

1 Objectif

Nouveaux outils pour l'ACP (Analyse en Composantes Principales) dans Tanagra : tests et composants pour la détection du nombre d'axes.

L'analyse en composantes principales (ACP) est une technique exploratoire très populaire. Selon les points de vue, on peut la considérer : comme une technique descriptive où l'on essaie de résumer les données dans ses dimensions les plus importantes ; comme une technique de visualisation où l'on essaie de préserver les proximités entre les individus dans un espace de représentation réduit ; comme une technique de compression de l'information ; etc.

Outre les excellents ouvrages en langue française qui les décrivent, les références sont suffisamment abondantes sur le web pour que chacun se fasse son idée. J'en ai moi-même beaucoup parlé dans plusieurs didacticiels¹ et, récemment², j'ai décrit la programmation sous R du test de Bartlett, de l'indice KMO (MSA – Measure of Sampling Adequacy), et des indicateurs pour la détermination du nombre de facteurs en ACP. On les trouve rarement sous une forme native dans les logiciels libres, je me suis dit qu'il était opportun de les intégrer dans **Tanagra 1.4.45**.

Dans ce tutoriel, nous décrivons la mise en œuvre de ces nouveaux outils. Nous mettrons en parallèle, quand cela est possible, les résultats de la PROC FACTOR de SAS. Nous avons choisi cette dernière plutôt que PRINCOMP parce que ses sorties sont plus complètes³.

2 Données

Le fichier « [beer_pca.xls](#) » décrit l'importance (une note allant de 0 à 100) accordée à différents critères par $n = 99$ consommateurs lors du choix de packs de bière. Les $p = 7$ variables (critères) sont : le coût (COST), la taille (SIZE), le pourcentage d'alcool (ALCOHOL), la réputation (REPUTAT), le couleur (COLOR), l'arôme (AROMA), le goût (TASTE).

Les données proviennent du site du Dr Wuensch⁴, elles sont traitées entre autres dans le tutoriel de Jacques Baillargeon dédié à l'ACP⁵. Nous n'avons pas exactement les mêmes résultats cependant parce que le nettoyage du fichier n'est pas le même. C'est la raison pour laquelle, même sur des bases ultra-connues et référencées comme celles de l'UCI, je m'applique toujours à fournir le jeu de données effectivement utilisé.

3 Traitement avec la PROC FACTOR de SAS 9.3

Deux programmes ont été rédigés pour l'analyse en composantes principales avec SAS. Nous avons tout d'abord lancé une analyse classique, en demandant l'affichage de plusieurs indicateurs statistiques que nous décrirons plus bas.

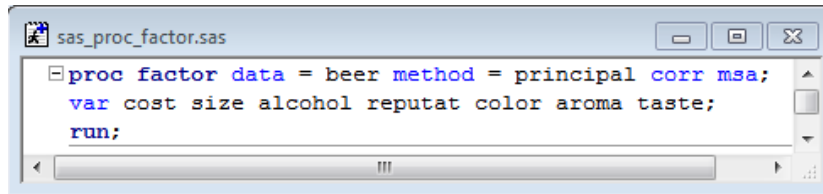
¹ [Analyse en composantes principales avec R](#) ; [ACP – Description de véhicules](#).

² [ACP sous R – Indice KMO et test de Bartlett](#) ; [ACP avec R – Détection du nombre d'axes](#).

³ SAS User's Guide, « [Comparison of the PRINCOMP and FACTOR Procedures](#) ».

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

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

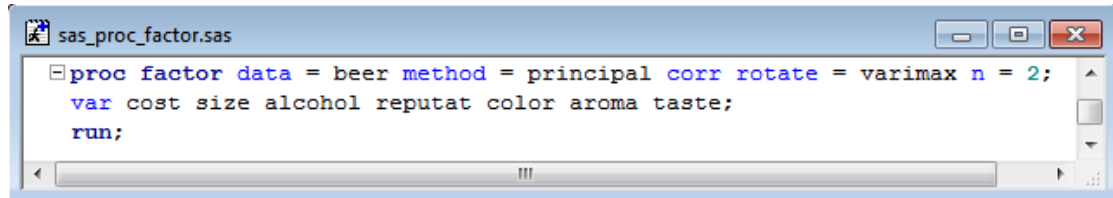


```

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

