1 Topic

Description of two alternative approaches to the PCA (Principal Component Analysis) available into Tanagra: Principal Factor Analysis and Harris Component Analysis (non-iterative algorithms). Comparison with the tools from SAS, R (package PSYCH) and SPSS.

PCA (Principal Component Analysis)[1] is a dimension reduction technique which enables to obtain a synthetic description of a set of quantitative variables. It produces latent variables called principal components (or factors) which are linear combinations of the original variables. The number of useful components is much lower than to the number of original variables because these last ones are (more or less) correlated. PCA enables also to reveal the internal structure of the data because the components are constructed in a manner as to explain optimally the variance of the data.

PFA (Principal Factor Analysis)[2] is often confused with PCA. There has been significant controversy about the equivalence or otherwise of the two techniques. One of the point of view which enables to distinguish them is to consider that the factors from the PCA account the maximal amount of variance of the available variables, while those from PFA account only the common variance in the data. The latter seems more appropriate if the goal of the analysis is to produce latent variables which highlight the underlying relation between the original variables. The influence of the variables which are not related to the other should be excluded.

They are thus different due to the nature of the information they make use. But the nuance is not obvious. Especially as they are often grouped in the same tool into some popular software (e.g. “PROC FACTOR” into SAS; “ANALYZE / DATA REDUCTION / FACTOR” into SPSS; etc.). In addition, their outputs and their interpretation are very similar.

In this tutorial, we present three approaches: Principal Component Analysis – PCA; non iterative Principal Factor Analysis - PFA; non iterative Harris Component Analysis - Harris. We highlight the differences by comparing the matrix (correlation matrix for the PCA) used for the diagonalization process. We detail the steps of the calculations using a program for R. We check our results by comparing them to those of SAS (PROC FACTOR). Thereafter, we implement these methods with Tanagra, with R using the PSYCH package, and with SPSS.

2 Dataset

The “beer_rnd.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[3]. 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.

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

[2] http://en.wikipedia.org/wiki/Factor_analysis

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

We have already processed a version of this dataset previously[4]. But, to make difficult the analysis, we add 7 randomly generated variables (rnd1…rnd7). Thus, we have **p = 14** variables in our dataset. Our aim is to check the ability of the various approaches to extract the useful information i.e. their ability to detect the relation between the variables knowing that there are noisy variables (generated randomly) in the database[5].

Below, we show the first 5 instances of the data file.

cost	size	alcohol	reputat	color	aroma	taste	rnd1	rnd2	rnd3	rnd4	rnd5	rnd6	rnd7
10	15	20	85	40	30	50	40	80	65	25	90	45	40
100	70	50	30	75	60	80	70	55	45	25	95	95	60
65	30	35	80	80	60	90	45	90	65	90	20	95	35
0	0	20	30	80	90	100	85	30	45	85	40	80	5
10	25	10	100	50	40	60	20	5	25	25	20	25	80

3 Steps for completing factor analysis using R

In this section, we detail the calculations for each approach using a program for R.

First, we import the "beer_rnd.txt" data file (text file format) and we display the first 5 instances.

```
R Console
> beer.data <- read.table(file="beer_rnd.txt",sep="\t",header=T)
> print(head(beer.data,n=5))
  cost size alcohol reputat color aroma taste rnd1 rnd2 rnd3 rnd4 rnd5 rnd6 rnd7
1   10   15      20      85    40    30    50   40   80   65   25   90   45   40
2  100   70      50      30    75    60    80   70   55   45   25   95   95   60
3   65   30      35      80    80    60    90   45   90   65   90   20   95   35
4    0    0      20      30    80    90   100   85   30   45   85   40   80    5
5   10   25      10     100    50    40    60   20    5   25   25   20   25   80
> |
```

3.1 Principal component analysis (PCA)

The correlation matrix (p x p) is the starting point of the PCA. Under R, we obtain this matrix with the **cor()** function.

```
beer.cor <- cor(beer.data)
print(round(beer.cor,2))
```

The matrix displays the correlation between each pair of variables (Figure 1). By rearranging it wisely, we observe groups of variables:

- (COST, SIZE and ALCOHOL) are highly correlated. They characterize the consumers which want to drink a lot of alcohol in cheap way.
- The second group consists of (COLOR, AROMA and TASTE). It corresponds to the consumers which are sensitive to the quality of the beer.
- REPUTAT is moderately negatively correlated to this second group i.e. the consumers sensitive to (COLOR, AROMA and TASTE) are not sensitive to the reputation.

[4] http://data-mining-tutorials.blogspot.fr/2013/01/new-features-for-pca-in-tanagra.html

[5] "Noise" variable is not really the appropriate term in the factor analysis context. These are variables which are not related to the others. It does not mean that they are not interesting.

- The random variables (rnd1…rnd7) are not correlated to any other variables of the dataset. This is not surprising.

Of course, the correlation of a variable with itself is 1. We observe it in the main diagonal of the correlation matrix. The PCA process makes use of this information when it diagonalizes the matrix. It treats all the variation of the variables by giving them the same importance.

	cost	size	alcohol	reputat	color	aroma	taste	rnd1	rnd2	rnd3	rnd4	rnd5	rnd6	rnd7
cost	1	0.88	0.88	-0.17	0.32	-0.03	0.05	0.17	-0.05	0.03	0.10	0.00	-0.02	-0.06
size	0.88	1	0.82	-0.06	0.01	-0.29	-0.31	0.21	-0.04	0.06	-0.02	-0.04	0.00	-0.03
alcohol	0.88	0.82	1	-0.36	0.40	0.10	0.06	0.18	-0.03	0.09	0.08	0.00	-0.08	-0.08
reputat	-0.17	-0.06	-0.36	1	-0.52	-0.52	-0.63	0.05	0.05	-0.10	-0.15	0.04	-0.05	0.09
color	0.32	0.01	0.40	-0.52	1	0.82	0.80	-0.01	0.11	0.06	0.25	0.02	-0.09	0.05
aroma	-0.03	-0.29	0.10	-0.52	0.82	1	0.87	-0.05	0.07	0.04	0.15	0.04	-0.05	-0.01
taste	0.05	-0.31	0.06	-0.63	0.80	0.87	1	-0.08	0.03	0.00	0.21	-0.01	0.03	-0.04
rnd1	0.17	0.21	0.18	0.05	-0.01	-0.05	-0.08	1	0.07	-0.04	-0.11	0.19	0.10	-0.04
rnd2	-0.05	-0.04	-0.03	0.05	0.11	0.07	0.03	0.07	1	-0.01	0.06	0.07	0.06	0.07
rnd3	0.03	0.06	0.09	-0.10	0.06	0.04	0.00	-0.04	-0.01	1	0.16	-0.07	0.07	0.01
rnd4	0.10	-0.02	0.08	-0.15	0.25	0.15	0.21	-0.11	0.06	0.16	1	0.09	-0.02	0.07
rnd5	0.00	-0.04	0.00	0.04	0.02	0.04	-0.01	0.19	0.07	-0.07	0.09	1	-0.08	0.01
rnd6	-0.02	0.00	-0.08	-0.05	-0.09	-0.05	0.03	0.10	0.06	0.07	-0.02	-0.08	1	-0.02
rnd7	-0.06	-0.03	-0.08	0.09	0.05	-0.01	-0.04	-0.04	0.07	0.01	0.07	0.01	-0.02	1

Figure 1 – Correlation matrix

Eigenvalues. We use the following commands to diagonalize the correlation matrix and display the eigenvalues:

```
#eigenvalues and eigenvectors of the correlation matrix
eig.pca <- eigen(beer.cor)
#print
print eigenvalues
print("eigenvalues")
print(eig.pca$values)
#screeplot
plot(1:14,eig.pca$values,type="b")
abline(a=1,b=0)
```

The results are consistent with those of SAS (PROC FACTOR[6]) (Figure 2). SAS shows that we used the full variability of the variables, i.e. we perform a PCA, by mentioning "Prior Communality Estimates: ONE" in the eigenvalues table.

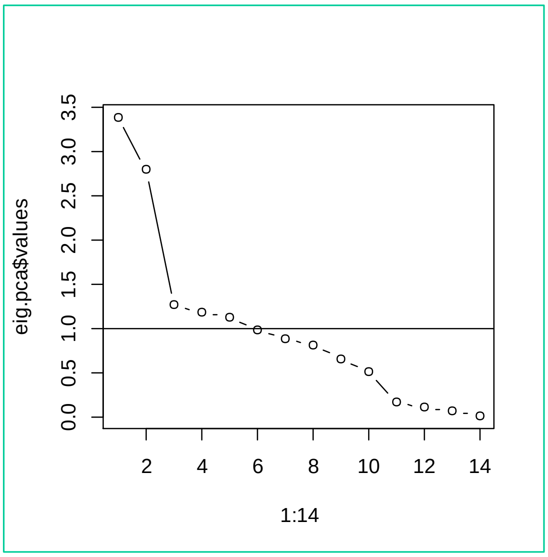
The determination of the right number of component is a difficult problem. According to the Kaiser-Guttman rule, we select 5 components here (even 6 because the 6-th eigenvalue is equal to 0.9927). This is not surprising. At least 7 variables among 14 are generated orthogonally. We need a large

[6] We use the following command:

```
proc factor data = mesdata.beer_rnd
method=principal
score
nfactors=3;
run;
```

number of components if we want to take into account all the observed variance of the variables. But, this choice is not really appropriate if we want to highlight the relations between the variables (the shared variance). The influence of the 7 variables generated randomly must be neglected.

```
> print(eig.pca$values)
 [1] 3.38655702 2.79466471 1.26759646 1.18217245
 [5] 1.12968654 0.99271966 0.88386983 0.81545409
 [9] 0.66464548 0.51059022 0.17321278 0.11239238
[13] 0.07083513 0.01560324
>
```



(R)

Prior Communality Estimates: ONE

	Eigenvalues of the Correlation Matrix: Total = 14 Average = 1			
	Eigenvalue	Difference	Proportion	Cumulative
1	3.38655702	0.59189231	0.2419	0.2419
2	2.79466471	1.52706826	0.1996	0.4415
3	1.26759646	0.08542400	0.0905	0.5321
4	1.18217245	0.05248591	0.0844	0.6165
5	1.12968654	0.13696688	0.0807	0.6972
6	0.99271966	0.10884983	0.0709	0.7681
7	0.88386983	0.06841574	0.0631	0.8312
8	0.81545409	0.15080861	0.0582	0.8895
9	0.66464548	0.15405526	0.0475	0.9370
10	0.51059022	0.33737744	0.0365	0.9734
11	0.17321278	0.06082040	0.0124	0.9858
12	0.11239238	0.04155726	0.0080	0.9938
13	0.07083513	0.05523189	0.0051	0.9989
14	0.01560324		0.0011	1.0000

(SAS)

Figure 2 – Eigenvalues – Principal Component Analysis

The solution is quite different if we consider the scree plot. The suggested solution is two factors if we take the components before the elbow into the graphical representation (3 factors if we include the elbow in the selection). That is rather a good solution in view of the correlation matrix above (Figure 1), where we had detected groups of variables.

Loadings or Factor pattern. This table describes the correlation of the variables with the factors. These values are useful for the interpretation. In practice, we obtain them by multiplying the eigenvectors with the square root of the eigenvalues.

```
#correlation of the variables with the factors
loadings.pca <- matrix(0,nrow=nrow(beer.cor),ncol=3)
for (j in 1:3){
  loadings.pca[,j] <- sqrt(eig.pca$values[j])*eig.pca$vectors[,j]
}
print("loadings for the 3 first factors")
rownames(loadings.pca) <- colnames(beer.data)
print(round(loadings.pca,5))
```

We found on the two first factors the groups detected above into the correlation matrix. On the first one, (color, aroma and taste) are highly correlated, and are moderately negatively correlated to (reputation). On the second factor, we observe that cost, size and alcohol are correlated.

By choosing adequately the right number of factors, the random variables have no influence of the reading of the results in this context. If we include the third factor in the analysis (eigenvalue = 1.268; 9% o f the total variance), the situation becomes difficult. We must interpret the correlation of RND1 and RND5 with this factor. Of course, we know that there is no relevant information here.

R Console

```
> print(round(loadings.pca,5))
            [,1]     [,2]     [,3]
cost    -0.49678  0.81407  0.01273
size    -0.21378  0.94733  0.04357
alcohol -0.58837  0.76160  0.01900
reputat  0.73682  0.11434 -0.16204
color   -0.90757 -0.18174 -0.09099
aroma   -0.78387 -0.49557 -0.09363
taste   -0.80783 -0.49864 -0.02177
rnd1    -0.01831  0.30272 -0.60610
rnd2    -0.04235 -0.08543 -0.45303
rnd3    -0.11864  0.04597  0.39347
rnd4    -0.30514 -0.08602  0.05735
rnd5    -0.01361 -0.00533 -0.69254
rnd6     0.04716 -0.01364  0.04225
rnd7     0.05046 -0.07406 -0.09641
```

Factor Pattern

		Factor1	Factor2	Factor3
cost	cost	0.49678	0.81407	-0.01273
size	size	0.21378	0.94733	-0.04357
alcohol	alcohol	0.58837	0.76160	-0.01900
reputat	reputat	-0.73682	0.11434	0.16204
color	color	0.90757	-0.18174	0.09099
aroma	aroma	0.78387	-0.49557	0.09363
taste	taste	0.80783	-0.49864	0.02177
rnd1	rnd1	0.01831	0.30272	0.60610
rnd2	rnd2	0.04235	-0.08543	0.45303
rnd3	rnd3	0.11864	0.04597	-0.39347
rnd4	rnd4	0.30514	-0.08602	-0.05735
rnd5	rnd5	0.01361	-0.00533	0.69254
rnd6	rnd6	-0.04716	-0.01364	-0.04225
rnd7	rnd7	-0.05046	-0.07406	0.09641

(R) (SAS)

Figure 3 – Factor pattern - PCA

Communalities. This table shows the proportion of the variance in each variable that is accounted for on the extracted factors. We obtain these values by computing the square of the loadings and by summing them.

```
#communalities for the three first factors
comm.pca <- apply(loadings.pca,1,function(x){sum(x^2)})
print("communalities for the 3 first factors")
names(comm.pca) <- colnames(beer.data)
print(round(comm.pca,5))
```

All the original variables (not randomly generated) are well accounted for on the three first factors.

R Console (R)

```
> print(round(comm.pca,5))
   cost    size alcohol reputat   color   aroma   taste    rnd1    rnd2    rnd3    rnd4    rnd5    rnd6    rnd7
0.90966 0.94503 0.92656 0.58223 0.86499 0.86881 0.90171 0.45933 0.21433 0.17101 0.10380 0.47983 0.00420 0.01733
```

(SAS)

Final Communality Estimates: Total = 7.448818

cost	size	alcohol	reputat	color	aroma	taste	rnd1	rnd2	rnd3	rnd4	rnd5	rnd6	rnd7
0.90966494	0.94502828	0.92656390	0.58223291	0.86499268	0.86880794	0.90171079	0.45933306	0.21433226	0.17100692	0.10379690	0.47982577	0.00419547	0.01732637

Factor scores – 1. This tables provides the coefficients which enables to calculate the coordinates of the individuals on the factors. Because we can apply them to standardized variables, these coefficients indicate also the relative importance of the variable for the determination of the factor. We obtain these coefficients by multiplying the inverse of the correlation matrix with the loadings.

```
#inversion of the correlation matrix
inv.beer.cor <- solve(beer.cor)
# factor scores
fscores.pca <- inv.beer.cor%*%loadings.pca
print(round(fscores.pca,5))
```

We have the same values, but in negative direction for some factors. This does not influence the interpretation of the results.

R Console

```
> print(round(fscores.pca,5))
            [,1]     [,2]     [,3]
cost    -0.14669  0.29130  0.01004
size    -0.06313  0.33898  0.03438
alcohol -0.17374  0.27252  0.01499
reputat  0.21757  0.04091 -0.12784
color   -0.26799 -0.06503 -0.07178
aroma   -0.23146 -0.17733 -0.07386
taste   -0.23854 -0.17843 -0.01717
rnd1    -0.00541  0.10832 -0.47815
rnd2    -0.01251 -0.03057 -0.35740
rnd3    -0.03503  0.01645  0.31041
rnd4    -0.09010 -0.03078  0.04524
rnd5    -0.00402 -0.00191 -0.54634
rnd6     0.01393 -0.00488  0.03333
rnd7     0.01490 -0.02650 -0.07605
```

(R)

Standardized Scoring Coefficients

		Factor1	Factor2	Factor3
cost	cost	0.14669	0.29130	-0.01004
size	size	0.06313	0.33898	-0.03438
alcohol	alcohol	0.17374	0.27252	-0.01499
reputat	reputat	-0.21757	0.04091	0.12784
color	color	0.26799	-0.06503	0.07178
aroma	aroma	0.23146	-0.17733	0.07386
taste	taste	0.23854	-0.17843	0.01717
rnd1	rnd1	0.00541	0.10832	0.47815
rnd2	rnd2	0.01251	-0.03057	0.35740
rnd3	rnd3	0.03503	0.01645	-0.31041
rnd4	rnd4	0.09010	-0.03078	-0.04524
rnd5	rnd5	0.00402	-0.00191	0.54634
rnd6	rnd6	-0.01393	-0.00488	-0.03333
rnd7	rnd7	-0.01490	-0.02650	0.07605

(SAS)

By applying these coefficients on the learning sample, we obtain the coordinates (scores) of the instances for each factor. They are standardized in order to obtain a unit variance for SAS and SPSS.

Factor scores – 2. Another way to compute the scores is to obtain a variance equal to the eigenvalue of the factor. We have this kind of behavior in Tanagra and some R procedures (princomp, prcomp, etc.). To obtain the appropriate coefficients, we multiply the preceding ones by the square root of the eigenvalues:

```
#factor scores – 2nd version
for (j in 1:3){
  fscores.pca[,j] <- sqrt(eig.pca$values[j])*fscores.pca[,j]
}
print(fscores.pca)
```

The factor scores are the same as those of Tanagra now.

R Console

```
> print(round(fscores.pca,7))
              [,1]       [,2]       [,3]
cost    -0.2699491  0.4869663  0.0113028
size    -0.1161680  0.5666762  0.0387028
alcohol -0.3197190  0.4555749  0.0168717
reputat  0.4003883  0.0683939 -0.1439279
color   -0.4931756 -0.1087115 -0.0808162
aroma   -0.4259543 -0.2964452 -0.0831614
taste   -0.4389765 -0.2982811 -0.0193368
rnd1    -0.0099492  0.1810839 -0.5383362
rnd2    -0.0230128 -0.0511029 -0.4023843
rnd3    -0.0644687  0.0274971  0.3494791
rnd4    -0.1658117 -0.0514555  0.0509394
rnd5    -0.0073931 -0.0031866 -0.6151126
rnd6     0.0256286 -0.0081575  0.0375269
rnd7     0.0274217 -0.0443045 -0.0856281
> |
```

(R)

Factor Scores

Attribute	Mean	Std-dev	Axis_1	Axis_2	Axis_3
cost	27.7777778	31.1903752	-0.2699491	0.4869663	-0.0113028
size	22.2222222	20.1537302	-0.1161680	0.5666762	-0.0387028
alcohol	23.8888889	12.1969436	-0.3197190	0.4555749	-0.0168717
reputat	55.5555556	25.7600514	0.4003883	0.0683939	0.1439279
color	63.8888889	18.0705066	-0.4931756	-0.1087115	0.0808162
aroma	56.1111111	19.6889391	-0.4259543	-0.2964452	0.0831614
taste	80.5555556	17.2311805	-0.4389765	-0.2982811	0.0193368
rnd1	42.7777778	28.7379507	-0.0099492	0.1810839	0.5383362
rnd2	52.4242424	27.8012756	-0.0230128	-0.0511029	0.4023843
rnd3	49.9494949	25.8833333	-0.0644687	0.0274971	-0.3494791
rnd4	46.5151515	27.6381246	-0.1658117	-0.0514555	-0.0509394
rnd5	46.8181818	25.8243342	-0.0073931	-0.0031866	0.6151126
rnd6	47.0202020	29.7796554	0.0256286	-0.0081575	-0.0375269
rnd7	51.6161616	29.0404480	0.0274217	-0.0443045	0.0856281

(TANAGRA)

Factor scores – Contributions to the factors. The factor scores coefficients enable to compute the coordinates of the individuals. But we can use them also for the interpretation of the factors. Indeed, because they are applied on standardized variables, the coefficients are comparable. Thus, we can detect the variables which have the most influence on each factor.

