

Subject

Implementing the Principal Component Analysis (PCA) with TANAGRA.

The PCA belongs to the factor analysis approaches. It is used to discover the underlying structure of a set of variables. It reduces attribute space from a larger number of variables to a smaller number of factors (dimensions). It is an unsupervised procedure i.e. it does not assume a dependent variable is specified¹.

In this tutorial, we show how to implement this approach and how to interpret the results with Tanagra.

Dataset

We use the AUTOS_ACP.XLS dataset from the famous SAPORTA's² book (Tableau 17.1, page 428). The interest of this dataset is that we can compare our results with those described in the book (pages 177 to 181). We simply show the sequence of operations and the reading of the results tables in this tutorial. About the detailed interpretation, it is best to refer to the book.

The data table is the following:

Modele	CYL	PUISS	LONG	LARG	POIDS	V-MAX	FINITION	PRIX	R-POID.PUIS
Alfasud TI	1350	79	393	161	870	165	B	30570	11.01
Audi 100	1588	85	468	177	1110	160	TB	39990	13.06
Simca 1300	1294	68	424	168	1050	152	M	29600	15.44
Citroen GS Club	1222	59	412	161	930	151	M	28250	15.76
Fiat 132	1585	98	439	164	1105	165	B	34900	11.28
Lancia Beta	1297	82	429	169	1080	160	TB	35480	13.17
Peugeot 504	1796	79	449	169	1160	154	B	32300	14.68
Renault 16 TL	1565	55	424	163	1010	140	B	32000	18.36
Renault 30	2664	128	452	173	1320	180	TB	47700	10.31
Toyota Corolla	1166	55	399	157	815	140	M	26540	14.82
Alfetta-1.66	1570	109	428	162	1060	175	TB	42395	9.72
Princess-1800	1798	82	445	172	1160	158	B	33990	14.15
Datsun-200L	1998	115	469	169	1370	160	TB	43980	11.91
Taunus-2000	1993	98	438	170	1080	167	B	35010	11.02
Rancho	1442	80	431	166	1129	144	TB	39450	14.11
Mazda-9295	1769	83	440	165	1095	165	M	27900	13.19
Opel-Rekord	1979	100	459	173	1120	173	B	32700	11.20
Lada-1300	1294	68	404	161	955	140	M	22100	14.04

The first column is the label of the examples. The active variables, used during the computation of the axes, are in green; the supplementary (illustrative) variables, used only for the interpretation of the results, are in blue. Compared to the original dataset, we create a new variable, R-POID.PUIS which is the ratio between the horsepower and the weight of the vehicles. Its values are low for sports cars.

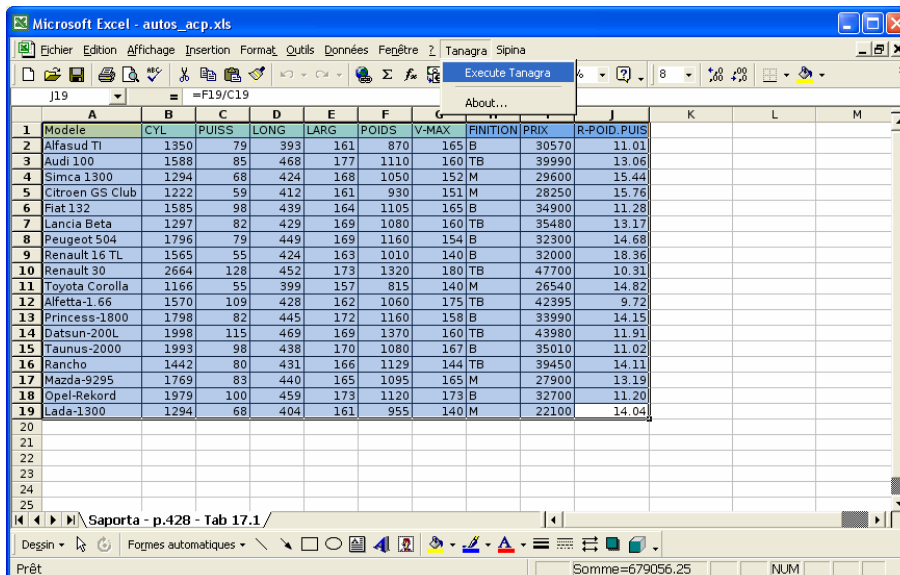
¹ <http://faculty.chass.ncsu.edu/garson/PA765/factor.htm>

² G. SAPORTA, « Probabilités, Analyse de données et Statistique », TECHNIP, 2006 (in French).

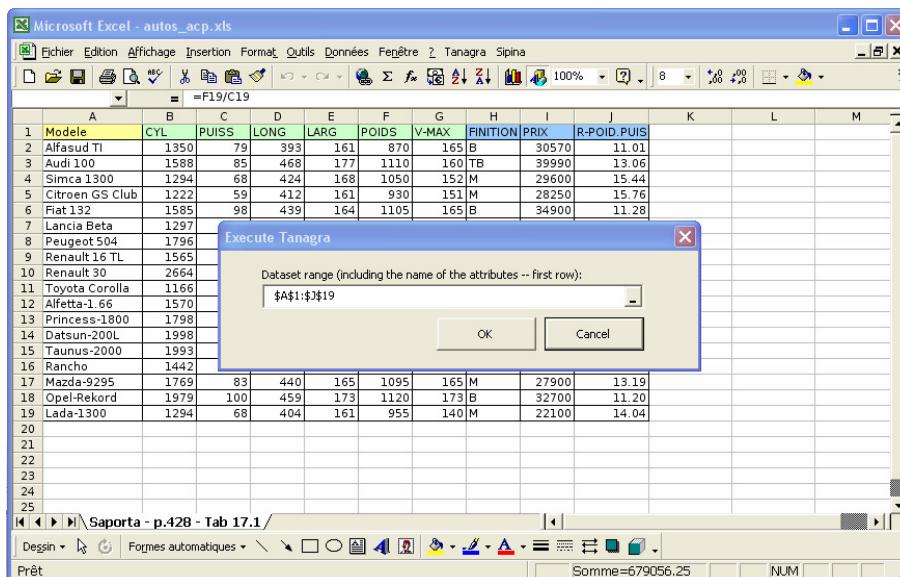
Principal Component Analysis with TANAGRA

Creating a diagram

We can launch Tanagra from Excel using an add-on³. We select the range of cells that contains the dataset. Then we click on the TANAGRA / EXECUTE TANAGRA menu.