```

Nous avons réitéré l'analyse en demandant une rotation VARIMAX pour rendre plus tranchées l'association des variables aux deux axes que nous avons choisis de conserver.



```

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

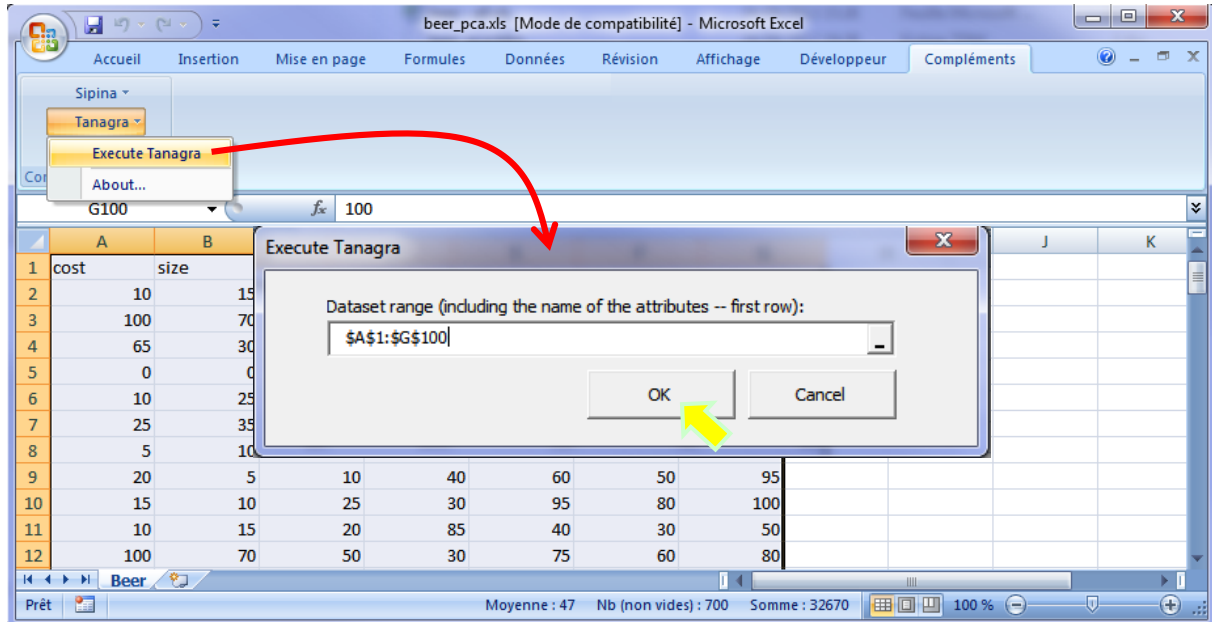
```

Les résultats seront repris dans la section suivante, nous les mettrons en miroir avec ceux de Tanagra. Je note avec un certain amusement que [les sorties de SAS sont par défaut au format HTML](#) dans la version 9.3. Solution que nous avons adoptée dès la première version de Tanagra en 2003.

4 Analyse en composantes principales avec Tanagra

4.1 Importation des données

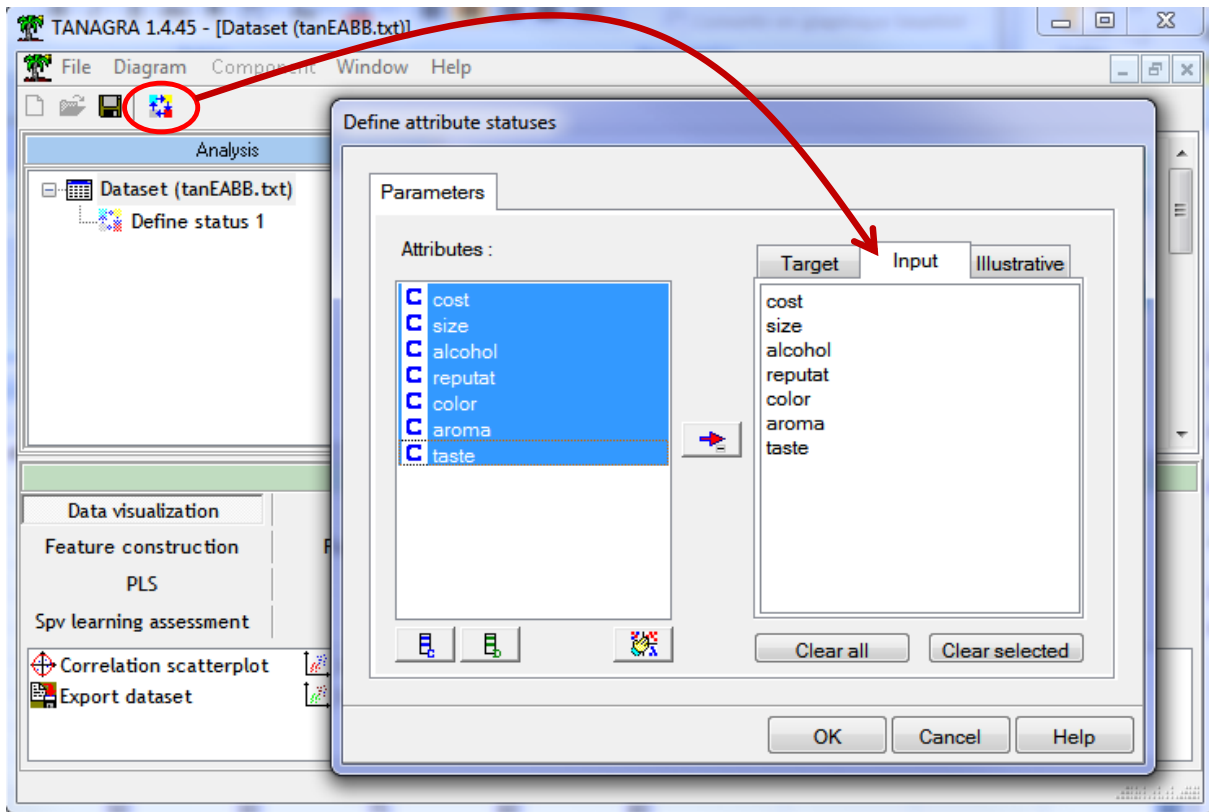
Le plus simple est de charger les données dans le tableur Excel⁶ (ou Libre/Open Office Calc⁷) et d'utiliser la macro-complémentaire (add-in) Tanagra pour lancer le logiciel et importer les données.



