1 Topic

New features for PCA (Principal Component Analysis) in Tanagra 1.4.45 and later: tools for the determination of the number of factors.

Principal Component Analysis (PCA)[1] is a very popular dimension reduction technique. The aim is to produce a few number of factors which summarizes as better as possible the amount of information in the data. The factors are linear combinations of the original variables. From a certain point a view, PCA can be seen as a compression technique.

The determination of the appropriate number of factors is a difficult problem in PCA. Various approaches are possible, it does not really exist a state-of-art method. The only way to proceed is to try different approaches in order to obtain a clear indication about the good solution. We had shown how to program them under R in a recent paper[2]. These techniques are now incorporated into **Tanagra (1.4.45)**. We have also added the KMO index (Measure of Sampling Adequacy – MSA) and the Bartlett's test of sphericity[3] in the Principal Component Analysis tool.

In this tutorial, we present these new features incorporated into Tanagra on a realistic example. To check our implementation, we compare our results with those of SAS PROC FACTOR when the equivalent is available.

2 Dataset

The "beer_pca.xls" data file describes what influences a consumer's choice behavior when he is shopping for beer. The dataset comes from the Dr. Wuensch SPSS-Data Page[4]. Consumers (n = 99) rate on a scale of 0-100 how important he considers each of seven qualities when deciding whether or not to buy the six pack: low COST of the six pack, high SIZE of the bottle (volume), high percentage of ALCOHOL in the beer, the REPUTATION of the brand, the COLOR of the beer, nice AROMA of the beer, and good TASTE of the beer.

This dataset is analyzed in some tutorials available online (e.g. Baillargeon's PCA case study[5]). We have not exactly the same results because the data preparation, here the handling of the missing values, is not the same. This is the reason for which the data really used for each case study is (must be) always distributed on our website.

3 PCA with SAS PROC FACTOR (SAS 9.3)

The analysis is performed in two steps with SAS. First, we perform a standard PCA. We incorporate the MSA index into the output.

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

[2] http://data-mining-tutorials.blogspot.fr/2013/01/choosing-number-of-components-in-pca.html

[3] http://data-mining-tutorials.blogspot.fr/2013/01/pca-using-r-kmo-index-and-bartletts-test.html

[4] Dr Karl Wuensch's SPSS-Data Page, http://core.ecu.edu/psyc/wuenschk/spss/spss-Data.htm

[5] Jacques Baillargeon, « L'analyse en composantes principales ».

```
sas_proc_factor.sas
proc factor data = beer method = principal corr msa;
  var cost size alcohol reputat color aroma taste;
  run;
```

Second, we perform the PCA with the VARIMAX rotation. The aim is to obtain a better association of the variables with the selected factors (we select 2 factors).

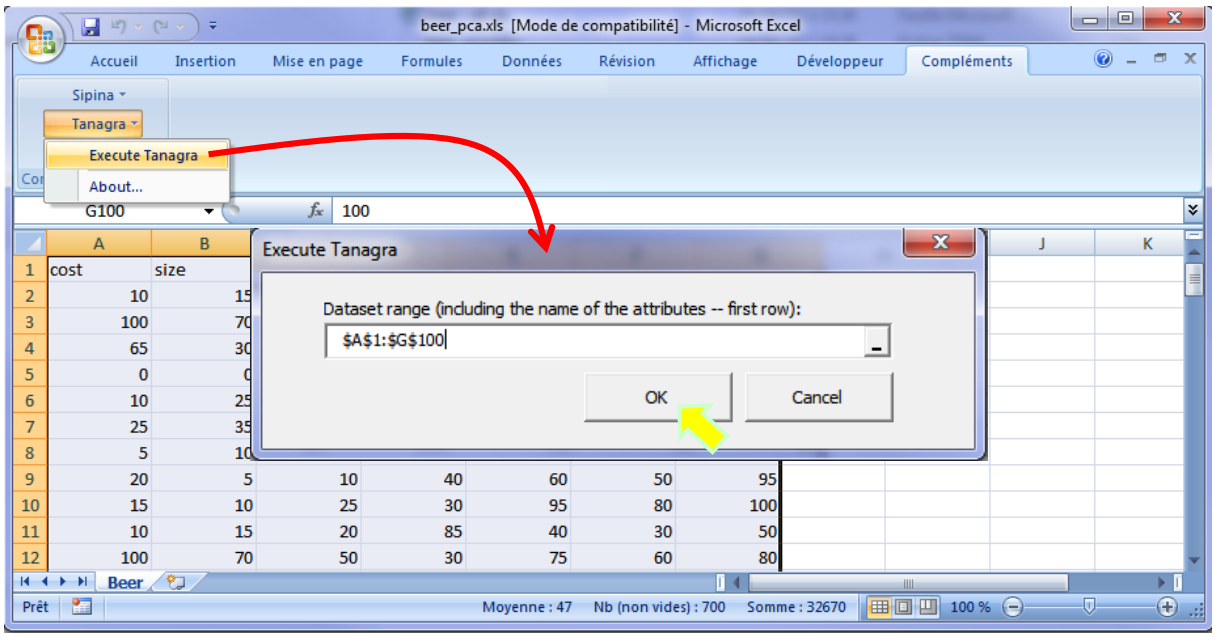
```
sas_proc_factor.sas
proc factor data = beer method = principal corr rotate = varimax n = 2;
  var cost size alcohol reputat color aroma taste;
  run;
```

We describe the results in the next section. We compare them with those of Tanagra. Both these tools use HTML for an attractive presentation of the various tables.

4 Principal Component Analysis with Tanagra

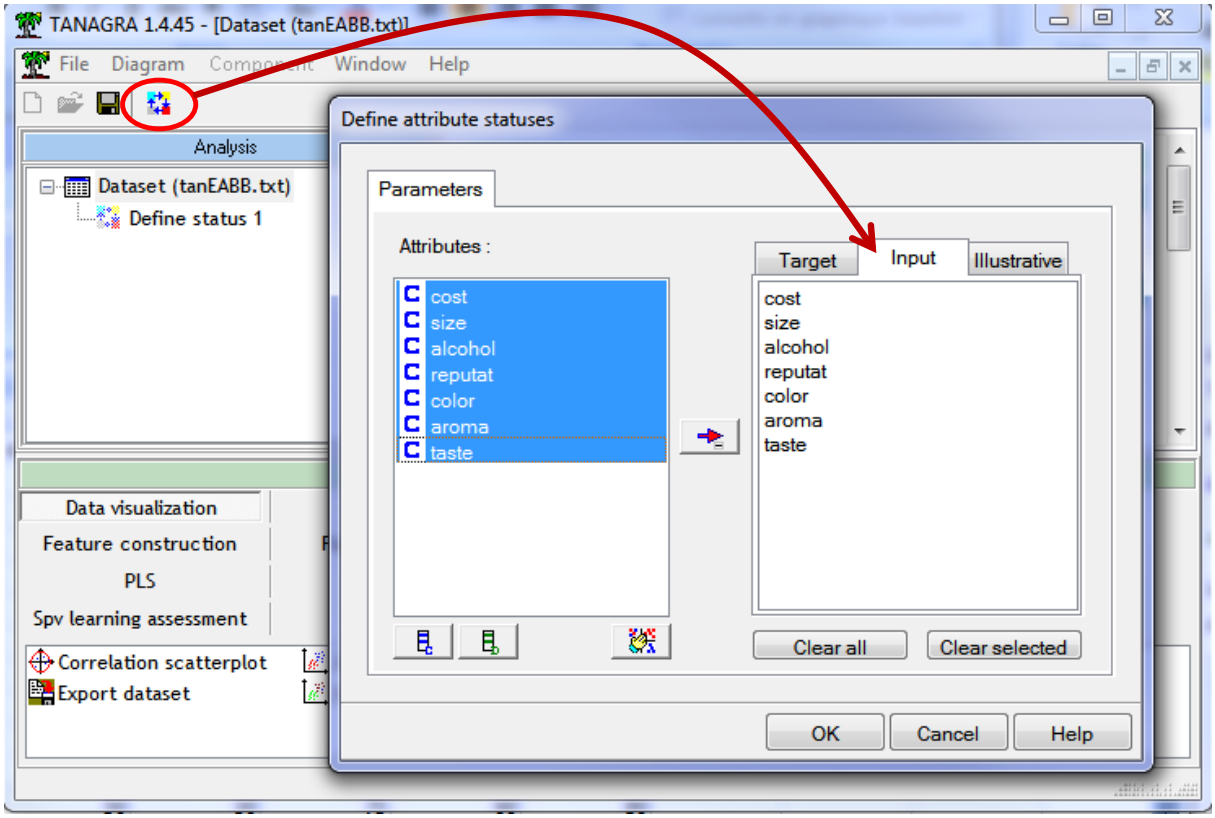
4.1 Importing the dataset

We use the **tanagra.xla**[6] add-in for sending the dataset from Excel to Tanagra.