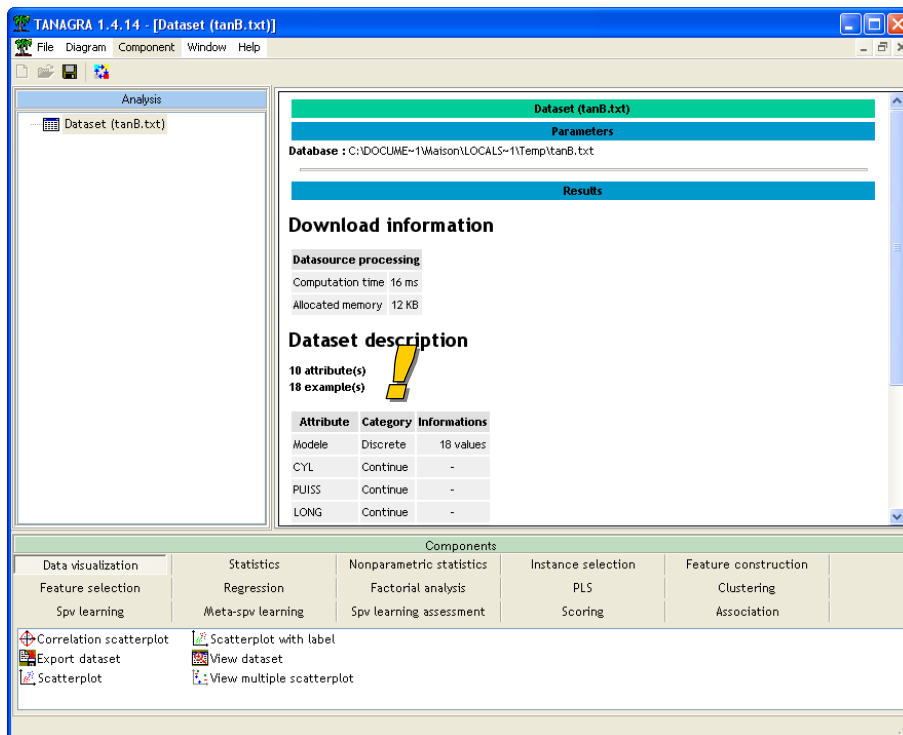
A dialog box appears. We click on OK if the selection is right.



TANAGRA is launched. We check that there are 18 examples and 10 variables⁴.

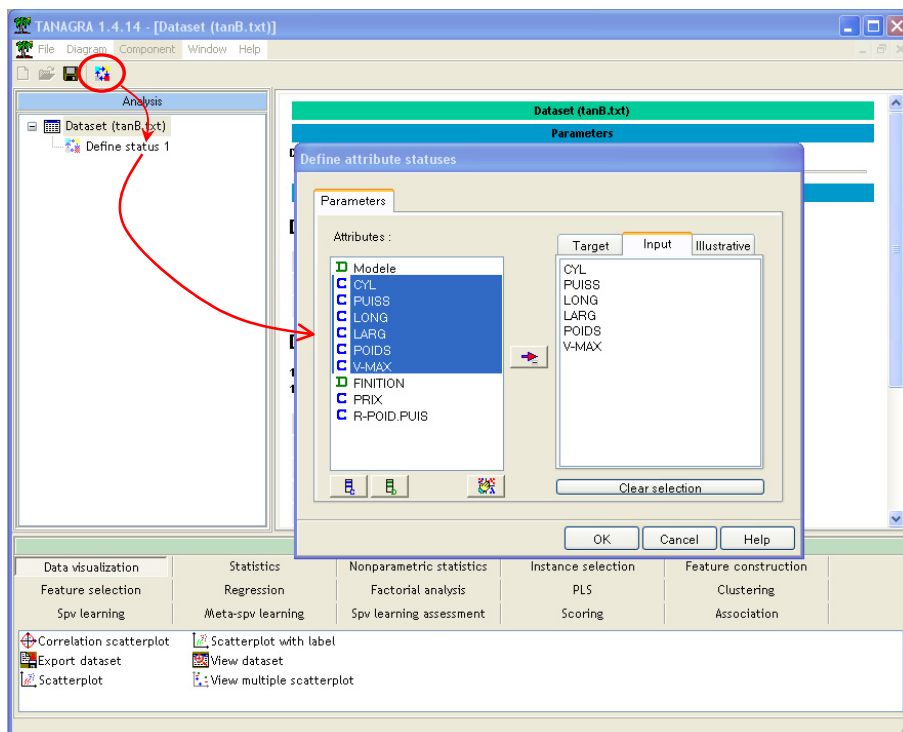
³ <http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html>; we can also import the XLS data file even if Excel is not installed on our computer, see <http://data-mining-tutorials.blogspot.com/2008/10/excel-file-format-direct-importation.html>.