Tanagra est automatiquement démarré, nous pouvons initier les traitements en choisissant les variables actives de l'analyse à l'aide du composant DEFINE STATUS.

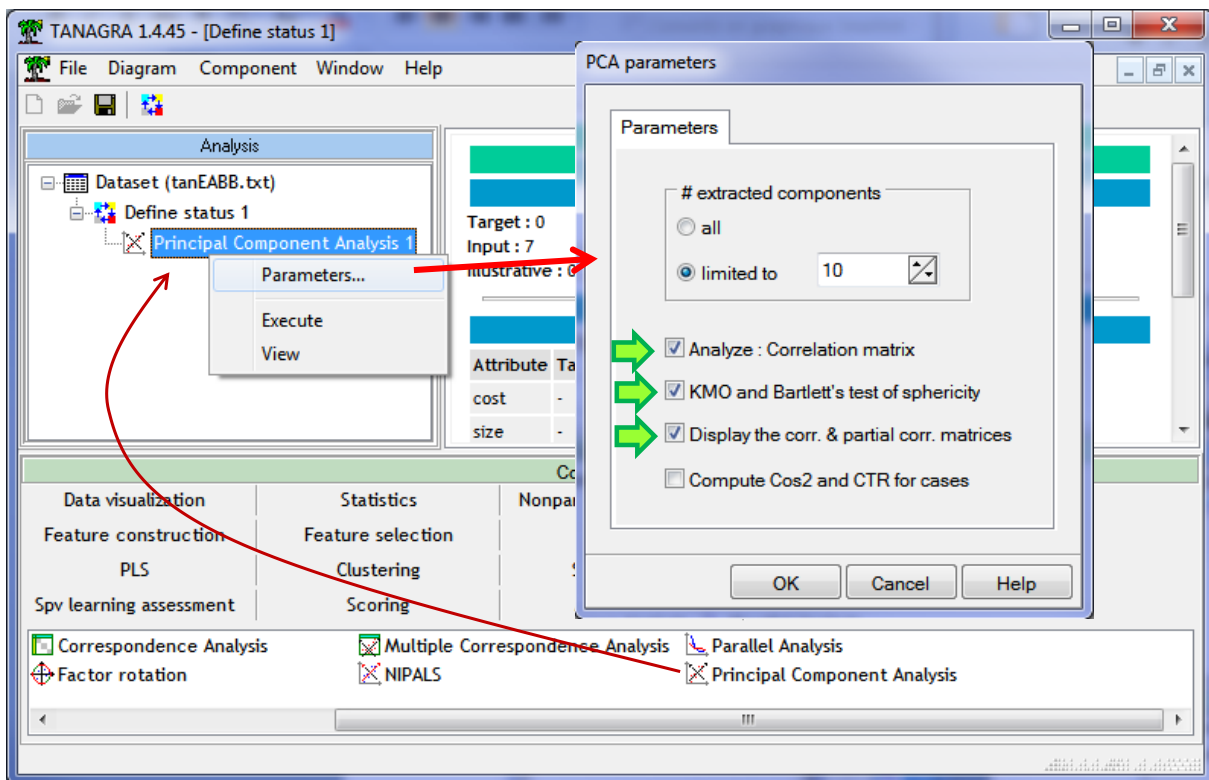
⁶ <http://tutoriels-data-mining.blogspot.fr/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html>

⁷ <http://tutoriels-data-mining.blogspot.fr/2011/07/tanagra-addon-pour-openoffice-33.html>



4.2 ACP avec Tanagra

Nous insérons le composant PRINCIPAL COMPONENT ANALYSIS (onglet FACTORIAL ANALYSIS) pour réaliser l'ACP. Nous le paramétrons comme suit :



Nous effectuons une ACP normée (ANALYZE : CORRELATION MATRIX). Nous demandons l’affichage : du test de sphéricité de Bartlett et de l’indice MSA de Kaiser-Mayer-Olkin (KMO) ; des matrices de corrélations brutes et partielles.