Starting from the table of factor scores, the contribution to the factor of a variable, for each factor, is the ratio between the squared factor scores and their sum. For instance, the coefficient of "cost" for the first factor is 0.14669; its squared is 0.02152. When we divide this value by the sum of the squared factor scores coefficient for the first factor, we obtain 0.02152/0.29528 = 7.287%. This is the relative contribution of the variable for the determination of the factor. We apply the same process to all the variables on the two first factors of the PCA.

	Standardized Scoring		Squared Coefficients		Contributions	
	Factor1	Factor2	Factor1	Factor2	Factor1	Factor2
cost	0.14669	0.2913	0.02152	0.08486	0.07287	0.23714
size	0.06313	0.33898	0.00399	0.11491	0.01350	0.32112
alcohol	0.17374	0.27252	0.03019	0.07427	0.10223	0.20755
reputat	-0.21757	0.04091	0.04734	0.00167	0.16031	0.00468
color	0.26799	-0.06503	0.07182	0.00423	0.24322	0.01182
aroma	0.23146	-0.17733	0.05357	0.03145	0.18143	0.08788
taste	0.23854	-0.17843	0.05690	0.03184	0.19270	0.08897
rnd1	0.00541	0.10832	0.00003	0.01173	0.00010	0.03279
rnd2	0.01251	-0.03057	0.00016	0.00093	0.00053	0.00261
rnd3	0.03503	0.01645	0.00123	0.00027	0.00416	0.00076
rnd4	0.0901	-0.03078	0.00812	0.00095	0.02749	0.00265
rnd5	0.00402	-0.00191	0.00002	0.00000	0.00005	0.00001
rnd6	-0.01393	-0.00488	0.00019	0.00002	0.00066	0.00007
rnd7	-0.01490	-0.02650	0.00022	0.00070	0.00075	0.00196
		Total	0.29528	0.35783	CTR(rnd) 3.37%	4.08%

The interpretation is consistent with those of the loadings. The sum of the contributions of the RND variables is negligible on the two first factors (3.37% for Factor 1; 4.08% for Factor 2).

Conclusion. These results of PCA are well-known in the literature. We recall them in order to better understand the results of the methods presented below.

3.2 Principal Factor Analysis (PFA)

The principal factor analysis (common factor analysis, principal axis factoring[7]) tries to identify latent variables which enable to structure and summarize the initial variables of the dataset. The approach deals exclusively with the shared variance between the variables.

The starting point is always the correlation matrix. But, for each variable, we replace 1 (the correlation of the variable with itself i.e. a variable is fully explained by itself) by the proportion of the variance explained by the others. Concretely, we use the coefficient of determination R_j^2 of the regression of the variable Xj on the (p-1) others. This is called "prior communalities" or "initial estimates of communalities".

Thus, we diagonalize the matrix F (Figure 4) in non-iterative principal factor analysis.

	cost	size	alcohol	reputat	color	aroma	taste	rnd1	rnd2	rnd3	rnd4	rnd5	rnd6	rnd7
cost	0.96	0.88	0.88	-0.17	0.32	-0.03	0.05	0.17	-0.05	0.03	0.10	0.00	-0.02	-0.06
size	0.88	0.94	0.82	-0.06	0.01	-0.29	-0.31	0.21	-0.04	0.06	-0.02	-0.04	0.00	-0.03
alcohol	0.88	0.82	0.91	-0.36	0.40	0.10	0.06	0.18	-0.03	0.09	0.08	0.00	-0.08	-0.08
reputat	-0.17	-0.06	-0.36	0.77	-0.52	-0.52	-0.63	0.05	0.05	-0.10	-0.15	0.04	-0.05	0.09
color	0.32	0.01	0.40	-0.52	0.85	0.82	0.80	-0.01	0.11	0.06	0.25	0.02	-0.09	0.05
aroma	-0.03	-0.29	0.10	-0.52	0.82	0.89	0.87	-0.05	0.07	0.04	0.15	0.04	-0.05	-0.01
taste	0.05	-0.31	0.06	-0.63	0.80	0.87	0.95	-0.08	0.03	0.00	0.21	-0.01	0.03	-0.04
rnd1	0.17	0.21	0.18	0.05	-0.01	-0.05	-0.08	0.14	0.07	-0.04	-0.11	0.19	0.10	-0.04
rnd2	-0.05	-0.04	-0.03	0.05	0.11	0.07	0.03	0.07	0.08	-0.01	0.06	0.07	0.06	0.07
rnd3	0.03	0.06	0.09	-0.10	0.06	0.04	0.00	-0.04	-0.01	0.07	0.16	-0.07	0.07	0.01
rnd4	0.10	-0.02	0.08	-0.15	0.25	0.15	0.21	-0.11	0.06	0.16	0.14	0.09	-0.02	0.07
rnd5	0.00	-0.04	0.00	0.04	0.02	0.04	-0.01	0.19	0.07	-0.07	0.09	0.11	-0.08	0.01
rnd6	-0.02	0.00	-0.08	-0.05	-0.09	-0.05	0.03	0.10	0.06	0.07	-0.02	-0.08	0.10	-0.02
rnd7	-0.06	-0.03	-0.08	0.09	0.05	-0.01	-0.04	-0.04	0.07	0.01	0.07	0.01	-0.02	0.09

Figure 4 – Matrix F for Principal Factor Analysis

The groups of variables are the same. But we note that (cost,…, taste) can be explained by the others, unlike (rnd1,.., rnd7) e.g. for the regression of "cost" on (size, alcohol, ..., rnd7), we have $R^2 = 0.96$; R^2(size / cost,alcohol, …, rnd7) = 0.94; …; R^2(rnd1/cost, alcohol, …) = 0.14; etc.

We do not need to perform explicitly 'p = 14' regressions to obtain these coefficients. We can compute them from the inverse (C^{-1}) of the correlation matrix (C).

$$R_j^2 = 1 - \frac{1}{c_{jj}^{-1}}$$

Where $\left(c_{jj}^{-1}\right)$ is the j[th] value on the diagonal of the matrix C^{-1}.

The quantity $u_j = 1 - R_j^2 = \frac{1}{c_{jj}^{-1}}$ is called "uniqueness". It corresponds to the unexplained variance of Xj. If its value is high (near 1), the variable is not related to the others.

We detail below the calculation of the main diagonal (the prior communalities) of the matrix F for principal factor analysis.

First we calculate the inverse of the correlation matrix.

[7] http://en.wikipedia.org/wiki/Principal_factor_analysis#Types_of_factoring

Inverse of the correlation matrix

	cost	size	alcohol	reputat	color	aroma	taste	rnd1	rnd2	rnd3	rnd4	rnd5	rnd6	rnd7
cost	25.67	-17.54	-10.39	-7.39	-0.55	9.10	-18.10	0.62	0.91	0.06	-0.74	-1.00	0.10	0.38
size	-17.54	17.82	1.77	4.62	0.86	-5.00	12.72	-0.54	-0.66	-0.01	0.57	0.87	-0.40	-0.51
alcohol	-10.39	1.77	11.41	4.48	-2.84	-4.04	8.98	-0.48	-0.17	-0.04	0.27	0.30	0.29	0.37
reputat	-7.39	4.62	4.48	4.39	-0.88	-2.60	7.21	-0.33	-0.31	0.15	0.23	0.28	0.06	-0.07
color	-0.55	0.86	-2.84	-0.88	6.82	-2.73	-3.15	0.13	-0.43	-0.13	-0.34	0.02	0.25	-0.63
aroma	9.10	-5.00	-4.04	-2.60	-2.73	8.83	-8.91	0.09	0.31	-0.20	0.11	-0.51	0.18	0.22
taste	-18.10	12.72	8.98	7.21	-3.15	-8.91	20.11	-0.48	-0.39	0.41	0.29	0.92	-0.50	0.14
rnd1	0.62	-0.54	-0.48	-0.33	0.13	0.09	-0.48	1.16	-0.05	0.04	0.11	-0.25	-0.15	0.03
rnd2	0.91	-0.66	-0.17	-0.31	-0.43	0.31	-0.39	-0.05	1.09	0.03	-0.08	-0.08	-0.08	-0.01
rnd3	0.06	-0.01	-0.04	0.15	-0.13	-0.20	0.41	0.04	0.03	1.08	-0.18	0.09	-0.11	0.00
rnd4	-0.74	0.57	0.27	0.23	-0.34	0.11	0.29	0.11	-0.08	-0.18	1.17	-0.11	0.00	-0.06
rnd5	-1.00	0.87	0.30	0.28	0.02	-0.51	0.92	-0.25	-0.08	0.09	-0.11	1.13	0.07	-0.01
rnd6	0.10	-0.40	0.29	0.06	0.25	0.18	-0.50	-0.15	-0.08	-0.11	0.00	0.07	1.11	0.01
rnd7	0.38	-0.51	0.37	-0.07	-0.63	0.22	0.14	0.03	-0.01	0.00	-0.06	-0.01	0.01	1.10

Figure 5 - Inverse of the correlation matrix

For 'cost', we obtain the uniqueness as follow $u_{cost} = \frac{1}{25.67} = 0.04$; and then the prior communality $R^2_{cost} = 1 - 0.04 = 0.96$. We use the following commands under R.

```
#uniqueness
d2 <- 1/diag(inv.beer.cor)
print(d2)
#prior communalities
init.comm <- 1-d2
print(init.comm)
```

The obtained values are:

R Console

```
> d2 <- 1/diag(inv.beer.cor)
> print(d2)
      cost       size    alcohol    reputat      color      aroma      taste 
0.03895001 0.05611313 0.08766377 0.22768073 0.14671970 0.11319932 0.04973029 
      rnd1       rnd2       rnd3       rnd4       rnd5       rnd6       rnd7 
0.86174167 0.91504656 0.92642532 0.85759860 0.88855600 0.90371677 0.91313601 
> #prior communalities
> init.comm <- 1-d2
> print(init.comm)
      cost       size    alcohol    reputat      color      aroma      taste 
0.96104999 0.94388687 0.91233623 0.77231927 0.85328030 0.88680068 0.95026971 
      rnd1       rnd2       rnd3       rnd4       rnd5       rnd6       rnd7 
0.13825833 0.08495344 0.07357468 0.14240140 0.11144400 0.09628323 0.08686399 
```

We insert these values into the main diagonal of the correlation matrix **C** to obtain the matrix **F**:

```
#new version of the correlation matrix
beer.cor.pfa <- beer.cor
#replace the values of the main diagonal
diag(beer.cor.pfa) <- init.comm
#the trace of the matrix F
print(sum(diag(beer.cor.pfa)))
```

Thus, the values of the matrix F are defined as follow:

$$f_{ij} = \begin{cases} c_{ij}\,, if\ i \neq j \\ R_j^2\,, if\ i = j \end{cases}$$

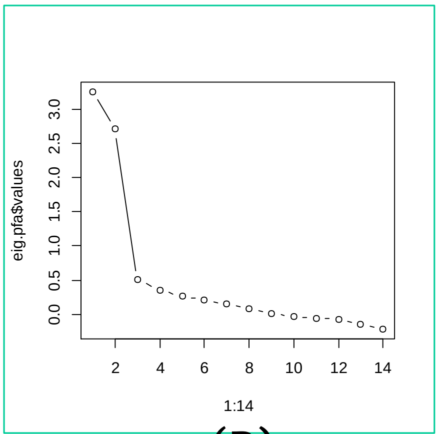
The trace of the matrix is $\sum_{j=1}^{p} R_j^2$ =**7.0137**. This is the total amount of information that we want to decompose in the principal factor analysis process.

Eigenvalues. We diagonalize the matrix **F** to obtain the eigenvalues[8].

```
#eigenvalues
eig.pfa <- eigen(beer.cor.pfa)
print("eigenvalues")
print(eig.pfa$values)
#screeplot
plot(1:14,eig.pfa$values,type="b")
```

Of course, we obtain the same values with SAS.

```
> print(eig.pfa$values)
 [1]  3.24993222  2.70605968  0.50552722
 [4]  0.35372255  0.26477240  0.21013567
 [7]  0.14965770  0.07535828  0.01049015
[10] -0.02720241 -0.05474559 -0.06991840
[13] -0.14661181 -0.21345552
```



(R)

Eigenvalues of the Reduced Correlation Matrix: Total = 7.01372212 Average = 0.50098015

	Eigenvalue	Difference	Proportion	Cumulative
1	3.24993222	0.54387254	0.4634	0.4634
2	2.70605968	2.20053246	0.3858	0.8492
3	0.50552722	0.15180466	0.0721	0.9213
4	0.35372255	0.08895015	0.0504	0.9717
5	0.26477240	0.05463673	0.0378	1.0095
6	0.21013567	0.06047797	0.0300	1.0394
7	0.14965770	0.07429943	0.0213	1.0608
8	0.07535828	0.06486813	0.0107	1.0715
9	0.01049015	0.03769256	0.0015	1.0730
10	-.02720241	0.02754317	-0.0039	1.0691
11	-.05474559	0.01517281	-0.0078	1.0613
12	-.06991840	0.07669341	-0.0100	1.0513
13	-.14661181	0.06684371	-0.0209	1.0304
14	-.21345552		-0.0304	1.0000

(SAS)

Some eigenvalues are negative. This is not surprising. Contrary to the correlation matrix **C**, **F** is not semi-definite positive. Up to the 4th one, the factors explain the shared variance because the sum of the eigenvalues does not exceed the matrix trace. From the 5th one, the intrinsic variance of the variables influences the factors. So, it is necessary to subtract eigenvalues (from the 10th factor) in order that the sum of all the eigenvalues is equal to the matrix trace (the total amount of information that we want analyze).

[8] With SAS, we set the following commands (the option "priors = smc" is essential):

```
proc factor data = mesdata.beer_rnd
method=principal
priors=smc
msa
nfactors=2
score;
run;
```

Clearly, selecting two factors is the right solution on our dataset. The gap between the 2nd eigenvalue and the 3rd one is very high in the scree plot. The first two factors explain 84.92% of the shared variance between the variables. This result was not as obvious for the principal component analysis (we hesitated between 2 and 3 factors; Figure 2).

Loadings or Factor pattern. Again, we calculate the loadings for the first two factors.

```
#loadings
loadings.pfa <- matrix(0,nrow=nrow(beer.cor.pfa),ncol=2)
for (j in 1:2){
  loadings.pfa[,j] <- sqrt(abs(eig.pfa$values[j]))*eig.pfa$vectors[,j]
}
rownames(loadings.pfa) <- colnames(beer.data)
print(round(loadings.pfa,5))
```

```
> print(round(loadings.pfa,5))
            [,1]     [,2]
cost    -0.52442 -0.80117
size    -0.24043 -0.93787
alcohol -0.60493 -0.73065
reputat  0.69728 -0.13038
color   -0.88243  0.20296
aroma   -0.76236  0.51145
taste   -0.80095  0.52573
rnd1    -0.02232 -0.20878
rnd2    -0.02930  0.06015
rnd3    -0.08501 -0.03166
rnd4    -0.22796  0.06342
rnd5    -0.00843  0.00856
rnd6     0.03627  0.01181
rnd7     0.04059  0.04624
```

Factor Pattern			
		Factor1	Factor2
cost	cost	0.52442	0.80117
size	size	0.24043	0.93787
alcohol	alcohol	0.60493	0.73065
reputat	reputat	-0.69728	0.13038
color	color	0.88243	-0.20296
aroma	aroma	0.76236	-0.51145
taste	taste	0.80095	-0.52573
rnd1	rnd1	0.02232	0.20878
rnd2	rnd2	0.02930	-0.06015
rnd3	rnd3	0.08501	0.03166
rnd4	rnd4	0.22796	-0.06342
rnd5	rnd5	0.00843	-0.00856
rnd6	rnd6	-0.03627	-0.01181
rnd7	rnd7	-0.04059	-0.04624

(R) (SAS)

Figure 6 - "Loadings" – Principal Factor Analysis

Loadings ≠ corrélation. Unlike the principal component analysis, the loadings do not correspond to the correlations between the variables and the factors in the principal factor analysis. These are rather the standardized coefficients of the regression of the factors on the variables[9]. Fortunately, the reading of the loadings is similar in practice. They enable to interpret the factors.

Communalities. The communalities allow to compare the amount of information reproduced for each variable on the selected factors with the amount of information initially workable (the shared variance for each variable).

```
#prior and estimated communalities for the 2 first factors
comm.pfa <- apply(loadings.pfa,1,function(x){sum(x^2)})
names(comm.pfa) <- colnames(beer.data)
print(round(cbind(init.comm,comm.pfa),5))
```

The quality of the representation for the "real" variables (cost,..., taste) is good on the two first factors. These factors are enough to understand the relations between the variables.

[9] Voir http://www.yorku.ca/ptryfos/f1400.pdf

```
> print(round(cbind(init.comm,comm.pfa),5))
        init.comm comm.pfa
cost      0.96105  0.91688
size      0.94389  0.93740
alcohol   0.91234  0.89979
reputat   0.77232  0.50319
color     0.85328  0.81988
aroma     0.88680  0.84278
taste     0.95027  0.91791
rnd1      0.13826  0.04409
rnd2      0.08495  0.00448
rnd3      0.07357  0.00823
rnd4      0.14240  0.05599
rnd5      0.11144  0.00014
rnd6      0.09628  0.00145
rnd7      0.08686  0.00379
```

Figure 7 – Initial and estimated communalities - PFA

We note that the sum of the two first eigenvalues is equal to the sum of the estimated communalities of the variables.

```
> print(sum(comm.pfa))
[1] 5.955992
> sum(eig.pfa$values[1:2])
[1] 5.955992
```

Factor scores. Again, the factor scores coefficients allow the calculation the coordinates of the individuals.

```
#factor scores
print("factor scores")
fscores.pfa <- inv.beer.cor%*%loadings.pfa
print(round(fscores.pfa,5))
```

Our results are consistent with those of SAS.

```
> print(round(fscores.pfa,5))
            [,1]     [,2]
cost     0.07718 -0.64741
size    -0.21226 -0.16184
alcohol -0.38278 -0.04766
reputat  0.04399  0.08779
color   -0.13617  0.05404
aroma   -0.12122 -0.00764
taste   -0.60210  0.52755
rnd1     0.01887 -0.01700
rnd2    -0.00141 -0.00859
rnd3    -0.02208  0.00835
rnd4    -0.02009  0.01793
rnd5    -0.02016  0.00531
rnd6     0.00542 -0.01042
rnd7    -0.01165  0.00673
```

(R)

Standardized Scoring Coefficients

		Factor1	Factor2
cost	cost	-0.07718	0.64741
size	size	0.21226	0.16184
alcohol	alcohol	0.38278	0.04766
reputat	reputat	-0.04399	-0.08779
color	color	0.13617	-0.05404
aroma	aroma	0.12122	0.00764
taste	taste	0.60210	-0.52755
rnd1	rnd1	-0.01887	0.01700
rnd2	rnd2	0.00141	0.00859
rnd3	rnd3	0.02208	-0.00835
rnd4	rnd4	0.02009	-0.01793
rnd5	rnd5	0.02016	-0.00531
rnd6	rnd6	-0.00542	0.01042
rnd7	rnd7	0.01165	-0.00673

(SAS)

Contributions of the variables "rnd". When we calculate the contribution of the variables RND on the factors, we note that they are considerably lowered (0.30% vs. 3.37 for the PCA for the 1st factor; 0.13% vs. 4.08% for the 2nd one). This is one of the main benefits of the PFA against the PCA in our context. The influence of the variables which are not related to the others is reduced.