⁴ Tanagra does not handle the label column. The first column is thus counted as a categorical variable in our dataset. It is not a problem. Tanagra considers that each label is a value of the variable... but in this case, the number of different values is limited to 255.



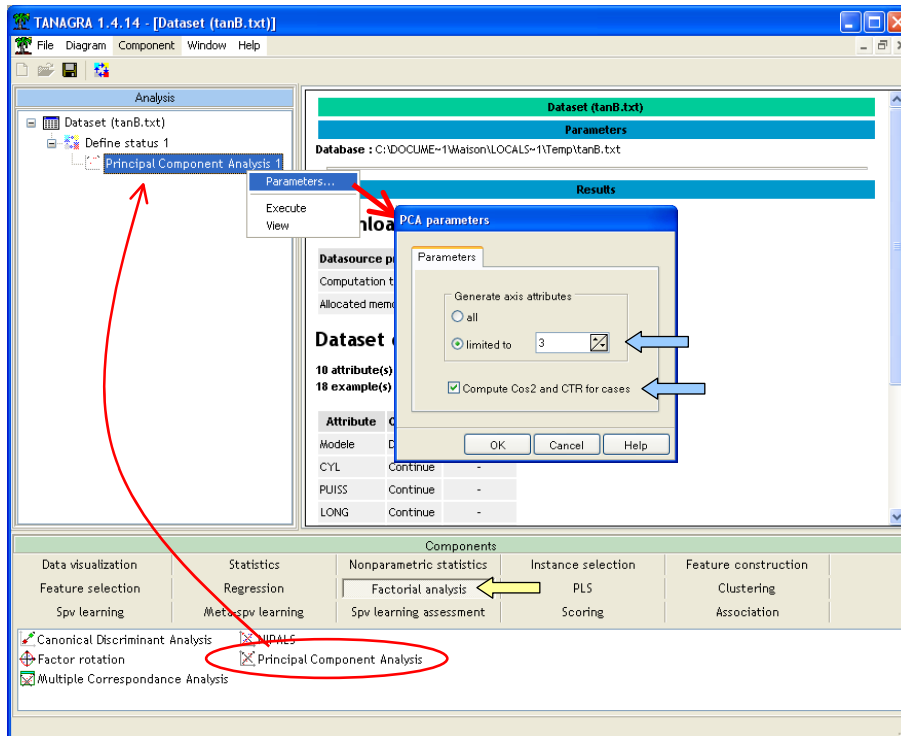
Principal component analysis

First, we must define the types of variables. We insert the DEFINE STATUS component into the diagram by clicking the shortcut in the toolbar. We set as INPUT the active variables. We see below how to use the illustrative variables.

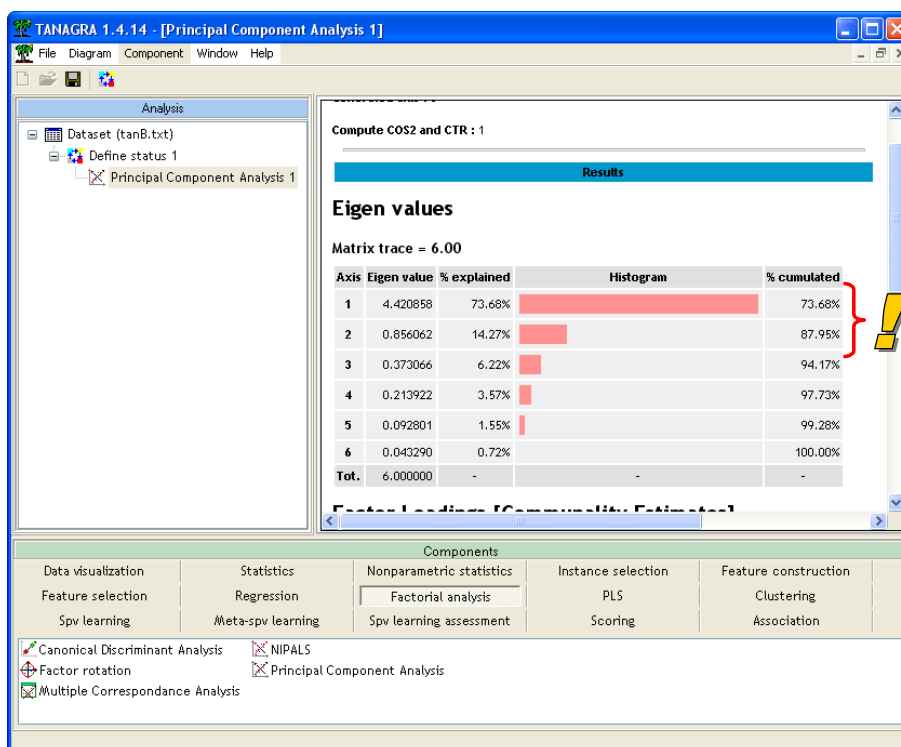


Then we add the PRINCIPAL COMPONENT ANALYSIS component (FACTORIAL ANALYSIS tab). We click on the PARAMETERS menu: we set the number of dimensions to calculate (3 factors); we want to compute the COS² (contribution of dimensions to points or squared correlations) and the CTR (contributions of points to dimensions).

