1 Topic

Determining the right number of components in PCA (Principal Component Analysis).

Principal Component Analysis (PCA)[1] is a dimension reduction technique. We obtain a set of factors which summarize, as well as possible, the information available in the data. The factors (or components) are linear combinations of the original variables.

Choosing the right number of factors is a crucial problem in PCA. If we select too much factors, we include noise from the sampling fluctuations in the analysis. If we choose too few factors, we lose relevant information, the analysis is incomplete. Unfortunately, there is not an indisputable approach for the determination of the number of factors. As a rule of thumb, we must select only the interpretable factors, knowing that the choice depends heavily on the domain expertise. And yet, this last one is not always available. We intend precisely to build on the data analysis to get a better knowledge on the studied domain.

In this tutorial, we present various approaches for the determination of the right number of factors for **PCA based on the correlation matrix**. Some of them, such as the Kaiser-Gutman rule or the scree plot method, are very popular even if they are not really statistically sound; others seems more rigorous, but seldom if ever used because they are not available in the popular statistical software suite.

In a first time, we use Tanagra and the Excel spreadsheet for the implementation of some methods; in a second time, especially for the resampling based approaches, we write programs for R from the results of the princomp() procedure.

2 Dataset – PCA using Tanagra

We use the “crime_dataset_pca.xls” data file. It contains **p = 14** variables and **n = 47** instances.

This data file comes from the DASL repository[2]. We had already processed it previously when we had presented the VARIMAX rotation in the context of the PCA[3]. Thus, we describe shortly the implementation and the reading of the output of the PCA in this tutorial. We are focusing on the determination of the right number of components.

2.1 Importing the data file

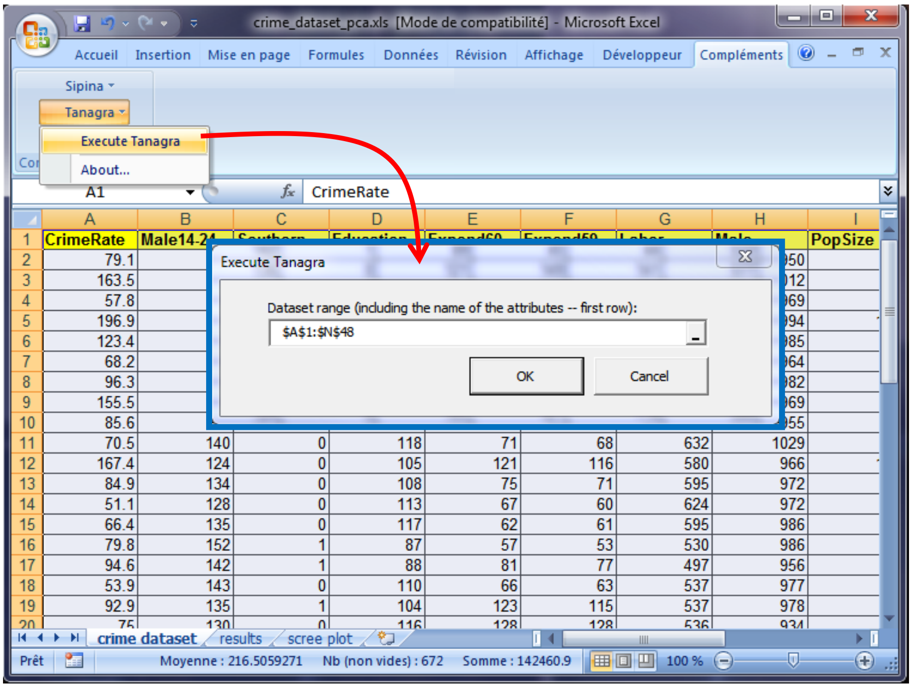
After we load the dataset into the Excel spreadsheet, we send it to Tanagra using the add-in “Tanagra.xla”[4].

[1] http://en.wikipedia.org/wiki/Principal_component_analysis

[2] http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html

[3] http://data-mining-tutorials.blogspot.fr/2009/12/varimax-rotation-in-principal-component.html

[4] http://data-mining-tutorials.blogspot.fr/2010/08/tanagra-add-in-for-office-2007-and.html; see http://data-mining-tutorials.blogspot.fr/2011/07/tanagra-add-on-for-openoffice-calc-33.html for OpenOffice and LibreOffice.



A1 CrimeRate

	A	B	C	D	E	F	G	H	I
1	CrimeRate	Male14-24							PopSize
2	79.1								
3	163.5								
4	57.8								
5	196.9								
6	123.4								
7	68.2								
8	96.3								
9	155.5								
10	85.6								
11	70.5	140	0	118	71	68	632	1029	
12	167.4	124	0	105	121	116	580	966	
13	84.9	134	0	108	75	71	595	972	
14	51.1	128	0	113	67	60	624	972	
15	66.4	135	0	117	62	61	595	986	
16	79.8	152	1	87	57	53	530	986	
17	94.6	142	1	88	81	77	497	956	
18	53.9	143	0	110	66	63	537	977	
19	92.9	135	1	104	123	115	537	978	

Execute Tanagra

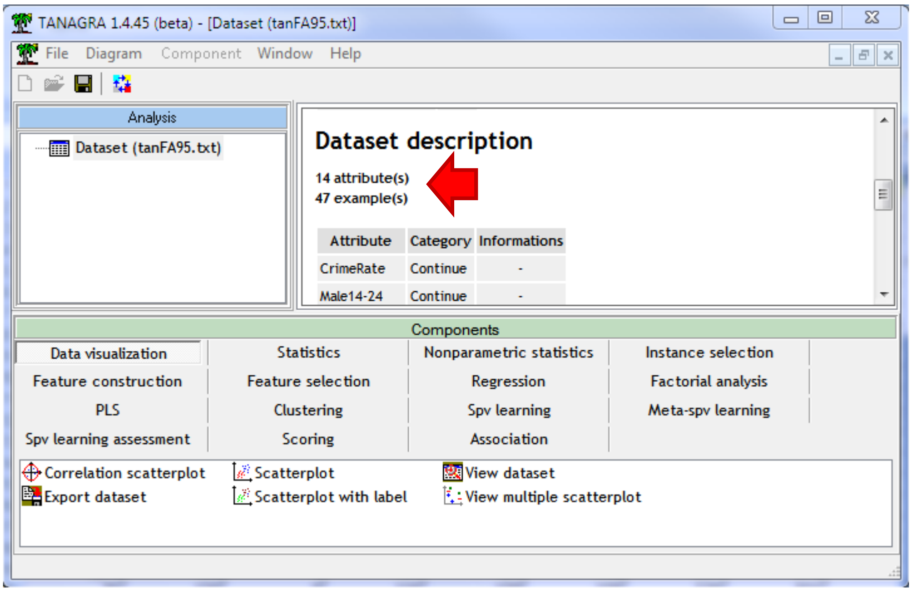
Dataset range (including the name of the attributes -- first row):

$A$1:$N$48

OK Cancel

Prêt Moyenne : 216.5059271 Nb (non vides) : 672 Somme : 142460.9 100 %

Tanagra is automatically launched, and the dataset is loaded. We check the number of instances (n = 47) and attributes (p = 14).

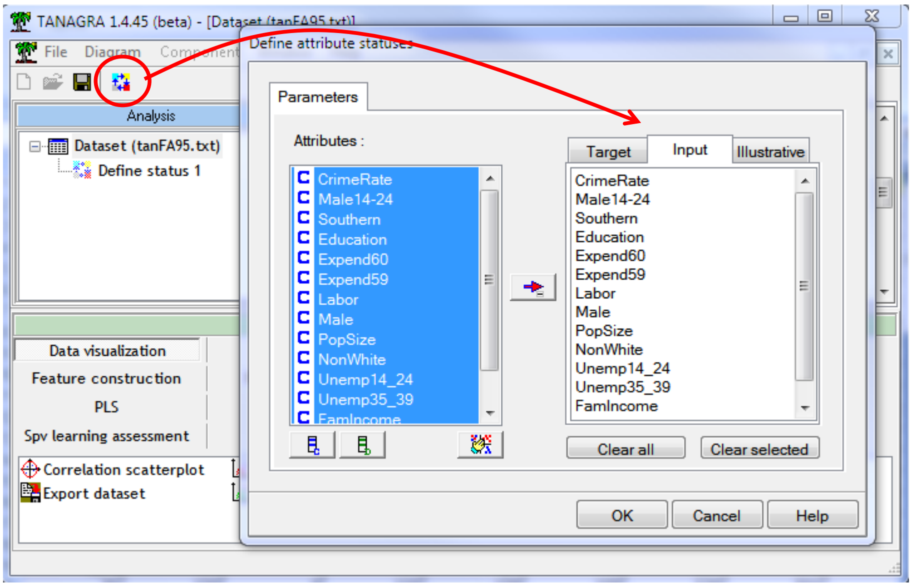


Dataset description

14 attribute(s)
47 example(s)

Attribute	Category	Informations
CrimeRate	Continue	-
Male14-24	Continue	-

We add the DEFINE STATUS component into the diagram in order to specify the variables used for the principal component analysis (CRIMERATE…INCUNDERMED).