Nous validons ce paramétrage. Nous cliquons sur VIEW pour lancer les calculs. Les résultats sont répartis sur plusieurs sections.

4.2.1 Tableau des valeurs propres

Tanagra fournit le tableau des valeurs propres « Eigenvalues », avec le pourcentage d’inertie expliquée (individuelle et cumulée) par les axes. Un histogramme permet de situer leur décroissance.

Eigen values

Matrix trace = 7.00

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

Prior Communality Estimates: ONE

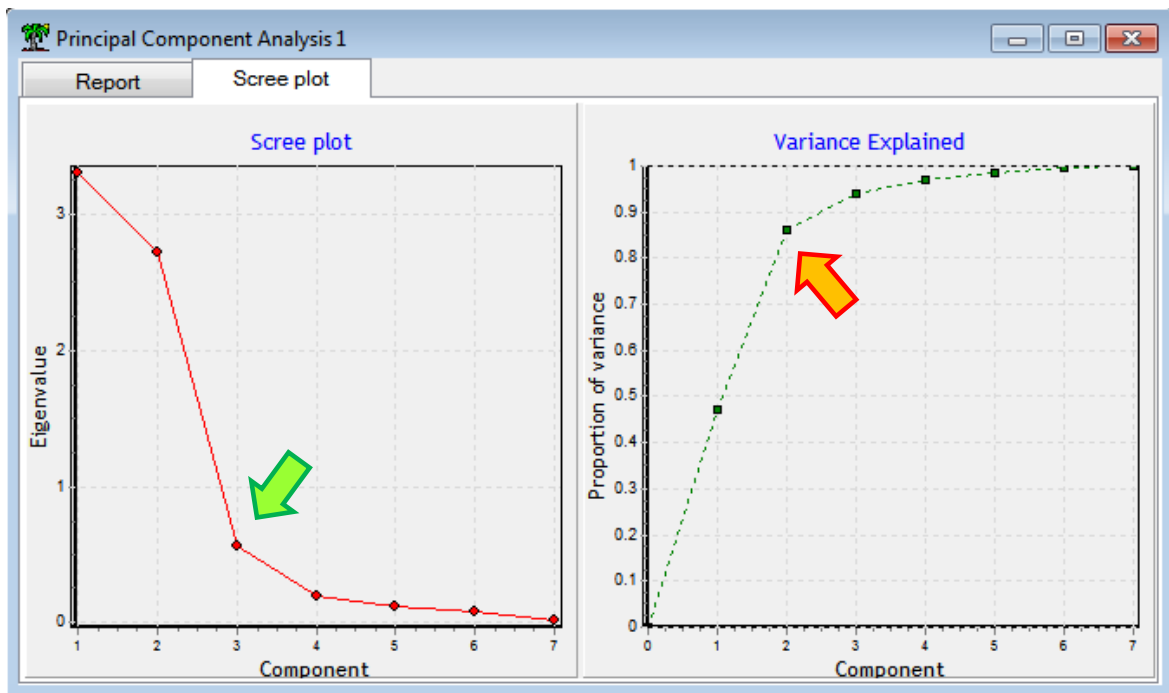
Eigenvalues of the Correlation Matrix: Total = 7 Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.30608184	0.58687618	0.4723	0.4723
2	2.71920565	2.15183494	0.3885	0.8608
3	0.56737072	0.37393939	0.0811	0.9418
4	0.19343133	0.07347757	0.0276	0.9694
5	0.11995376	0.04321918	0.0171	0.9866
6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

TANAGRA

SAS - PROC FACTOR

4.2.2 Eboulis des valeurs propres (Scree plot) et inertie expliquée

Pour compléter ce résultat, un graphique Scree plot est proposé dans le second onglet de la fenêtre de visualisation.



Il y a manifestement un « coude » au niveau de la 3^{ème} composante. Nous n'en sélectionnerons que deux néanmoins car la valeur propre associée est relativement faible. Le graphique des inerties expliquées cumulées « Variance explained » confirme cette idée. Le gain d'inertie en passant du 2nd facteur au 3^{ème} ne semble pas décisif, contrairement aux passages de 0 à 1, puis de 1 à 2.