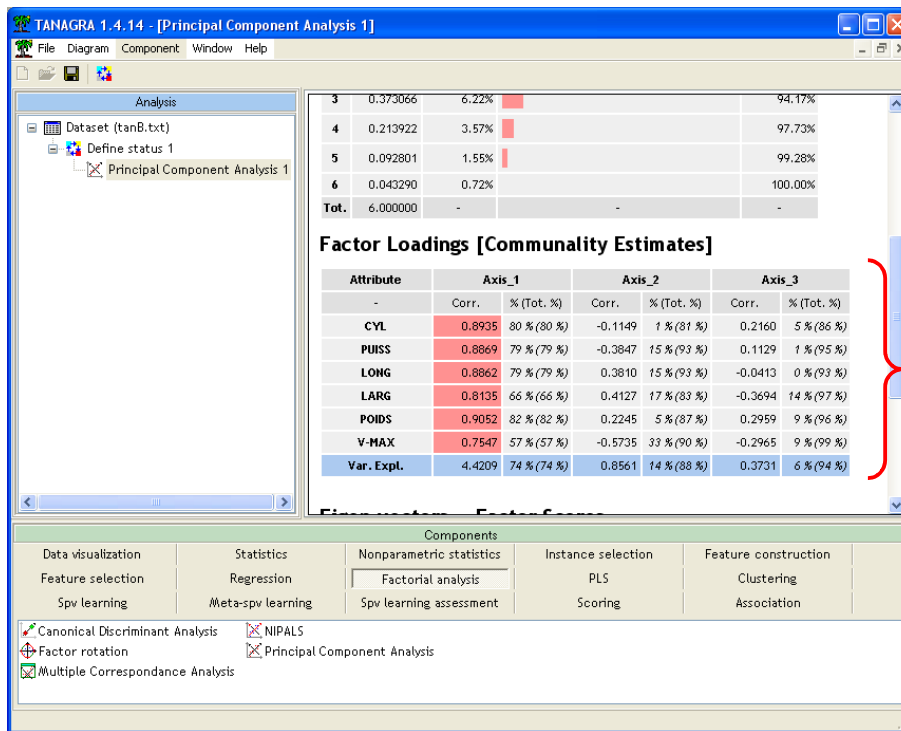
We click on the VIEW menu in order to obtain the results.



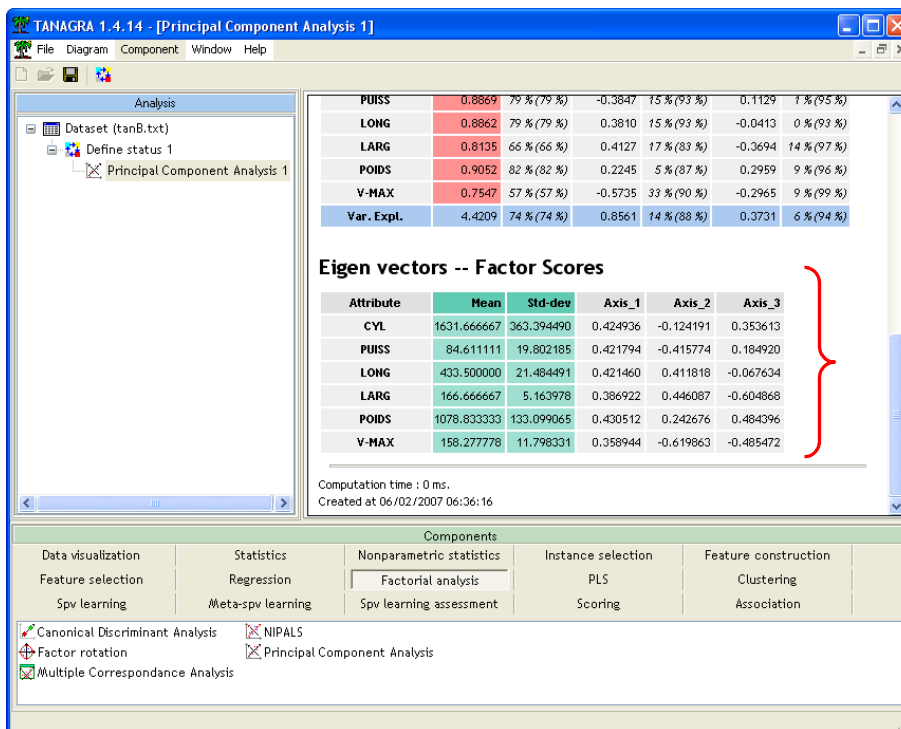
Eigenvalues. The first table describes the eigenvalues. They reflect the importance of each dimension. We see that the two first factors reflect the 87.95% of the available information.



Correlations between factors and active variables. The second part describes the correlations and the COS2 (squared correlations) in percentage and cumulated percentage between the variables and the factors.