Attributes :

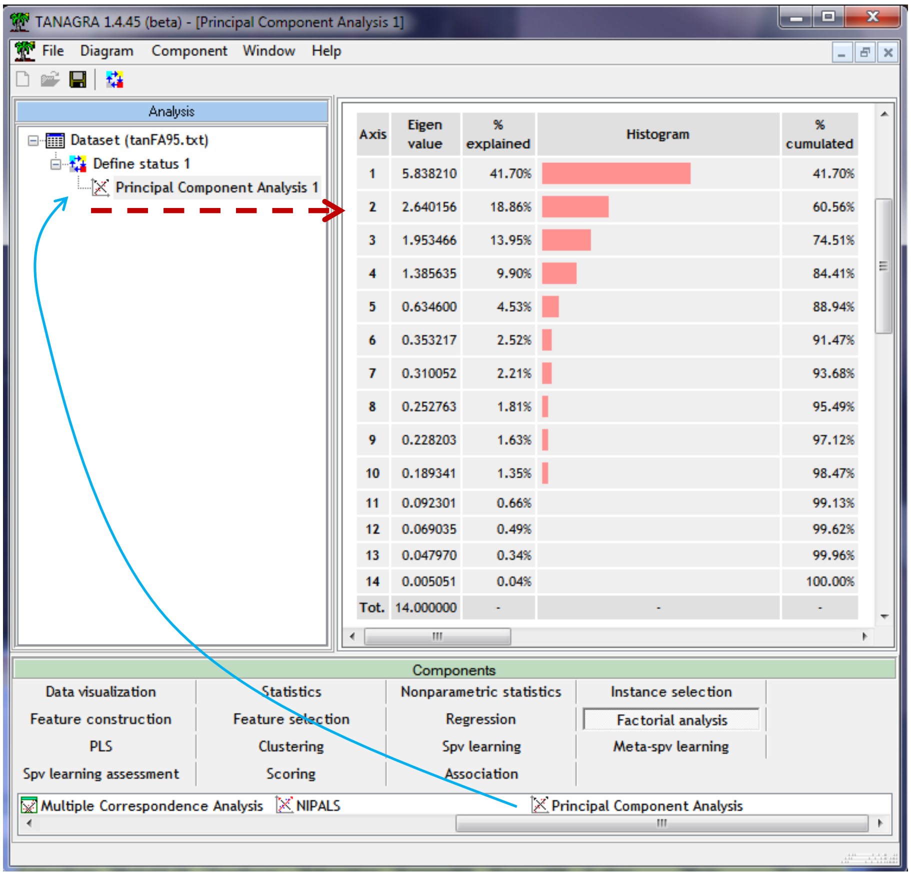
C CrimeRate
C Male14-24
C Southern
C Education
C Expend60
C Expend59
C Labor
C Male
C PopSize
C NonWhite
C Unemp14_24
C Unemp35_39
C FamIncome

Input:

CrimeRate
Male14-24
Southern
Education
Expend60
Expend59
Labor
Male
PopSize
NonWhite
Unemp14_24
Unemp35_39
FamIncome

2.2 Performing the PCA

We add the PRINCIPAL COMPONENT ANALYSIS component (FACTORIAL ANALYSIS tab). By default, Tanagra performs a PCA based on the correlation matrix.



Axis	Eigen value	% explained	Histogram	% cumulated
1	5.838210	41.70%		41.70%
2	2.640156	18.86%		60.56%
3	1.953466	13.95%		74.51%
4	1.385635	9.90%		84.41%
5	0.634600	4.53%		88.94%
6	0.353217	2.52%		91.47%
7	0.310052	2.21%		93.68%
8	0.252763	1.81%		95.49%
9	0.228203	1.63%		97.12%
10	0.189341	1.35%		98.47%
11	0.092301	0.66%		99.13%
12	0.069035	0.49%		99.62%
13	0.047970	0.34%		99.96%
14	0.005051	0.04%		100.00%
Tot.	14.000000	-	-	-

Tanagra provides, among others[5], the eigenvalues table. We know that the eigenvalue associated to a factor corresponds to its variance. Thus, the eigenvalue indicates the importance of the factor. The

[5] See http://data-mining-tutorials.blogspot.fr/2009/04/principal-component-analysis-pca.html for a detailed presentation of the output of the PCA component.

higher is the value, the higher is the importance of the factor. The challenge is determining the number of relevant factors that we need to keep, on the basis of their eigenvalue. This is not really easy. Several aspects should be considered (Jackson, 1993): the number of instances ‘n’; the number of variables ‘p’; the ratio between the instances and the variables ‘n:p’; the correlation between the variables; the possible existence of groups of correlated variables.

The ratio 'n:p' is an important thing. It determines the stability of the factors. Some references indicate that 'n:p' must be higher 3 in order to obtain reliable results (Grossman et al., 1991). We have 47/14 = 3.36 for our dataset.

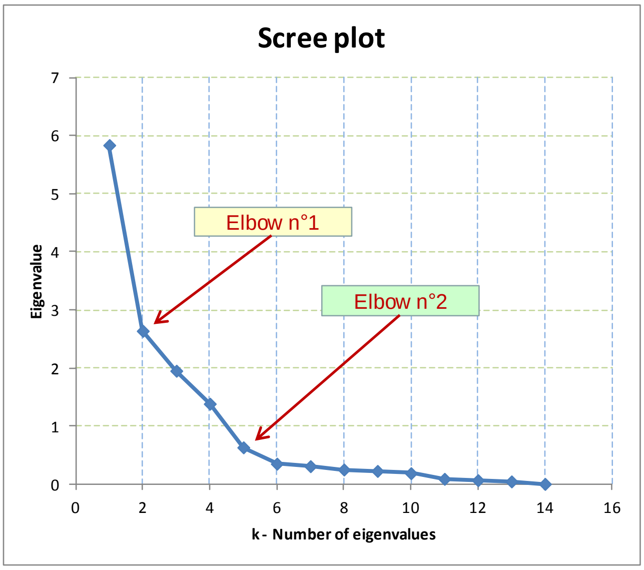
3 Scree plot

3.1 Scree plot

Cattell (1966, 1977) proposes to study the plotting of the eigenvalues (λ_k) according to the number of factors. The idea is to detect the "elbow" in the scree plot, highlighting a modification of the structure of the data. This approach is interesting because it is nuanced. It enables to go beyond the purely arbitrary numerical criterion. But it is complicated to implement because it can be subjective. The detection is not always obvious. We must answer several questions: where is located the elbow? Is it unique? Do we include the factor associated with the elbow in the selection?

Usually, the elbow is pronounced when we handle highly correlated variables. When the correlations are low or when there are blocks of correlated variables, rather than a single "obvious" solution, we face several possible solutions. Concerning the integration of the elbow in the selection, Cattel was hesitant. Originally (1966), he advised to select only the factors which are before the elbow; then, in a second time (1977), he advocates to integrate it. Actually, it depends on the value of the elbow. If the corresponding eigenvalue is high, we must include it into the selection. If it is low, the factor can be neglected.

About the CRIME dataset, we obtain:



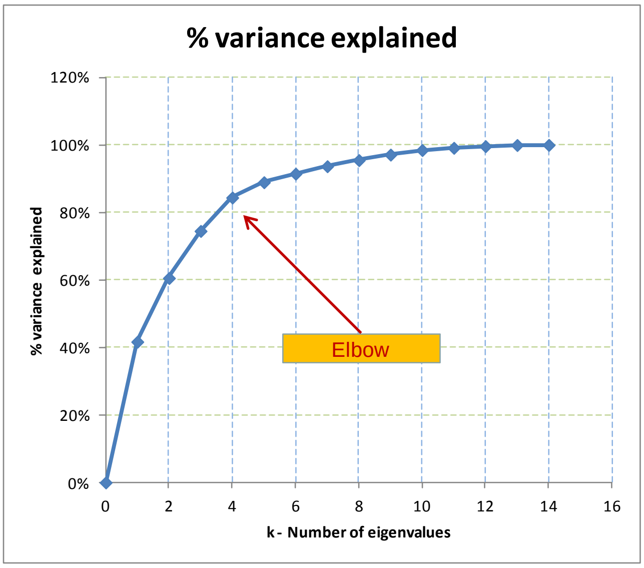
It seems there are 2 elbows. Which one to choose?

