Topic

Implementing the VARIMAX rotation in a Principal Component Analysis.

A VARIMAX rotation is a change of coordinates used in principal component analysis[1] (PCA) that maximizes the sum of the variances of the squared loadings. Thus, all the coefficients (squared correlation with factors) will be either large or near zero, with few intermediate values.

The goal is to associate each variable to at most one factor. The interpretation of the results of the PCA will be simplified. Then each variable will be associated to one and one only factor, they are split (as much as possible) into disjoint sets[2].

In this tutorial, we show how to perform this kind of rotation from the results of a standard PCA in Tanagra.

Dataset

We use the US CRIME DATAFILE[3] from the DASL website[4]. These data are crime-related and demographic statistics for 47 US states in 1960. The data were collected from the FBI's Uniform Crime Report and other government agencies to determine how the variable crime rate depends on the other variables measured in the study.

VARIMAX rotation

Importing the dataset

First of all, we must create a diagram and import the dataset. Tanagra can handle directly the XLS (Excel) file format[5]. We select the CRIME_DATASET_FROM_DASL.XLS data file.

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

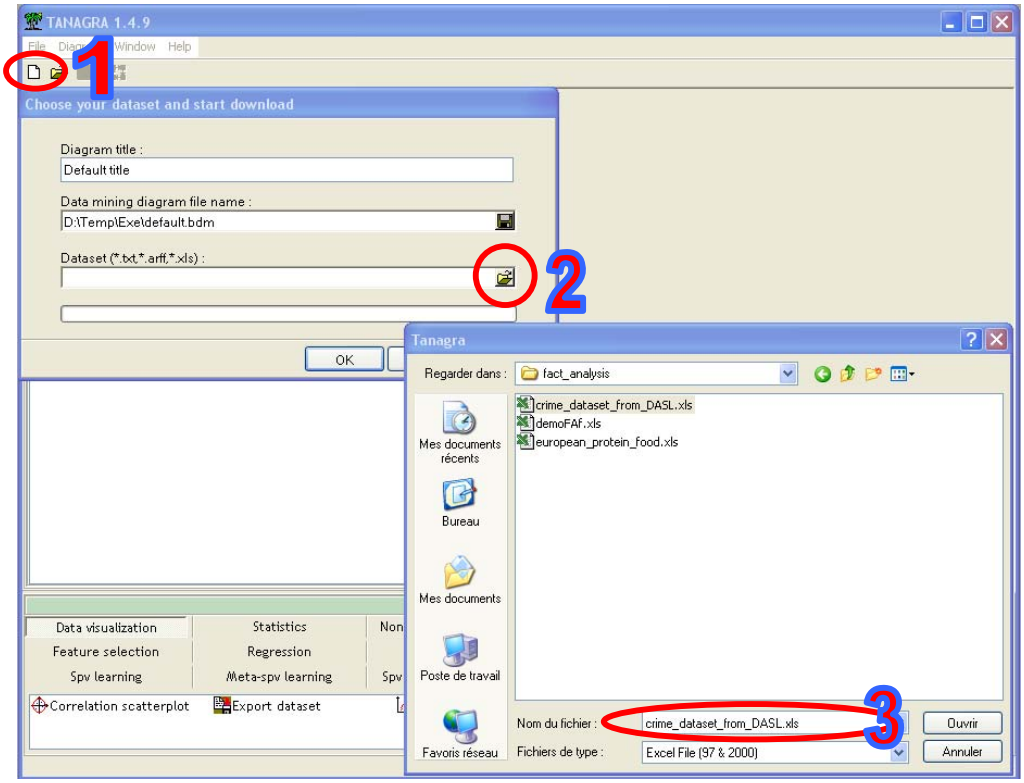
[2] See http://en.wikipedia.org/wiki/Varimax_rotation

See also http://www.utd.edu/~herve/Abdi-rotations-pretty.pdf

[3] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/crime_dataset_from_DASL.xls