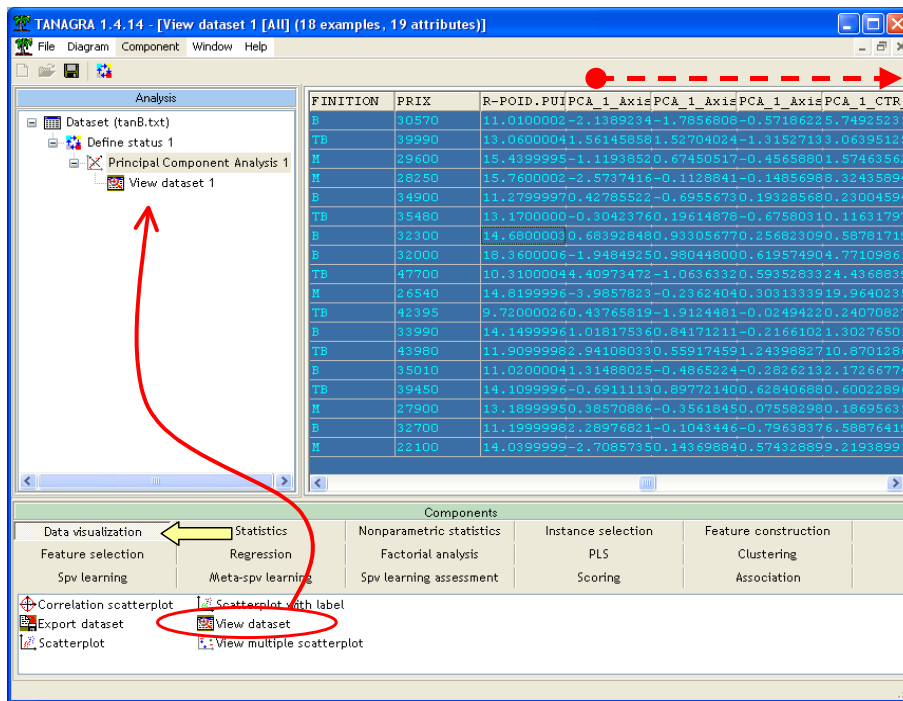


Factor scores. A third table displays the coefficients for the computation of the factor scores of the individuals. The mean and the standard deviation used during the computation of the correlation matrix are given. You must perform the transformation before applying the coefficients on a new instance.

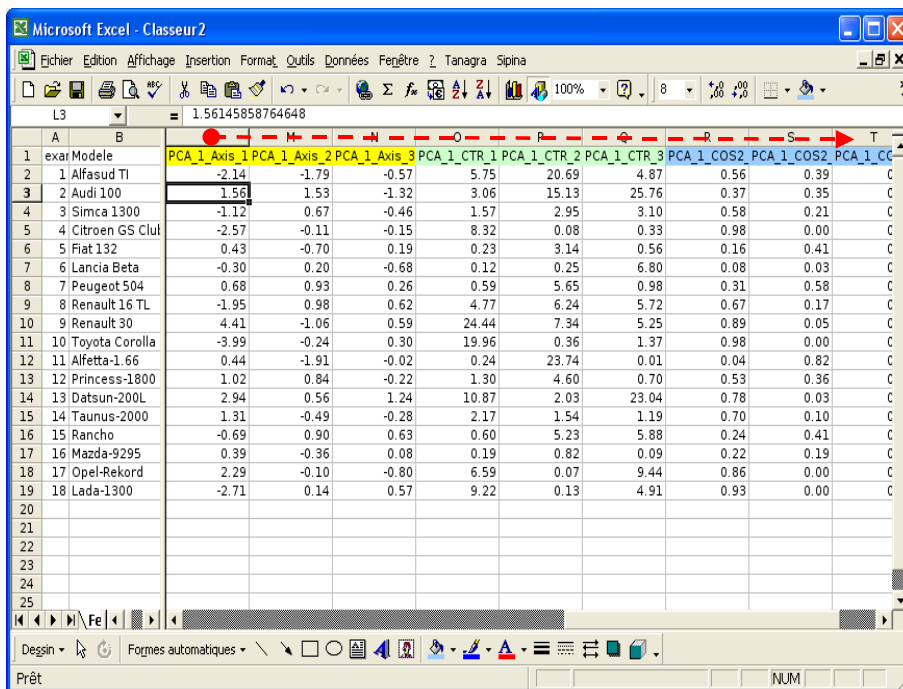


But we can also get the computed factor scores for all the individuals. Indeed, the PCA component adds automatically new columns to the current dataset. They are available in the subsequent part of the diagram. According our settings (see above), the COS2 and contributions are also computed.

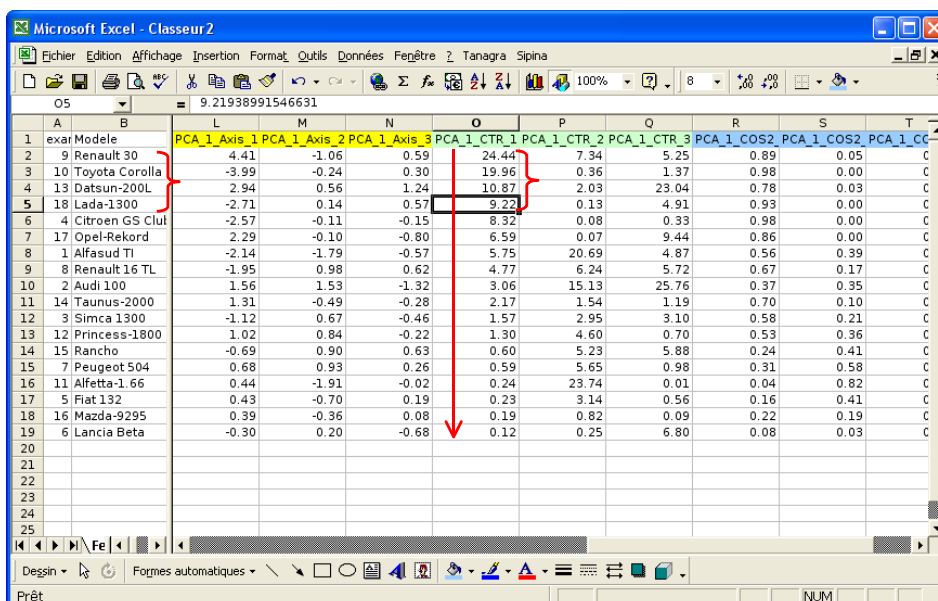
In order to view the related values, we add the VIEW DATASET component (DATA VISUALIZATION tab) into the diagram. We click on the VIEW menu.