For the first one, we must include the elbow into the selection i.e. we select k = 2 factors. Indeed, the eigenvalue (λ_2 = 2.64) associated with the 2nd factor is high. It corresponds to 18.86% of the variance.

If we prefer the second solution, we must neglect the 5th factor. Indeed, the corresponding eigenvalue is too low (λ_5 = 0.6346, 4.53% of the total variance). We select k = 4 factors in this case.

As we note, the determination of the right solution is not obvious. To improve the reading of this graph, we can use a second graph where we represent the evolution of the cumulative variance explained by the first k factors. We can detect also another elbow. It indicates that the remaining factors correspond to a too low proportion of variance and can be neglected.

For our dataset, we have the following graph:



It seems that the solution with k = 4 factors is the appropriate one. Clearly, the choice of k = 2 factors is not good. The proportion of variance after the 2nd factor cannot be neglected.

3.2 Proportion of total variance

Some references advise to explicitly use the proportion of total variance to determine the number of factors. The rule would be: "we select enough factors in order to explain at least x % of the total variance". This strategy is not really reliable because it does not take account for the correlations between the variables. For our dataset, if we select the factors which reproduce 95% of the total variance, we must select 8 (!) factors. This is clearly excessive. From the 5th factor, the eigenvalues associated to the factors are too low. They correspond to the sampling fluctuations.

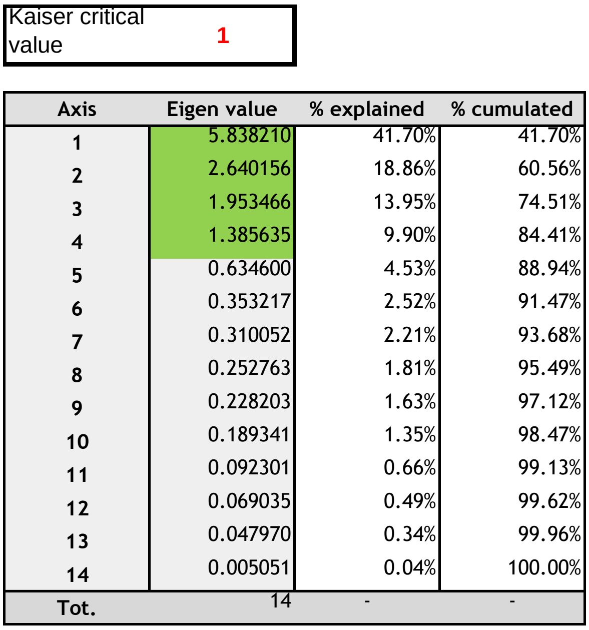
But, after the fact, when we have determined the right number of factors, it can be interesting to analyze the proportion of variance that they reproduce. This is important in the graphical representation when we want to evaluate the proximities between the instances.

4 Kaiser-Guttman rule and its variant

4.1 Kaiser – Guttman rule

The Kaiser-Guttman rule is based on a very simplistic idea. If the variables are independent, the eigenvalues for all the factors is 1. Thus, we select the factors which have an eigenvalue higher to 1.

For the CRIME dataset, we select k = 4 factors if we use the Kaiser-Guttman rule.



Kaiser critical value	1

Axis	Eigen value	% explained	% cumulated
1	5.838210	41.70%	41.70%
2	2.640156	18.86%	60.56%
3	1.953466	13.95%	74.51%
4	1.385635	9.90%	84.41%
5	0.634600	4.53%	88.94%
6	0.353217	2.52%	91.47%
7	0.310052	2.21%	93.68%
8	0.252763	1.81%	95.49%
9	0.228203	1.63%	97.12%
10	0.189341	1.35%	98.47%
11	0.092301	0.66%	99.13%
12	0.069035	0.49%	99.62%
13	0.047970	0.34%	99.96%
14	0.005051	0.04%	100.00%
Tot.	14	-	-

A numerical threshold is always comforting, even if it sometimes appears as lacking nuance. We note however that this rule confirms the reading of the scree plot above.

4.2 Karlis – Saporta - Spinaki rule (2003)

The threshold "1" is a too permissive criterion in the majority of cases, especially when the data are weakly correlated. Above all, it does not take into consideration the characteristics of the dataset (number of instances n, number of variables p, the ratio n:p).

A more restrictive rule overcomes this drawback. It is based on a statistical process. We select the factors for which their eigenvalue is **significantly higher than 1** (Saporta, 2006)[6]. At the 5% level (approximately), the critical region is defined as follows:

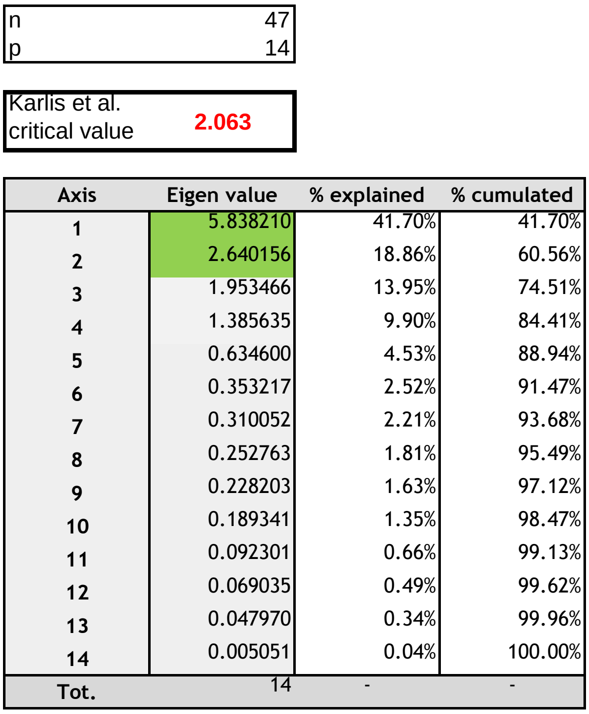
$$\lambda > 1 + 2\sqrt{\frac{p-1}{n-1}}$$

This rule has a good behavior. It is more restrictive when the number of variables p increases against to the number of instances n. For the CRIME dataset, the critical value is:

$$1 + 2\sqrt{\frac{p-1}{n-1}} = 1 + 2\sqrt{\frac{14-1}{47-1}} = 2.063$$

With this new threshold (**2.063)**, we keep only the (**k= 2)** first components. We remind that this is one of the possible solutions identified in the scree plot above. After all, this solution is perhaps not so bad. We observe also that the third eigenvalue (λ_3 = 1.953466) is not far from the critical value.

[6] See also http://cedric.cnam.fr/fichiers/RC489.pdf



n	47
p	14

Karlis et al. critical value	2.063

Axis	Eigen value	% explained	% cumulated
1	5.838210	41.70%	41.70%
2	2.640156	18.86%	60.56%
3	1.953466	13.95%	74.51%
4	1.385635	9.90%	84.41%
5	0.634600	4.53%	88.94%
6	0.353217	2.52%	91.47%
7	0.310052	2.21%	93.68%
8	0.252763	1.81%	95.49%
9	0.228203	1.63%	97.12%
10	0.189341	1.35%	98.47%
11	0.092301	0.66%	99.13%
12	0.069035	0.49%	99.62%
13	0.047970	0.34%	99.96%
14	0.005051	0.04%	100.00%
Tot.	14	-	-

5 Bartlett's test

5.1 Detecting the existence of relevant factors

The Bartlett's sphericity test enables to check if the correlation matrix is significantly different to the identity matrix. In our context, we can use it to determine if one of the eigenvalues at least is significantly different to 1. That means that there is at least one relevant factor. But we cannot determine the number of relevant factors.

The test statistic is defined as follows:

$$C = -\left(n - 1 - \frac{2p+5}{6}\right) \times \ln|R| = -\left(n - \frac{2p+11}{6}\right) \times \ln|R|$$

Under the null hypothesis, it follows a χ^2 distribution with [p x (p-1) / 2] degree of freedom.