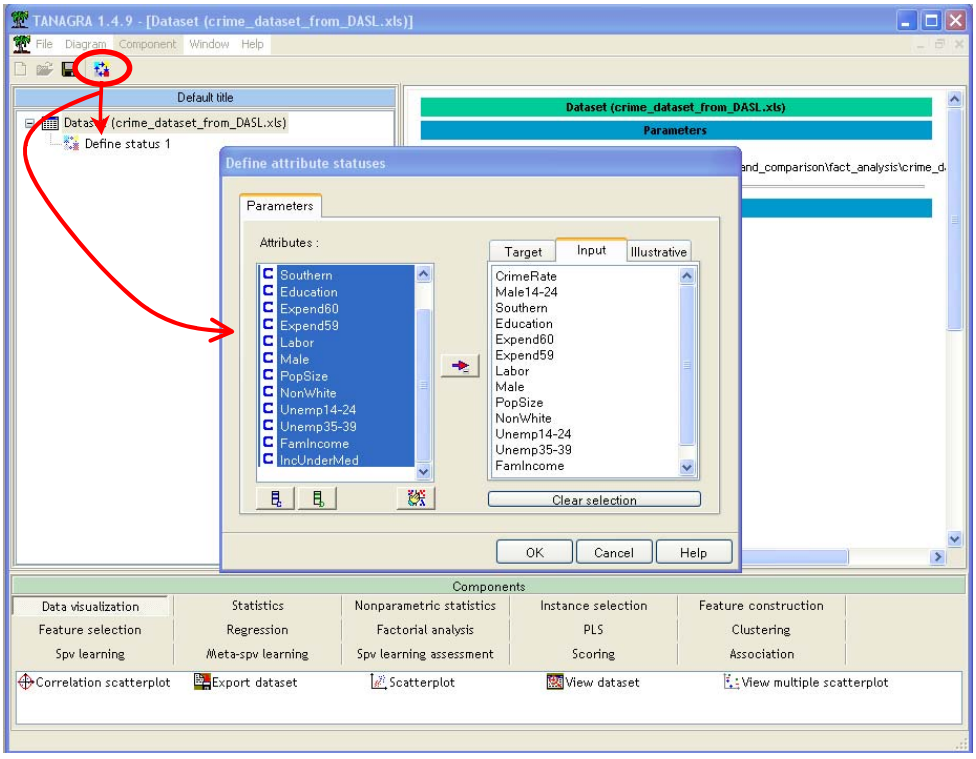
[4] http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html

[5] Tanagra can handle various file formats (see http://data-mining-tutorials.blogspot.com/search/label/Data%20file%20handling). About the spreadsheet file format, Tanagra can import without external library the XLS file format (http://data-mining-tutorials.blogspot.com/2008/10/excel-file-format-direct-importation.html); we can also use an add-in in order to send the dataset from Excel to Tanagra (http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html).



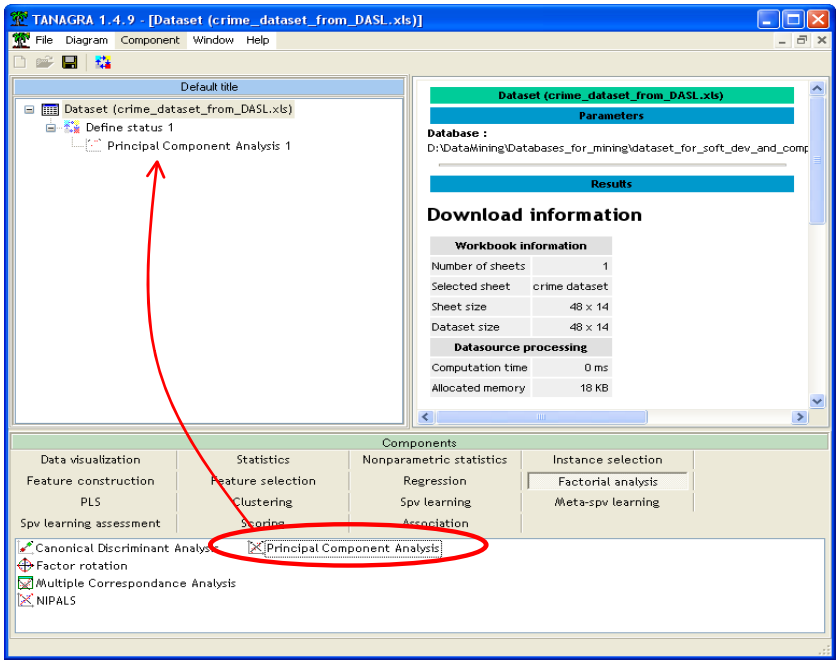
Defining the variables of the study

The next step is to specify the variables used in the study. We add the DEFINE STATUS component into the diagram. We set all the available variables as INPUT.