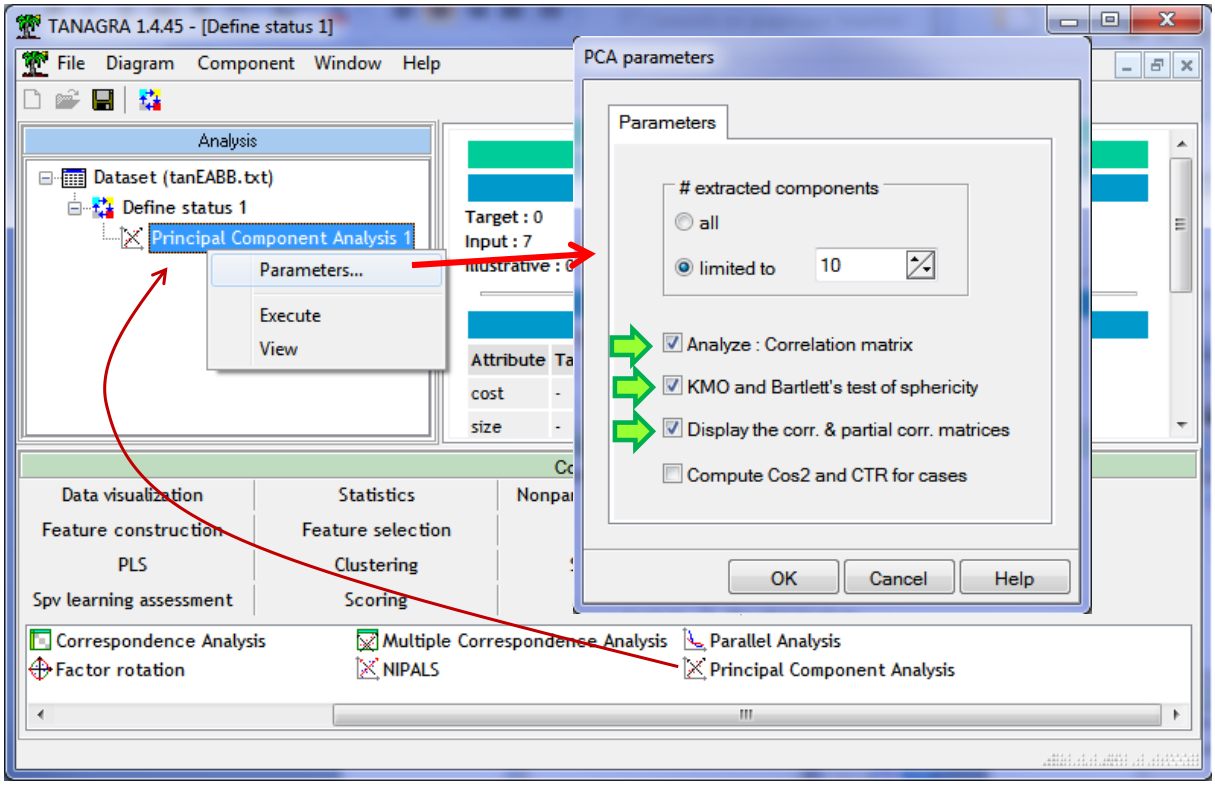
Tanagra is automatically launched and the dataset is loaded. We use the DEFINE STATUS component to define the role on the variables for the analysis.

[6] http://data-mining-tutorials.blogspot.fr/2010/08/tanagra-add-in-for-office-2007-and.html; we can use also other spreadsheet program such as LibreOffice or OpenOffice: http://data-mining-tutorials.blogspot.fr/2011/07/tanagra-add-on-for-openoffice-calc-33.html



4.2 PCA with Tanagra

We add the PRINCIPAL COMPONENT ANALYSIS (FACTORIAL ANALYSIS tab) component into the diagram. We set the following settings:



We perform a PCA based on the correlation matrix (ANALYZE: CORRELATION MATRIX), we ask the display of: the Bartlett's test of sphericity; the KMO index (MSA: measure of sampling adequacy); the correlation and partial correlation matrices. We confirm these settings and we click on VIEW menu. The output is subdivided in several subsections.

4.2.1 Eigenvalues table

This table describes the eigenvalues associated to the factors. We have also the percentage of the total variance (individual and cumulative).

Eigen values

Matrix trace = 7.00

Axis	Eigen value	% explained	Histogram	% cumulated
1	3.306082	47.23%		47.23%
2	2.719206	38.85%		86.08%
3	0.567371	8.11%		94.18%
4	0.193431	2.76%		96.94%
5	0.119954	1.71%		98.66%
6	0.076735	1.10%		99.75%
7	0.017222	0.25%		100.00%
Tot.	7.000000	-	-	-

TANAGRA

Prior Communality Estimates: ONE

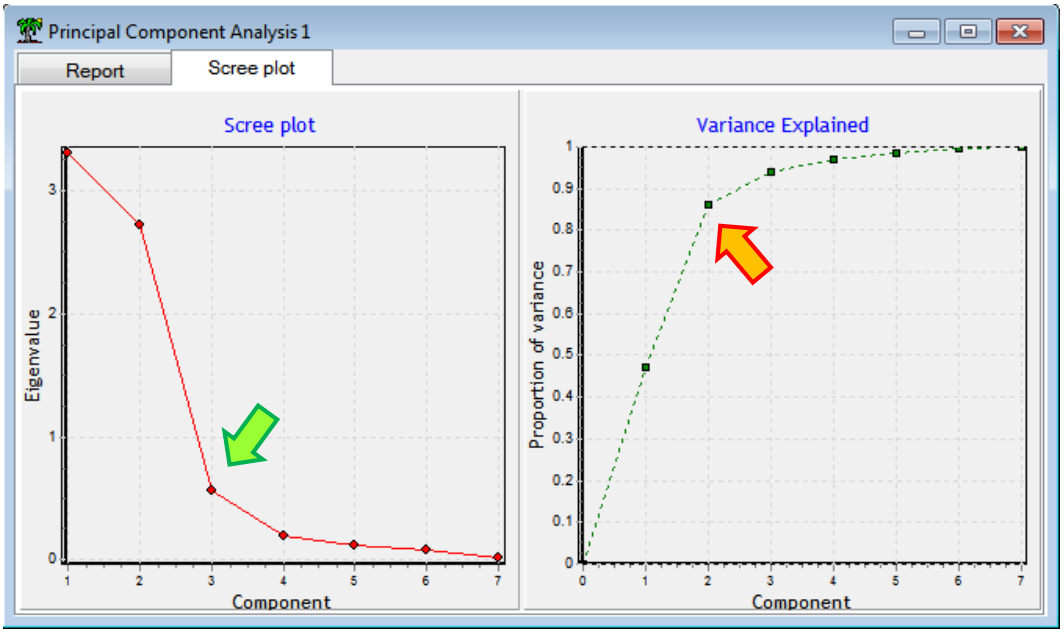
Eigenvalues of the Correlation Matrix: Total = 7 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	3.30608184	0.58687618	0.4723	0.4723
2	2.71920565	2.15183494	0.3885	0.8608
3	0.56737072	0.37393939	0.0811	0.9418
4	0.19343133	0.07347757	0.0276	0.9694
5	0.11995376	0.04321918	0.0171	0.9866
6	0.07673458	0.05951246	0.0110	0.9975
7	0.01722212		0.0025	1.0000

SAS - PROC FACTOR

4.2.2 Scree plot and percentage of the total variable

To complete the previous table, the scree plot and the graph of the cumulative percentage of the variance according the factors are displayed in the "Scree plot" tab of the visualization window.



According to the scree plot, there is an elbow at the third eigenvalue. But we select only two factors. Indeed, when we consider the graph of the cumulative percentage, we observe that the additional variance explained by the third factor and the following ones can be neglected.