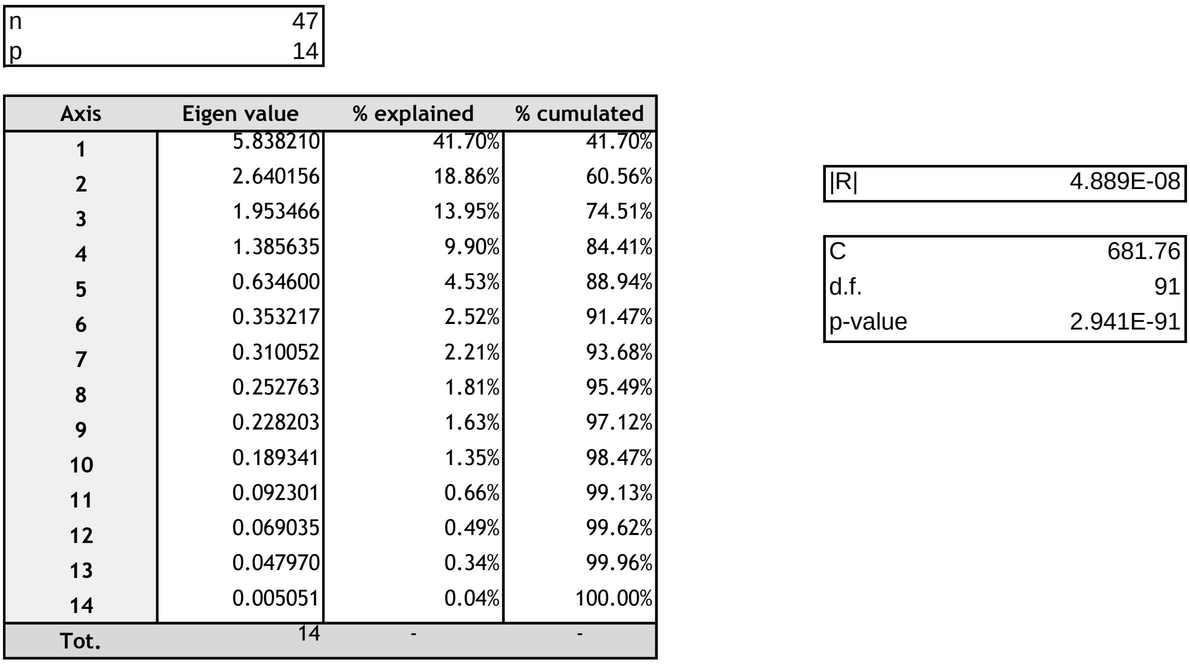
The determinant of the correlation matrix is equal to the product of the eigenvalues. We can use the results from the PCA for the calculations:

$$C = -\left(n - 1 - \frac{2p+5}{6}\right) \times \ln|R| = -\left(n - \frac{2p+11}{6}\right) \times \sum_{k=1}^{p} \ln(\lambda_k)$$

For the CRIME dataset, we obtain

$$|R| = \prod_k \lambda_k = (5.838210 \times \cdots \times 0.005051) = 4.889 \times 10^{-8}$$

Thus, the test statistic and the corresponding p-value are computed below:



n	47
p	14

Axis	Eigen value	% explained	% cumulated
1	5.838210	41.70%	41.70%
2	2.640156	18.86%	60.56%
3	1.953466	13.95%	74.51%
4	1.385635	9.90%	84.41%
5	0.634600	4.53%	88.94%
6	0.353217	2.52%	91.47%
7	0.310052	2.21%	93.68%
8	0.252763	1.81%	95.49%
9	0.228203	1.63%	97.12%
10	0.189341	1.35%	98.47%
11	0.092301	0.66%	99.13%
12	0.069035	0.49%	99.62%
13	0.047970	0.34%	99.96%
14	0.005051	0.04%	100.00%
Tot.	14	-	-

|R|	4.889E-08

C	681.76
d.f.	91
p-value	2.941E-91

Clearly, there is at least one relevant factor for our analysis.

Note: There is however a restriction on the use of this test, it tends to be always significant when the sample size increases.

5.2 Detection of the number of relevant factors

A variant of the Bartlett's test for the determination of the right number of factors exists. It was originally developed for the PCA based on the covariance matrix (Saporta, 2006; Grossman and al., 1991). But, we can use it in our context. It seems that the test becomes conservative in this case i.e. it tends to select too few numbers of factors (Jackson, 1993; Neto et al., 2004). We will study its behavior on our data file.

The process is based on the following idea: we select "k" factors because the (p-k) eigenvalues of the remaining factors are equals. These eigenvalues correspond to the horizontal part of the scree plot, after the "elbow". Here is the null hypothesis of the test:

$$H_0 : \lambda_{k+1} = \cdots = \lambda_p$$

Unfortunately, the references used vary about the formula of the test statistic. I decided to use the one presented by Saporta (2006), because it is consistent to the Bartlett's test of sphericity when we check the equality of all the factors.

Under the null hypothesis, the (p-k) last eigenvalues are equals. Their arithmetic mean ($\bar{\lambda}$) is equal to their geometrical mean ($\tilde{\lambda}$). We use the following test statistic to compare them:

$$c_k = \left(n - \frac{2p+11}{6} \right) \times (p-k) \times \ln\left(\frac{\bar{\lambda}}{\tilde{\lambda}} \right)$$

Where $\bar{\lambda} = \frac{1}{p-k} \sum_{i=k+1}^{p} \lambda_i$

And $\ln\left(\tilde{\lambda}\right) = \frac{1}{p-k}\sum_{i=k+1}^{p}\ln(\lambda_i)$

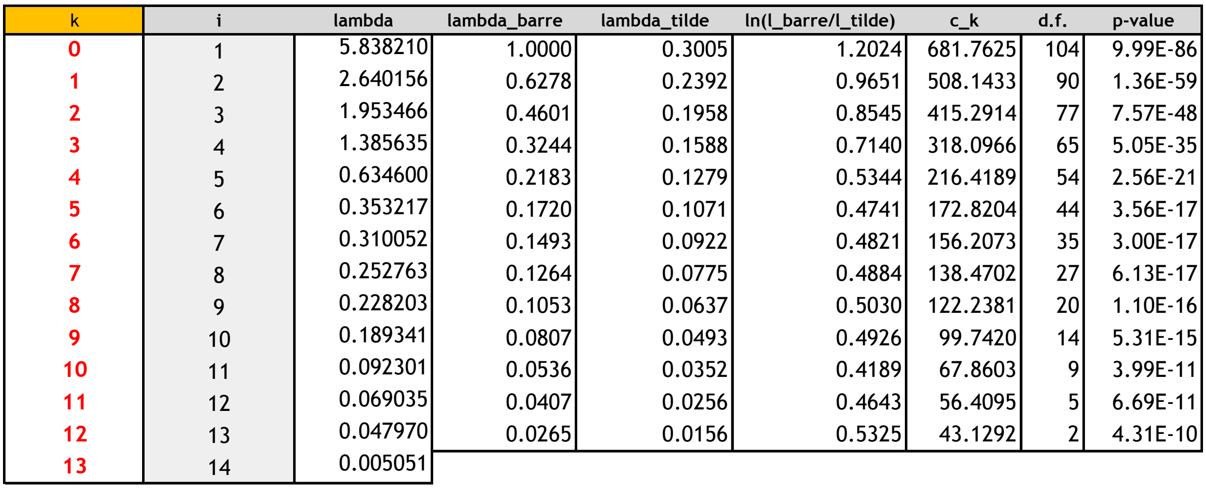
Under H0, it follows a χ^2 distribution with $\frac{(p-k+2)(p-k-1)}{2}$ degree of freedom.

This test statistic is consistent to the Bartlett's sphericity test. Indeed, for k = 0, we test the equality of all the eigenvalues i.e. λ_k=1, ∀k and $\bar{\lambda} = 1$. We verify that:

$$
\begin{aligned}
c_0 &= \left(n - \frac{2p+11}{6}\right) \times p \times \left[-\frac{1}{p}\ln\left(\prod_{i=1}^{p}\lambda_i\right)\right] \\
&= \left(n - \frac{2p+11}{6}\right) \times p \times \left[-\frac{1}{p}\sum_{i=1}^{p}\ln(\lambda_i)\right] \\
&= -\left(n - \frac{2p+11}{6}\right) \times \sum_{i=1}^{p}\ln(\lambda_i) \\
&= C
\end{aligned}
$$

Unfortunately, the degrees of freedom are not consistent. When k = 0, we have d.f. = (p+2)(p-1)/2, this is different to d.f. = p x (p - 1) / 2 of the Bartlett's test of sphericity. However, all references are agreed upon the formula above. It remains a mystery.