	Standardized Scoring Coefficients		Squared Coefficients		Contributions	
	Factor1	Factor2	Factor1	Factor2	Factor1	Factor2
cost	-0.07718	0.64741	0.00596	0.41914	0.00998	0.56830
size	0.21226	0.16184	0.04505	0.02619	0.07546	0.03551
alcohol	0.38278	0.04766	0.14652	0.00227	0.24541	0.00308
reputat	-0.04399	-0.08779	0.00194	0.00771	0.00324	0.01045
color	0.13617	-0.05404	0.01854	0.00292	0.03106	0.00396
aroma	0.12122	0.00764	0.01469	0.00006	0.02461	0.00008
taste	0.60210	-0.52755	0.36252	0.27831	0.60719	0.37735
rnd1	-0.01887	0.01700	0.00036	0.00029	0.00060	0.00039
rnd2	0.00141	0.00859	0.00000	0.00007	0.00000	0.00010
rnd3	0.02208	-0.00835	0.00049	0.00007	0.00082	0.00009
rnd4	0.02009	-0.01793	0.00040	0.00032	0.00068	0.00044
rnd5	0.02016	-0.00531	0.00041	0.00003	0.00068	0.00004
rnd6	-0.00542	0.01042	0.00003	0.00011	0.00005	0.00015
rnd7	0.01165	-0.00673	0.00014	0.00005	0.00023	0.00006
		Total	0.59705	0.73753		
				CTR(rnd)	0.30%	0.13%

Accuracy of the factors. The factors have a theoretical unit variance. But because we work on a sample, we have no guarantee to obtain the unit variance on the dataset. The computed variances of the factors indicate their reliability. A sample variance near to 1 is desirable.

For the two first factors, we obtain theses variance by summing the product between the factor scores coefficients and the loadings. We use the following program for R:

```
#variance of the scores
vscores <- numeric(2)
for (j in 1:2){
 vscores[j] <- sum(fscores.pfa[,j]*loadings.pfa[,j])
}
print(round(vscores,5))
```

We obtain for R and SAS:

(R)

```
> print(round(vscores,5))
[1] 0.97357 0.98239
```

(SAS)

Squared Multiple Correlations of the Variables with Each Factor	
Factor1	Factor2
0.97357476	0.98238932

SAS calls these values "squared multiple correlations of the variables with each factors" because they correspond also to the squared correlations between the theoretical latent variable defined on the population and the factors estimated on the sample.

A high value reveals a good reliability of the factor (≥ 0.7 according to some references). We observe that we can have confident in the two first factors from the PFA on our dataset.

If we compute the first 5 factors, we note that starting from the third factor, the results are not really convincing. Obviously, two factors is the right solution for our dataset.

Scoring Coefficients Estimated by Regression

Squared Multiple Correlations of the Variables with Each Factor				
Factor1	Factor2	Factor3	Factor4	Factor5
0.97357476	0.98238932	0.65231496	0.41287901	0.32996949

Figure 8 - Variance of the 5 first factors - PFA

3.3 An iterative approach for Principal Factor Analysis

There is an iterative method for the principal factor analysis. We specify the number of factors used for the analysis. The previous approach is the first step of the algorithm. Then, we replace the initial communalities with the estimated communalities in the matrix F. We compute again the factors. The process is stopped when estimated communalities is stable (SAS) or when we reach a certain number of iterations (SPSS).

Sometimes, the estimated communality of a variable can exceed 1 is some circumstances. This is the "Heywood problem". That means that there are inconsistencies in the process. There are many reasons for that, among other things because we have selected a wrong number of factors[10].

3.4 Harris principal factor analysis (Harris)

The Harris' approach works also on a modified version of the correlation matrix. We are concerned with the shared variance also. We increase the correlations between the variables when they (either or both) are highly related to the others. In concrete terms, we start from the matrix F for the principal factor analysis (Figure 1), we weight the values with the uniqueness of the variables:

$$h_{ij} = \frac{f_{ij}}{\sqrt{u_i \times u_j}}$$

For our dataset, the computed matrix H is (Figure 9):

	cost	size	alcohol	reputat	color	aroma	taste	rnd1	rnd2	rnd3	rnd4	rnd5	rnd6	rnd7
cost	24.67	18.79	15.01	-1.86	4.24	-0.42	1.23	0.91	-0.27	0.17	0.55	-0.01	-0.12	-0.33
size	18.79	16.82	11.74	-0.54	0.16	-3.59	-5.82	0.94	-0.16	0.26	-0.10	-0.17	0.00	-0.12
alcohol	15.01	11.74	10.41	-2.55	3.51	0.98	0.85	0.66	-0.10	0.32	0.28	0.00	-0.28	-0.29
reputat	-1.86	-0.54	-2.55	3.39	-2.87	-3.25	-5.89	0.12	0.12	-0.21	-0.34	0.09	-0.10	0.20
color	4.24	0.16	3.51	-2.87	5.82	6.39	9.42	-0.04	0.29	0.17	0.69	0.07	-0.23	0.15
aroma	-0.42	-3.59	0.98	-3.25	6.39	7.83	11.54	-0.14	0.21	0.13	0.49	0.12	-0.16	-0.04
taste	1.23	-5.82	0.85	-5.89	9.42	11.54	19.11	-0.40	0.16	-0.02	1.00	-0.06	0.13	-0.19
rnd1	0.91	0.94	0.66	0.12	-0.04	-0.14	-0.40	0.16	0.08	-0.05	-0.12	0.21	0.12	-0.04
rnd2	-0.27	-0.16	-0.10	0.12	0.29	0.21	0.16	0.08	0.09	-0.01	0.07	0.07	0.06	0.08
rnd3	0.17	0.26	0.32	-0.21	0.17	0.13	-0.02	-0.05	-0.01	0.08	0.18	-0.08	0.08	0.02
rnd4	0.55	-0.10	0.28	-0.34	0.69	0.49	1.00	-0.12	0.07	0.18	0.17	0.10	-0.02	0.08
rnd5	-0.01	-0.17	0.00	0.09	0.07	0.12	-0.06	0.21	0.07	-0.08	0.10	0.13	-0.09	0.01
rnd6	-0.12	0.00	-0.28	-0.10	-0.23	-0.16	0.13	0.12	0.06	0.08	-0.02	-0.09	0.11	-0.02
rnd7	-0.33	-0.12	-0.29	0.20	0.15	-0.04	-0.19	-0.04	0.08	0.02	0.08	0.01	-0.02	0.10

Figure 9 - Matrix H for Harris Principal Factor Analysis

For instance, the correlation between cost and size is rather high: 0.88. In addition, the proportion of the variance of cost (size) explained by the other varibles is R^2_{cost} = 0.961 (R^2_{size}=0.944). Both are

[10] See http://v8doc.sas.com/sashtml/stat/chap26/sect21.htm

highly related to the other variables. We calculate the uniqueness: $u_{cost} = 0.039$ and $u_{size} = 0.056$. Thus, the relation between 'cost' and 'size' is more intense in the matrix H:

$$h_{cost,size} = \frac{0.88}{\sqrt{0.039 \times 0.056}} = 18.79$$

We observe the same groups as above in the matrix H. But here, the discrepancy between the values is higher, especially the values of relations between the original variables compared with those of relations with and between variables generated randomly. The analysis should exploit this property during the calculation of the factors.

For R, we use the formulas available online (SPSS, *"Image (Kaiser, 1963)"*; SAS, *"Harris, 1962")*[11] :

```
#see SPSS and SAS online documentation
S <- matrix(0,nrow=nrow(beer.cor),ncol=ncol(beer.cor))
diag(S) <- sqrt(1/diag(inv.beer.cor))
inv.S <- solve(S)
beer.cor.harris <- beer.cor
diag(beer.cor.harris) <- init.comm
beer.cor.harris <- inv.S%*%beer.cor.harris%*%inv.S
print("matrix to diagonalize")
print(round(beer.cor.harris,2))
print("trace of the matrix")
print(sum(diag(beer.cor.harris)))
```

The trace of the matrix is [**Tr(H) = 88.87841**].

```
[1] "matrix to diagonalize"
> print(round(beer.cor.harris,2))
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11] [,12] [,13] [,14]
 [1,] 24.67 18.79 15.01 -1.86  4.24 -0.42  1.23  0.91 -0.27  0.17  0.55 -0.01 -0.12 -0.33
 [2,] 18.79 16.82 11.74 -0.54  0.16 -3.59 -5.82  0.94 -0.16  0.26 -0.10 -0.17  0.00 -0.12
 [3,] 15.01 11.74 10.41 -2.55  3.51  0.98  0.85  0.66 -0.10  0.32  0.28  0.00 -0.28 -0.29
 [4,] -1.86 -0.54 -2.55  3.39 -2.87 -3.25 -5.89  0.12  0.12 -0.21 -0.34  0.09 -0.10  0.20
 [5,]  4.24  0.16  3.51 -2.87  5.82  6.39  9.42 -0.04  0.29  0.17  0.69  0.07 -0.23  0.15
 [6,] -0.42 -3.59  0.98 -3.25  6.39  7.83 11.54 -0.14  0.21  0.13  0.49  0.12 -0.16 -0.04
 [7,]  1.23 -5.82  0.85 -5.89  9.42 11.54 19.11 -0.40  0.16 -0.02  1.00 -0.06  0.13 -0.19
 [8,]  0.91  0.94  0.66  0.12 -0.04 -0.14 -0.40  0.16  0.08 -0.05 -0.12  0.21  0.12 -0.04
 [9,] -0.27 -0.16 -0.10  0.12  0.29  0.21  0.16  0.08  0.09 -0.01  0.07  0.07  0.06  0.08
[10,]  0.17  0.26  0.32 -0.21  0.17  0.13 -0.02 -0.05 -0.01  0.08  0.18 -0.08  0.08  0.02
[11,]  0.55 -0.10  0.28 -0.34  0.69  0.49  1.00 -0.12  0.07  0.18  0.17  0.10 -0.02  0.08
[12,] -0.01 -0.17  0.00  0.09  0.07  0.12 -0.06  0.21  0.07 -0.08  0.10  0.13 -0.09  0.01
[13,] -0.12  0.00 -0.28 -0.10 -0.23 -0.16  0.13  0.12  0.06  0.08 -0.02 -0.09  0.11 -0.02
[14,] -0.33 -0.12 -0.29  0.20  0.15 -0.04 -0.19 -0.04  0.08  0.02  0.08  0.01 -0.02  0.10
> print("trace of the matrix")
[1] "trace of the matrix"
> print(sum(diag(beer.cor.harris)))
[1] 88.87841
```

Eigenvalues. We diagonalize H:

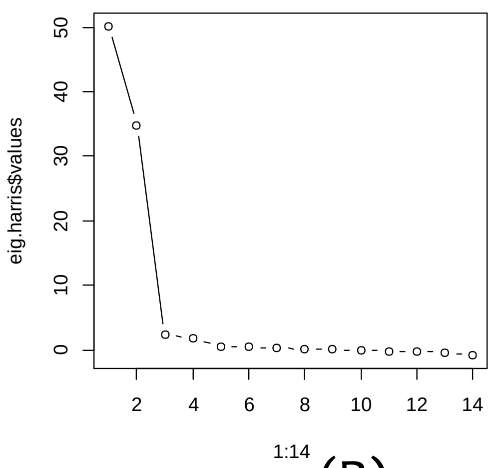
[11] We submit the following commands under SAS:

```
proc factor data = mesdata.beer_rnd
method=harris
msa
nfactors=2
score;
run;
```

```
#diagonalization
Eig.harris <- eigen(beer.cor.harris)
print("eigenvalues")
print(eig.harris$values)
```

Here also, we know why we can obtain negative eigenvalues (see section 3.2). The most interesting information is that the gap between the 2nd and the 3rd eigenvalues is really high. Undoubtedly, the choice of two factors is the right solution for our dataset. We dispose of 95.52% of the available information (shared between the variables) on the two first factors [(50.078 + 34.815) / 88.878 = 0.9552].

```
> print(eig.harris$values)
 [1] 50.07820931 34.81447959  2.36157964  1.81539252
 [5]  0.52336728  0.41672692  0.33417301  0.12974938
 [9]  0.04105762 -0.04830416 -0.20580390 -0.24486590
[13] -0.38182102 -0.75553010
```



(R)

Eigenvalues of the Weighted Reduced Correlation Matrix: Total = 88.8784102 Average = 6.34845787

	Eigenvalue	Difference	Proportion	Cumulative
1	50.0782093	15.2637297	0.5634	0.5634
2	34.8144796	32.4529000	0.3917	0.9552
3	2.3615796	0.5461871	0.0266	0.9817
4	1.8153925	1.2920252	0.0204	1.0022
5	0.5233673	0.1066404	0.0059	1.0080
6	0.4167269	0.0825539	0.0047	1.0127
7	0.3341730	0.2044236	0.0038	1.0165
8	0.1297494	0.0886918	0.0015	1.0179
9	0.0410576	0.0893618	0.0005	1.0184
10	-0.0483042	0.1574997	-0.0005	1.0179
11	-0.2058039	0.0390620	-0.0023	1.0156
12	-0.2448659	0.1369551	-0.0028	1.0128
13	-0.3818210	0.3737091	-0.0043	1.0085
14	-0.7555301		-0.0085	1.0000

(SAS)

Loadings or Factor pattern. This table is again used for the interpretation of the factors. The formula is slightly modified because we must take into account the uniqueness u_i of the variables:

```
#loadings
loadings.harris <- matrix(0,nrow=nrow(beer.cor.harris),ncol=2)
for (j in 1:2){
loadings.harris[,j] <- sqrt(eig.harris$values[j])*eig.harris$vectors[,j]*sqrt(d2)
}
print("loadings for the 2 first factors")
rownames(loadings.harris) <- colnames(beer.data)
print(round(loadings.harris,5))
```

The two groups of the variables are strongly highlighted with the Harris approach. Definitely, the randomly generated variables (RND) are not relevant.

The association of the variables with the factors is more clear, without need to rotate the factors (we will see below the factor rotation techniques, section 4).

(R)

```
> print(round(loadings.harris,5))
           [,1]     [,2]
cost    -0.96686  0.09576
size    -0.93749 -0.24530
alcohol -0.91821  0.15672
reputat  0.18924 -0.64742
color   -0.25172  0.87165
aroma    0.08793  0.91231
taste    0.06418  0.96662
rnd1    -0.19090 -0.06900
rnd2     0.04254  0.04813
rnd3    -0.05841  0.02748
rnd4    -0.06224  0.21959
rnd5     0.01283  0.00801
rnd6     0.02993 -0.01323
rnd7     0.05548 -0.02875
```

(SAS)

Factor Pattern			
		Factor1	Factor2
cost	cost	0.96686	0.09576
size	size	0.93749	-0.24530
alcohol	alcohol	0.91821	0.15672
reputat	reputat	-0.18924	-0.64742
color	color	0.25172	0.87165
aroma	aroma	-0.08793	0.91231
taste	taste	-0.06418	0.96662
rnd1	rnd1	0.19090	-0.06900
rnd2	rnd2	-0.04254	0.04813
rnd3	rnd3	0.05841	0.02748
rnd4	rnd4	0.06224	0.21959
rnd5	rnd5	-0.01283	0.00801
rnd6	rnd6	-0.02993	-0.01323
rnd7	rnd7	-0.05548	-0.02875

Unweighted variance. The weighted variance corresponds to the eigenvalue. SAS provides also the unweighted variance. This is the sum of the squared values of the loadings. Here, we obtain **2.81752** and **3.09661** for the first and the second factor.

Variance Explained by Each Factor		
Factor	Weighted	Unweighted
Factor1	50.0782093	2.81752174
Factor2	34.8144796	3.09660636

We can calculate easily these values with R.

```
#unweighted variance explained
unweighted.var.harris <- apply(loadings.harris,2,function(x){sum(x^2)})
print(round(unweighted.var.harris,5))
```

Communalities. We add up the squared values of loadings per variable on the selected factors to obtain the communalities.

```
#communalities
print("communalities for the 2 first factors")
comm.harris <- apply(loadings.harris,1,function(x){sum(x^2)})
print(round(cbind(init.comm,comm.harris),5))
```

We can compare these values with the initial communalities to evaluate the quality of representation of each variable.

```
> print(round(cbind(init.comm,comm.harris),5))
        init.comm comm.harris
cost      0.96105     0.94399
size      0.94389     0.93907
alcohol   0.91234     0.86767
reputat   0.77232     0.45497
color     0.85328     0.82313
aroma     0.88680     0.84004
taste     0.95027     0.93847
rnd1      0.13826     0.04120
rnd2      0.08495     0.00413
rnd3      0.07357     0.00417
rnd4      0.14240     0.05209
rnd5      0.11144     0.00023
rnd6      0.09628     0.00107
rnd7      0.08686     0.00390
```

Factor scores. The factor scores are computed like for the principal factor analysis.

```
#factor scores
print("factor scores")
fscores.harris <- inv.beer.cor%*%loadings.harris
print(round(fscores.harris,5))
#variance of the scores
vscores.harris <- numeric(2)
for (j in 1:2){
 vscores.harris[j] <- sum(fscores.harris[,j]*loadings.harris[,j])
}
print(round(vscores.harris,5))
```

R and SAS are also consistent here.

(R)

```
> print(round(fscores.harris,5))
            [,1]     [,2]
cost    -0.48598  0.06864
size    -0.32709 -0.12206
alcohol -0.20506  0.04992
reputat  0.01627 -0.07940
color   -0.03359  0.16588
aroma    0.01521  0.22503
taste    0.02527  0.54272
rnd1    -0.00434 -0.00224
rnd2     0.00091  0.00147
rnd3    -0.00123  0.00083
rnd4    -0.00142  0.00715
rnd5     0.00028  0.00025
rnd6     0.00065 -0.00041
rnd7     0.00119 -0.00088
```

```
> print(round(vscores.harris,5))
[1] 0.98042 0.97208
```

(SAS)

Standardized Scoring Coefficients

		Factor1	Factor2
cost	cost	0.48598	0.06864
size	size	0.32709	-0.12206
alcohol	alcohol	0.20506	0.04992
reputat	reputat	-0.01627	-0.07940
color	color	0.03359	0.16588
aroma	aroma	-0.01521	0.22503
taste	taste	-0.02527	0.54272
rnd1	rnd1	0.00434	-0.00224
rnd2	rnd2	-0.00091	0.00147
rnd3	rnd3	0.00123	0.00083
rnd4	rnd4	0.00142	0.00715
rnd5	rnd5	-0.00028	0.00025
rnd6	rnd6	-0.00065	-0.00041
rnd7	rnd7	-0.00119	-0.00088

Squared Multiple Correlations of the Variables with Each Factor

Factor1	Factor2
0.98042218	0.97207833

Contribution of the variables to the factors. The factor scores coefficients allows to obtain the relative influence of the variables on the factors.