4.2.3 Tools for the determination of the right number of components

Tools based on statistical approaches enable to detect the right number of factors (it does not mean that they are more efficient). We have presented them in a previous tutorial[7]. They are incorporated into Tanagra. We can use them only for PCA based on the correlation matrix.

Three rules for the determination of the right number of components are combined:

Significance of Principal Components

Global critical values	
Kaiser-Guttman	1
Karlis-Saporta-Spinaki	1.49487

Eigenvalue table - Test for significance

Eigenvalues - Significance		
Axis	Eigenvalue	Broken-stick critical values
1	3.306082	2.592857
2	2.719206	1.592857
3	0.567371	1.092857
4	0.193431	0.759524
5	0.119954	0.509524
6	0.076735	0.309524
7	0.017222	0.142857

1. Kaiser-Guttman rule. A factor is relevant if its eigenvalue is higher than 1.
2. Karlis-Saporta-Spinaki. This is a more restrictive variant of the Kaiser-Guttman rule. It considers the dataset characteristics i.e. the n:p ratio, n is the number of instance, p is the number of variables.
3. Legendre-Legendre, the broken stick method. It defines a critical value according the number of factors that we want to select.

The intensity of the red color depends on the number of detection rules activated. Here, we observe that selecting two components seems to be the right choice. From the third factor, no detection rules are activated.

4.2.4 Bartlett's test of sphericity

Bartlett's test of sphericity

Bartlett's test	
|CORR.MATRIX|	0.0001564015
CHISQ	831.0325
d.f.	21
p-value	2.376082E-162

The Bartlett's test compares the observed correlation matrix to the identity matrix[8]. It enables to check if there is at least one relevant factor. The Bartlett's test has a strong drawback. It tends to be always statistically significant when the number of instances 'n' increases. But we incorporate this test into Tanagra because it is widely referenced in the literature.

For our dataset, we observe that we reject the null hypothesis. It means that the correlation matrix differs significantly from the identity matrix. We can extract at least one interesting factor from the PCA. But we cannot determine the right number of components with this tool.

4.2.5 KMO (Kaiser-Mayer-Olkin) index – Measure of Sampling Adequacy (MSA)

The KMO index (or MSA) measures the ability of the PCA to summarize the information provided by the original variables in a few number of factors. It compares the correlation matrix with the partial correlation matrix (that we describe later).

[7] http://data-mining-tutorials.blogspot.fr/2013/01/choosing-number-of-components-in-pca.html; a detailed description of the approaches is available.

[8] http://data-mining-tutorials.blogspot.fr/2013/01/pca-using-r-kmo-index-and-bartletts-test.html

Kaiser's Measure of Sampling Adequacy (MSA) TANAGRA

Overall MSA = 0.503618									
cost	0.4035957	size	0.5245393	alcohol	0.5511566	reputat	0.3689421	color	0.847379
aroma	0.5606907	taste	0.427865						

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.50361804

cost	size	alcohol	reputat	color	aroma	taste
0.40359574	0.52453925	0.55115660	0.36894206	0.84737899	0.56069066	0.42786498

SAS - PROC FACTOR

We obtain MSA = 0.503618. It seems disappointing if we refer to the interpretation tables that we can read online[9]. But this is not really surprising. We cannot obtain a good compression of the information because we have initially 7 variables and the PCA provides 2 factors in order to summarize them. **It does not mean that our results are not interesting**.

About the MSA per variable, we observe that MSA(REPUTAT) = 0.3689 is the lowest. This variable seems to be the least correlated to the others.

4.2.6 Loadings and communalities

The factor loadings correspond to the correlation of the variables with the components. The squared correlation is the percentage of variance of the variable explained by the factor. By cumulating them (communality), we get the quality of representation of variables on selected factors. In the last row, we have the percentage of the total variance that accounted for the factors. Here, the two first factors explain 86% of the total variance.

Factor Loadings [Communality Estimates]

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
cost	0.5051	26 % (26 %)	0.8164	67 % (92 %)
size	0.2258	5 % (5 %)	0.9507	90 % (95 %)
alcohol	0.5977	36 % (36 %)	0.7568	57 % (93 %)
reputat	-0.7411	55 % (55 %)	0.1238	2 % (56 %)
color	0.9061	82 % (82 %)	-0.1877	4 % (86 %)
aroma	0.7861	62 % (62 %)	-0.5135	26 % (88 %)
taste	0.8090	65 % (65 %)	-0.5115	26 % (92 %)
Var. Expl.	3.3061	47 % (47 %)	2.7192	39 % (86 %)

TANAGRA

2 factors will be retained by the MINEIGEN criterion.

Factor Pattern

		Factor1	Factor2
cost	cost	0.50507	0.81643
size	size	0.22575	0.95075
alcohol	alcohol	0.59771	0.75683
reputat	reputat	-0.74110	0.12380
color	color	0.90611	-0.18768
aroma	aroma	0.78614	-0.51354
taste	taste	0.80901	-0.51154

Variance Explained by Each Factor

Factor1	Factor2
3.3060818	2.7192057

Final Communality Estimates: Total = 6.025287

cost	size	alcohol	reputat	color	aroma	taste
0.92165047	0.95488671	0.93004894	0.56454880	0.85624989	0.88173669	0.91616599

SAS - PROC FACTOR

The first factor indicates the preference of the people for the taste, color and aroma. These are the esthetes of the beer. The second factor shows the people which want to drink a lot (alcohol, size) at the lower cost. We note that the REPUTAT is not really related to one of these two factors

[9] http://peoplelearn.homestead.com/Topic20-FACTORanalysis3a.html

(Communality[Reputat] = 56%, while other communalities are greater than 85 %). It is not correlated to the other variables as we see later. We could guess it when we had calculated its MSA above.

4.2.7 Correlation and partial correlation matrices

The correlation matrix reports the correlation between each couple of variables. Because the number of variables is low (p = 7) for our dataset, we can study it easily. We distinguish essentially two groups of variables: (cost, size alcohol) and (color, aroma, taste). The correlations confirm that 'reputation' (REPUTAT) is a bit apart.

Matrices

Correlations

	cost	size	alcohol	reputat	color	aroma	taste
cost	1.00000	0.87839	0.87702	-0.17478	0.32089	-0.02764	0.05398
size	0.87839	1.00000	0.82367	-0.06123	0.01441	-0.28624	-0.30751
alcohol	0.87702	0.82367	1.00000	-0.36051	0.39770	0.09768	0.05580
reputat	-0.17478	-0.06123	-0.36051	1.00000	-0.52380	-0.52151	-0.62650
color	0.32089	0.01441	0.39770	-0.52380	1.00000	0.82324	0.80487
aroma	-0.02764	-0.28624	0.09768	-0.52151	0.82324	1.00000	0.86607
taste	0.05398	-0.30751	0.05580	-0.62650	0.80487	0.86607	1.00000

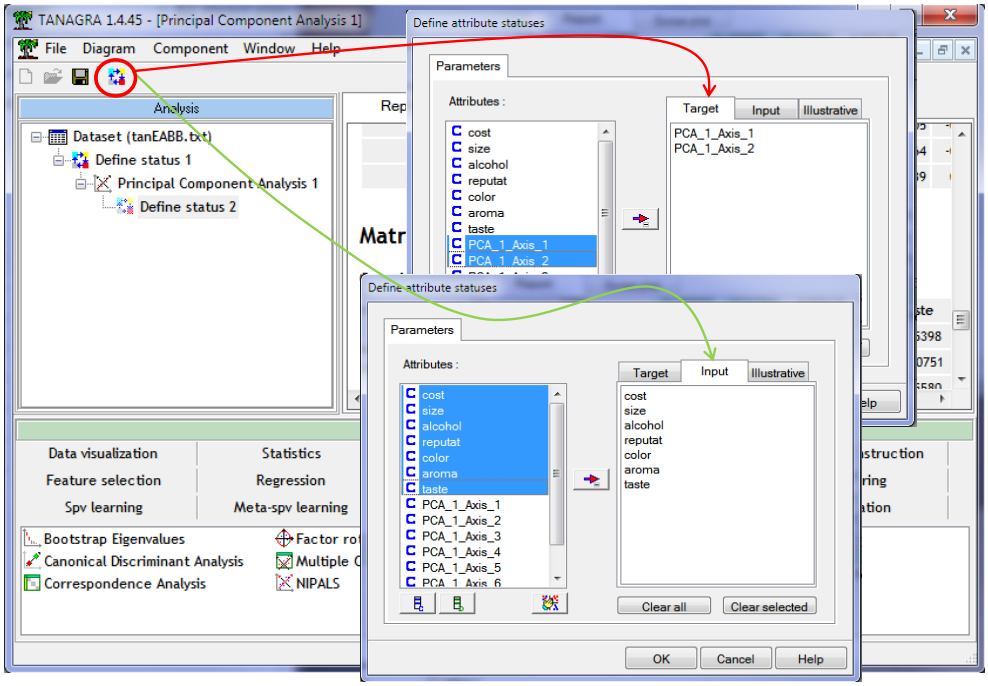
Partial Correlations Controlling all other Variables