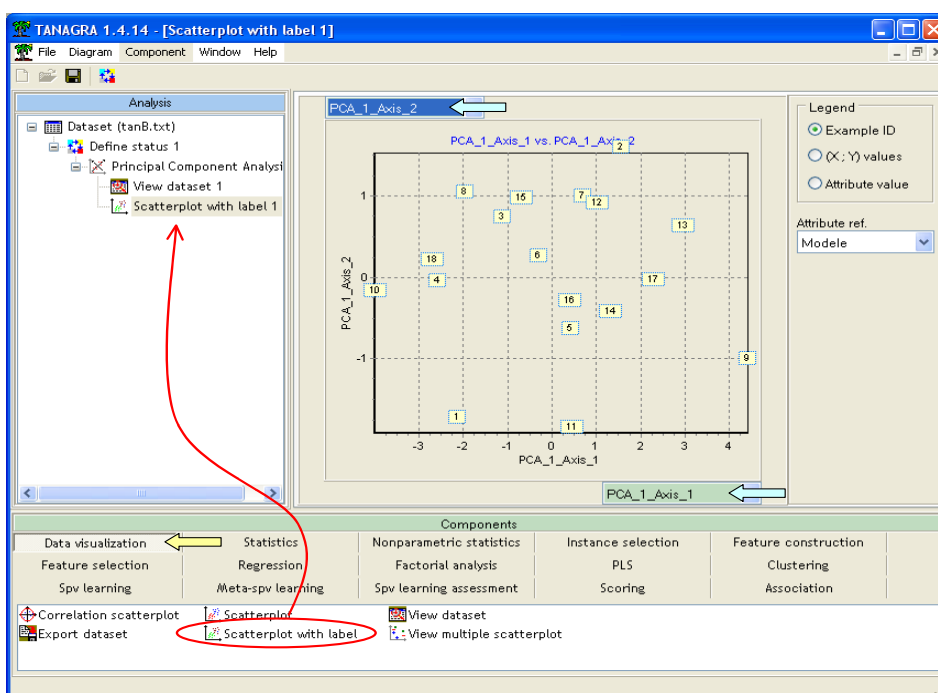
TANAGRA uses a scientific format. A simple way to obtain a more convenient presentation is to copy/paste the values into a spreadsheet (COMPONENT / COPY RESULTS menu) as the follows.



In addition, with the spreadsheet, we have multiple sorting options that allow to highlight the relevant information. For instance, according the contributions, we note that the first dimension is mainly defined by the opposition between RENAULT 30 TS + DATSUN-200 L ("big" cars) and TOYOTA COROLLA + LADA-1300 ('small" cars).

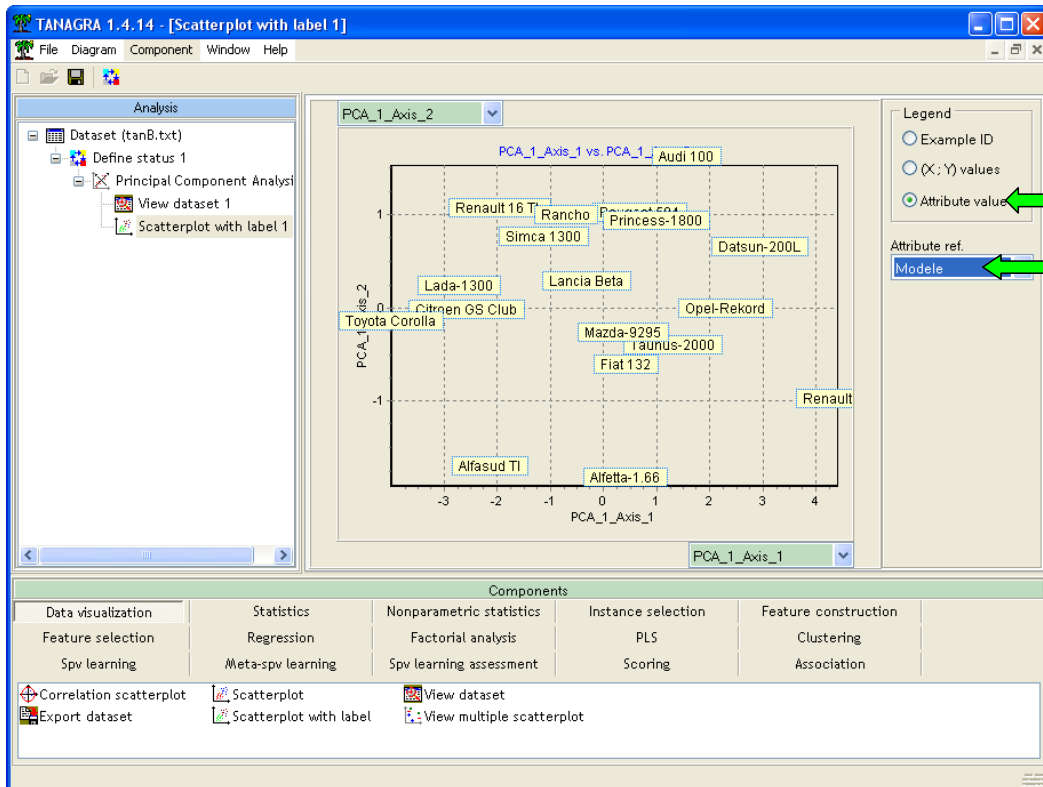


Scatter plots. The popularity of factorial methods is based largely on graphical representations. They allow us to visually evaluate the proximity between observations. In our case, we project the observations in the first two dimensions. We can associate a label to each point. We insert the SCATTERPLOT WITH LABEL component (DATA VISUALIZATION tab) into the diagram. We set the first dimension for the horizontal axis and the second dimension for the vertical one. We note that we can easily modify the axes.



