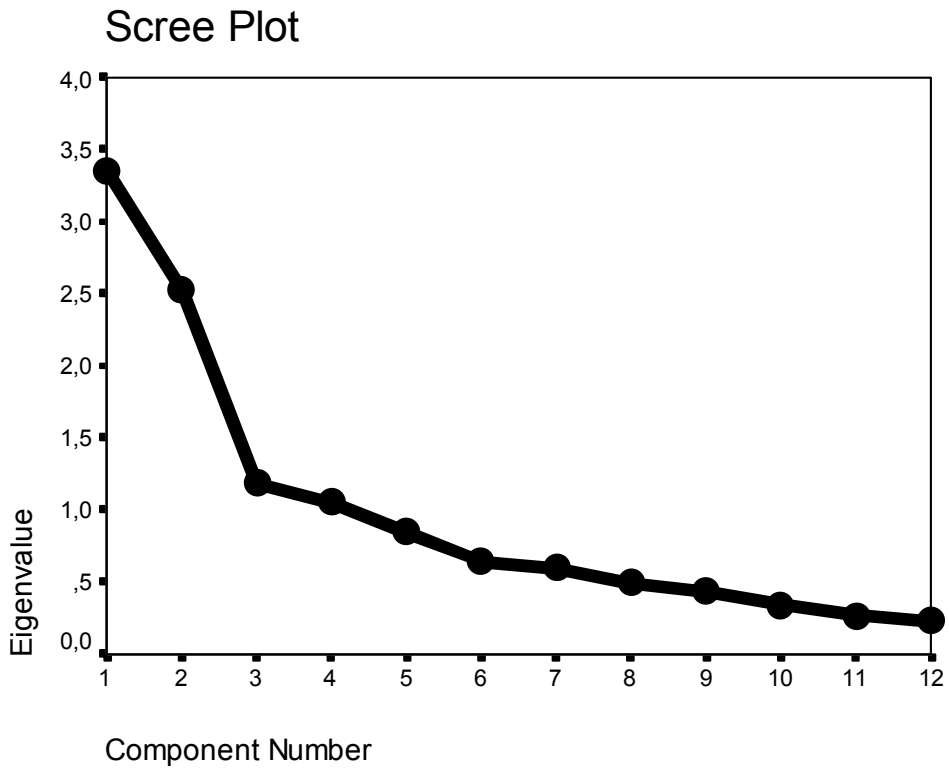
The population consists of small companies employing at least 1 and up to 50 employees in all sectors of the economy except agriculture in Slovenia. A random sample was drown from the data-bank of the Chamber of Commerce of Slovenia and from the Crafts Chamber. Out of the 200 companies selected, 151 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.

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)?

PROD-MET	– improvement of productive methods
MARK-MET	– improvement of marketing methods
PRODUCT	– improvement of products
RELATION	– good relations among employees
SKIL-EMP	– skilled employees
SKIL-MAN	– skilled managers
FAMILY	– support of the family
ECON-ASO	– support of economic associations
POL-CON	– political connections
LOC-AUT	– support of local authorities
STATE	– support of the state
COMPANY	– support of other companies

Let us first see how many dimensions are hidden in the 12–dimensional space:



There are 2 dimensions.

Here are the PCA results for the first two components:

	F_1	F_2
PROD-MET	.46	.47
MARK-MET	.51	.36
PRODUCT	.48	.47
RELATION	.65	.42
SKIL-EMP	.67	.41
SKIL-MAN	.55	.36
FAMILY	.27	.19
ECON-ASO	.50	-.49
POL-CON	.49	-.55
LOC-AUT	.59	-.63
STATE	.54	-.59
COMPANY	.53	-.41
eigenvalue	3.38	2.53
% p.v.	28.0	21.1

The first component is a general one, the second is a bipolar one.

Larger values on the first component have those the owners that meant that all factors are important. The high value on the second component have those companies owners that agreed that everthing inside the company has to be improved and did not agree that the outside support is important for the company success.

Let us remember the obtained clusters by hierarchical clustering obtained by the same variables:

cluster	number of cases	cluster description
1	62	AVERAGE
2	32	YES - SAYERS
3	18	BAD GUYS
4	24	NO - SAYERS
5	15	GOOD GUYS

Let us present the labeled units by the cluster membership in the 2–dimensional space determined by the obtained components.