	cost	size	alcohol	reputat	color	aroma	taste
cost	1.00000	0.80374	0.61583	0.67853	0.03276	-0.59860	0.79655
size	0.80374	1.00000	-0.10712	-0.49420	-0.07244	0.37208	-0.65765
alcohol	0.61583	-0.10712	1.00000	-0.63063	0.31559	0.41932	-0.59998
reputat	0.67853	-0.49420	-0.63063	1.00000	0.17771	0.40573	-0.76699
color	0.03276	-0.07244	0.31559	0.17771	1.00000	0.35445	0.26851
aroma	-0.59860	0.37208	0.41932	0.40573	0.35445	1.00000	0.66426
taste	0.79655	-0.65765	-0.59998	-0.76699	0.26851	0.66426	1.00000

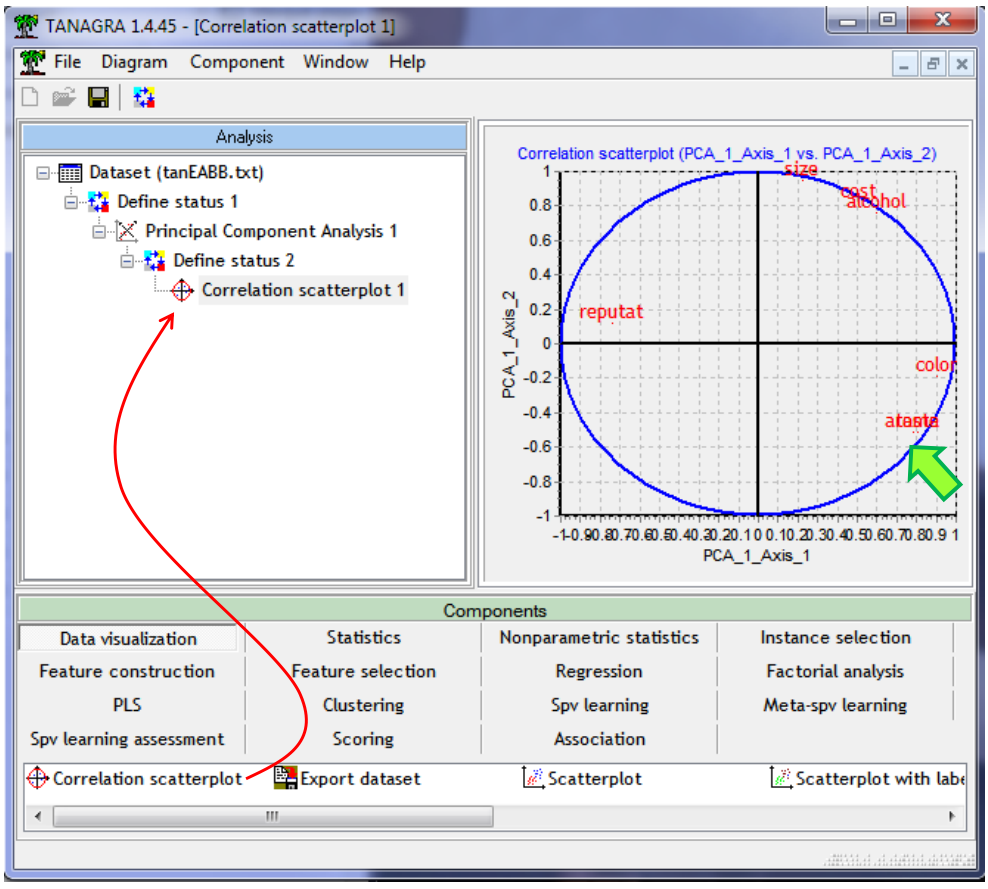
The partial correlation matrix reports the correlation between each couple of variables by controlling the influence of the others. We can thus detect the misleading nature of certain correlations. For instance, the correlation between 'size' and 'alcohol' seems very significant (r = 0.8237). But, when we compute the partial correlation, we obtain a low value (-0.10712), which is not statistically significant. It means that the correlation is in fact influenced by the other variables, mainly COST if we further analyze the relation.

4.2.8 Correlation circle

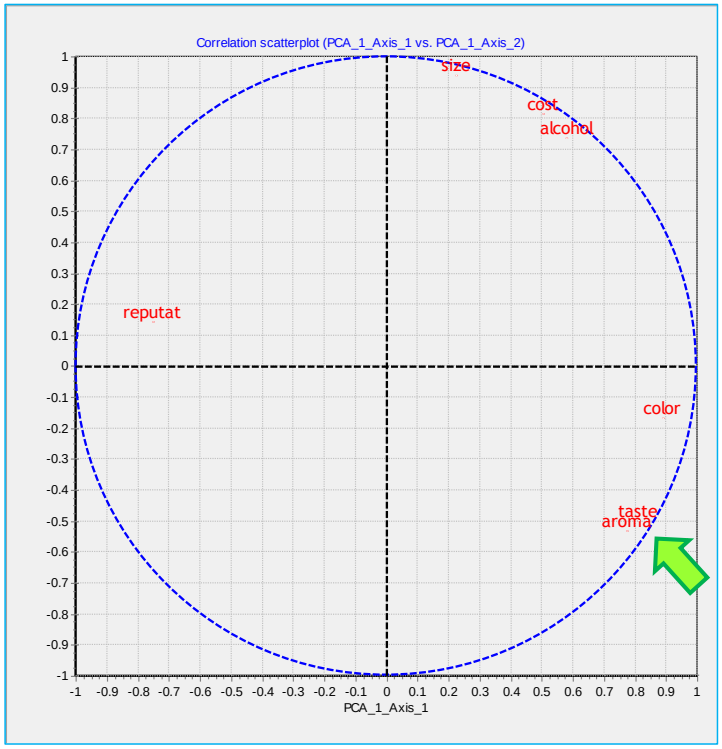
The correlation circle shows graphically the correlation of the variables with a couple of factors. We add the DEFINE STATUS component into the diagram. We set the two first factors as TARGET, the initial variables as INPUT.



Then we add the CORRELATION SCATTERPLOT component (DATA VISUALIZATION tab).



We observe the influence of the variables on the various factors. We use jittering in order to overcome the overlapping of some variables (COMPONENT / JITTER menu).



5 Additional tools for PCA

5.1 Parallel analysis

Parallel analysis is a method for determining the number of factors to retain from PCA. Essentially, the program works by creating a random dataset with the same numbers of observations (n) and variables (p) as the original data. A correlation matrix is computed from the randomly generated dataset and then eigenvalues (q_k) of the correlation matrix are computed. When the k^{th} (λ_k) eigenvalue from the PCA is significantly larger than the eigenvalue from the random data (e.g. the quantile $q_k^{0.95}$ at the 95% level), we can conclude that the k^{th} component is valid[10].