We observe that the influence of the randomly generated variables (RND) on the first two factors is near zero. This is the desirable result that we expect since the beginning of this tutorial.

	Standardized Scoring Coefficients		Squared Coefficients		Contributions	
	Factor1	Factor2	Factor1	Factor2	Factor1	Factor2
cost	0.48598	0.06864	0.23618	0.00471	0.60948	0.01174
size	0.32709	-0.12206	0.10699	0.01490	0.27610	0.03714
alcohol	0.20506	0.04992	0.04205	0.00249	0.10851	0.00621
reputat	-0.01627	-0.0794	0.00026	0.00630	0.00068	0.01572
color	0.03359	0.16588	0.00113	0.02752	0.00291	0.06859
aroma	-0.01521	0.22503	0.00023	0.05064	0.00060	0.12623
taste	-0.02527	0.54272	0.00064	0.29454	0.00165	0.73422
rnd1	0.00434	-0.00224	0.00002	0.00001	0.00005	0.00001
rnd2	-0.00091	0.00147	0.00000	0.00000	0.00000	0.00001
rnd3	0.00123	0.00083	0.00000	0.00000	0.00000	0.00000
rnd4	0.00142	0.00715	0.00000	0.00005	0.00001	0.00013
rnd5	-0.00028	0.00025	0.00000	0.00000	0.00000	0.00000
rnd6	-0.00065	-0.00041	0.00000	0.00000	0.00000	0.00000
rnd7	-0.00119	-0.00088	0.00000	0.00000	0.00000	0.00000
Total			0.38750	0.40117		
CTR(rnd)					0.01%	0.01%

3.5 Comparison of the three approaches

The tables of loadings and contributions are the tools that we use to compare the approaches studied in this paper. We observe that they provide similar results (Figure 10).

	Factor Pattern - PCA		Factor Pattern - PFA		Factor Pattern - Harris	
	Factor1	Factor2	Factor1	Factor2	Factor1	Factor2
cost	0.49678	0.81407	0.52442	0.80117	0.96686	0.09576
size	0.21378	0.94733	0.24043	0.93787	0.93749	-0.2453
alcohol	0.58837	0.7616	0.60493	0.73065	0.91821	0.15672
reputat	-0.73682	0.11434	-0.69728	0.13038	-0.18924	-0.64742
color	0.90757	-0.18174	0.88243	-0.20296	0.25172	0.87165
aroma	0.78387	-0.49557	0.76236	-0.51145	-0.08793	0.91231
taste	0.80783	-0.49864	0.80095	-0.52573	-0.06418	0.96662
rnd1	0.01831	0.30272	0.02232	0.20878	0.1909	-0.069
rnd2	0.04235	-0.08543	0.0293	-0.06015	-0.04254	0.04813
rnd3	0.11864	0.04597	0.08501	0.03166	0.05841	0.02748
rnd4	0.30514	-0.08602	0.22796	-0.06342	0.06224	0.21959
rnd5	0.01361	-0.00533	0.00843	-0.00856	-0.01283	0.00801
rnd6	-0.04716	-0.01364	-0.03627	-0.01181	-0.02993	-0.01323
rnd7	-0.05046	-0.07406	-0.04059	-0.04624	-0.05548	-0.02875

Figure 10 – Comparison of methods - "Loadings" – Unrotated factors

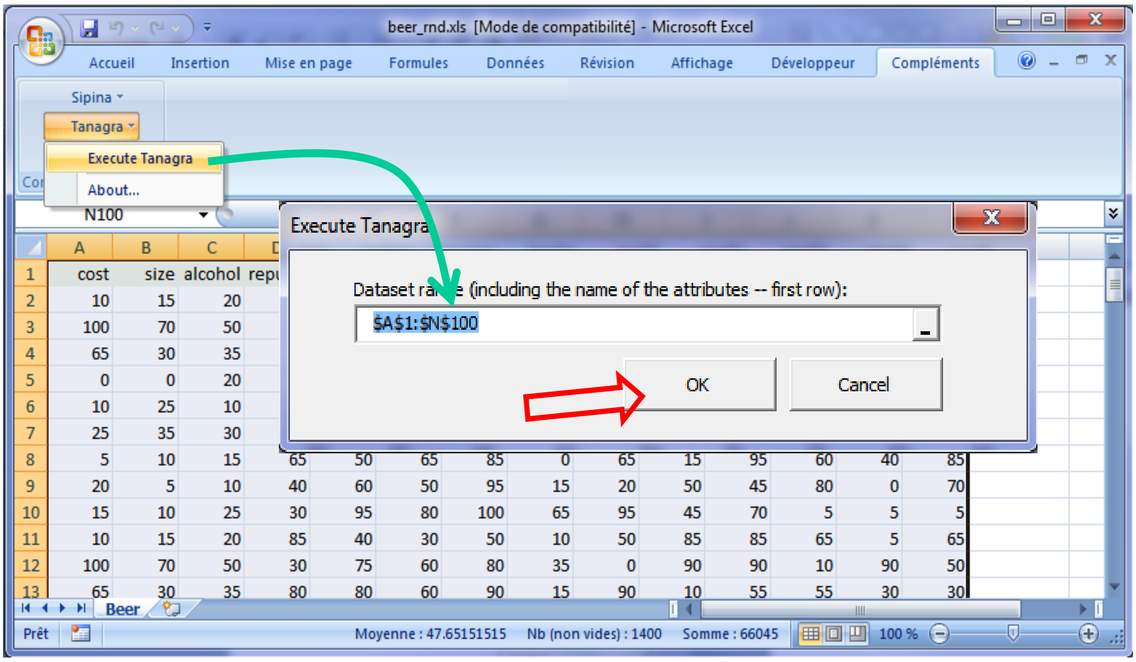
Perhaps, Harris is the more interesting in our context because the contribution of the RND variables is near zero on the selected factors (the two first ones). The groups are immediately identified. However, as we will see in the following section, all the methods are equivalent after the factor rotation process.

4 Factor analysis with Tanagra

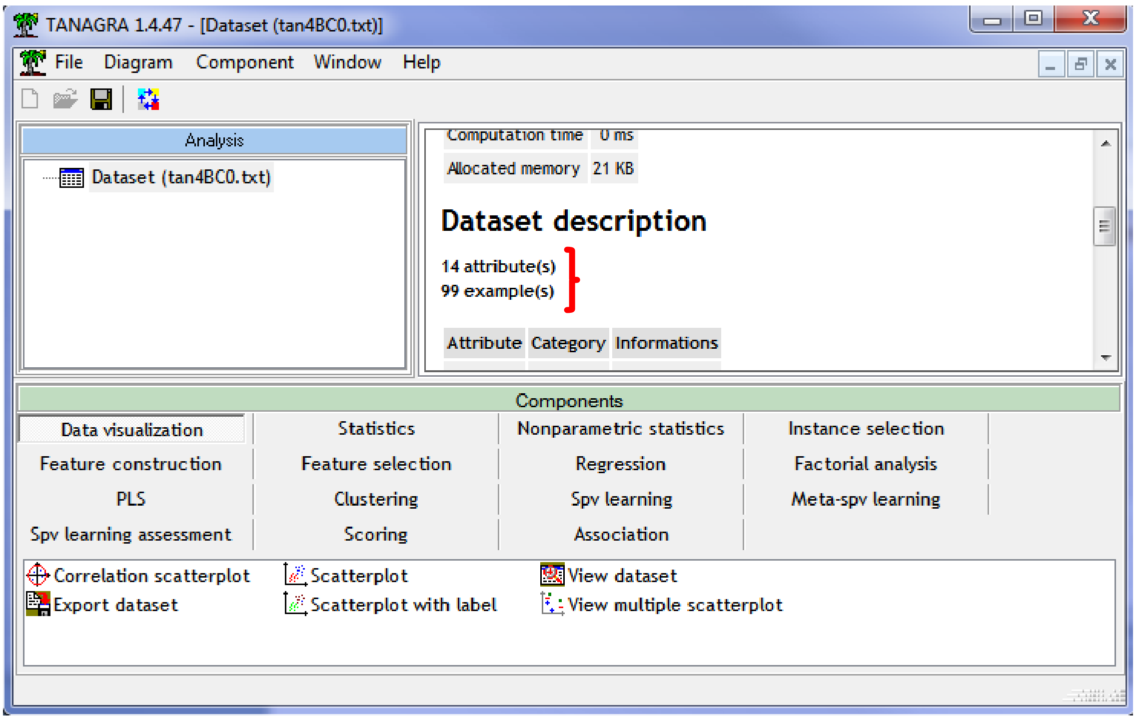
The principal factor analysis and the Harris approach described above are implemented in Tanagra 1.4.47. In this section, we show how to use them on the "beer_rnd.xls" dataset. Of course, the results are strictly identical to those R and SAS. Tanagra stands appart from the others by the formatting of the reports. We use also the VARIMAX[12] orthogonal rotation in this section.

4.1 Importing the dataset

We use the add-in "tanagra.xla" to send the dataset from the Excel spreadsheet to Tanagra[13].



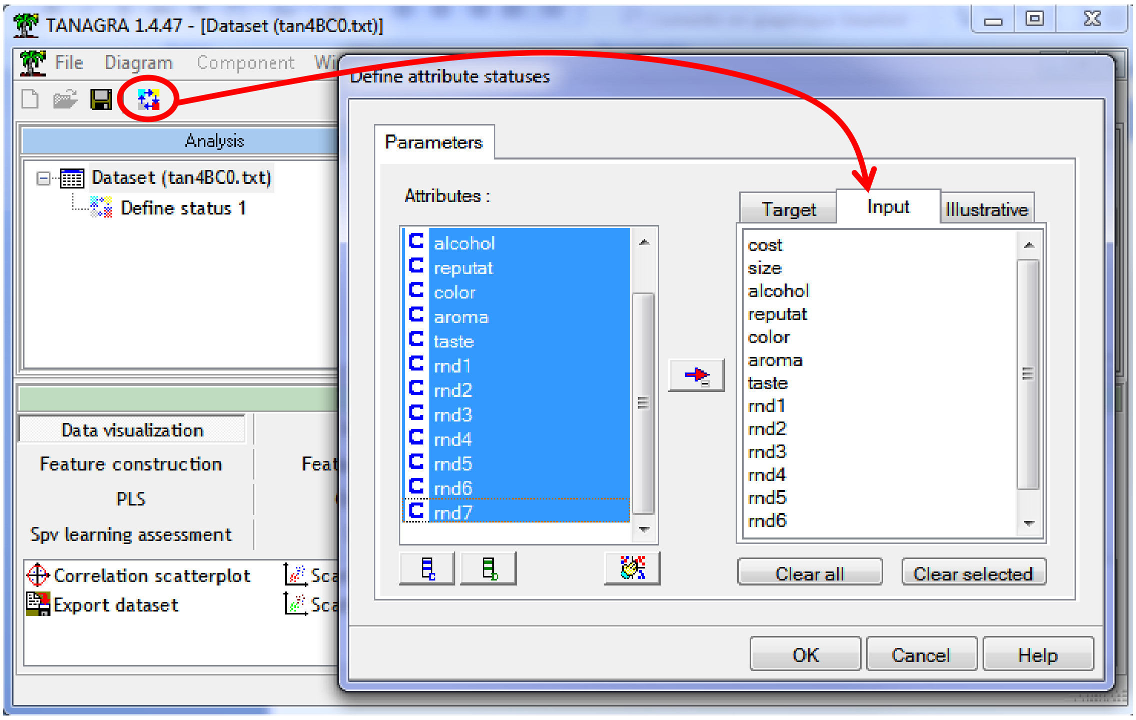
Tanagra is launched and the dataset loaded. We have n = 99 instances and p = 14 variables.



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

[13] http://data-mining-tutorials.blogspot.fr/2010/08/tanagra-add-in-for-office-2007-and.html

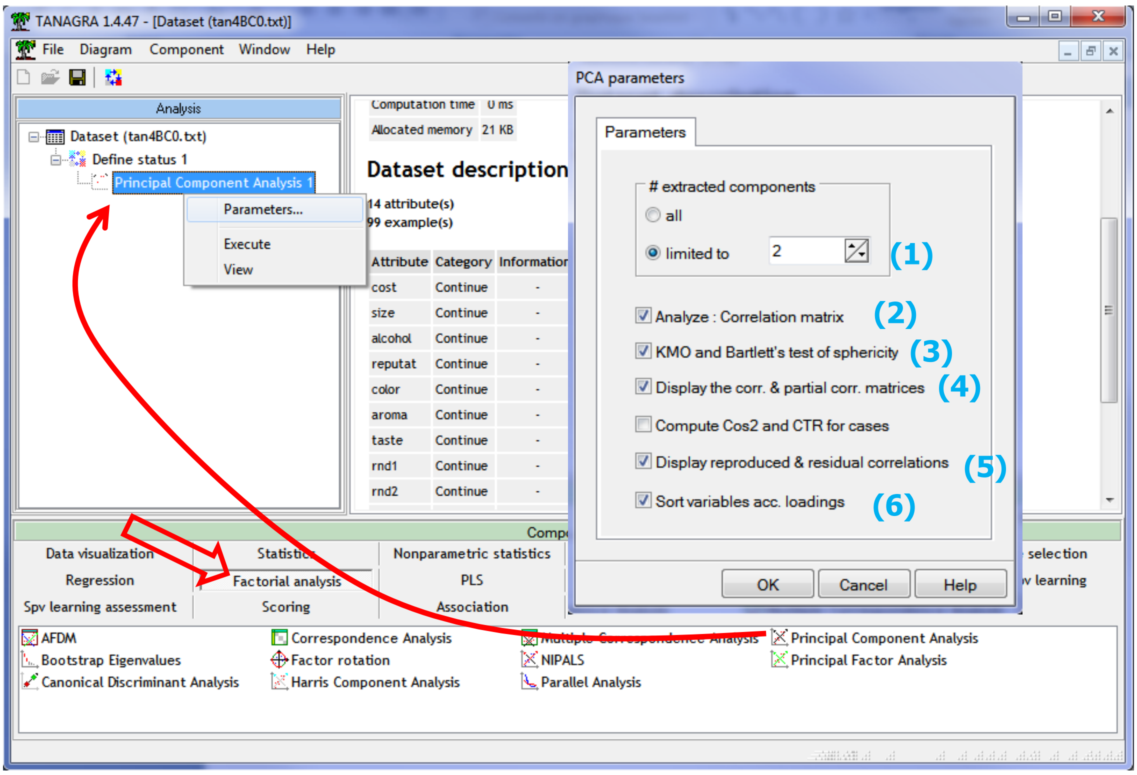
To start the analysis, we must define the role of the variables. We add the DEFINE STATUS component into the diagram. We set all the variables as INPUT.



4.2 Principal component analysis and VARIMAX rotation

4.2.1 Principal component analysis

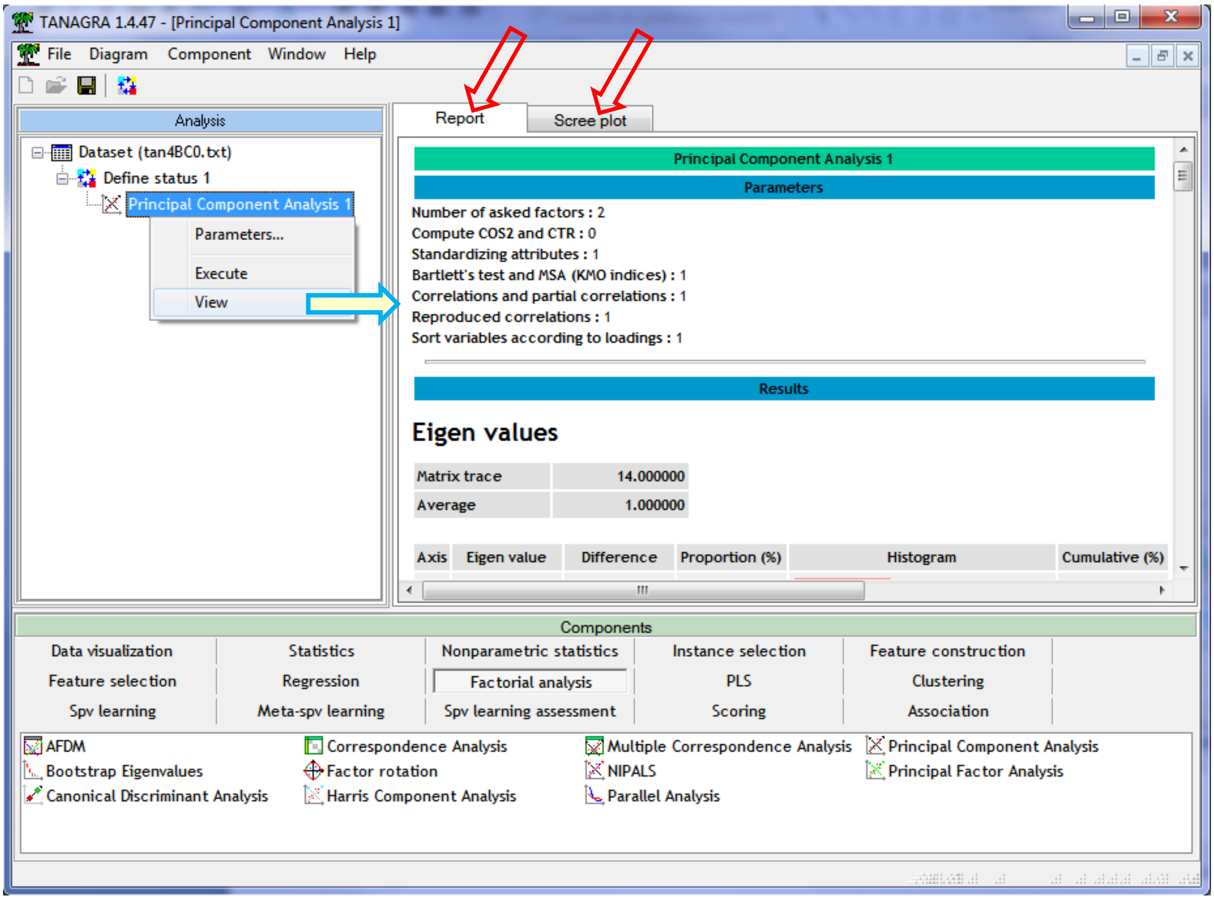
We insert the tool PRINCIPAL COMPONENT ANALYSIS (Factorial Analysis tab) to perform the PCA. We click on the contextual menu PARAMETERS to set the settings.



Here are the selected options for our study:

1. We select 2 factors.
2. We perform a PCA based on the correlation matrix.
3. The MSA (measure of sampling adequacy of Kaiser-Mayer-Olkin) and the Bartlett's test for sphericity are computed.
4. The correlation matrix and the partial correlation matrix are displayed.
5. The reproduced correlations by the selected factors of PCA and the residuals are displayed.
6. The variables are sorted according to the loadings into the table. It enables to better identify the group of variables. It is especially useful when the number of variables is large.