We elaborate a worksheet under Excel. We test from k = 0 (equality of all the eigenvalues) to k = 12 (equality of 13th and 14th eigenvalues).



k	i	lambda	lambda_barre	lambda_tilde	ln(l_barre/l_tilde)	c_k	d.f.	p-value
0	1	5.838210	1.0000	0.3005	1.2024	681.7625	104	9.99E-86
1	2	2.640156	0.6278	0.2392	0.9651	508.1433	90	1.36E-59
2	3	1.953466	0.4601	0.1958	0.8545	415.2914	77	7.57E-48
3	4	1.385635	0.3244	0.1588	0.7140	318.0966	65	5.05E-35
4	5	0.634600	0.2183	0.1279	0.5344	216.4189	54	2.56E-21
5	6	0.353217	0.1720	0.1071	0.4741	172.8204	44	3.56E-17
6	7	0.310052	0.1493	0.0922	0.4821	156.2073	35	3.00E-17
7	8	0.252763	0.1264	0.0775	0.4884	138.4702	27	6.13E-17
8	9	0.228203	0.1053	0.0637	0.5030	122.2381	20	1.10E-16
9	10	0.189341	0.0807	0.0493	0.4926	99.7420	14	5.31E-15
10	11	0.092301	0.0536	0.0352	0.4189	67.8603	9	3.99E-11
11	12	0.069035	0.0407	0.0256	0.4643	56.4095	5	6.69E-11
12	13	0.047970	0.0265	0.0156	0.5325	43.1292	2	4.31E-10
13	14	0.005051						

Clearly, this test is not appropriate for our dataset. It claims that all the factors are relevant. This is not consistent with the results of the other approaches. It seems that the Bartlett's test is usable only for the determination of the existence of relevant factors, only if the number of instances is not too high. We cannot use it for the determination of the right number of factors (Neto and al., 2004).

6 Broken-stick method

This approach is based on the following idea: under the null hypothesis where the total variance is randomly allocated on the factors, the eigenvalues is distributed according to the "broken-stick" distribution (Frontier, 1976; Legendre-Legendre, 1983).

One of the main interests of the approach is that the critical values are very easily to calculate. For the evaluation of a solution with "k" components, the critical value is:

$$b_k = \sum_{i=k}^{p} \frac{1}{i}$$

If we want to reason in terms of proportion of total variance, the critical value becomes:

$$b'_k = \frac{1}{p} \sum_{i=k}^{p} \frac{1}{i}$$

We can calculate the threshold for each value of k in a worksheet:

k	Eigen value	1/i	b_k
1	5.838210	1.000	3.252
2	2.640156	0.500	2.252
3	1.953466	0.333	1.752
4	1.385635	0.250	1.418
5	0.634600	0.200	1.168
6	0.353217	0.167	0.968
7	0.310052	0.143	0.802
8	0.252763	0.125	0.659
9	0.228203	0.111	0.534
10	0.189341	0.100	0.423
11	0.092301	0.091	0.323
12	0.069035	0.083	0.232
13	0.047970	0.077	0.148
14	0.005051	0.071	0.071

If we test the relevance of the first factor (k=1), we have the following critical value:

$$b_1 = \left(\frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{14} \right) = 3.252$$

λ_1 = 5.83821, the first factor is accepted.

If we want to test the relevance of the two first factors (k = 2), we have:

$$b_2 = \left(\frac{1}{2} + \cdots + \frac{1}{14} \right) = 2.252$$

λ_2 = 2.640156, the two first factors are accepted.

According to the broken-stick method, k = 3 components is the right solution. We note however that the 4th factor is removed narrowly.

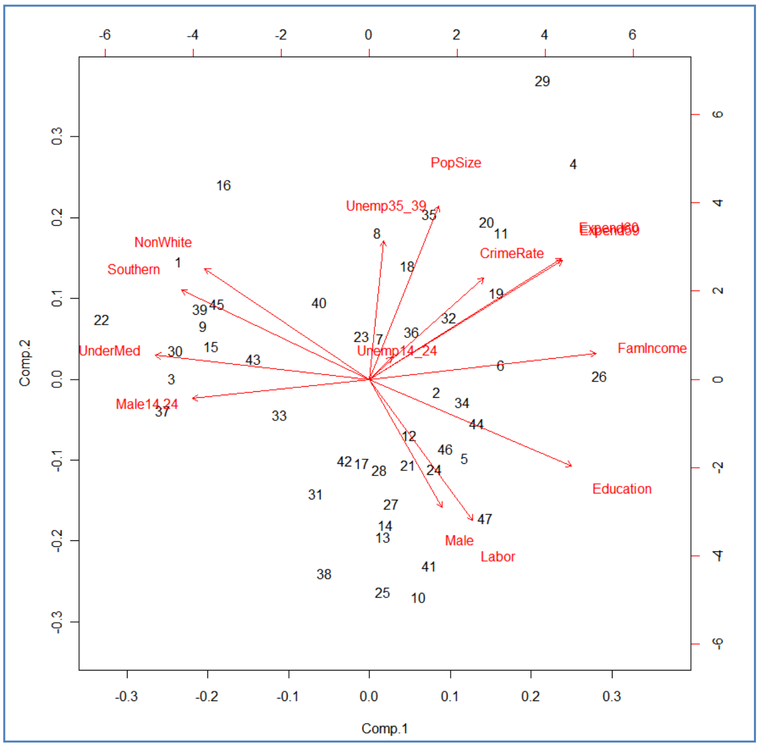
The broken-stick approach is often efficient (Jackson, 1993). Its advantage is that it use the number of variables 'p' for the calculation of the critical value. Its drawback is that it does not take into account the size of the sample 'n' and the ratio 'n:p' (Franklin and al., 1995).

7 Principal component analysis with R

We perform the same analysis with R to prepare the presentation of resampling techniques in the next section. We used the following program:

```
rm(list=ls())
#importing the data file
library(xlsx)
crime.data <- read.xlsx(file="crime_dataset_pca.xls",sheetIndex=1,header=T)
#performing the pca with princomp
crime.pca <- princomp(crime.data,cor=T)
eig.val <- crime.pca$sdev^2
print(eig.val)
biplot(crime.pca)
```

We obtain exactly the same results. We show here the biplot graph for the first two factors.



8 Resampling approaches

8.1 Parallel Analysis Method

The parallel analysis enables to calculate the critical values for the eigenvalues without determination of their distribution under the null hypothesis (H0: the variables are independent). For that, it uses a Monte Carlo approach (Neto et al, 2004). The idea is to calculate the many versions of eigenvalues on artificial datasets with the same characteristics (n and p) than the studied dataset, but where the variables are independent. A factor is considered relevant if its observed eigenvalue is higher than the mean or the quantile at 95% level of the simulated eigenvalues under the null hypothesis. Here are the main steps of the process:

1. Generate randomly a dataset with n instances and p variables. Each variable is distributed as a Gaussian distribution N(0, 1). The variables are generated in independent way.
2. Perform a PCA on this dataset. We obtain 'p' eigenvalues (λ_k, $k = 1, \ldots, p$).
3. Repeat T times (e.g. T = 1000) the steps (1) and (2).

4. Compute the mean (μ_k) of the eigenvalues for each factor as critical value.
5. We consider that a factor is relevant if its eigenvalue computed on the original dataset is higher than the corresponding critical value ($\lambda_k > \mu_k$).

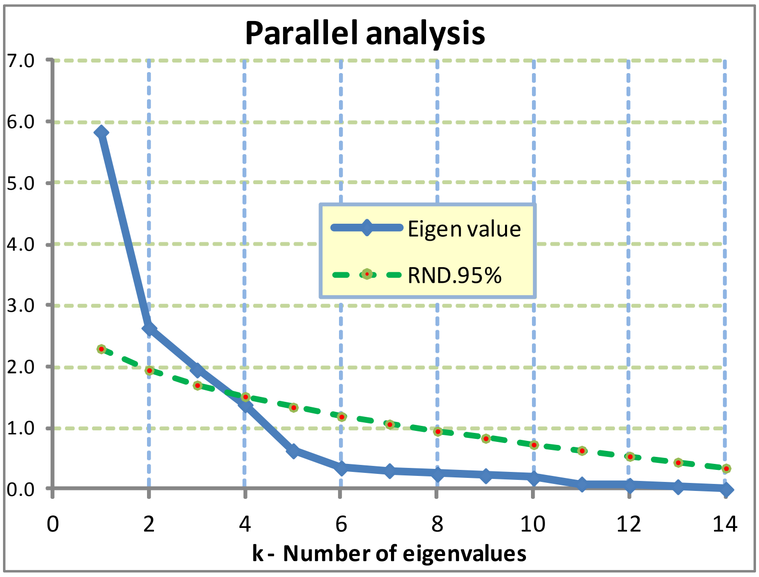
A more restrictive rule is to use the quantile at 95% [$q_k^{0.95}$]. We accept the factor if ($\lambda_k > q_k^{0.95}$) at the step 5 of the process.