4.2.3 Outils pour la détection du nombre d'axes

Pour les matheux qui ont du mal à se contenter d'un raisonnement basé sur des « impressions », d'autres outils⁸ sont disponibles dans la section « Significance of Principal Components » du rapport. Attention, **ils ne sont accessibles que pour l'ACP normée**.

3 règles de détection sont proposées :

Significance of Principal Components		
Global critical values		
Kaiser-Guttman	1	
Kartis-Saporta-Spinaki	1.49487	
Eigenvalue table - Test for significance		
Eigenvalues - Significance		
Axis	Eigenvalue	Broken-stick critical values
1	3.306082	2.592857
2	2.719206	1.592857
3	0.567371	1.092857
4	0.193431	0.759524
5	0.119954	0.509524
6	0.076735	0.309524
7	0.017222	0.142857

1. Kaiser-Guttman, un axe est pertinent si la valeur propre est supérieure à 1.
2. Karlis-Saporta-Spinaki, il corrige la règle précédente en introduisant une contrainte supplémentaire tenant compte du ratio « nombre d'observations / nombre de variables », très important pour la fiabilité des résultats de l'ACP.
3. Legendre-Legendre, le test des bâtons brisés (broken stick) qui, à la différence des précédents, définit une valeur seuil spécifique pour chaque axe.

L'intensité de la couleur de fond rouge des cellules du tableau dépend du nombre de règles

de détection déclenchée. S'il n'y en a aucune, elle est simplement grisée. Ici, nous constatons qu'une ACP en 2 axes semble être la bonne solution. A partir du 3^{ème}, aucune règle n'est activée.

4.2.4 Test de sphéricité de Bartlett

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

Le test de Bartlett indique si un axe au moins est significatif. L'hypothèse nulle correspond à : toutes les valeurs propres sont identiques, elles valent 1. Il est peu connu dans le monde francophone, car il ne paraît pas très décisif à bien y regarder. En effet, il a tendance à être toujours significatif dès que l'effectif « n » augmente. Je l'ai intégré dans Tanagra néanmoins car les auteurs en langue anglaise y font souvent référence dans leurs tutoriels.

Dans notre cas, les données contredisent l'hypothèse nulle. On sait qu'un axe au moins est pertinent, mais nous ne savons pas quel est leur nombre avec cet outil.

⁸ Voir <http://tutoriels-data-mining.blogspot.fr/2012/06/acp-avec-r-detection-du-nombre-daxes.html> pour la description des techniques et des formules associées.

4.2.5 Indice MSA de Kaiser-Mayer-Olkin

L'indice MSA (Measure of Sampling Adequacy, ou indice KMO si l'on se réfère aux initiales des auteurs) indique la capacité de l'ACP à résumer l'information en un nombre réduit de facteurs. Il oppose les corrélations brutes et partielles que nous décrirons plus loin.

Kaiser's Measure of Sampling Adequacy (MSA) TANAGRA									
Overall MSA = 0.503618									
cost	0.4035957	size	0.5245393	alcohol	0.5511566	reputat	0.3689421	color	0.847379
aroma	0.5606907	taste	0.427865						

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.50361804						
cost	size	alcohol	reputat	color	aroma	taste
0.40359574	0.52453925	0.55115660	0.36894206	0.84737899	0.56069066	0.42786498

SAS - PROC FACTOR

Nous avons pour notre part le MSA global = 0.503618. Ce qui semble décevant si l'on se réfère aux grilles de lecture que l'on veut bien nous donner ici ou là⁹. En réduisant la description de 7 variables en 2 facteurs, 14% de l'inertie reste non expliquée. Mais cela ne veut pas dire pour autant que nous n'avons pas une meilleure compréhension des relations qui sous-tendent les données. La nuance est énorme. Il ne faut pas que l'indice nous induise en erreur.

Nous disposons par la suite de l'indice MSA par variable. L'idée est de détecter les variables mal modélisées dans le processus de compression de l'information. Ici, il semble que la variable REPUTAT (réputation) ne soit pas assimilable aux autres puisqu'elle présente le plus faible MSA (0.3689).

4.2.6 Corrélation des variables avec les facteurs et cos²

Le tableau « Factor Loadings [Communalities] » indique les corrélations et les cos² des variables avec les facteurs. Dans la dernière ligne, nous observons les inerties exprimées. Les deux facteurs cumulés restituent 86% de l'information disponible.