PCA

We want to implement a principal component analysis. The idea is not to perform a detailed analysis in this tutorial but rather to explain very briefly why in some cases the rotation of the factors resulting from the PCA can be beneficial when we want to read the results of the PCA, and how to proceed with Tanagra.



We click on the VIEW menu. The first 4 factors summarize 84% of the available information.

Eigen values

Matrix trace = 14.00

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	-	-	-

For the interpretation of the factors, we inspect the factor loadings table (see **Factor Loadings and Communality Estimates –** We show the results for the 4 first factors here).

Factor Loadings [Communality Estimates]

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 %)

Some results draw our attention:

- « CORR » is the correlation between the variables and the factors. When the absolute value is higher than 0.7, it is highlighted.
- « % » is the squared cosine COS^2, it is the squared of CORR. We have the cumulative values into the brackets. If we select all the factors, we obtain 100% in the last column.
- In the last row, we have the variance associated to each factor (into the brackets the cumulative values). We have the same values than the EIGENVALUES table above.

One of the main difficulties of the CPA is the interpretation of factors. In this example, if we consider the first two factors, we find that:

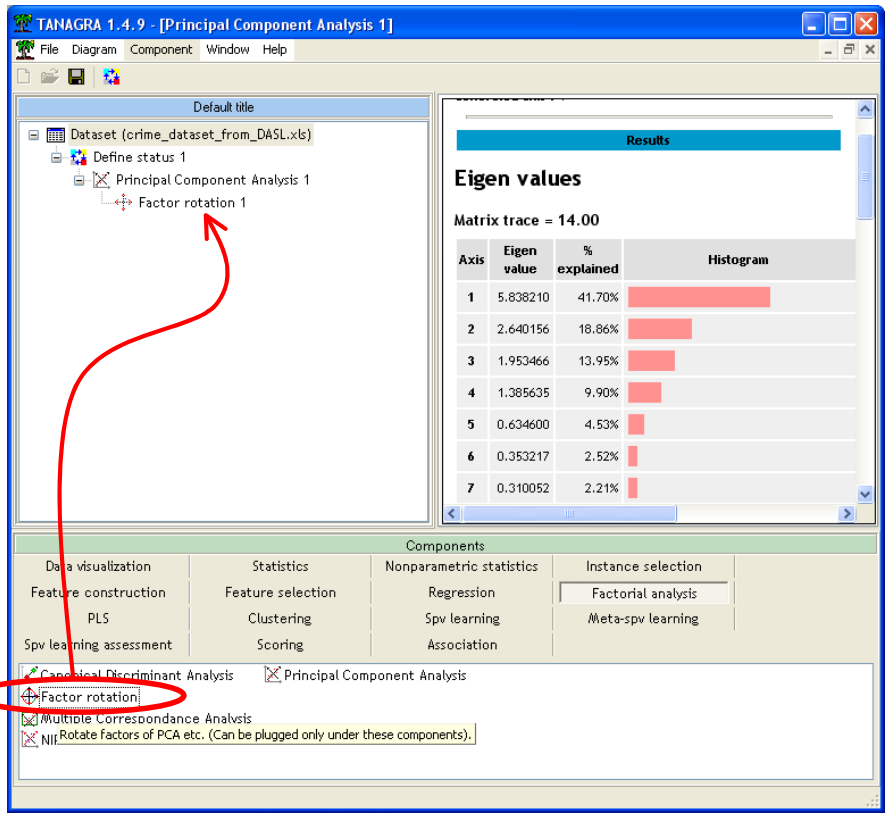
- The first factor contrasts, on the one hand, the southern states (Southern) with a high proportion of people on low wages (IncUnderMed) and predominantly young male (Male14-24) with, on the other hand, states composed of families with high income (FamilyIncome) with a high level of education (Education), for which the expenditure related to security are high (Expend).
- The second factor is not easy to interpret.

Here is a strange result. It suggests that the safety expenditure is high for the states populated by people with high incomes who studied longer. In the same time, the crime rate is not dramatically higher elsewhere. What is the meaning of these results?