We confirm these settings. We obtain the results by clicking on the VIEW menu.



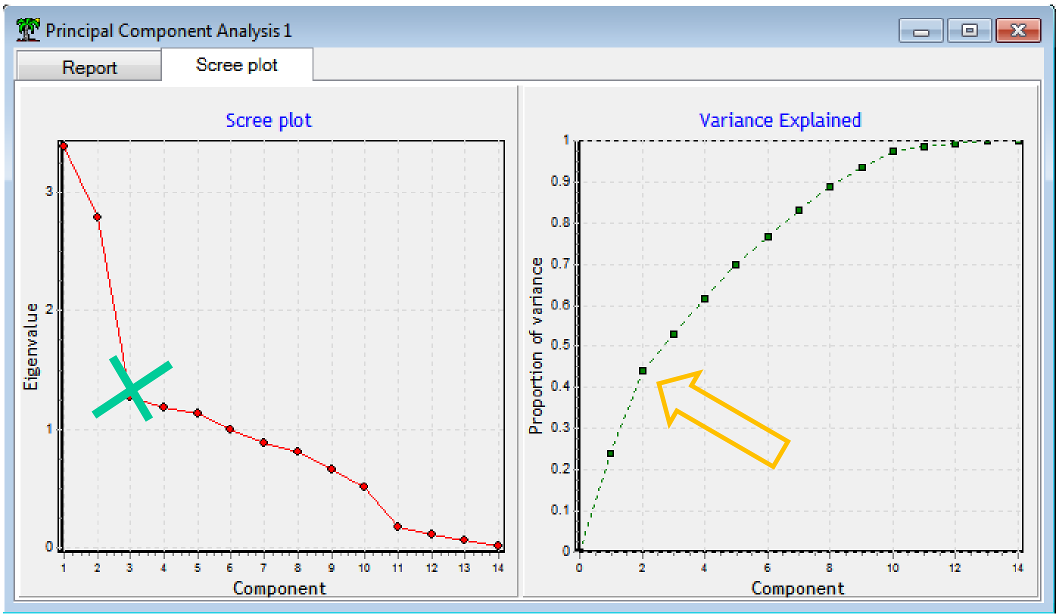
Eigenvalues. The table of eigenvalues shows also the proportion of explained variance by the factors.

Eigen values

Matrix trace	14.000000
Average	1.000000

Axis	Eigen value	Difference	Proportion (%)	Histogram	Cumulative (%)
1	3.386557	0.591892	24.19 %		24.19 %
2	2.794665	1.527068	19.96 %		44.15 %
3	1.267596	0.085424	9.05 %		53.21 %
4	1.182172	0.052486	8.44 %		61.65 %
5	1.129687	0.136967	8.07 %		69.72 %
6	0.992720	0.108850	7.09 %		76.81 %
7	0.883870	0.068416	6.31 %		83.12 %
8	0.815454	0.150809	5.82 %		88.95 %
9	0.664645	0.154055	4.75 %		93.70 %
10	0.510590	0.337377	3.65 %		97.34 %
11	0.173213	0.060820	1.24 %		98.58 %
12	0.112392	0.041557	0.80 %		99.38 %
13	0.070835	0.055232	0.51 %		99.89 %
14	0.015603	-	0.11 %		100.00 %
Tot.	14.000000	-	-	-	-

Scree plot. The scree plot shows the decreasing of the eigenvalues according to the number of the factors. Tanagra provides also the cumulative fraction of total variance explained by the factors. These plots are useful for the selection of the factors to retain for the interpretation of the results. Here, the choice of two factors seems the most appropriate.



Other tools for the detection of the right number of factors. Tanagra incorporates other tools for the determination of the right number of factors. Clearly, the Kaiser-Guttman rule (selecting the factors for which the corresponding eigenvalue is higher to 1) is not appropriate here. It leads us to retain 5 or 6 factors.

The Karlis-Saporta-Spinaki test (A) is better, among other things, because it takes into account the sample size (n), the number of variables (p), and the ratio p/n. It recommends two factors for our dataset.

The broken-stick test (B) (Legendre) detects also two relevant factors[14].

Significance of Principal Components

Global critical values	
Kaiser-Guttman	1
Karlis-Saporta-Spinaki	1.72843

(A)

Eigenvalue table - Test for significance

(B)

Eigenvalues - Significance		
Axis	Eigenvalue	Broken-stick critical values
1	3.386557	3.251562
2	2.794665	2.251562
3	1.267596	1.751562
4	1.182172	1.418229
5	1.129687	1.168229
6	0.992720	0.968229
7	0.883870	0.801562
8	0.815454	0.658705
9	0.664645	0.533705
10	0.510590	0.422594
11	0.173213	0.322594
12	0.112392	0.231685
13	0.070835	0.148352
14	0.015603	0.071429

Bartlett's test of sphericity. It enables to check the existence of at least one factor. Its main drawback is that it is always significant when the dataset size (n) increases.

Bartlett's test of sphericity

Bartlett's test	
|CORR.MATRIX|	8.370766E-5
CHISQ	868.4067
d.f.	91
p-value	4.000073E-127

MSA - Measure of Sampling Adequacy (KMO index). The MSA indicates the redundancy between the variables, advertising the possibility to obtain an efficient factorization. Here, the global value is not really good (MSA = 0.491). But, it corresponds mainly to the existence of the variables generated randomly into the dataset. That does not mean that we cannot obtain interesting results in the PCA.

Kaiser's Measure of Sampling Adequacy (MSA)

(Tanagra)

Overall MSA = 0.4910682									
cost	0.3962305	size	0.4987689	alcohol	0.5549174	reputat	0.3635211	color	0.8160946
aroma	0.5523418	taste	0.4255714	rnd1	0.5366791	rnd2	0.2554571	rnd3	0.5098051
rnd4	0.6441655	rnd5	0.215428	rnd6	0.3770795	rnd7	0.2774695		

(SAS)

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.49106818

cost	size	alcohol	reputat	color	aroma	taste	rnd1	rnd2	rnd3	rnd4	rnd5	rnd6	rnd7
0.39623047	0.49876893	0.55491737	0.36352108	0.81609457	0.55234179	0.42557140	0.53667911	0.25545708	0.50980512	0.64416554	0.21542800	0.37707955	0.27746946

[14] See http://data-mining-tutorials.blogspot.fr/2013/01/choosing-number-of-components-in-pca.html ; and http://data-mining-tutorials.blogspot.fr/2013/01/new-features-for-pca-in-tanagra.html

Factor loadings. The variables can be sorted according to the absolute value of the loadings in Tanagra. The variables with loadings higher than 0.5 are sorted in decreasing order for the first factor. Then, for the remaining variables, those for which the loadings are higher than 0.5 for the second factor are sorted. Etc. The goal is to distinguish the group of variables associated to the factors. For our dataset, we observe that (color, taste, aroma and reputat) are related to the first factor; (alcohol, size and cost) to the second factor[15].

Factor Loadings [Communality Estimates]

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
color	-0.90757	82 % (82 %)	-0.18174	3 % (86 %)
taste	-0.80783	65 % (65 %)	-0.49864	25 % (90 %)
aroma	-0.78387	61 % (61 %)	-0.49557	25 % (86 %)
reputat	0.73682	54 % (54 %)	0.11434	1 % (56 %)
alcohol	-0.58837	35 % (35 %)	0.76160	58 % (93 %)
size	-0.21378	5 % (5 %)	0.94733	90 % (94 %)
cost	-0.49678	25 % (25 %)	0.81407	66 % (91 %)
rnd1	-0.01831	0 % (0 %)	0.30272	9 % (9 %)
rnd4	-0.30514	9 % (9 %)	-0.08602	1 % (10 %)
rnd2	-0.04235	0 % (0 %)	-0.08543	1 % (1 %)
rnd7	0.05046	0 % (0 %)	-0.07406	1 % (1 %)
rnd3	-0.11864	1 % (1 %)	0.04597	0 % (2 %)
rnd6	0.04716	0 % (0 %)	-0.01364	0 % (0 %)
rnd5	-0.01361	0 % (0 %)	-0.00533	0 % (0 %)
Var. Expl.	3.38656	24 % (24 %)	2.79466	20 % (44 %)

Factor scores. The factor scores coefficient enables to compute the coordinates of the individuals.

Factor Scores

Attribute	Mean	Std-dev	Axis_1	Axis_2
cost	27.7777778	31.1903752	-0.2699491	0.4869663
size	22.2222222	20.1537302	-0.1161680	0.5666762
alcohol	23.8888889	12.1969436	-0.3197190	0.4555749
reputat	55.5555556	25.7600514	0.4003883	0.0683939
color	63.8888889	18.0705066	-0.4931756	-0.1087115
aroma	56.1111111	19.6889391	-0.4259543	-0.2964452
taste	80.5555556	17.2311805	-0.4389765	-0.2982811
rnd1	42.7777778	28.7379507	-0.0099492	0.1810839
rnd2	52.4242424	27.8012756	-0.0230128	-0.0511029
rnd3	49.9494949	25.8833333	-0.0644687	0.0274971
rnd4	46.5151515	27.6381246	-0.1658117	-0.0514555
rnd5	46.8181818	25.8243342	-0.0073931	-0.0031866
rnd6	47.0202020	29.7796554	0.0256286	-0.0081575
rnd7	51.6161616	29.0404480	0.0274217	-0.0443045

According to the French school of principal component analysis, the variance of the scores corresponds to the eigenvalue associated to the factor (this variance is 1 for the other tools). Tanagra uses the original order of the variables in this table.

[15] The correlation is highlighted in light red if the absolute value is higher than 0.5, dark red if it is higher than 0.7.

Correlation matrix. Tanagra can display the correlation matrix. To better identify the group of variables, they are sorted in the same way that the "Factor Loadings" table. The cell is highlighted if the absolute value of the correlation is higher than 0.5 (darker color if higher than 0.7).

Correlations

	color	taste	aroma	reputat	alcohol	size	cost	rnd1	rnd4	rnd2	rnd7	rnd3	rnd6	rnd5
color	1.00000	0.80487	0.82324	-0.52380	0.39770	0.01441	0.32089	-0.01448	0.24506	0.10690	0.05395	0.06197	-0.08546	0.02435
taste	0.80487	1.00000	0.86607	-0.62650	0.05580	-0.30751	0.05398	-0.08216	0.20609	0.03409	-0.04116	-0.00333	0.02734	-0.01248
aroma	0.82324	0.86607	1.00000	-0.52151	0.09768	-0.28624	-0.02764	-0.04518	0.15190	0.06705	-0.01286	0.04372	-0.05120	0.03874
reputat	-0.52380	-0.62650	-0.52151	1.00000	-0.36051	-0.06123	-0.17478	0.05420	-0.15086	0.05383	0.09095	-0.09729	-0.04722	0.03872
alcohol	0.39770	0.05580	0.09768	-0.36051	1.00000	0.82367	0.87702	0.18243	0.07691	-0.02855	-0.08120	0.09021	-0.08003	0.00080
size	0.01441	-0.30751	-0.28624	-0.06123	0.82367	1.00000	0.87839	0.20604	-0.02101	-0.03576	-0.02685	0.05976	-0.00075	-0.03833
cost	0.32089	0.05398	-0.02764	-0.17478	0.87702	0.87839	1.00000	0.16606	0.10116	-0.05174	-0.06239	0.03302	-0.02290	-0.00188
rnd1	-0.01448	-0.08216	-0.04518	0.05420	0.18243	0.20604	0.16606	1.00000	-0.10640	0.06711	-0.03806	-0.04395	0.10498	0.18715
rnd4	0.24506	0.20609	0.15190	-0.15086	0.07691	-0.02101	0.10116	-0.10640	1.00000	0.06358	0.06680	0.15684	-0.01967	0.08672
rnd2	0.10690	0.03409	0.06705	0.05383	-0.02855	-0.03576	-0.05174	0.06711	0.06358	1.00000	0.07021	-0.01317	0.05661	0.06702
rnd7	0.05395	-0.04116	-0.01286	0.09095	-0.08120	-0.02685	-0.06239	-0.03806	0.06680	0.07021	1.00000	0.01422	-0.02159	0.00652
rnd3	0.06197	-0.00333	0.04372	-0.09729	0.09021	0.05976	0.03302	-0.04395	0.15684	-0.01317	0.01422	1.00000	0.07188	-0.07391
rnd6	-0.08546	0.02734	-0.05120	-0.04722	-0.08003	-0.00075	-0.02290	0.10498	-0.01967	0.05661	-0.02159	0.07188	1.00000	-0.07734
rnd5	0.02435	-0.01248	0.03874	0.03872	0.00080	-0.03833	-0.00188	0.18715	0.08672	0.06702	0.00652	-0.07391	-0.07734	1.00000

Partial correlation matrix. The partial correlation measures the association between a pair of variables, by removing the influence of the (p-2) other variables of the dataset. For instance, the correlation between "color" and "taste" seems high (r = 0.80487). When we remove the influence of the other variables, we note that the correlation is not really high ultimately (partial r = 0.26931).

Partial Correlations Controlling all other Variables

	color	taste	aroma	reputat	alcohol	size	cost	rnd1	rnd4	rnd2	rnd7	rnd3	rnd6	rnd5
color	1.00000	0.26931	0.35225	0.16033	0.32208	-0.07819	0.04164	-0.04609	0.11999	0.15729	0.23054	0.04811	-0.09071	-0.00591
taste	0.26931	1.00000	0.66857	-0.76740	-0.59295	-0.67220	0.79647	0.09987	-0.05932	0.08274	-0.02899	-0.08828	0.10559	-0.19395
aroma	0.35225	0.66857	1.00000	0.41675	0.40248	0.39883	-0.60429	-0.02879	-0.03478	-0.10066	-0.07066	0.06435	-0.05617	0.16120
reputat	0.16033	-0.76740	0.41675	1.00000	-0.63251	-0.52231	0.69590	0.14477	-0.10151	0.14230	0.03222	-0.06704	-0.02766	-0.12578
alcohol	0.32208	-0.59295	0.40248	-0.63251	1.00000	-0.12422	0.60690	0.13322	-0.07289	0.04776	-0.10471	0.01231	-0.08237	-0.08363
size	-0.07819	-0.67220	0.39883	-0.52231	-0.12422	1.00000	0.82016	0.11772	-0.12588	0.14953	0.11456	0.00238	0.08943	-0.19381
cost	0.04164	0.79647	-0.60429	0.69590	0.60690	0.82016	1.00000	-0.11283	0.13497	-0.17157	-0.07135	-0.01195	-0.01955	0.18541
rnd1	-0.04609	0.09987	-0.02879	0.14477	0.13322	0.11772	-0.11283	1.00000	-0.09119	0.04247	-0.02403	-0.03168	0.13235	0.21998
rnd4	0.11999	-0.05932	-0.03478	-0.10151	-0.07289	-0.12588	0.13497	-0.09119	1.00000	0.06700	0.05631	0.16150	-0.00340	0.09565
rnd2	0.15729	0.08274	-0.10066	0.14230	0.04776	0.14953	-0.17157	0.04247	0.06700	1.00000	0.00949	-0.02803	0.07720	0.07202
rnd7	0.23054	-0.02899	-0.07066	0.03222	-0.10471	0.11456	-0.07135	-0.02403	0.05631	0.00949	1.00000	-0.00123	-0.00663	0.01176
rnd3	0.04811	-0.08828	0.06435	-0.06704	0.01231	0.00238	-0.01195	-0.03168	0.16150	-0.02803	-0.00123	1.00000	0.09829	-0.08039
rnd6	-0.09071	0.10559	-0.05617	-0.02766	-0.08237	0.08943	-0.01955	0.13235	-0.00340	0.07720	-0.00663	0.09829	1.00000	-0.06659
rnd5	-0.00591	-0.19395	0.16120	-0.12578	-0.08363	-0.19381	0.18541	0.21998	0.09565	0.07202	0.01176	-0.08039	-0.06659	1.00000

Original, reproduced and residual correlations. This table shows the ability of the PCA to reproduce the correlations between the variables using the selected factors.

We observe: (1) the correlation obtained from the correlation matrix underlying the PCA; (2) the correlation reproduced by the selected factors, obtained from the factor loadings; (3) the difference between the measured correlation and the reproduced correlation.

Here, Tanagra highlights the high correlation which are well reproduced i.e. the measured correlation is higher than '0.5' in absolute value, the residual correlation is lower than '0.05' in absolute value.