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

TANAGRA

2 factors will be retained by the MINEIGEN criterion.			
Factor Pattern			
		Factor1	Factor2
cost	cost	0.50507	0.81643
size	size	0.22575	0.95075
alcohol	alcohol	0.59771	0.75683
reputat	reputat	-0.74110	0.12380
color	color	0.90611	-0.18768
aroma	aroma	0.78614	-0.51354
taste	taste	0.80901	-0.51154

Variance Explained by Each Factor		
	Factor1	Factor2
	3.3060818	2.7192057

Final Communality Estimates: Total = 6.025287							
	cost	size	alcohol	reputat	color	aroma	taste
	0.92165047	0.95488671	0.93004894	0.56454880	0.85624989	0.88173669	0.91616599

SAS - PROC FACTOR

⁹ <http://peoplelearn.homestead.com/Topic20-FACTORanalysis3a.html>

Une interprétation rapide permettrait de dire que le 1^{er} axe est essentiellement défini par la conjonction de couleur, arôme et goût. Il caractérise « l'esthétisme » des consommateurs. Le 2nd correspond plutôt à leur attrait pour le coût (*payer le moins cher possible j'imagine*), la taille (*plus il y en a, mieux c'est*) et l'alcool (*boivons, boivons, y a que ça de vrai*). Bref, il exprime l'inclination de l'homme à la soulographie (dixit Jacques Brel, « [La chanson de Jacky](#) »).

On note que la réputation (REPUTAT) est mal représentée sur les 2 premiers axes avec un COS2 = 56%. En scrutant le tableau (non visible dans la copie d'écran), nous constatons qu'elle est associée principalement au 3^{ème} axe... que nous avons choisi de ne pas sélectionner puisqu'il ne semble pas significatif.

4.2.7 Matrice des corrélations brutes et partielles

La matrice des corrélations brutes montre la relation entre les variables.

Assez difficile à lire lorsque nous traitons un grand nombre de variables, il est plutôt instructif en ce qui nous concerne ($p = 7$). Nous distinguons 2 blocs de respectivement trois variables (COST, SIZE, ALCOHOL) vs. (COLOR, AROMA, TASTE). REPUTAT semble occuper une position à part, la variable qui lui est la plus liée, négativement, est le goût (TASTE). Les puristes sensibles au goût, le sont moins par rapport à la réputation ($r = -0.62650$), *parce qu'ils font confiance à leur propre jugement j'imagine*.

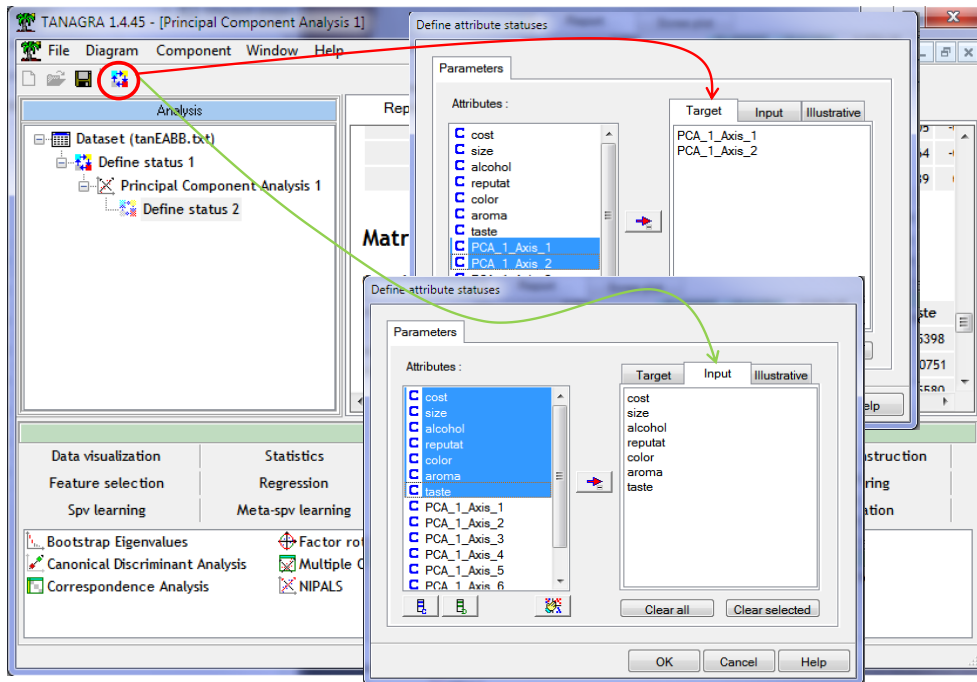
Matrices							
Correlations							
	cost	size	alcohol	reputat	color	aroma	taste
cost	1.00000	0.87839	0.87702	-0.17478	0.32089	-0.02764	0.05398
size	0.87839	1.00000	0.82367	-0.06123	0.01441	-0.28624	-0.30751
alcohol	0.87702	0.82367	1.00000	-0.36051	0.39770	0.09768	0.05580
reputat	-0.17478	-0.06123	-0.36051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