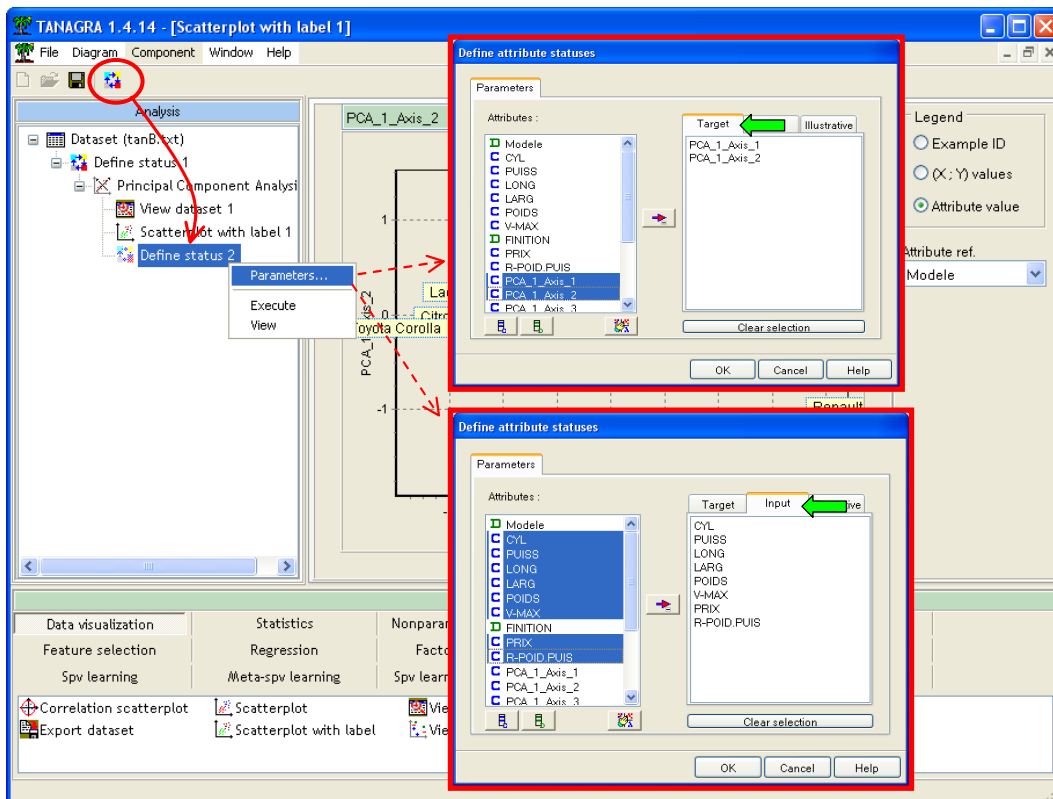
The points are automatically labeled by their number. We can modify this by clicking on LEGEND / ATTRIBUTE VALUE option. We select MODELE as reference attribute. We obtain the scatter plot on the two first dimensions. Of course, this option is practical as long as the number of points remains reasonable. Beyond of a certain number of observations, the graph would be unreadable.

Sometimes, some points are superposed. In this case, copying the coordinates in a spreadsheet and ranking the examples according the dimensions is the best way to identify precisely each example. We can also modify the size of the labels with the shortcuts CTRL + W and CTRL + Q.

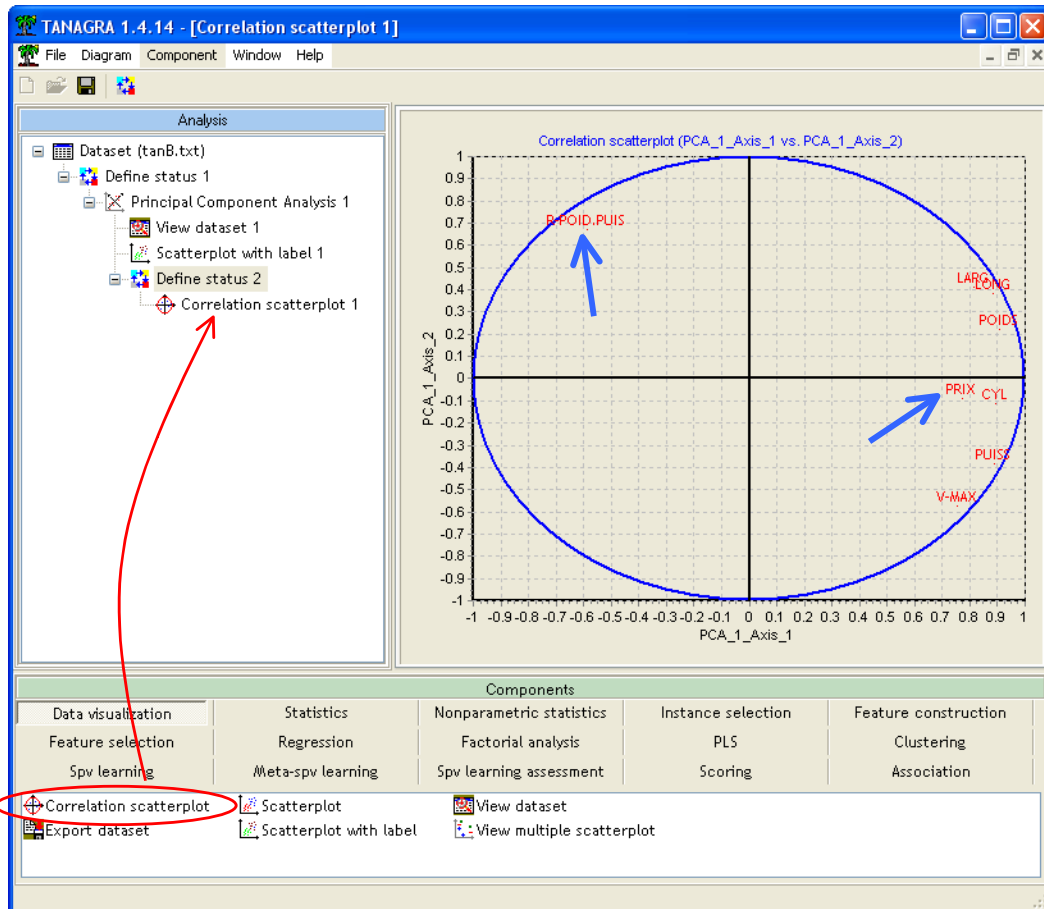


Correlation circle and illustrative variables. The correlation circle (or correlation scatter plot) is a graphical tool which allows to enhance the interpretation of the factors. Their correlations with the active and the illustrative continuous variables are computed. These are the coordinates of the variables into the scatter plot.