VARIMAX rotation

We note especially that there are many variables with medium correlations on these first two factors (around 0.5 in absolute value), making the interpretation of factors complicated. The aim of the rotation methods is precisely to make the values of these correlations more contrasted by rotating the factors. Indeed, their interpretation will be easier.

The VARIMAX rotation is an orthogonal rotation i.e. the factors remain orthogonal after the rotation, preserving an essential property of the PCA. We insert the VARIMAX component into the diagram. The default number of analyzed factors is 2, but we can modify this.



We click on the VIEW menu. We obtain the table of loadings, before and after the rotation.

Rotated Factor Loadings

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
CrimeRate	0.0784	1 % (1 %)	0.6269	39 % (40 %)
Male14-24	-0.5000	25 % (25 %)	-0.5420	29 % (54 %)
Southern	-0.8283	69 % (69 %)	-0.2365	6 % (74 %)
Education	0.8665	75 % (75 %)	0.2819	8 % (83 %)
Expend60	0.2684	7 % (7 %)	0.9003	81 % (88 %)
Expend59	0.2771	8 % (8 %)	0.8963	80 % (88 %)
Labor	0.7068	50 % (50 %)	-0.1567	2 % (52 %)
Male	0.5755	33 % (33 %)	-0.2014	4 % (37 %)
PopSize	-0.2551	7 % (7 %)	0.7274	53 % (59 %)
NonWhite	-0.8142	66 % (66 %)	-0.1056	1 % (67 %)
Unemp14-24	0.0098	0 % (0 %)	0.1332	2 % (2 %)
Unemp35-39	-0.3329	11 % (11 %)	0.4706	22 % (33 %)
FamIncome	0.6344	40 % (40 %)	0.6989	49 % (89 %)
IncUnderMed	-0.7316	54 % (54 %)	-0.5100	26 % (80 %)
Var. Expl.	4.4492	32 % (32 %)	4.0292	29 % (61 %)

vs. Unrotated Factor Loadings

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
CrimeRate	0.4721	22 % (22 %)	-0.4198	18 % (40 %)
Male14-24	-0.7332	54 % (54 %)	0.0781	1 % (54 %)
Southern	-0.7788	61 % (61 %)	-0.3680	14 % (74 %)
Education	0.8375	70 % (70 %)	0.3591	13 % (83 %)
Expend60	0.7952	63 % (63 %)	-0.5002	25 % (88 %)
Expend59	0.7991	64 % (64 %)	-0.4915	24 % (88 %)
Labor	0.4283	18 % (18 %)	0.5836	34 % (52 %)
Male	0.3001	9 % (9 %)	0.5307	28 % (37 %)
PopSize	0.2875	8 % (8 %)	-0.7152	51 % (59 %)
NonWhite	-0.6819	47 % (47 %)	-0.4572	21 % (67 %)
Unemp14-24	0.0952	1 % (1 %)	-0.0937	1 % (2 %)
Unemp35-39	0.0598	0 % (0 %)	-0.5733	33 % (33 %)
FamIncome	0.9378	88 % (88 %)	-0.1075	1 % (89 %)
IncUnderMed	-0.8864	79 % (79 %)	-0.0986	1 % (80 %)
Var. Expl.	5.8382	42 % (42 %)	2.6402	19 % (61 %)

We note that:

- The cumulated variance (61%) associated to the two first factors is the same; it is a very important result. But the repartition between the factors is not the same.
- The southern states where people have low income are contrasted to the other states where people have high income and a high level of education. It seems that it described mostly the opposition between southern states and northern states (mostly urban).
- On the second factor, we note now that crime is mostly related to the population size and a low level of labor.

It should nevertheless be wary of these techniques. Such a difference in interpretation before and after rotation can also be a reflection of insidious anomalies. We often found two possible causes: either a very important variable missing from the dataset, in our case, the proportion of people living in urban and rural areas can play an important role, we do not have this information; either few outliers completely distorts the calculations.

Conclusion

PCA is a very popular tool to summarize a dataset into a few factors which highlight the most important information. But the factors are sometimes difficult to interpret. By using the rotation methods such as VARIMAX, we have additional tools which make easier the interpretation of the factors, and which can thus improve the relevance of the results.