Original, reproduced and residual correlations

	color	taste	aroma	reputat	alcohol	size	cost	rnd1	rnd4	rnd2	rnd7	rnd3	rnd6	rnd5
color	-	0.8049 0.8238 (-0.0189)	0.8232 0.8015 (0.0218)	-0.5238 -0.6895 (0.1657)	0.3977 0.3956 (0.0021)	0.0144 0.0219 (-0.0074)	0.3209 0.3029 (0.0180)	-0.0145 -0.0384 (0.0239)	0.2451 0.2926 (-0.0475)	0.1069 0.0540 (0.0529)	0.0539 -0.0323 (0.0863)	0.0620 0.0993 (-0.0374)	-0.0855 -0.0403 (-0.0451)	0.0244 0.0133 (0.0110)
taste	0.8049 0.8238 (-0.0189)	-	0.8661 0.8803 (-0.0143)	-0.6265 -0.6522 (0.0257)	0.0558 0.0955 (-0.0397)	-0.3075 -0.2997 (-0.0078)	0.0540 -0.0046 (0.0586)	-0.0822 -0.1362 (0.0540)	0.2061 0.2894 (-0.0833)	0.0341 0.0768 (-0.0427)	-0.0412 -0.0038 (-0.0373)	-0.0033 0.0729 (-0.0763)	0.0273 -0.0313 (0.0586)	-0.0125 0.0136 (-0.0261)
aroma	0.8232 0.8015 (0.0218)	0.8661 0.8803 (-0.0143)	-	-0.5215 -0.6342 (0.1127)	0.0977 0.0838 (0.0139)	-0.2862 -0.3019 (0.0157)	-0.0276 -0.0140 (-0.0136)	-0.0452 -0.1357 (0.0905)	0.1519 0.2818 (-0.1299)	0.0670 0.0755 (-0.0085)	-0.0129 -0.0029 (-0.0100)	0.0437 0.0702 (-0.0265)	-0.0512 -0.0302 (-0.0210)	0.0387 0.0133 (0.0254)
reputat	-0.5238 -0.6895 (0.1657)	-0.6265 -0.6522 (0.0257)	-0.5215 -0.6342 (0.1127)	-	-0.3605 -0.3464 (-0.0141)	-0.0612 -0.0492 (-0.0120)	-0.1748 -0.2730 (0.0982)	0.0542 0.0211 (0.0331)	-0.1509 -0.2347 (0.0838)	0.0538 -0.0410 (0.0948)	0.0910 0.0287 (0.0622)	-0.0973 -0.0822 (-0.0151)	-0.0472 0.0332 (-0.0804)	0.0387 -0.0106 (0.0494)
alcohol	0.3977 0.3956 (0.0021)	0.0558 0.0955 (-0.0397)	0.0977 0.0838 (0.0139)	-0.3605 -0.3464 (-0.0141)	-	0.8237 0.8473 (-0.0236)	0.8770 0.9123 (-0.0353)	0.1824 0.2413 (-0.0589)	0.0769 0.1140 (-0.0371)	-0.0285 -0.0401 (0.0116)	-0.0812 -0.0861 (0.0049)	0.0902 0.1048 (-0.0146)	-0.0800 -0.0381 (-0.0419)	0.0008 0.0039 (-0.0031)
size	0.0144 0.0219 (-0.0074)	-0.3075 -0.2997 (-0.0078)	-0.2862 -0.3019 (0.0157)	-0.0612 -0.0492 (-0.0120)	0.8237 0.8473 (-0.0236)	-	0.8784 0.8774 (0.0010)	0.2060 0.2907 (-0.0847)	-0.0210 -0.0163 (-0.0047)	-0.0358 -0.0719 (0.0361)	-0.0268 -0.0810 (0.0541)	0.0598 0.0689 (-0.0092)	-0.0007 -0.0230 (0.0223)	-0.0383 -0.0021 (-0.0362)
cost	0.3209 0.3029 (0.0180)	0.0540 -0.0046 (0.0586)	-0.0276 -0.0140 (-0.0136)	-0.1748 -0.2730 (0.0982)	0.8770 0.9123 (-0.0353)	0.8784 0.8774 (0.0010)	-	0.1661 0.2555 (-0.0895)	0.1012 0.0816 (0.0196)	-0.0517 -0.0485 (-0.0032)	-0.0624 -0.0854 (0.0230)	0.0330 0.0964 (-0.0633)	-0.0229 -0.0345 (0.0116)	-0.0019 0.0024 (-0.0043)
rnd1	-0.0145 -0.0384 (0.0239)	-0.0822 -0.1362 (0.0540)	-0.0452 -0.1357 (0.0905)	0.0542 0.0211 (0.0331)	0.1824 0.2413 (-0.0589)	0.2060 0.2907 (-0.0847)	0.1661 0.2555 (-0.0895)	-	-0.1064 -0.0205 (-0.0859)	0.0671 -0.0251 (0.0922)	-0.0381 -0.0233 (-0.0147)	-0.0439 0.0161 (-0.0600)	0.1050 -0.0050 (0.1100)	0.1871 -0.0014 (0.1885)
rnd4	0.2451 0.2926 (-0.0475)	0.2061 0.2894 (-0.0833)	0.1519 0.2818 (-0.1299)	-0.1509 -0.2347 (0.0838)	0.0769 0.1140 (-0.0371)	-0.0210 -0.0163 (-0.0047)	0.1012 0.0816 (0.0196)	-0.1064 -0.0205 (-0.0859)	-	0.0636 0.0203 (0.0433)	0.0668 -0.0090 (0.0758)	0.1568 0.0322 (0.1246)	-0.0197 -0.0132 (-0.0065)	0.0867 0.0046 (0.0821)
rnd2	0.1069 0.0540 (0.0529)	0.0341 0.0768 (-0.0427)	0.0670 0.0755 (-0.0085)	0.0538 -0.0410 (0.0948)	-0.0285 -0.0401 (0.0116)	-0.0358 -0.0719 (0.0361)	-0.0517 -0.0485 (-0.0032)	0.0671 -0.0251 (0.0922)	0.0636 0.0203 (0.0433)	-	0.0702 0.0042 (0.0660)	-0.0132 0.0011 (-0.0143)	0.0566 -0.0008 (0.0574)	0.0670 0.0010 (0.0660)
rnd7	0.0539 -0.0323 (0.0863)	-0.0412 -0.0038 (-0.0373)	-0.0129 -0.0029 (-0.0100)	0.0910 0.0287 (0.0622)	-0.0812 -0.0861 (0.0049)	-0.0268 -0.0810 (0.0541)	-0.0624 -0.0854 (0.0230)	-0.0381 -0.0233 (-0.0147)	0.0668 -0.0090 (0.0758)	0.0702 0.0042 (0.0660)	-	0.0142 -0.0094 (0.0236)	-0.0216 0.0034 (-0.0250)	0.0065 -0.0003 (0.0068)
rnd3	0.0620 0.0993 (-0.0374)	-0.0033 0.0729 (-0.0763)	0.0437 0.0702 (-0.0265)	-0.0973 -0.0822 (-0.0151)	0.0902 0.1048 (-0.0146)	0.0598 0.0689 (-0.0092)	0.0330 0.0964 (-0.0633)	-0.0439 0.0161 (-0.0600)	0.1568 0.0322 (0.1246)	-0.0132 0.0011 (-0.0143)	0.0142 -0.0094 (0.0236)	-	0.0719 -0.0062 (0.0781)	-0.0739 0.0014 (-0.0753)
rnd6	-0.0855 -0.0403 (-0.0451)	0.0273 -0.0313 (0.0586)	-0.0512 -0.0302 (-0.0210)	-0.0472 0.0332 (-0.0804)	-0.0800 -0.0381 (-0.0419)	-0.0007 -0.0230 (0.0223)	-0.0229 -0.0345 (0.0116)	0.1050 -0.0050 (0.1100)	-0.0197 -0.0132 (-0.0065)	0.0566 -0.0008 (0.0574)	-0.0216 0.0034 (-0.0250)	0.0719 -0.0062 (0.0781)	-	-0.0773 -0.0006 (-0.0768)
rnd5	0.0244 0.0133 (0.0110)	-0.0125 0.0136 (-0.0261)	0.0387 0.0133 (0.0254)	0.0387 -0.0106 (0.0494)	0.0008 0.0039 (-0.0031)	-0.0383 -0.0021 (-0.0362)	-0.0019 0.0024 (-0.0043)	0.1871 -0.0014 (0.1885)	0.0867 0.0046 (0.0821)	0.0670 0.0010 (0.0660)	0.0065 -0.0003 (0.0068)	-0.0739 0.0014 (-0.0753)	-0.0773 -0.0006 (-0.0768)	-

The reproduced correlation is obtained from the factor loadings. We detail the calculations for "color" and "aroma".

Factor Loadings [Communality Estimates]

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
color	-0.90757	*82 % (82 %)*	-0.18174	*3 % (86 %)*
taste	-0.80783	*65 % (65 %)*	-0.49864	*25 % (90 %)*
aroma	-0.78387	*61 % (61 %)*	-0.49557	*25 % (86 %)*
reputat	0.73682	*54 % (54 %)*	0.11434	*1 % (56 %)*
alcohol	-0.58837	*35 % (35 %)*	0.7616	*58 % (93 %)*
size	-0.21378	*5 % (5 %)*	0.94733	*90 % (94 %)*
cost	-0.49678	*25 % (25 %)*	0.81407	*66 % (91 %)*
rnd1	-0.01831	*0 % (0 %)*	0.30272	*9 % (9 %)*
rnd4	-0.30514	*9 % (9 %)*	-0.08602	*1 % (10 %)*
rnd2	-0.04235	*0 % (0 %)*	-0.08543	*1 % (1 %)*
rnd7	0.05046	*0 % (0 %)*	-0.07406	*1 % (1 %)*
rnd3	-0.11864	*1 % (1 %)*	0.04597	*0 % (2 %)*
rnd6	0.04716	*0 % (0 %)*	-0.01364	*0 % (0 %)*
rnd5	-0.01361	*0 % (0 %)*	-0.00533	*0 % (0 %)*
Var. Expl.	3.38656	*24 % (24 %)*	2.79466	*20 % (44 %)*

corr.	0.82324
axis 1	0.71142
axis 2	0.09006
reprod. corr.	0.80148
residual corr.	0.02176

The measured correlation is **0.82324**. Using the factor loadings table, we calculate:

Cor. Reproduced (color, aroma) = (-0.90757 x -0.78387) + (-0.18174 x -0.49557) = **0.80148**

We calculate the difference to obtain the residual:

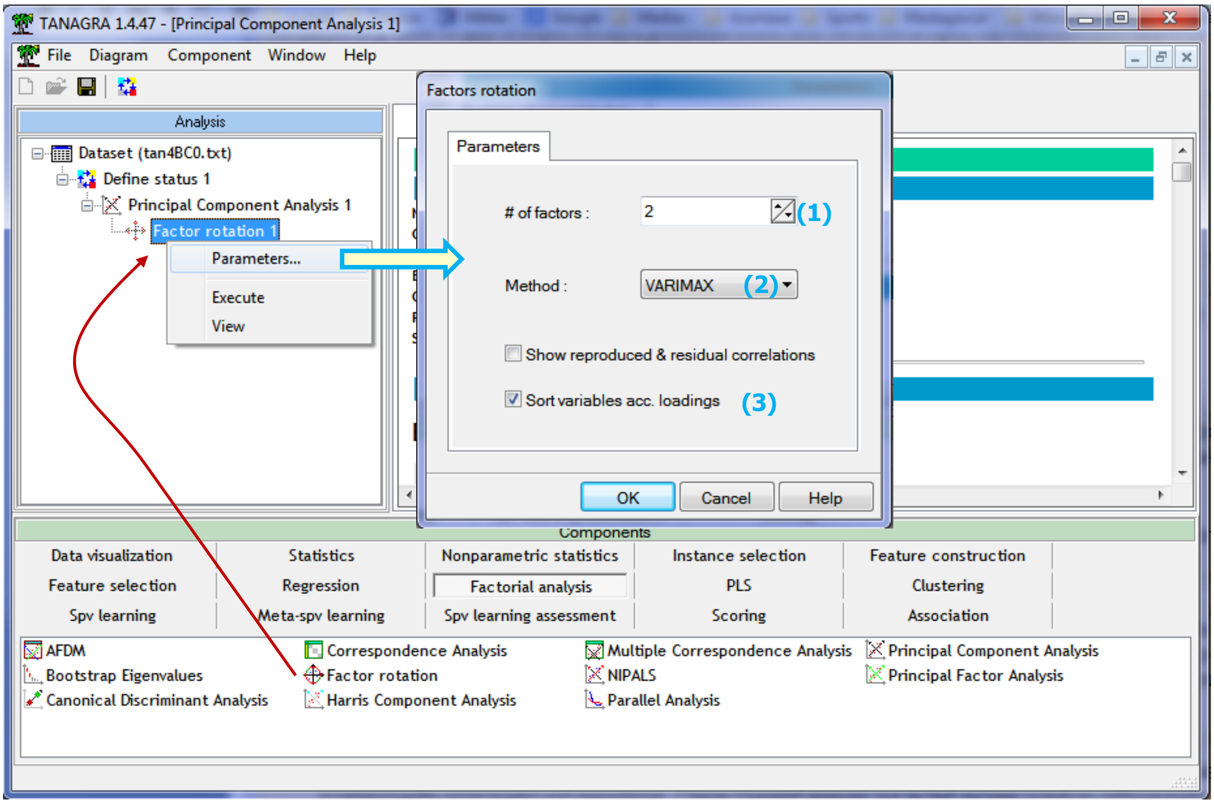
Cor. Residual (color, aroma) = 0.82324 – 0.80148 = **0.02176**

We note that if we include all the factors (14) in our analysis, the original correlation is perfectly reproduced by the PCA for all pairs of variables.

4.2.2 VARIMAX rotation based on two factors

The VARIMAX approach rotates the factors in order to obtain stronger associations between each variable and one of the selected factors. The goal is to make easier the interpretation of the results. The factors remain orthogonal.

We insert the FACTOR ROTATION tool (FACTORIAL ANALYSIS tab) into the diagram. We set the following settings: (1) we deal with two factors from the PCA; (2) we use the VARIMAX approach[16]; (3) the variables are sorted according to their loadings in the results table.



We confirm these options and we click on the VIEW menu.

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

Rotated Factor Loadings

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
taste	0.93638	88 % (88 %)	-0.15630	2 % (90 %)
aroma	0.91303	83 % (83 %)	-0.16252	3 % (86 %)
color	0.90893	83 % (83 %)	0.17480	3 % (86 %)
reputat	-0.72537	53 % (53 %)	-0.17266	3 % (56 %)
size	-0.16016	3 % (3 %)	0.95785	92 % (94 %)
cost	0.15221	2 % (2 %)	0.94145	89 % (91 %)
alcohol	0.25684	7 % (7 %)	0.92749	86 % (93 %)
rnd1	-0.09747	1 % (1 %)	0.28718	8 % (9 %)
rnd7	-0.01872	0 % (0 %)	-0.08764	1 % (1 %)
rnd3	0.09246	1 % (1 %)	0.08740	1 % (2 %)
rnd2	0.07150	1 % (1 %)	-0.06308	0 % (1 %)
rnd4	0.31501	10 % (10 %)	0.03570	0 % (10 %)
rnd6	-0.03851	0 % (0 %)	-0.03045	0 % (0 %)
rnd5	0.01461	0 % (0 %)	0.00021	0 % (0 %)
Var. Expl.	3.30199	24 % (24 %)	2.87923	21 % (44 %)

vs. Unrotated Factor Loadings

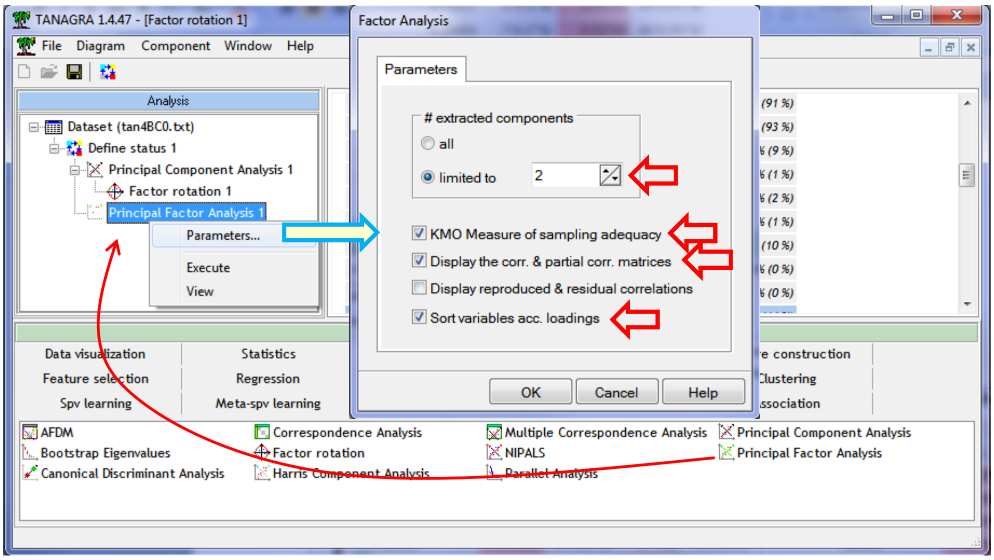
Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
taste	-0.80783	65 % (65 %)	-0.49864	25 % (90 %)
aroma	-0.78387	61 % (61 %)	-0.49557	25 % (86 %)
color	-0.90757	82 % (82 %)	-0.18174	3 % (86 %)
reputat	0.73682	54 % (54 %)	0.11434	1 % (56 %)
size	-0.21378	5 % (5 %)	0.94733	90 % (94 %)
cost	-0.49678	25 % (25 %)	0.81407	66 % (91 %)
alcohol	-0.58837	35 % (35 %)	0.76160	58 % (93 %)
rnd1	-0.01831	0 % (0 %)	0.30272	9 % (9 %)
rnd7	0.05046	0 % (0 %)	-0.07406	1 % (1 %)
rnd3	-0.11864	1 % (1 %)	0.04597	0 % (2 %)
rnd2	-0.04235	0 % (0 %)	-0.08543	1 % (1 %)
rnd4	-0.30514	9 % (9 %)	-0.08602	1 % (10 %)
rnd6	0.04716	0 % (0 %)	-0.01364	0 % (0 %)
rnd5	-0.01361	0 % (0 %)	-0.00533	0 % (0 %)
Var. Expl.	3.38656	24 % (24 %)	2.79466	20 % (44 %)

Tanagra shows the loadings after and before the rotation. We observe that the global variance explained by the selected factors is almost the same. But we have not the same repartition (3.30199 vs. 3.38656 for the 1st factor; 2.87923 vs. 2.79466 for the 2nd).

We note above all that the association of each original variable of the dataset (cost,…, taste) with one of the factors is very strong. The interpretation of the result becomes easier. The results are comparable to those of the Harris approach described in the previous section.

4.3 Principal factor analysis and varimax rotation

Principal factor analysis. We insert the PRINCIPAL FACTOR ANALYSIS tool into the diagram (FACTORIAL ANALYSIS tab). We set the following parameters (menu PARAMETERS).



We confirm and we click on the VIEW menu to obtain the results.

Compared to the PCA, some distinctive features can be noted. Into the loadings table, Tanagra displays the initial (prior) and the estimated communalities for the selected factors.

Factor Loadings [Communality Estimates]

Attribute	Communality Estimates		Axis_1		Axis_2	
-	Prior	Final	Corr.	Sq. (Cumul.)	Corr.	Sq. (Cumul.)
color	0.85328	0.81988	-0.88243	0.78 (0.78)	-0.20296	0.04 (0.82)
taste	0.95027	0.91791	-0.80095	0.64 (0.64)	-0.52573	0.28 (0.92)
aroma	0.88680	0.84278	-0.76236	0.58 (0.58)	-0.51145	0.26 (0.84)
reputat	0.77232	0.50319	0.69728	0.49 (0.49)	0.13038	0.02 (0.50)
alcohol	0.91234	0.89979	-0.60493	0.37 (0.37)	0.73065	0.53 (0.90)
cost	0.96105	0.91688	-0.52442	0.28 (0.28)	0.80117	0.64 (0.92)
size	0.94389	0.93740	-0.24043	0.06 (0.06)	0.93787	0.88 (0.94)
rnd1	0.13826	0.04409	-0.02232	0.00 (0.00)	0.20878	0.04 (0.04)
rnd4	0.14240	0.05599	-0.22796	0.05 (0.05)	-0.06342	0.00 (0.06)
rnd2	0.08495	0.00448	-0.02930	0.00 (0.00)	-0.06015	0.00 (0.00)
rnd7	0.08686	0.00379	0.04059	0.00 (0.00)	-0.04624	0.00 (0.00)
rnd3	0.07357	0.00823	-0.08501	0.01 (0.01)	0.03166	0.00 (0.01)
rnd6	0.09628	0.00145	0.03627	0.00 (0.00)	-0.01181	0.00 (0.00)
rnd5	0.11144	0.00014	-0.00843	0.00 (0.00)	-0.00856	0.00 (0.00)
Var. Expl.	7.01372	5.95599	3.24993	46 % (46 %)	2.70606	39 % (85 %)

The variance of the scores in the "factor scores" coefficients table enables to check the reliability of the factors. As we mentioned above, it corresponds to the squared multiple correlation of the variables with the factors. Tanagra shows also the mean and the variance used for the standardization of the variables when we want to apply the coefficients for the calculation of the coordinates of new instances.

Factor Scores

Squared Multiple Corr. of the Variables with Each Fact			0.9735748	0.9823893
Attribute	Mean	Std-dev	Axis_1	Axis_2
cost	27.7777778	31.1903752	0.0771794	0.6474088
size	22.2222222	20.1537302	-0.2122558	0.1618406
alcohol	23.8888889	12.1969436	-0.3827776	0.0476624
reputat	55.5555556	25.7600514	0.0439872	-0.0877897
color	63.8888889	18.0705066	-0.1361719	-0.0540378
aroma	56.1111111	19.6889391	-0.1212157	0.0076416
taste	80.5555556	17.2311805	-0.6020989	-0.5275486
rnd1	42.7777778	28.7379507	0.0188726	0.0170036
rnd2	52.4242424	27.8012756	-0.0014051	0.0085949
rnd3	49.9494949	25.8833333	-0.0220836	-0.0083483
rnd4	46.5151515	27.6381246	-0.0200868	-0.0179266
rnd5	46.8181818	25.8243342	-0.0201605	-0.0053109
rnd6	47.0202020	29.7796554	0.0054159	0.0104204
rnd7	51.6161616	29.0404480	-0.0116456	-0.0067328

VARIMAX rotation. The VARIMAX rotation enables also to rotate the factors in principal factor analysis. We deactivate the sorting of the variables in order to compare the results of Tanagra with those of SAS[17].