First we insert the DEFINE STATUS component. We set as TARGET the two first factors; then we set as INPUT all the continuous variables, including the illustrative ones.



Then, we add the CORRELATION SCATTERPLOT component. We click on the VIEW menu.



We see that PRICE (PRIX), which is an illustrative variable, is highly correlated to the first factor. It is associated to "big" cars (with high horsepower, high weight, etc.).

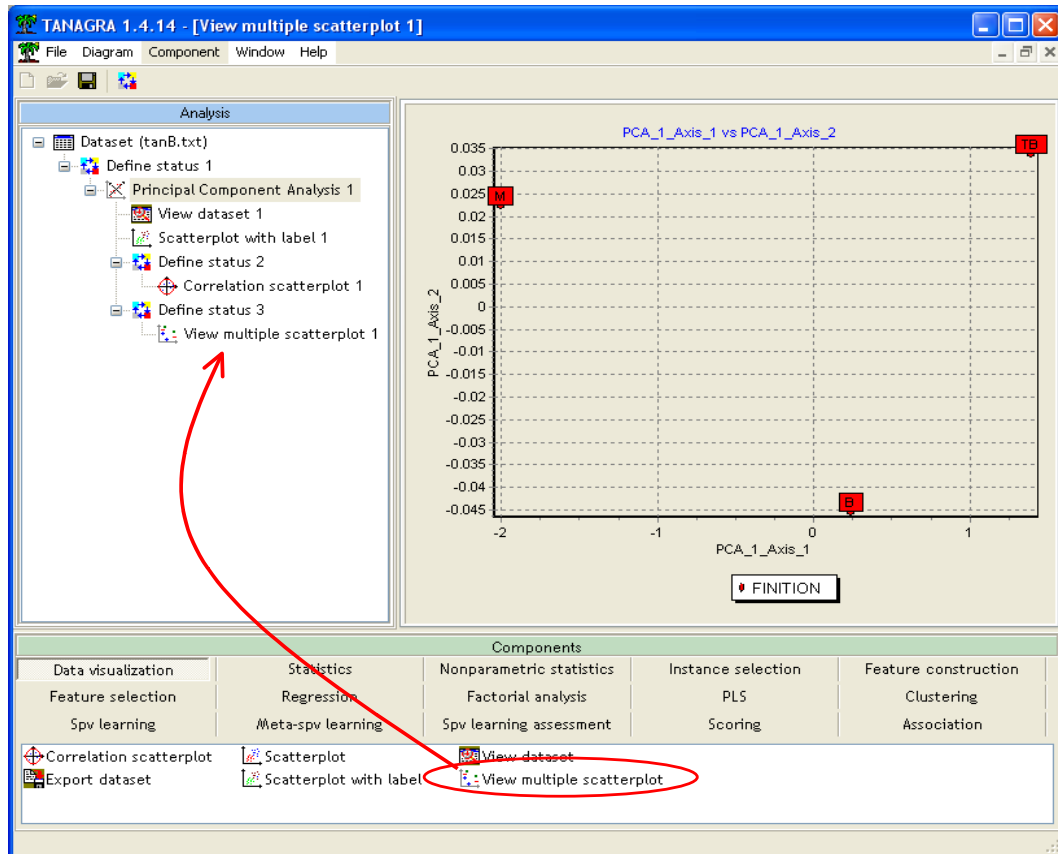
The second factor is associated to the "sports" characteristic of the vehicles. The location of the R.POIS-PUIS in the scatter plot confirms this analysis.

Categorical illustrative variables. Active variables must be continuous one for PCA. But we can use categorical illustrative variables in order to improve the interpretation of the factors.

Like for the correlation scatter plot, we must first define the types of the variables using the DEFINE STATUS component. We set as TARGET the two first factors; then we set as INPUT the categorical illustrative variable e.g. FINITION (finishing touches of the cars: M → mediocre; B → good; TB → very good).

Then we insert the VIEW MULTIPLE SCATTERPLOT component.

The coordinates correspond to the conditional average of the categories of the variable on each dimension.



We note that "big" cars have also very good finishing touches.

Illustrative examples. We do not use this option in our tutorial, but we can also subdivide the dataset in a learning sample and an illustrative sample. The first is used for the computation of the new dimensions. The second corresponds to new instances that we want to locate into this new representation space, for instance when we want to apply the results on other sub-population.

Conclusion

The principal component analysis, and in general the factor analysis, is useful to understand the underlying structure of a tabular dataset. In this tutorial, we show how to implement this approach with Tanagra.

The opportunity to copy/paste the results into a spreadsheet is certainly one of the most interesting functionalities of the software. Indeed, it gives us access to tools (sorting, formatting, etc.) in a well-known environment of the experts of the data processing. For example, the possibility to sorting the various tables according to the contributions and the COS2 is really an interesting functionality when we wish to interpret the dimensions.