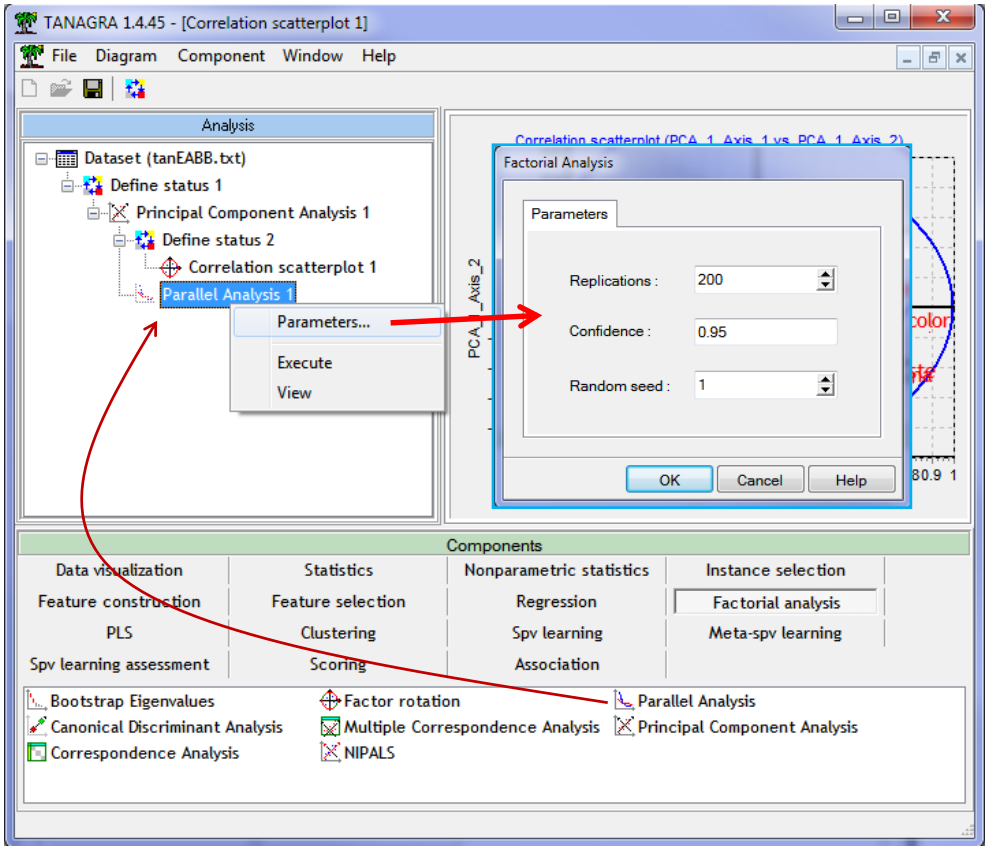
Instead of generating random data, Tanagra uses a randomization approach. The protocol is the following: we randomize the values within the variables in the dataset; we perform the PCA on the randomized dataset; we repeat T times this process. So if the observed eigenvalue (from the PCA of the original dataset) is significantly larger than the eigenvalue from the randomization process (e.g. the quantile at the 95% level), we validate the factor.

The advantage of this strategy is that the same tool can be used for the factor analysis for qualitative variables i.e. the Multiple Correspondence Analysis[11].

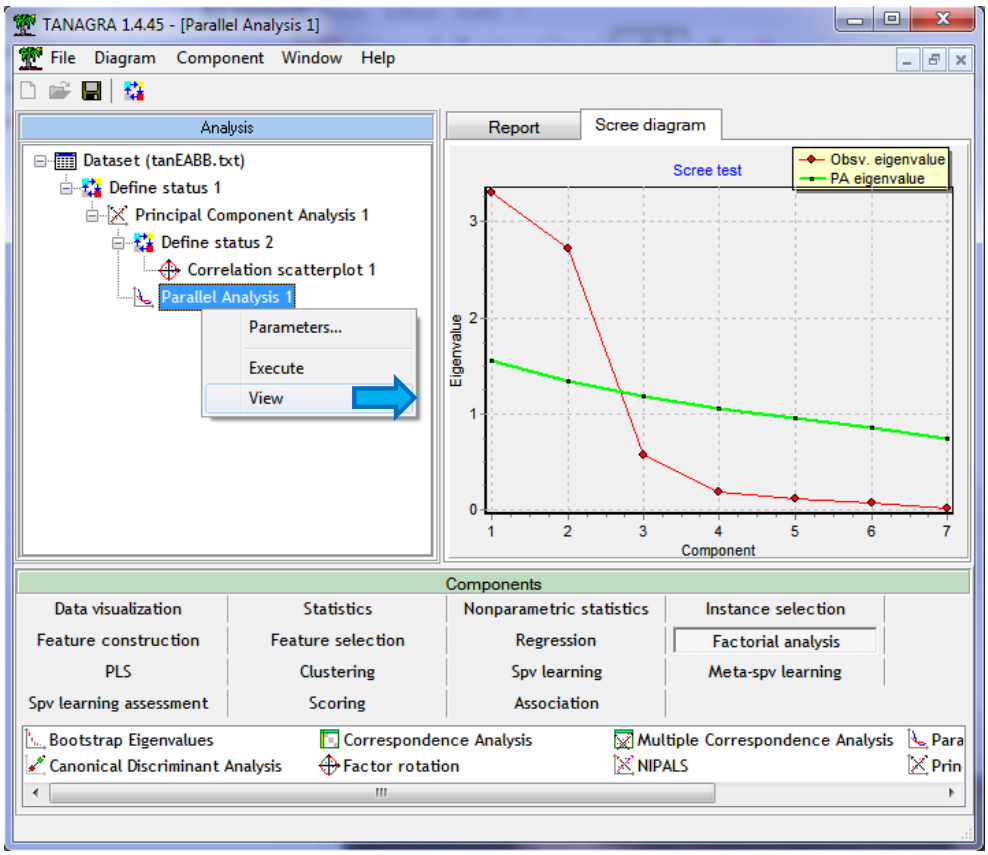
We insert the PARALLEL ANALYISIS tool (FACTORIAL ANALYSIS tab) after the principal component analysis into the diagram. We click on the PARAMETERS menu.

[10] http://www.ats.ucla.edu/stat/stata/faq/parallel.htm

[11] http://data-mining-tutorials.blogspot.fr/2009/04/multiple-correspondence-analysis-mca.html



The process is repeated T = 200 times (Replications) for the calculation of the critical value, this last one corresponds to the quantile at the 95% level (confidence), the random seed used for the random number generator is 1. We confirm and we click on the VIEW menu.



In the scree diagram above, the observed eigenvalue are in red, the critical values in green. Clearly, for our dataset, we can retain 2 factors. The values are reported into the REPORT tab.

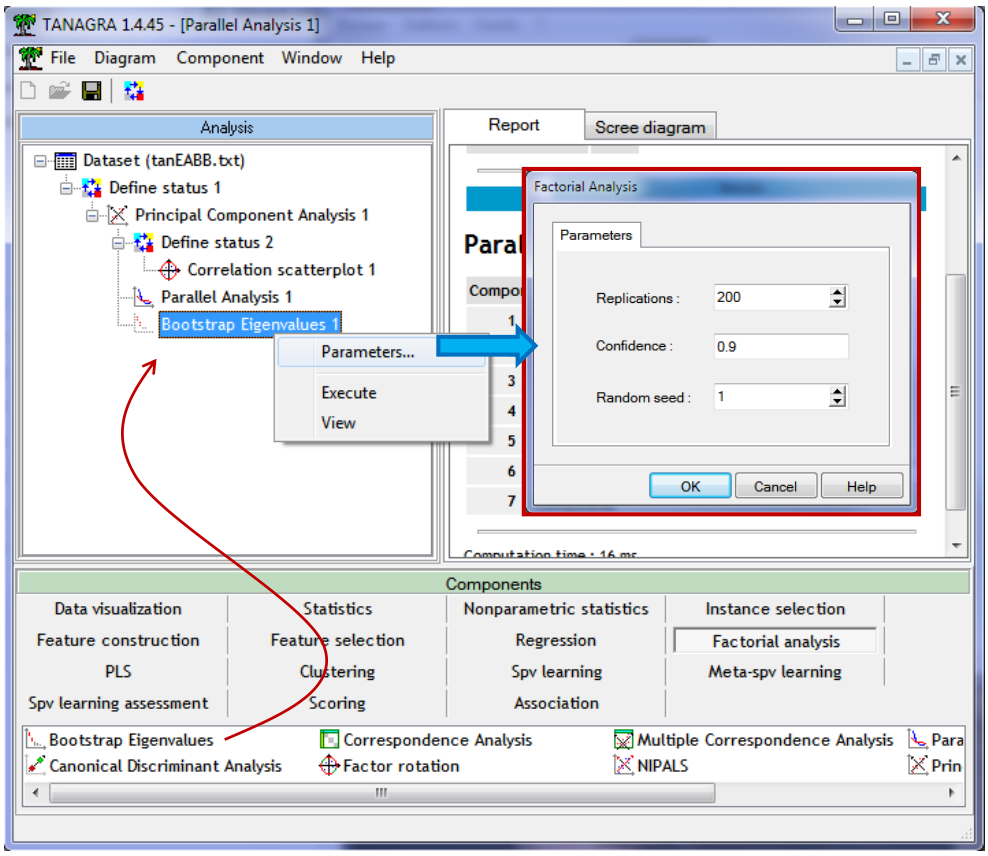
Parallel Analysis

Component	Eigenvalue	(0.95) Critical value
1	3.306082	1.552704
2	2.719206	1.339758
3	0.567371	1.185519
4	0.193431	1.050049
5	0.119954	0.957981
6	0.076735	0.851874
7	0.017222	0.738887