Here is the R source code for the calculation of the critical values (T = 1000).

```
#*****************
#PARALLEL ANALYSIS
#*****************
#n : number of instance, p : number of variables
n <- nrow(crime.data)
p <- ncol(crime.data)
#generation of a dataset
gendata <- function(n,p){
  df <- list()
  for (k in 1:p){
    x <- rnorm(n)
    df[[k]] <- x
  }
  df <- data.frame(df)
  colnames(df) <- 1:p
  return(df)
}
#pca on gendata
pca.gendata <- function(n,p){
  data.gen <- gendata(n,p)
  pca <- princomp(data.gen,cor=T)
  eig <- pca$sd^2
  return(eig)
}
set.seed(1)
#repeating T times the analysis
T <- 1000
res <- replicate(T, pca.gendata(n,p))
#computing the mean of the eigenvalues
rnd.mean <- apply(res,1,mean)
print(rnd.mean)
#computing the 0.95 percentile
rnd.95 <- apply(res,1,quantile,probs=(0.95))
print(rnd.95)
```

We show in the table below the mean and the quantile from the parallel analysis. We insert into the graph the observed eigenvalue on the CRIME dataset and the quantile at 95% level.

Axis	Eigen value	RND.Mean	RND.95%
1	5.838210	2.051060	2.303778
2	2.640156	1.771630	1.952445
3	1.953466	1.563366	1.708953
4	**1.385635**	**1.388553**	**1.521725**
5	0.634600	1.239614	1.347275
6	0.353217	1.098187	1.195431
7	0.310052	0.978249	1.071068
8	0.252763	0.862912	0.955985
9	0.228203	0.752236	0.840423
10	0.189341	0.649536	0.737475
11	0.092301	0.551690	0.642607
12	0.069035	0.457871	0.545664
13	0.047970	0.366953	0.448322
14	0.005051	0.268144	0.351839



The first three factors are clearly approved. The situation is more complicated for the 4th factor. The observed eigenvalue (λ_4=1.3856) is very close to the mean of the eigenvalues obtained on the generated datasets (μ_4 = 1.3885). We find again here the uncertainty observed for the broken stick method. The 5th and the remaining factors can be clearly neglected.

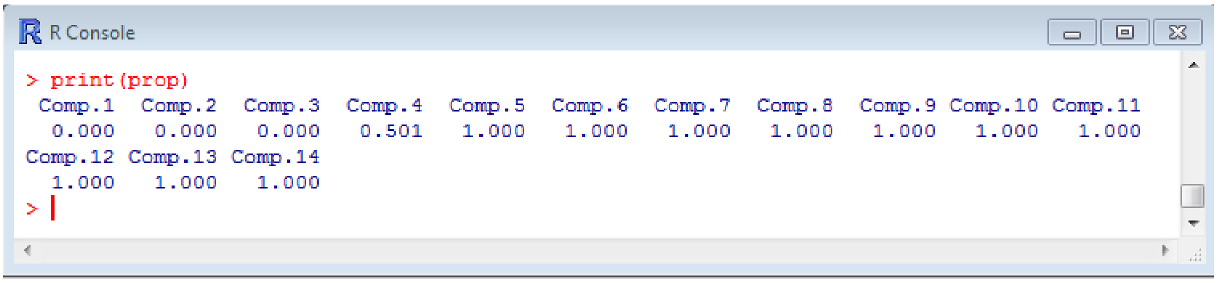
Statistical tables. Some references give statistical tables at the 95% level for various values of n and p. We can thus directly compare the observed eigenvalue with the provided critical values.

P-value. We can obtain an approximation of the p-value by calculating the proportion of simulated eigenvalues higher than the observed eigenvalue (λ_k). If this p-value is lower than 5%, we can claim that the factor is significant.

Here is the source code:

```
#computing the proportion of values upper than eig.val
prop <- rep(0,length(eig.val))
names(prop) <- names(eig.val)
for (k in 1:length(eig.val)){
 prop[k] <- length(which(res[k,] > eig.val[k]))/T
}
print(prop)
```

We obtain:



```
> print(prop)
 Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9 Comp.10 Comp.11 
  0.000   0.000   0.000   0.501   1.000   1.000   1.000   1.000   1.000   1.000   1.000 
Comp.12 Comp.13 Comp.14 
  1.000   1.000   1.000 
> 
```

The first 3 factors are highly significant, the 5th and the remaining are clearly not significant.

Parallel analysis for PCA based on covariance matrix. The parallel analysis can be easily extended to the PCA on covariance matrix. For each variable, we use the observed mean and standard deviation during the generation of the column according to the Gaussian distribution.

8.2 Randomization approach

The parallel analysis is robust against departures from normality used during the generation of the artificial dataset. But the process appears uselessly restrictive. We can use the available dataset in order to create the artificial dataset by using randomization approach.

The protocol is the following: we randomize the values within the variables in the dataset; we perform the PCA on the randomized dataset; we repeat T times this process (Neto and al., 2004). Thus, the correlations between the variables, if they exist, are degraded. The main advantage is that we use the available values, the characteristics of each variable, independently to the others, are preserved. We use the collected eigenvalues to define the critical values (the quantile or the mean).

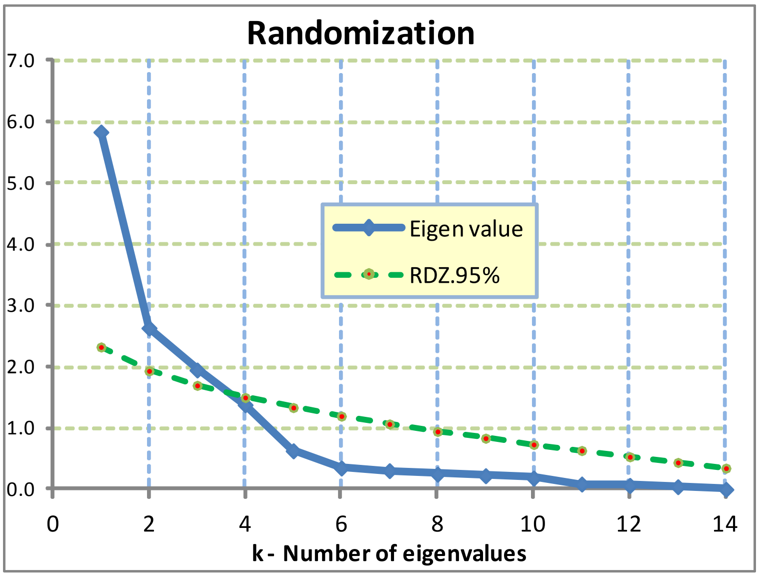
```
#**************
# RANDOMIZATION
#**************
set.seed(1)
one.randomization <- function(dataset){
 dataset.rdz                                                          <-
data.frame(lapply(dataset,function(x){sample(x,length(x),replace=F)}))
 pca.rdz <- princomp(dataset.rdz,cor=T)
 eig.rdz <- pca.rdz$sd^2
 return(eig.rdz)
}

#repeat the procedure
res.rdz <- replicate(T,one.randomization(crime.data))

#mean
rdz.mean <- apply(res.rdz,1,mean)
print(rdz.mean)
#quantile
rdz.95 <- apply(res.rdz,1,quantile,probs=(0.95))
print(rdz.95)
```

We obtain a new version of the table from the parallel analysis. It is very similar to the previous one actually. But now, for the fourth factor, the observed eigenvalue is slightly higher than the mean of the simulated values (but remains lower than the quantile at 95% level).

Axis	Eigen value	RDZ.Mean	RDZ.95%
1	5.838210	2.062149	2.330069
2	2.640156	1.767845	1.940688
3	1.953466	1.561557	1.703605
4	**1.385635**	1.385119	1.505952
5	0.634600	1.236817	1.341472
6	0.353217	1.100492	1.202589
7	0.310052	0.974988	1.073624
8	0.252763	0.860179	0.951342
9	0.228203	0.752086	0.842079
10	0.189341	0.648836	0.739747
11	0.092301	0.552711	0.640388
12	0.069035	0.458954	0.537746
13	0.047970	0.367645	0.445864
14	0.005051	0.270622	0.352655



8.3 Bootstrap method (1)

The bootstrap approach enables to obtain an estimation of the standard error of the eigenvalue calculated on the dataset. We repeat T times the following process in order to obtain various estimates of each eigenvalue λ_k: we create a bootstrapped sample of size n; we perform the PCA on this sample; we collect the values of λ_k.

Then, we use a variant of Kaiser-Guttman rule. We approve a factor if the quantile at 5% level ($\lambda_k^{0.05}$) of the bootstrapped eigenvalues is higher than 1. This quantile corresponds to the lower limit of bootstrap confidence interval according to the percentile method.