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

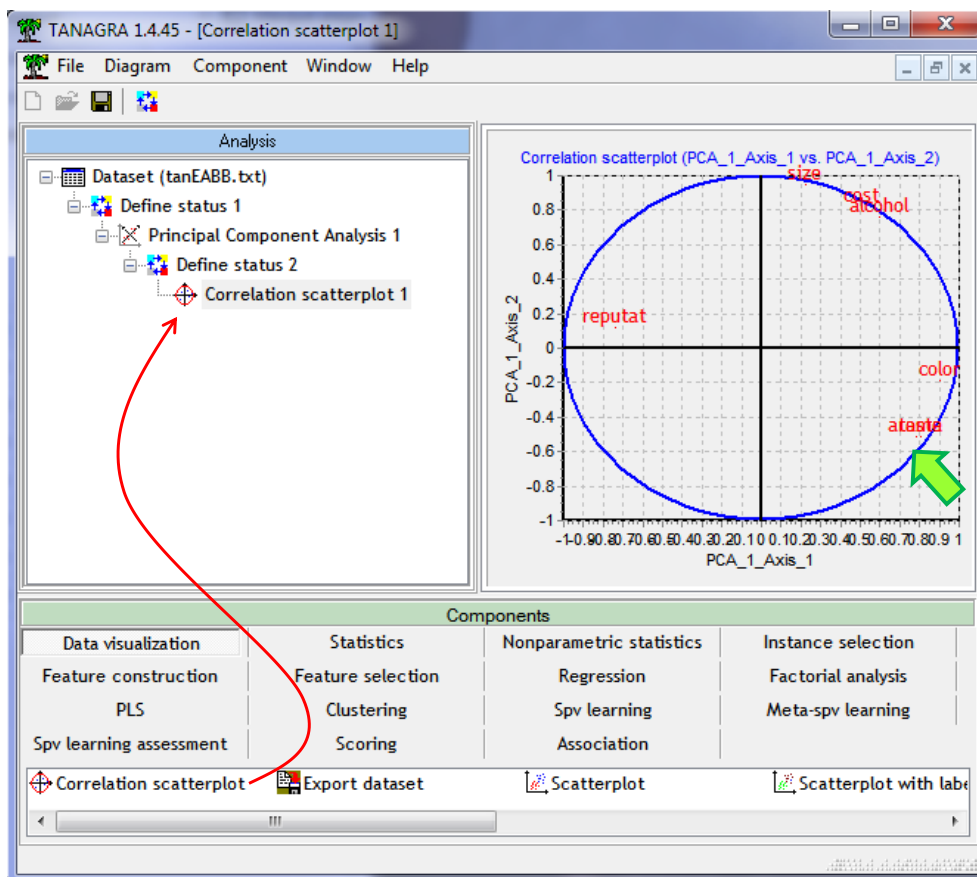
La matrice des corrélations partielles indique les relations entre les variables en contrôlant l'impact de toutes les autres. Nous pouvons ainsi déceler la nature factice de certaines corrélations. Par exemple, $r[\text{SIZE}, \text{ALCOHOL}] = 0.82367$ laisse à penser que l'intempérance est un trait de caractère fort chez les consommateurs. Mais, en calculant $r[\text{SIZE}, \text{ALCOHOL} / (p-2) \text{ autres variables}] = -0.10712$ (non significatif), on se rend compte que la liaison est en réalité déterminée par les autres préférences (principalement le coût si on approfondit l'exploration).

4.2.8 Cercle des corrélations

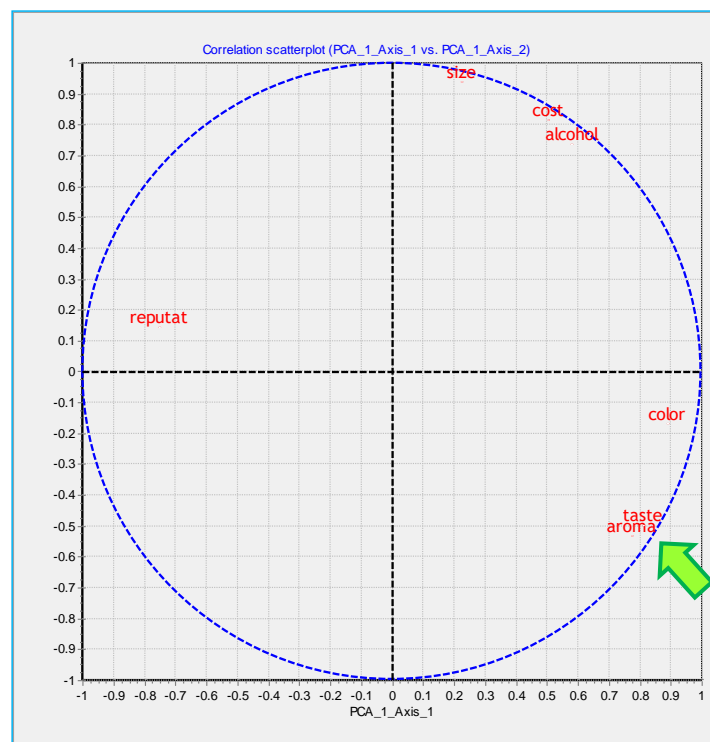
Le cercle de corrélation sert à obtenir une vue synthétique du positionnement des variables par rapport aux axes. Nous insérons le composant DEFINE STATUS dans le diagramme, nous plaçons les deux premiers axes générés par l'ACP en TARGET, les variables de l'analyse en INPUT.