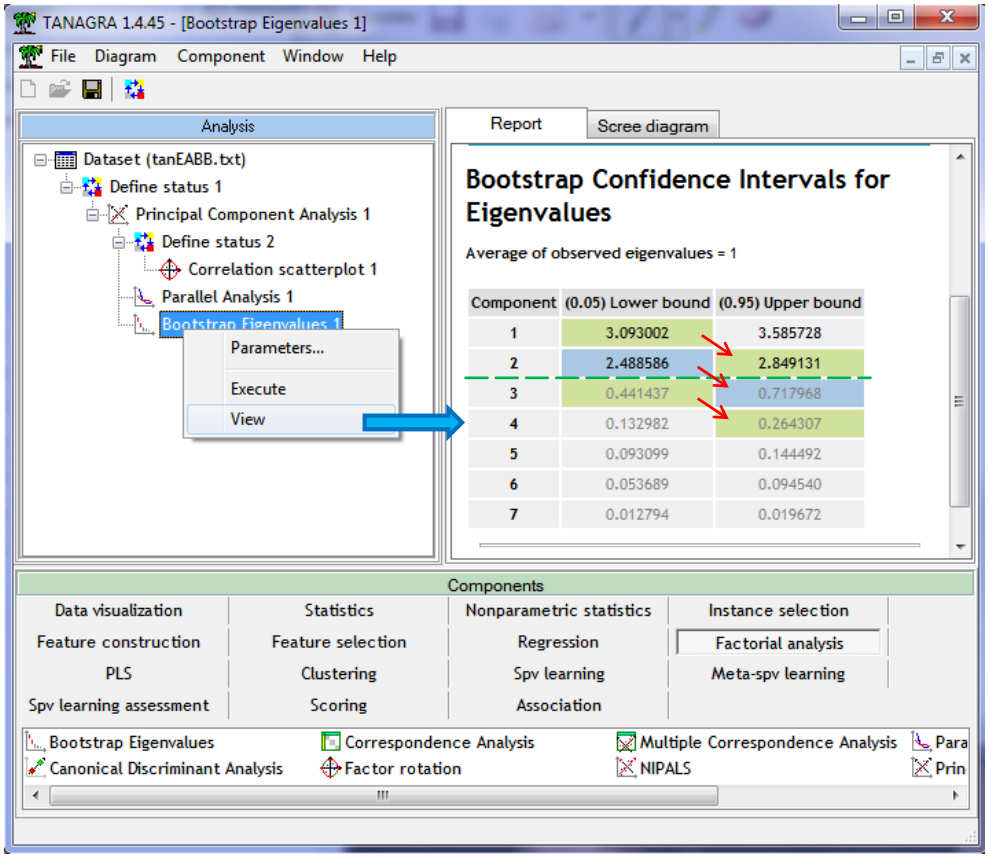
5.2 Bootstrap confidence interval for eigenvalues

A tool enables to compute the percentile bootstrap confidence intervals of the eigenvalues. Two rules are used for the detection of the relevant factors: the lower limit of the confidence interval is higher than 1; the lower limit of the k^{th} eigenvalue is higher than the upper limit of the $(k+1)^{th}$ eigenvalue i.e. two successive confidence intervals are not overlapped. This tool can be applied to both principal component analysis and multiple correspondence analysis.

We insert the BOOTSTRAP EIGENVALUES tool (FACTORIAL ANALYSIS tab) after the PCA component. We set the following settings: REPLICATIONS = 200, CONFIDENCE = 0.90 (confidence level for the calculation of the confidence intervals), RANDOM SEED = 1.



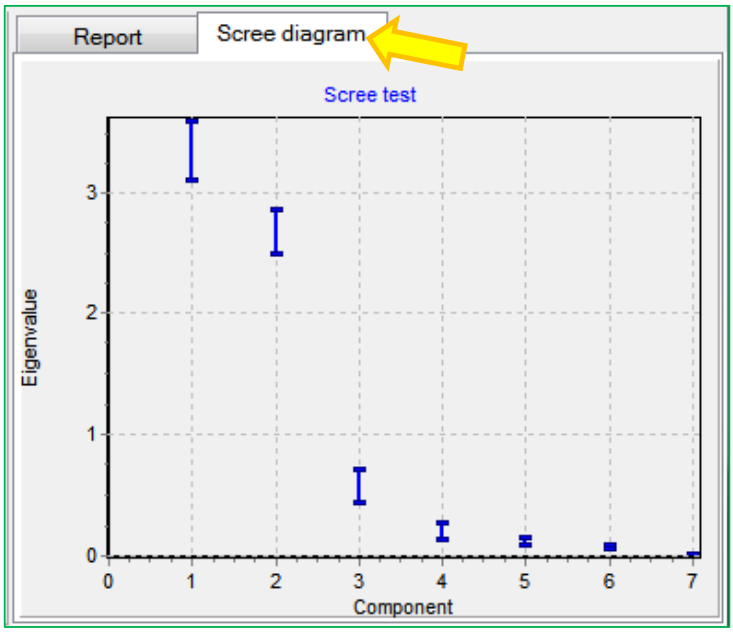
We confirm and we click on the VIEW menu.



First result, the lower limits of the two first factors are higher than 1 (3.09 and 2.49). The right solution is to retain two factors according to this detection rule.

Second result, confidence intervals of the 1st and the 2nd eigenvalues are not overlapped, we have the same situation for the 2nd and the 3rd eigenvalues, for the 3rd and the 4th factors. Selecting three factors seems better here. But we observe that the third eigenvalue is low. The upper limit of its confidence interval is lower than 1. Finally, we can omit the third factor.