We use the following program for our dataset:

```
#**********
# BOOTSTRAP
#**********

#creating one replication of the dataset
one.replication <- function(dataset){
 n <- nrow(dataset)
 index <- sort(sample.int(n,replace=T))
 out.dataset <- dataset[index,]
 return(out.dataset)
}

#performing a pca on a replication of the dataset
pca.replication <- function(dataset){
 one.dataset <- one.replication(dataset)
 pca <- princomp(one.dataset,cor=T)
 eig <- pca$sd^2
 return(eig)
}
```

```
#bootstraping pca
res.boot <- replicate(T,pca.replication(crime.data))

#quantile 0.05
boot.05 <- apply(res.boot,1,quantile,probs=(0.05))
print(boot.05)
```

The results confirm those obtained with the other approaches. There is always a doubt about the 4th factor ($\lambda_4^{0.05} = 0.966$).

Axis	Boot 0.05
1	5.176478
2	2.297267
3	1.649235
4	0.966243
5	0.459941
6	0.309183
7	0.228693
8	0.169048
9	0.120807
10	0.078729
11	0.047379
12	0.030524
13	0.016199
14	0.001778

8.4 Bootstrap method (2)

This approach checks the equality of the successive eigenvalues ($\lambda_k > \lambda_{k+1}$) using the bootstrap scheme. In practice, we checks if there are overlap between the percentile confidence interval of the successive eigenvalues. Thus, we adopt the kth factor if $\lambda_k^{0.05} > \lambda_{k+1}^{0.95}$ i.e. the lower limit of the confidence interval of the kth factor is higher than the upper limit of the following factor.

We can use the results of the bootstrap process from the previous section to calculate the upper limit of the confidence intervals.

```
#quantile 0.95
boot.95 <- apply(res.boot,1,quantile,probs=(0.95))
print(boot.95)
```

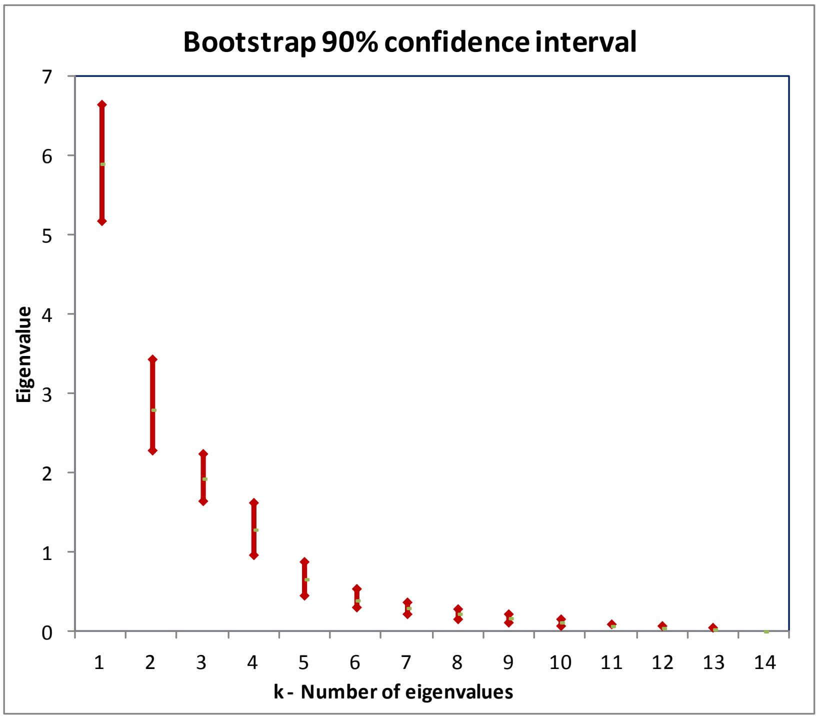
We show the values of the lower and upper limits of the confidence intervals into the following table.

Besides the 3 first factors, we observe that the 4th factor is considered relevant with this approach. Indeed: $\lambda_4^{0.05} = 0.9662 > \lambda_5^{0.95} = 0.8790$.

Axis	Boot 0.05
1	5.1765
2	2.2973
3	1.6492
4	0.9662
5	0.4599
6	0.3092
7	0.2287
8	0.1690
9	0.1208
10	0.0787
11	0.0474
12	0.0305
13	0.0162
14	0.0018

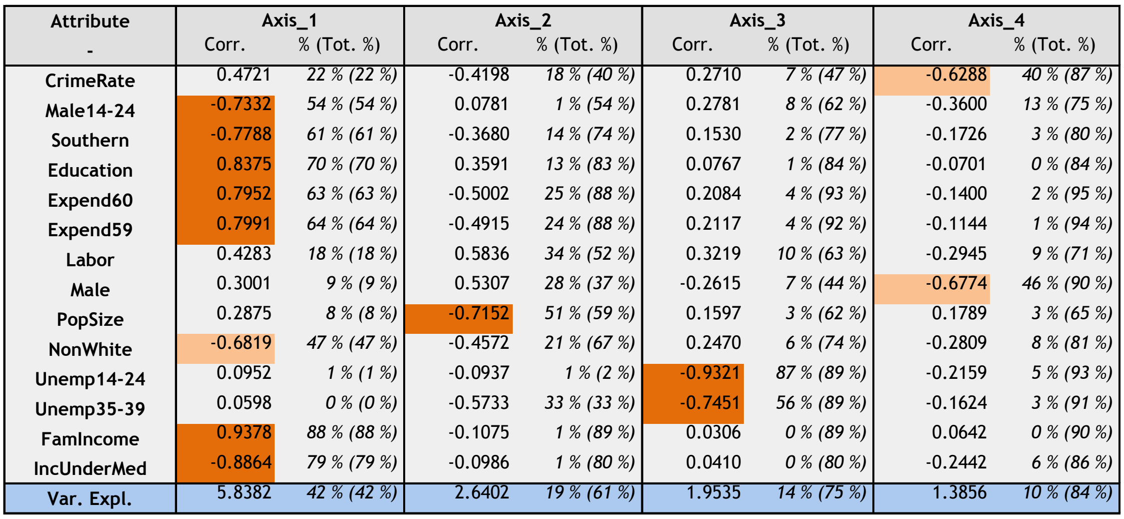
Boot 0.95	Axis
6.6547	1
3.4381	2
2.2418	3
1.6303	4
0.8790	5
0.5374	6
0.3787	7
0.2945	8
0.2263	9
0.1652	10
0.0975	11
0.0653	12
0.0426	13
0.0053	14

A graphical representation of the confidence intervals is more intuitive. Obviously, the gap between the confidence intervals disappears starting from the 5th factor.



9 Interpretation of the 4th factor

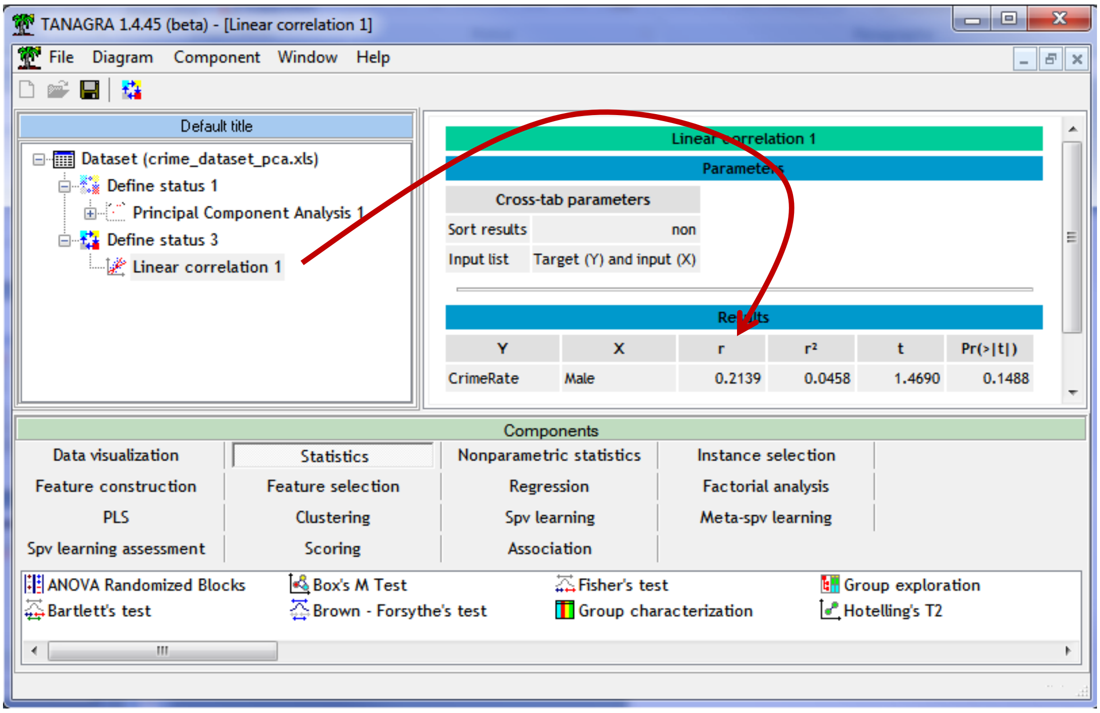
The first three factors are indisputable. The fifth and the remaining factors are clearly irrelevant. The doubt concerns the fourth factor. Another way to validate a factor is to check if we can extract a valuable interpretation. Let us analyze the loadings from the output of Tanagra.