[17]
```
proc factor data = mesdata.beer_rnd method=principal priors=smc nfactors=2
rotate=varimax;
run;
```

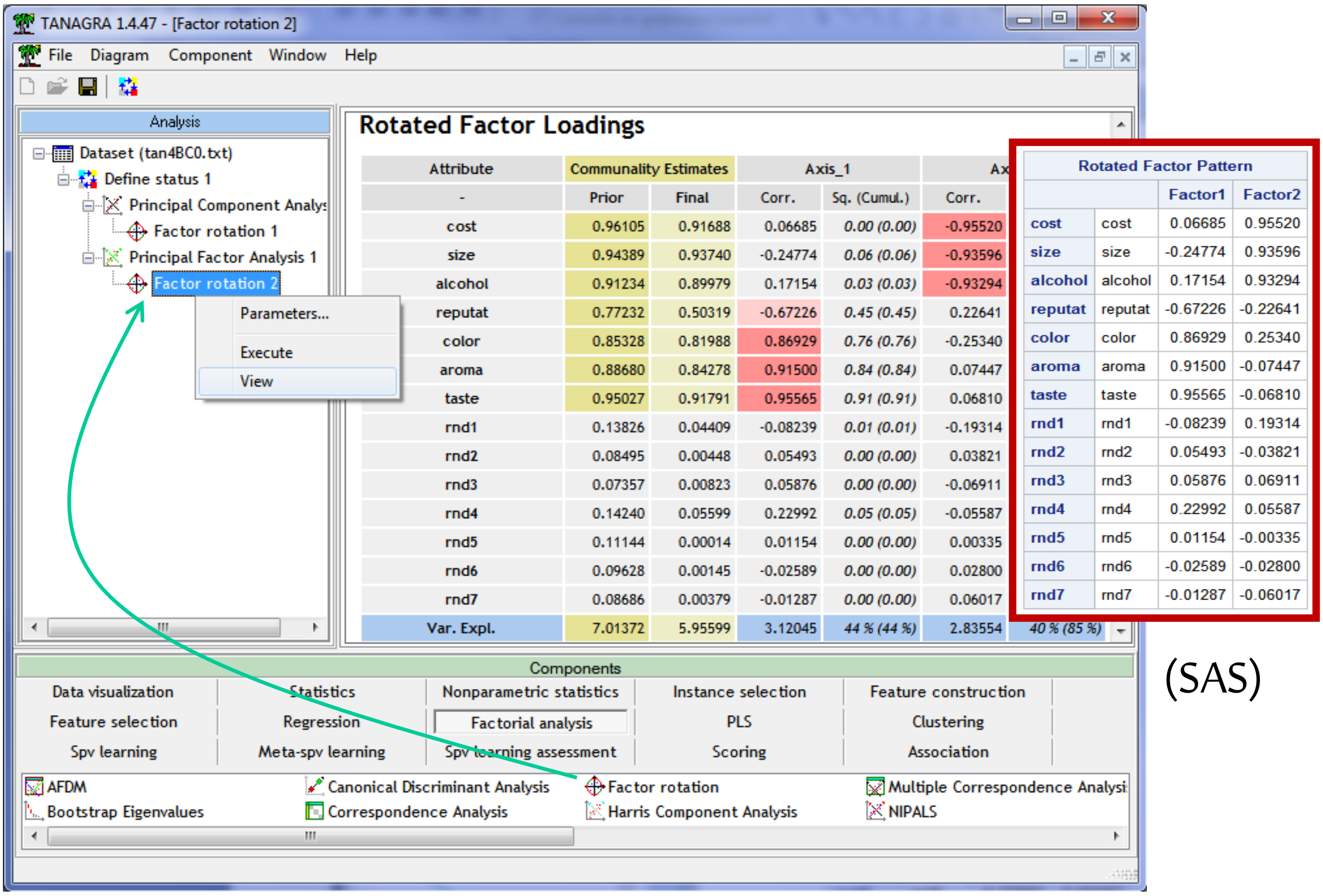
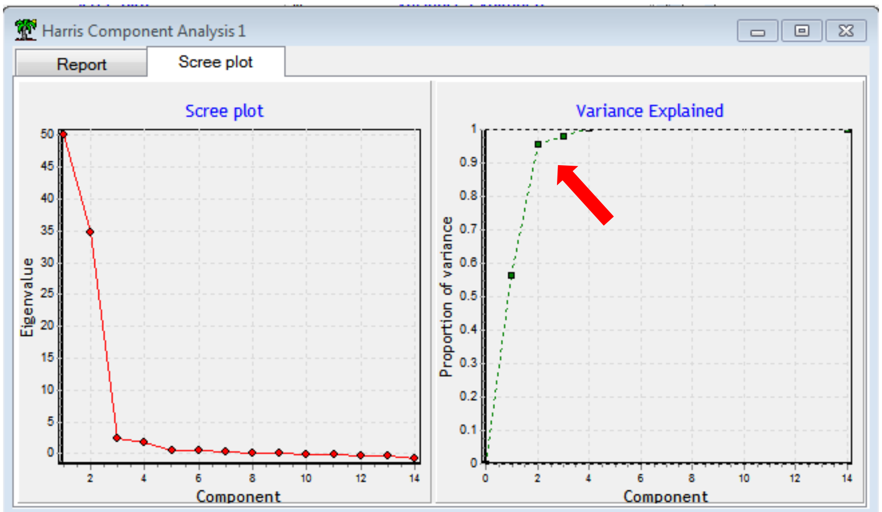


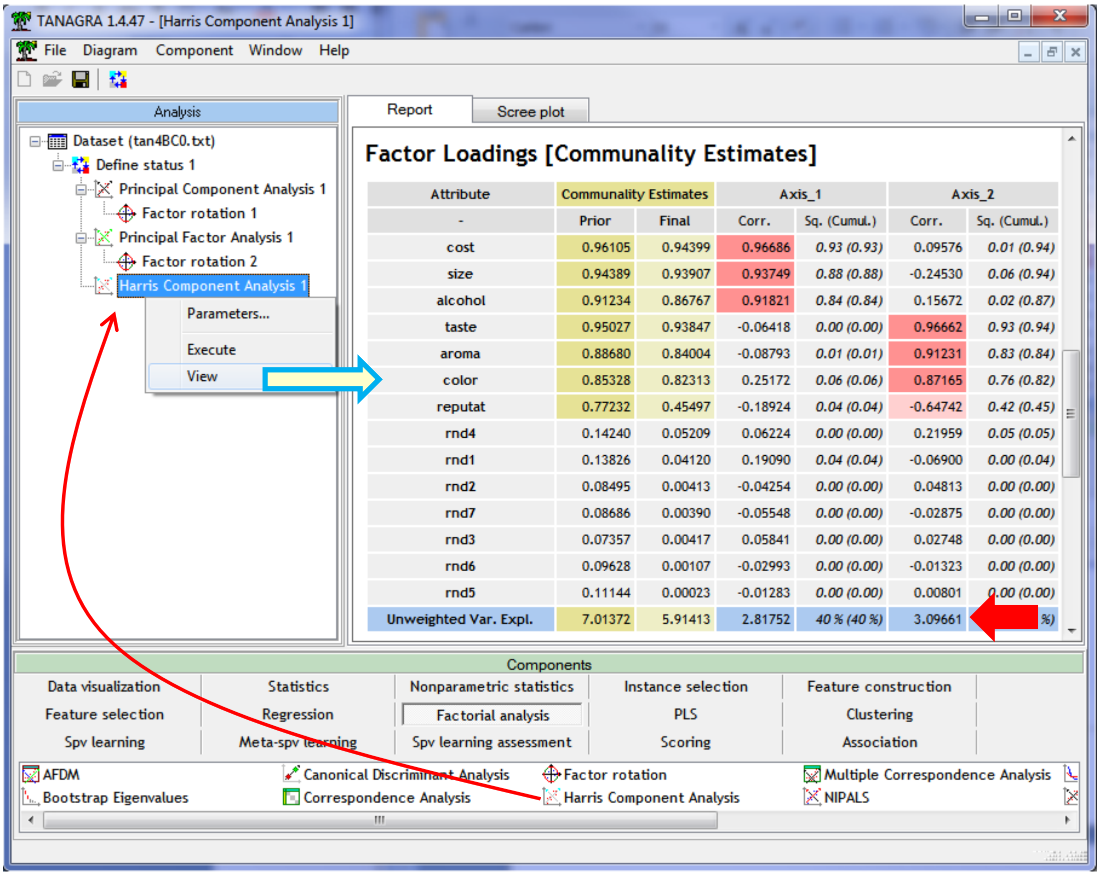
Figure 11 - "Loadings" after the varimax rotation – Principal Factor Analysis

4.4 Harris Component Analysis and varimax rotation

Harris approach. We add the HARRIS COMPONENT ANALYSIS tool (Factorial Analysis tab) into the diagram. We select 2 factors for the analysis. The scree plot and the plot of the cumulative variance show clearly that the selection of 2 factors is the right solution.



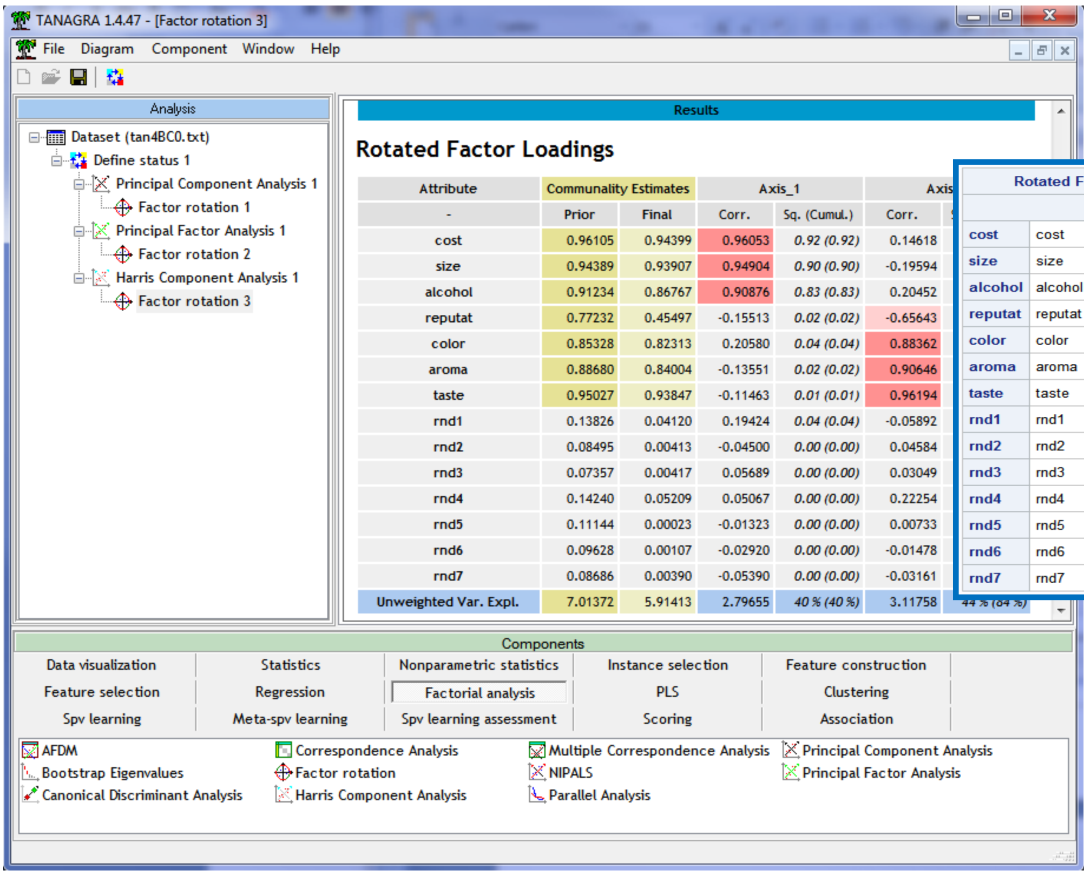
Into the loadings table, Tanagra displays the unweighted variance into the last row of the table for each factor.



Factor Loadings [Communality Estimates]

Attribute	Communality Estimates		Axis_1		Axis_2	
-	Prior	Final	Corr.	Sq. (Cumul.)	Corr.	Sq. (Cumul.)
cost	0.96105	0.94399	0.96686	0.93 (0.93)	0.09576	0.01 (0.94)
size	0.94389	0.93907	0.93749	0.88 (0.88)	-0.24530	0.06 (0.94)
alcohol	0.91234	0.86767	0.91821	0.84 (0.84)	0.15672	0.02 (0.87)
taste	0.95027	0.93847	-0.06418	0.00 (0.00)	0.96662	0.93 (0.94)
aroma	0.88680	0.84004	-0.08793	0.01 (0.01)	0.91231	0.83 (0.84)
color	0.85328	0.82313	0.25172	0.06 (0.06)	0.87165	0.76 (0.82)
reputat	0.77232	0.45497	-0.18924	0.04 (0.04)	-0.64742	0.42 (0.45)
rnd4	0.14240	0.05209	0.06224	0.00 (0.00)	0.21959	0.05 (0.05)
rnd1	0.13826	0.04120	0.19090	0.04 (0.04)	-0.06900	0.00 (0.04)
rnd2	0.08495	0.00413	-0.04254	0.00 (0.00)	0.04813	0.00 (0.00)
rnd7	0.08686	0.00390	-0.05548	0.00 (0.00)	-0.02875	0.00 (0.00)
rnd3	0.07357	0.00417	0.05841	0.00 (0.00)	0.02748	0.00 (0.00)
rnd6	0.09628	0.00107	-0.02993	0.00 (0.00)	-0.01323	0.00 (0.00)
rnd5	0.11144	0.00023	-0.01283	0.00 (0.00)	0.00801	0.00 (0.00)
Unweighted Var. Expl.	7.01372	5.91413	2.81752	40 % (40 %)	3.09661	%)

Varimax rotation. The association of the variables with one of the two factors is already strong for the Harris Analysis. Thus, the varimax rotation does not really modify the loadings.



Rotated Factor Loadings

Attribute	Communality Estimates		Axis_1		Axis_2
-	Prior	Final	Corr.	Sq. (Cumul.)	Corr.
cost	0.96105	0.94399	0.96053	0.92 (0.92)	0.14618
size	0.94389	0.93907	0.94904	0.90 (0.90)	-0.19594
alcohol	0.91234	0.86767	0.90876	0.83 (0.83)	0.20452
reputat	0.77232	0.45497	-0.15513	0.02 (0.02)	-0.65643
color	0.85328	0.82313	0.20580	0.04 (0.04)	0.88362
aroma	0.88680	0.84004	-0.13551	0.02 (0.02)	0.90646
taste	0.95027	0.93847	-0.11463	0.01 (0.01)	0.96194
rnd1	0.13826	0.04120	0.19424	0.04 (0.04)	-0.05892
rnd2	0.08495	0.00413	-0.04500	0.00 (0.00)	0.04584
rnd3	0.07357	0.00417	0.05689	0.00 (0.00)	0.03049
rnd4	0.14240	0.05209	0.05067	0.00 (0.00)	0.22254
rnd5	0.11144	0.00023	-0.01323	0.00 (0.00)	0.00733
rnd6	0.09628	0.00107	-0.02920	0.00 (0.00)	-0.01478
rnd7	0.08686	0.00390	-0.05390	0.00 (0.00)	-0.03161
Unweighted Var. Expl.	7.01372	5.91413	2.79655	40 % (40 %)	3.11758

Rotated Factor Pattern

		Factor1	Factor2
cost	cost	0.14618	0.96053
size	size	-0.19594	0.94904
alcohol	alcohol	0.20452	0.90876
reputat	reputat	-0.65643	-0.15513
color	color	0.88362	0.20580
aroma	aroma	0.90646	-0.13551
taste	taste	0.96194	-0.11463
rnd1	rnd1	-0.05892	0.19424
rnd2	rnd2	0.04584	-0.04500
rnd3	rnd3	0.03049	0.05689
rnd4	rnd4	0.22254	0.05067
rnd5	rnd5	0.00733	-0.01323
rnd6	rnd6	-0.01478	-0.02920
rnd7	rnd7	-0.03161	-0.05390

(SAS)

Both SAS and Tanagra provide the same results. But because SAS sorts the factors according to the unweighted variance, the first factor for SAS[18] corresponds to the 2nd of Tanagra and vice versa.

4.5 Comparison of the approaches after varimax rotation

All the methods provide very similar results after factor rotation (Figure 12).

Rotated Factor Loadings - PCA

Attribute	Axis_1	Axis_2
-	Corr.	Corr.
cost	0.15221	0.94145
size	-0.16016	0.95785
alcohol	0.25684	0.92749
reputat	-0.72537	-0.17266
color	0.90893	0.1748
aroma	0.91303	-0.16252
taste	0.93638	-0.1563
rnd1	-0.09747	0.28718
rnd2	0.0715	-0.06308
rnd3	0.09246	0.0874
rnd4	0.31501	0.0357
rnd5	0.01461	0.00021
rnd6	-0.03851	-0.03045
rnd7	-0.01872	-0.08764
Var. Expl.	3.30199	2.87923

Rotated Factor Loadings - PFA

Attribute	Axis_1	Axis_2
-	Corr.	Corr.
cost	0.06685	-0.9552
size	-0.24774	-0.93596
alcohol	0.17154	-0.93294
reputat	-0.67226	0.22641
color	0.86929	-0.2534
aroma	0.915	0.07447
taste	0.95565	0.0681
rnd1	-0.08239	-0.19314
rnd2	0.05493	0.03821
rnd3	0.05876	-0.06911
rnd4	0.22992	-0.05587
rnd5	0.01154	0.00335
rnd6	-0.02589	0.028
rnd7	-0.01287	0.06017
Var. Expl.	3.12045	2.83554

Rotated Factor Loadings - Harris

Attribute	Axis_1	Axis_2
-	Corr.	Corr.
cost	0.96053	0.14618
size	0.94904	-0.19594
alcohol	0.90876	0.20452
reputat	-0.15513	-0.65643
color	0.2058	0.88362
aroma	-0.13551	0.90646
taste	-0.11463	0.96194
rnd1	0.19424	-0.05892
rnd2	-0.045	0.04584
rnd3	0.05689	0.03049
rnd4	0.05067	0.22254
rnd5	-0.01323	0.00733
rnd6	-0.0292	-0.01478
rnd7	-0.0539	-0.03161
Unw.Var.Exp.	2.79655	3.11758

Figure 12 - "Loadings" of the approaches after VARIMAX rotation

This is probably the reason for which the principal component analysis (PCA) remains the most popular method in the case studies, even if it seem to suffer some theoretical restrictions for the analysis of the relations between the variables (it treat all the variance and not the shared variance).