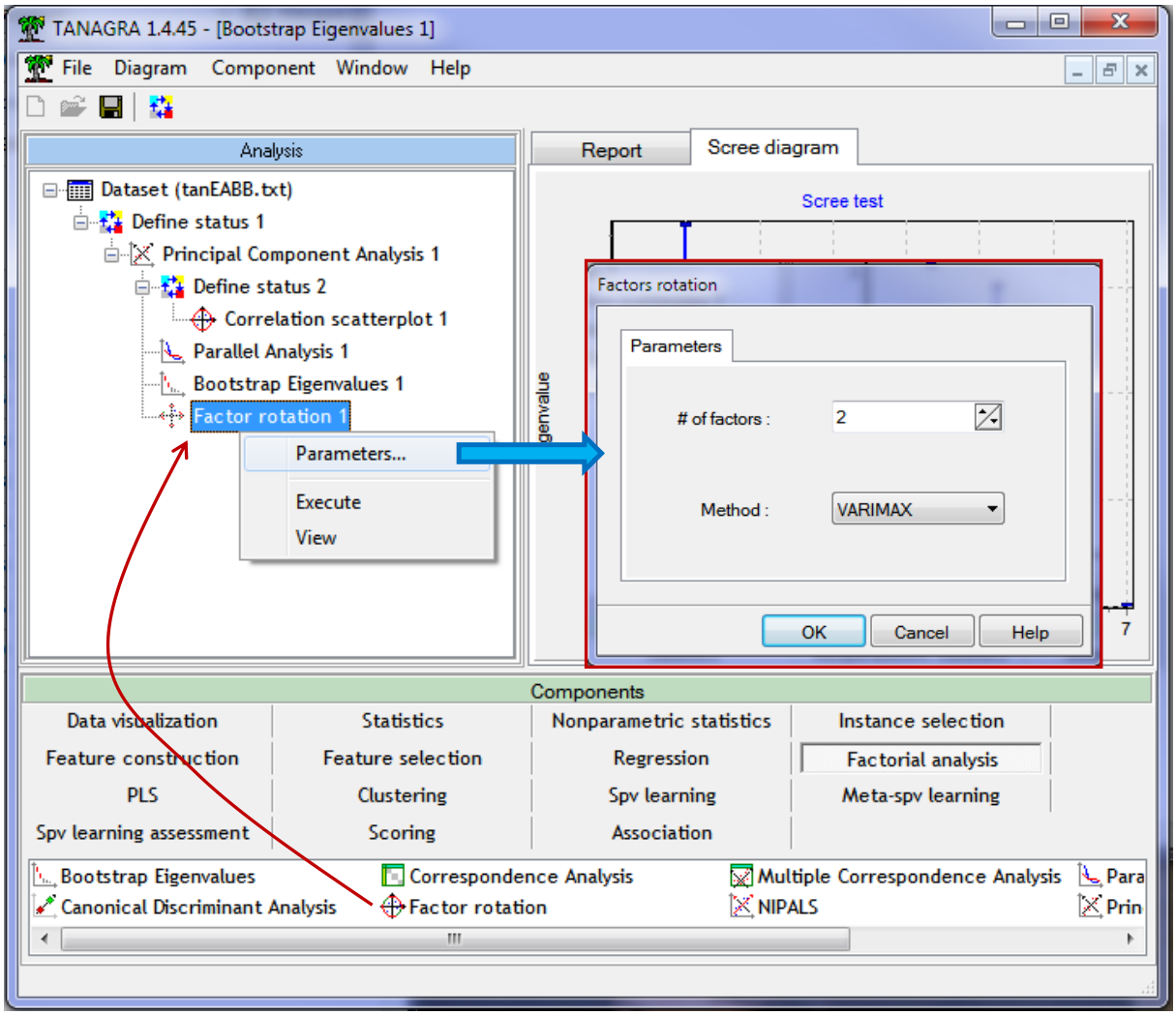
The confidence intervals can be visualized into the SCREE DIAGRAM tab.



5.3 VARIMAX rotation

The factor rotation[12] method enables to strengthen the association of the variables with one of the factors. The goal is to make easier the interpretation of the results. We use the VARIMAX[13] orthogonal rotation for our dataset. With this approach: (1) the percentage of variance explained by the selected factors is not modified; (2) the factors remain uncorrelated.

We add the FACTOR ROTATION tool (FACTORIAL ANALYSIS tab) into our diagram. We perform the rotation for the two first factors.



We obtain the loadings and the communalities after and before the rotation. The two first factors still explain 86% of the total variance. The interpretation provided previously (from unrotated factors) is reinforced. We observe that REPUTAT is now more associated to the first factor here. The consumers which are concerned to (COLOR, AROMA, and TASTE) are less sensitive to the (REPUTATION).

[12] http://www.utd.edu/~herve/Abdi-rotations-pretty.pdf

[13] http://en.wikipedia.org/wiki/Varimax_rotation

Rotated Factor Loadings

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
cost	0.1101	1 % (1 %)	0.9537	91 % (92 %)
size	-0.1998	4 % (4 %)	0.9565	91 % (95 %)
alcohol	0.2193	5 % (5 %)	0.9391	88 % (93 %)
reputat	-0.7234	52 % (52 %)	-0.2030	4 % (56 %)
color	0.8999	81 % (81 %)	0.2153	5 % (86 %)
aroma	0.9299	86 % (86 %)	-0.1306	2 % (88 %)
taste	0.9497	90 % (90 %)	-0.1191	1 % (92 %)
Var. Expl.	3.2000	46 % (46 %)	2.8253	40 % (86 %)

vs. Unrotated Factor Loadings

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
cost	0.5051	26 % (26 %)	0.8164	67 % (92 %)
size	0.2258	5 % (5 %)	0.9507	90 % (95 %)
alcohol	0.5977	36 % (36 %)	0.7568	57 % (93 %)
reputat	-0.7411	55 % (55 %)	0.1238	2 % (56 %)
color	0.9061	82 % (82 %)	-0.1877	4 % (86 %)
aroma	0.7861	62 % (62 %)	-0.5135	26 % (88 %)
taste	0.8090	65 % (65 %)	-0.5115	26 % (92 %)
Var. Expl.	3.3061	47 % (47 %)	2.7192	39 % (86 %)

TANAGRA

Rotated Factor Pattern

		Factor1	Factor2
cost	cost	0.11011	0.95369
size	size	-0.19982	0.95653
alcohol	alcohol	0.21929	0.93913
reputat	reputat	-0.72343	-0.20298
color	color	0.89994	0.21530
aroma	aroma	0.92988	-0.13064
taste	taste	0.94973	-0.11911

Variance Explained by Each Factor

Factor1	Factor2
3.2000307	2.8252567

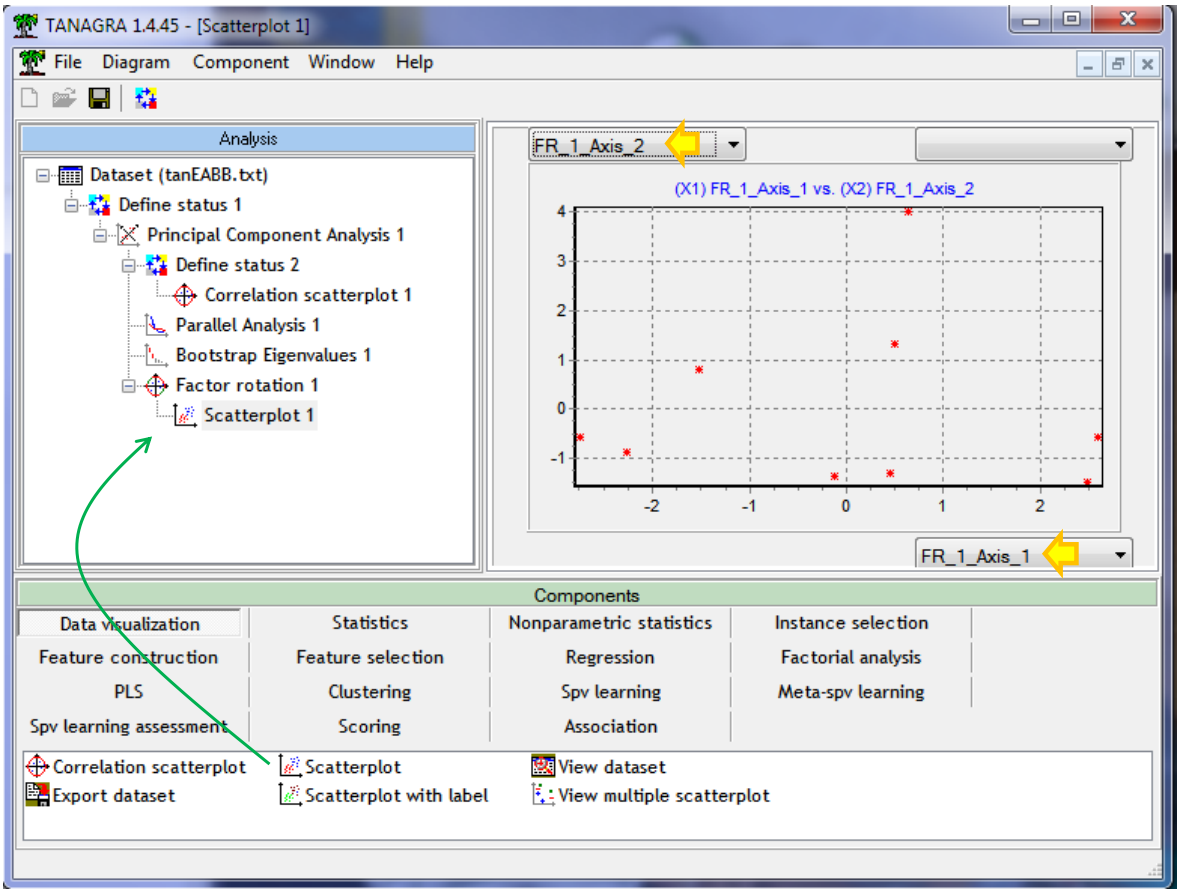
Final Communality Estimates: Total = 6.025287

cost	size	alcohol	reputat	color	aroma	taste
0.92165047	0.95488671	0.93004894	0.56454880	0.85624989	0.88173669	0.91616599