Attribute	Axis_1		Axis_2		Axis_3		Axis_4	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)
CrimeRate	0.4721	22 % (22 %)	-0.4198	18 % (40 %)	0.2710	7 % (47 %)	-0.6288	40 % (87 %)
Male14-24	-0.7332	54 % (54 %)	0.0781	1 % (54 %)	0.2781	8 % (62 %)	-0.3600	13 % (75 %)
Southern	-0.7788	61 % (61 %)	-0.3680	14 % (74 %)	0.1530	2 % (77 %)	-0.1726	3 % (80 %)
Education	0.8375	70 % (70 %)	0.3591	13 % (83 %)	0.0767	1 % (84 %)	-0.0701	0 % (84 %)
Expend60	0.7952	63 % (63 %)	-0.5002	25 % (88 %)	0.2084	4 % (93 %)	-0.1400	2 % (95 %)
Expend59	0.7991	64 % (64 %)	-0.4915	24 % (88 %)	0.2117	4 % (92 %)	-0.1144	1 % (94 %)
Labor	0.4283	18 % (18 %)	0.5836	34 % (52 %)	0.3219	10 % (63 %)	-0.2945	9 % (71 %)
Male	0.3001	9 % (9 %)	0.5307	28 % (37 %)	-0.2615	7 % (44 %)	-0.6774	46 % (90 %)
PopSize	0.2875	8 % (8 %)	-0.7152	51 % (59 %)	0.1597	3 % (62 %)	0.1789	3 % (65 %)
NonWhite	-0.6819	47 % (47 %)	-0.4572	21 % (67 %)	0.2470	6 % (74 %)	-0.2809	8 % (81 %)
Unemp14-24	0.0952	1 % (1 %)	-0.0937	1 % (2 %)	-0.9321	87 % (89 %)	-0.2159	5 % (93 %)
Unemp35-39	0.0598	0 % (0 %)	-0.5733	33 % (33 %)	-0.7451	56 % (89 %)	-0.1624	3 % (91 %)
FamIncome	0.9378	88 % (88 %)	-0.1075	1 % (89 %)	0.0306	0 % (89 %)	0.0642	0 % (90 %)
IncUnderMed	-0.8864	79 % (79 %)	-0.0986	1 % (80 %)	0.0410	0 % (80 %)	-0.2442	6 % (86 %)
Var. Expl.	5.8382	42 % (42 %)	2.6402	19 % (61 %)	1.9535	14 % (75 %)	1.3856	10 % (84 %)

It seems that the presence of male positively affects the criminality (MALE: "The number of males per 1000 females"; CRIMERATE: "# of offenses reported to police per million populations"). This is a mysterious result. We need to better understand the context of the study and the situation of the USA in the 1960s to better understand this result.

Let us not forget an important element for the reading of the fourth factor. It measures the association between the variables by controlling the influence of the three first factors. Indeed, if we measure the correlation between MALE and CRIME, we obtain **r(MALE,CRIME) = 0.2139**. It is not significant at the 5% level.



Linear correlation 1

Parameters

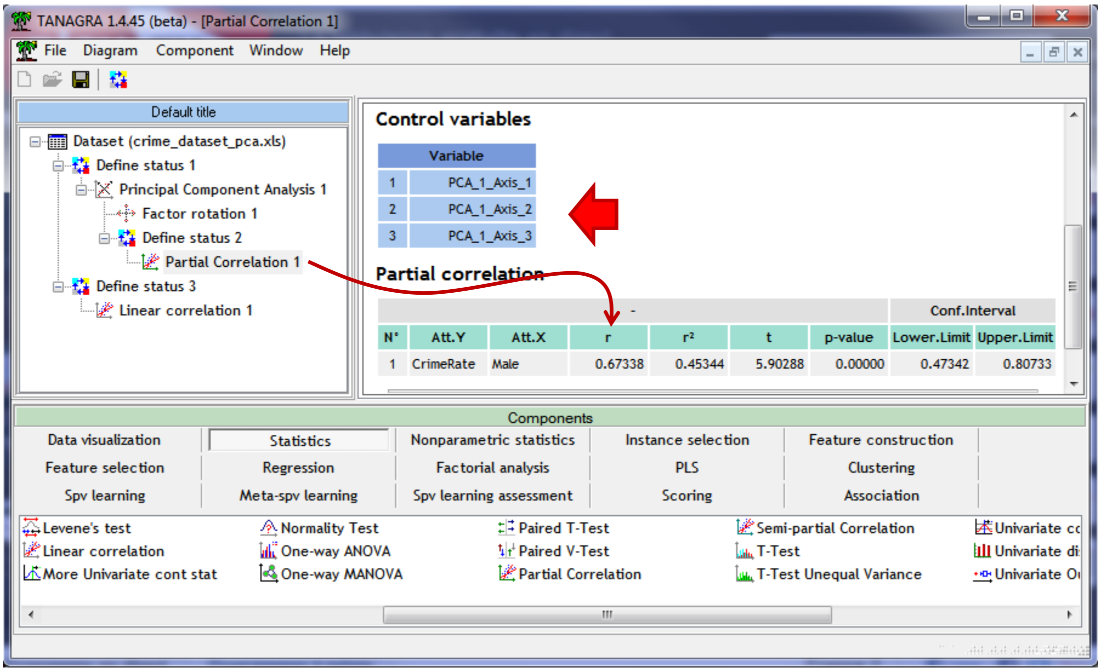
Cross-tab parameters	
Sort results	non
Input list	Target (Y) and input (X)

Results

Y	X	r	r²	t	Pr(>|t|)
CrimeRate	Male	0.2139	0.0458	1.4690	0.1488

But, if we measure the partial correlation, by controlling the three first factors, we have **r(MALE, CRIME / FACT.1, FACT.2, FACT.3) = 0.67338**. It becomes significant[7].

[7] The calculation of the degree of freedom is a bit complicated here because the factors are linear combination of the original variables.



Control variables

	Variable
1	PCA_1_Axis_1
2	PCA_1_Axis_2
3	PCA_1_Axis_3

Partial correlation

			-				Conf.Interval	
N°	Att.Y	Att.X	r	r²	t	p-value	Lower.Limit	Upper.Limit
1	CrimeRate	Male	0.67338	0.45344	5.90288	0.00000	0.47342	0.80733

Undoubtedly, the fourth factor is informative. But its interpretation is not easy.

10 Conclusion

This tutorial relies heavily on the articles cited in the bibliography, especially the Jackson's paper (1993). Our contributions are: (1) the approaches are detailed on a dataset; (2) all the tools used are available (Excel worksheet, R programs), the reader can reproduce all the calculations, he can also apply the programs on another dataset.

Finally, apart from Bartlett's test, all approaches are roughly equivalent for the detection of the number of components on our dataset. But, they bring a different perspective for the same problem. This aspect is really interesting.

11 References

A. Crawford, S. Green, R. Levy, W. Lo, L. Scott, D. Svetina, M. Thompson, "Evaluation of Parallel Analysis for Determining the Number of Factors", in Educational and Psychological Measurement, 70(6), pp. 885-901, 2010.

S. Franklin, D. Gibson, P. Robertson, J. Pohlmann, F. Fralish, "Parallel Analysis: A Method for Determining Significant Principal Components", in Journal of Vegetation Science, Vol. 6, Issue 1, pp. 99-106, 1995.

G. Grossman, D. Nickerson, M. Freeman, "Principal Component Anayses of Assemblage Structure: Utility of Tests based on Eigenvalues", in Ecology, 72(1), pp. 341-347, 1991.

D. Jackson, "Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches", in Ecology, 74(8), pp. 2204-2214, 1993.

P. Neto, D. Jackson, K. Somers, "How Many Principal Components? Stopping Rules for Determining the Number of Aon-trivial Axes Revisited", in Computational Statistics & Data Analysis, 49(2005), pp. 974-997, 2004.

G. Saporta, « Probabilités, analyses des données et Statistiques », Dunod, 2006.