But the main pitfall of PCA is the choice of the number of factors. We saw that this is not obvious when we have noisy variables in the dataset. If we select 3 factors in our study (this choice is possible if we consider the scree plot), the results provided by PCA become less readable.

5 Analysis under R with the PSYCH package

The principal component analysis is available in numerous packages for R. This is less true for the principal factor analysis and the Harris approach. But, as we seen above, we can program them if it is necessary. I have look around on the net. I found the PSYCH[19] package which can perform the principal factor analysis.

18
```
proc factor data = mesdata.beer_rnd
method=harris
msa
nfactors=2
score
rotate=varimax;
run;
```

19 http://cran.r-project.org/web/packages/psych/index.html

5.1 Principal component analysis

We can perform the principal component analysis with many tools under R (e.g. princomp or prcomp from the STAT package). Here, we use the **principal()** procedure from the PSYCH package.

```
#load the libraries
library(psych)
library(GPArotation)
#PCA
pca.unrotated <- principal(beer.data, nfactors=2, rotate="none")
print(pca.unrotated$loadings[,])
```

We obtain the same loadings as SAS or Tanagra:

```
> print(pca.rotated$loadings[,])
                PC1          PC2
cost     0.15221148  0.941453487
size    -0.16015848  0.957851137
alcohol  0.25684201  0.927488683
reputat -0.72537206 -0.172654920
color    0.90893314  0.174797696
aroma    0.91303327 -0.162516787
taste    0.93637979 -0.156300062
rnd1    -0.09747445  0.287184337
rnd2     0.07149913 -0.063084338
rnd3     0.09246219  0.087401320
rnd4     0.31501305  0.035700067
rnd5     0.01460936  0.000210707
rnd6    -0.03850968 -0.030452578
rnd7    -0.01872358 -0.087644503
```

When we perform the VARIMAX rotation

```
#PCA + varimax
pca.rotated <- principal(beer.data, nfactors=2, rotate="varimax")
print(pca.rotated$loadings[,])
```

The results are also consistent (**Erreur ! Source du renvoi introuvable.**):

```
> print(pca.rotated$loadings[,])
                PC1          PC2
cost     0.15221148  0.941453487
size    -0.16015848  0.957851137
alcohol  0.25684201  0.927488683
reputat -0.72537206 -0.172654920
color    0.90893314  0.174797696
aroma    0.91303327 -0.162516787
taste    0.93637979 -0.156300062
rnd1    -0.09747445  0.287184337
rnd2     0.07149913 -0.063084338
rnd3     0.09246219  0.087401320
rnd4     0.31501305  0.035700067
rnd5     0.01460936  0.000210707
rnd6    -0.03850968 -0.030452578
rnd7    -0.01872358 -0.087644503
```

5.2 Principal factor analysis

The **fa()** procedure enables to launch the principal factor analysis. We must set the option "**max.iter=1**" to perform the non iterative approach.

```
#Non-iterative PFA (principal factor analysis)
pfa.unrotated <- fa(beer.data,nfactors=2,rotate="none",SMC=T,fm="pa",max.iter=1)
print(pfa.unrotated$loadings[,])
```

We obtain the following loadings, consistent with those of SAS and Tanagra:

```
> print(pfa.unrotated$loadings[,])
                 PA1         PA2
cost      0.524418623  0.80116519
size      0.240428244  0.93786565
alcohol   0.604929383  0.73065402
reputat  -0.697277292  0.13037872
color     0.882431136 -0.20295513
aroma     0.762359004 -0.51145452
taste     0.800948130 -0.52572609
rnd1      0.022321664  0.20878400
rnd2      0.029304179 -0.06015002
rnd3      0.085007045  0.03165822
rnd4      0.227959211 -0.06341560
rnd5      0.008434735 -0.00855580
rnd6     -0.036266480 -0.01180995
rnd7     -0.040589907 -0.04624353
```

We modify the option "rotate" in order to perform the VARIMAX rotation.

```
#PFA + varimax
pfa.varimax <- fa(beer.data,nfactors=2,rotate="varimax",SMC=T,fm="pa",max.iter=1)
print(pfa.varimax$loadings[,])
```

Here also, the results are consistent (**Erreur ! Source du renvoi introuvable.**) :

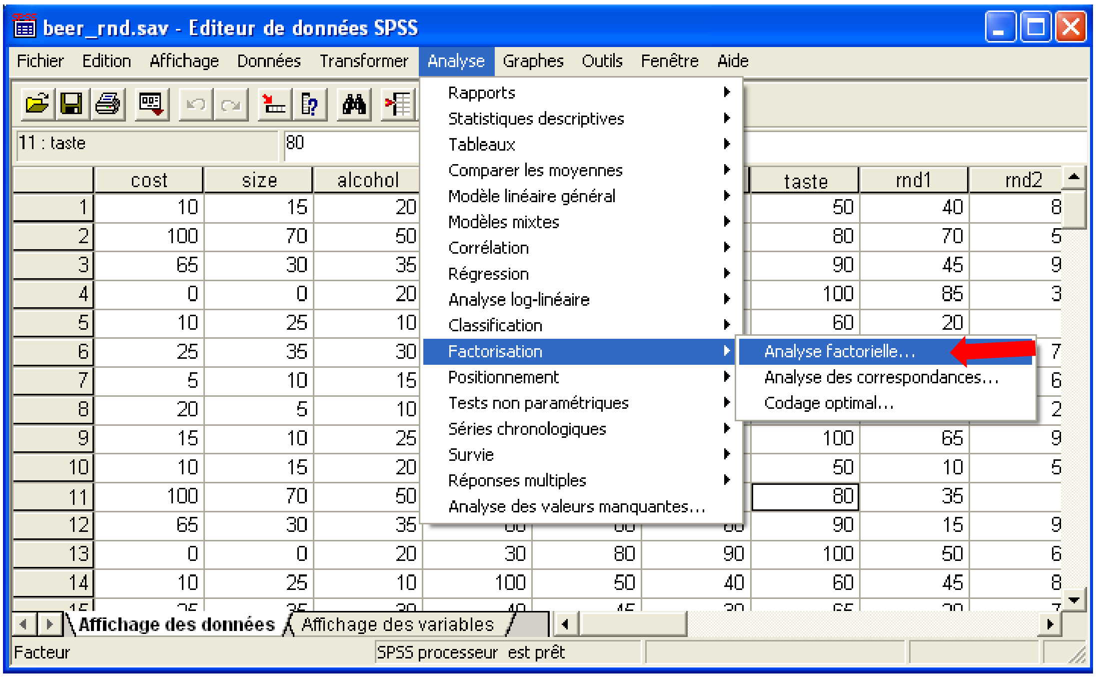
```
> print(pfa.varimax$loadings[,])
                PA1         PA2
cost      0.06686663  0.95520124
size     -0.24772440  0.93596492
alcohol   0.17154705  0.93293433
reputat  -0.67226251 -0.22640086
color     0.86929298  0.25338749
aroma     0.91500223 -0.07448421
taste     0.95564970 -0.06811382
rnd1     -0.08238309  0.19313736
rnd2      0.05492718 -0.03820686
rnd3      0.05875586  0.06910998
rnd4      0.22992540  0.05586817
rnd5      0.01153709 -0.00335293
rnd6     -0.02589463 -0.02800358
rnd7     -0.01286809 -0.06016990
```

5.3 Harris approach

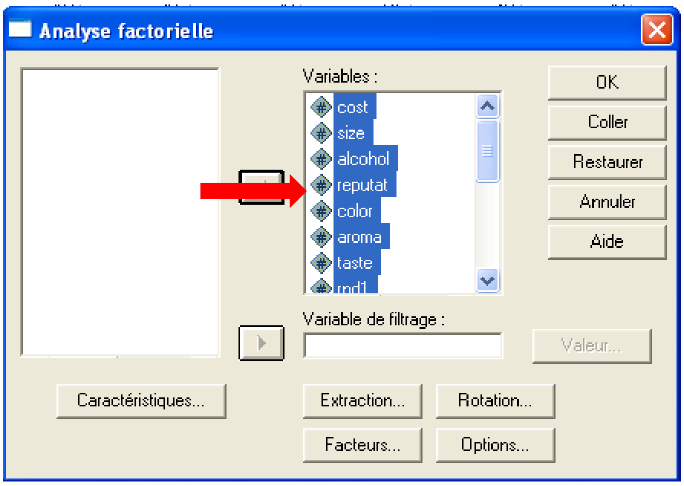
I have not found a package which implements the Harris approach. It does not matter. We saw above (section 3.4) that we can write a program for R which enables to perform the approach on a dataset. This is one of the main attractive features of R.

6 Principal factor analysis with SPSS

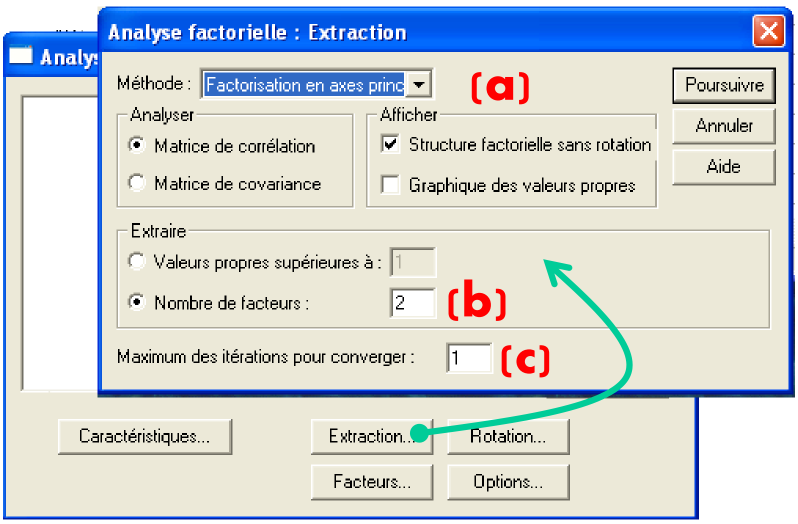
We use the French version of SPSS (12.0.1) in this section. After we import the dataset, we activate the menu ANALYSE / FACTORISATION / ANALYSE FACTORIELLE. A dialog box enables to set the parameters of the study.



We specify the variables for the analysis.

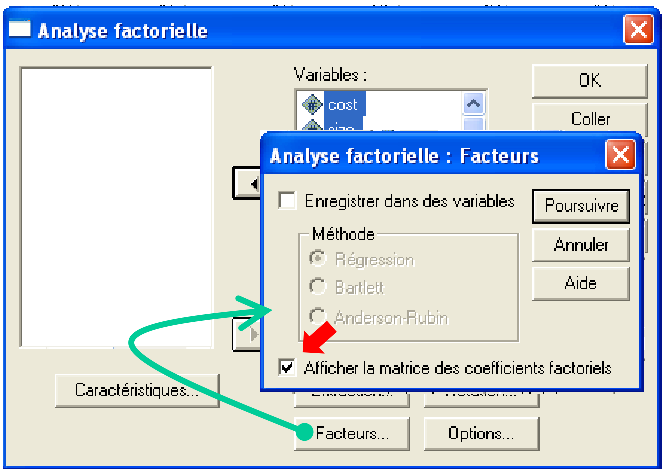


We choose the factorial method by clicking on the "Extraction" button.

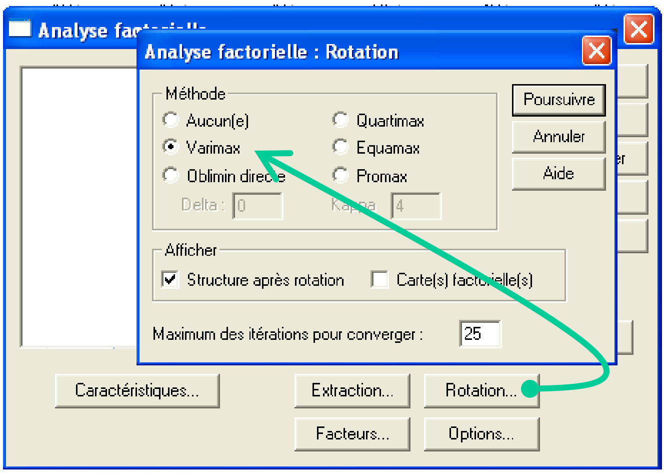


We select the principal factor analysis (a) with 2 factors (b), by limiting the number of iterations to 1 (c). This last option is important. By default, such as the **fa()** procedure of the PSYCH package, SPSS performs the iterative approach. When we set 'iterations = 1', we obtain the same results as Tanagra and SAS.

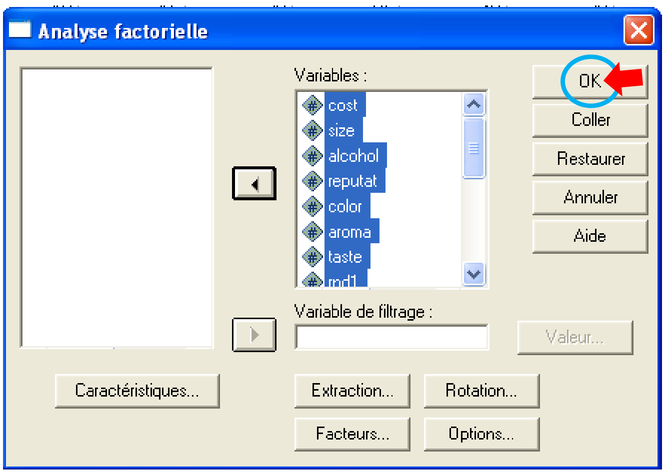
Then, we select the "Facteurs" button. We ask the displaying of the factor score coefficients.



Last, we ask the varimax factor rotation.



We confirm these options. We click on the OK button to launch the analysis.



SPSS generates a report which describes the results of the analysis.

Initial and estimated communalities. The quality of the representation is obtained by comparing the initial and the estimated communalities (in comparison, see Figure 7).

Qualité de représentation

	Initial	Extraction
cost	.96105	.91688
size	.94389	.93740
alcohol	.91234	.89979
reputat	.77232	.50319
color	.85328	.81988
aroma	.88680	.84278
taste	.95027	.91791
rnd1	.13826	.04409
rnd2	.08495	.00448
rnd3	.07357	.00823
rnd4	.14240	.05599
rnd5	.11144	.00014
rnd6	.09628	.00145
rnd7	.08686	.00379

Loadings (Factor Pattern) before and after rotation. Then, we have the loadings, before [**a**] (see Figure 6) and after [**b**] (see Figure 11) the varimax factor rotation.

Matrice factorielle[a]

(a)	Facteur	
	1	2
cost	.52442	.80117
size	.24043	.93787
alcohol	.60493	.73065
reputat	-.69728	.13038
color	.88243	-.20296
aroma	.76236	-.51145
taste	.80095	-.52573
rnd1	.02232	.20878
rnd2	.02930	-.06015
rnd3	.08501	.03166
rnd4	.22796	-.06342
rnd5	.00843	-.00856
rnd6	-.03627	-.01181
rnd7	-.04059	-.04624

Matrice factorielle après rotation[a]

(b)	Facteur	
	1	2
cost	.06685	.95520
size	-.24774	.93596
alcohol	.17154	.93294
reputat	-.67226	-.22641
color	.86929	.25340
aroma	.91500	-.07447
taste	.95565	-.06810
rnd1	-.08239	.19314
rnd2	.05493	-.03821
rnd3	.05876	.06911
rnd4	.22992	.05587
rnd5	.01154	-.00335
rnd6	-.02589	-.02800
rnd7	-.01287	-.06017

Factor Scores. SPSS provides the factor scores coefficients after the factor rotation. We compare here the results of SPSS with those of Tanagra.

Not surprisingly, we have exactly the same values. We have also the same results with SAS.

Factor Scores

Squared Multiple Corr. of the Variables with Each Factor			0.9758792	0.9800848
Attribute	Mean	Std-dev	Axis_1	Axis_2
cost	27.7777778	31.1903752	-0.3832525	-0.5274584
size	22.2222222	20.1537302	0.1063105	-0.2448325
alcohol	23.8888889	12.1969436	0.3108668	-0.2283686
reputat	55.5555556	25.7600514	0.0044384	0.0980929
color	63.8888889	18.0705066	0.1452290	-0.0192720
aroma	56.1111111	19.6889391	0.1020793	-0.0658137
taste	80.5555556	17.2311805	0.7829666	0.1667154
rnd1	42.7777778	28.7379507	-0.0247701	-0.0056339
rnd2	52.4242424	27.8012756	-0.0029671	-0.0081879
rnd3	49.9494949	25.8833333	0.0233498	-0.0034879
rnd4	46.5151515	27.6381246	0.0262803	0.0058471
rnd5	46.8181818	25.8243342	0.0201892	-0.0052009
rnd6	47.0202020	29.7796554	-0.0098118	-0.0064534
rnd7	51.6161616	29.0404480	0.0134504	0.0001949

(Tanagra)

Matrice des coordonnées factorielles

	Facteur	
	1	2
cost	-.38325	.52746
size	.10631	.24483
alcohol	.31087	.22837
reputat	.00444	-.09809
color	.14523	.01927
aroma	.10208	.06581
taste	.78297	-.16672
rnd1	-.02477	.00563
rnd2	-.00297	.00819
rnd3	.02335	.00349
rnd4	.02628	-.00585
rnd5	.02019	.00520
rnd6	-.00981	.00645
rnd7	.01345	-.00019

(SPSS)

Variance and covariance of the factors. As we say previously, the factors have theoretically a unit variance. But, the observed variance is not equal to 1. The discrepancy between the observed variance and the theoretical variance is an indication about the reliability of the factor. In a similar process, the factors have theoretically a null covariance. But the covariance measured on the sample can be slightly different to zero. SPSS provides the observed covariance matrix of the factors.

Matrice de covariance factorielle

Facteur	1	2
1	.97588	-.00388
2	-.00388	.98008

Here, the variances of the selected factors are near to 1. In addition, their covariance is near to 0. These factors are relevant.

Variance and covariance when selecting 5 factors. When we perform the same analysis by setting 5 factors (see SAS, Figure 8), we note that starting from the third factor: the variance becomes largely different than 1; the covariance with the other factors becomes largely different than 0.

Matrice de covariance factorielle

Facteur	1	2	3	4	5
1	.97357	-.00024	-.00453	-.03402	-.02313
2	-.00024	.98239	.01396	.03588	-.01047
3	-.00453	.01396	.65231	.12348	-.10014
4	-.03402	.03588	.12348	.41288	-.01229
5	-.02313	-.01047	-.10014	-.01229	.32997

Méthode d'extraction : Factorisation en axes principaux.

The last 3 factors are clearly unstable. They do not correspond to relevant information from the data.

7 Conclusion

In this tutorial, we present various factor analysis approaches. They differ in the matrix used for the diagonalization process. The principal component analysis uses the standard correlation matrix; the principal factor analysis replaces the main diagonal of the correlation matrix with the proportion of the variance explained by the others for each variables; the Harris component analysis intensifies the correlation with the uniqueness of the variables.

Despite these differences, we note that they provide similar results on our dataset. The PCA in particular is enough for performing the analysis the relations of the variables, even if there are many noisy variables (a half of the variables in our dataset). In this context, the main challenge is to determine the adequate number of factors to retain in the analysis.

These methods fall within the same framework into Tanagra. Thus, we can apply the factor rotation tool (FACTOR ROTATION) to any approaches. We can also apply the tools based on a resampling scheme for the detection of the right number of factors (BOOTSTRAP EIGENVALUES, PARALLEL ANALYSIS[20]).

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