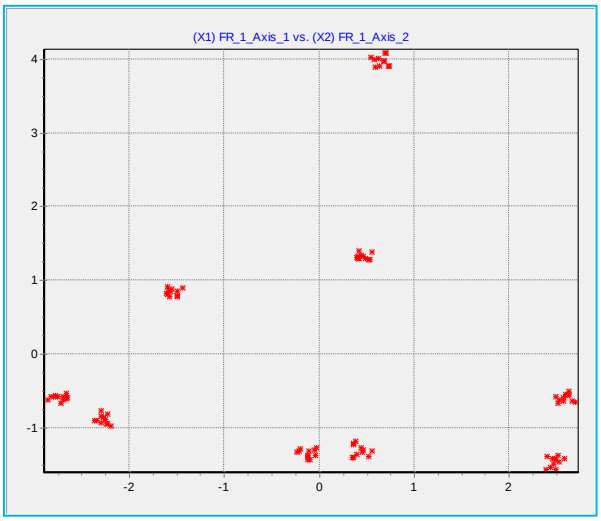
SAS - PROC FACTOR

5.4 Plotting of individuals

Plotting of individuals enables to detect the individuals which are similar (or dissimilar) according to their characteristics. We add the SCATTERPLOT tool (DATA VISUALIZATION tab).

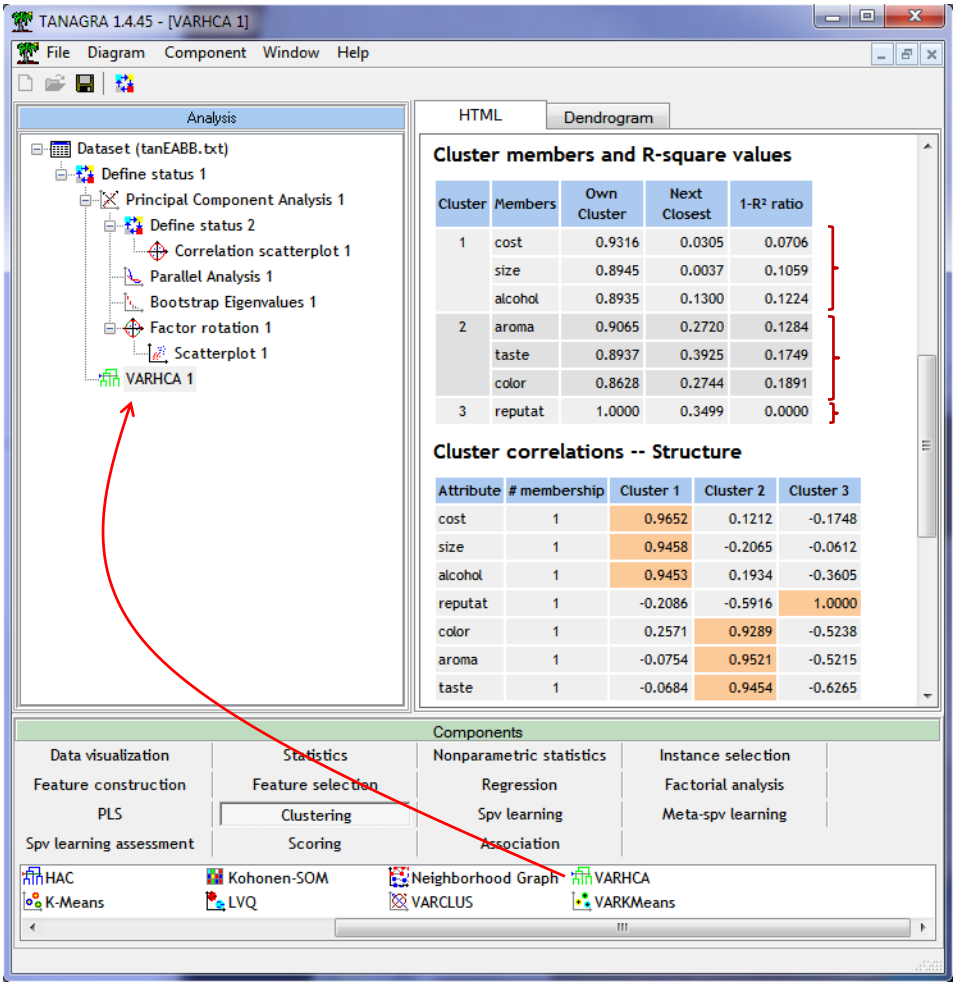


Each point is located according to its coordinates on the rotated factors [FR_1_AXIS_1, FR_2_AXIS_1] (after rotation). We activate the COMPONENT / JITTER menu to overcome the overplotting.



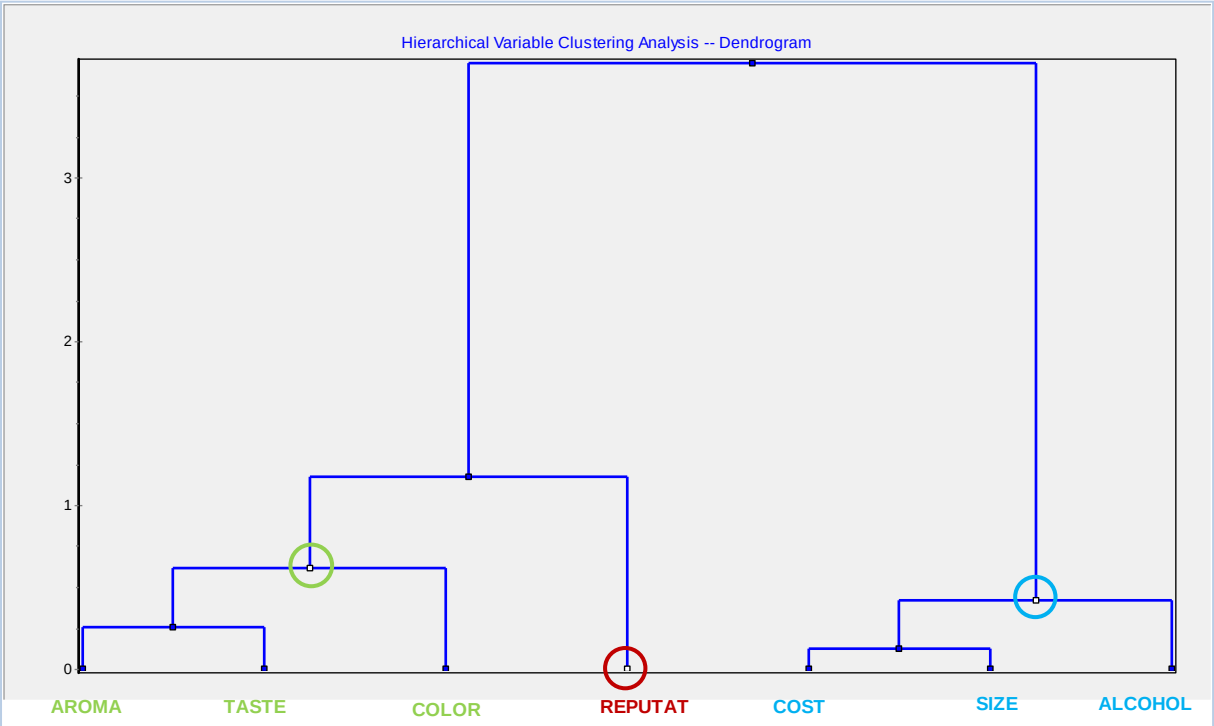
6 Clustering variables

To confirm (or to refute) the previous analysis, we perform a clustering of variables using the VARHCA tool. Unlike the PCA, we have not the orthogonality constraint. We will see if this modifies the content of the results.



We add the VARHCA[14] tool (CLUSTERING tab) into the diagram. We click on the VIEW menu.

Indeed, three dimensions appear from the clustering process. The following dendrogram enables to visualize the associations (the calculations are based on the squared correlation) between the variables.



7 Conclusion

In this tutorial, we describe some new features intended for the principal components analysis in **Tanagra** (**1.4.45 and later**). Of course, they are available in various free tools. Our main contribution is of have incorporated them in a unified and coherent way in a single software package.

[14] http://data-mining-tutorials.blogspot.fr/2008/11/variable-clustering-varclus.html