Puis nous ajoutons le composant CORRELATION SCATTERPLOT.



Nous visualisons immédiatement les attractions et oppositions entre les variables. Nous devinons la superposition entre AROMA et TASTE. Pour les distinguer, nous pouvons les déplacer très légèrement avec le menu COMPONENT / JITTER¹⁰.



5 Outils supplémentaires pour l'ACP

5.1 Analyse parallèle

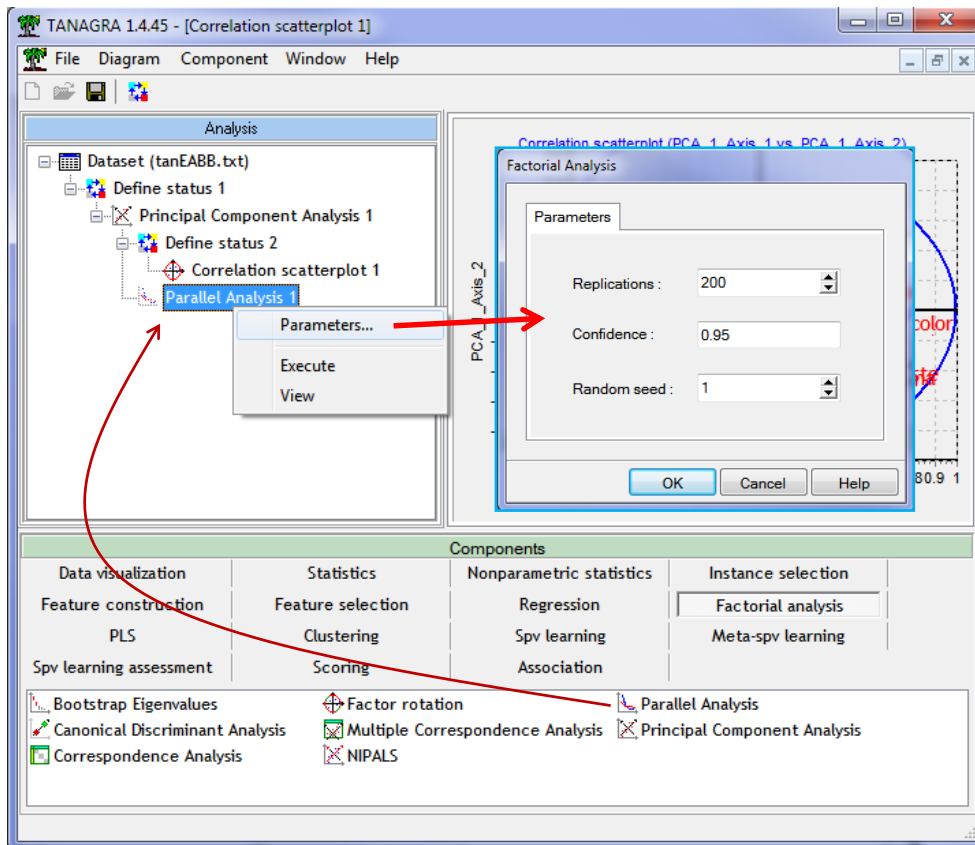
L'analyse parallèle est une technique de Monte-Carlo destinée à déterminer les seuils critiques pour tester la significativité des axes. Des données de mêmes caractéristiques « n » et « p » que notre fichier sont générées aléatoirement, les valeurs propres (q_k) sont collectées. En répétant le processus, nous obtenons la distribution sous H_0 (les variables sont deux à deux indépendantes) de chaque q_k . Nous considérons qu'un axe est significatif si sa valeur propre observée (λ_k) est supérieure au quantile d'ordre 0.95 (pour un test à 5%) de q_k sous H_0 .

Tanagra, plutôt que de générer de nouvelles données, s'appuie sur la randomisation. Les valeurs sont mélangées aléatoirement dans chaque colonne, indépendamment les unes des autres. Les résultats sont équivalents à ceux de l'analyse parallèle originelle et, surtout, le procédé est applicable à la fois à l'analyse en composantes principales et l'analyse factorielle des correspondances multiples (composant MULTIPLE CORRESPONDANCE ANALYSIS de Tanagra)¹¹.

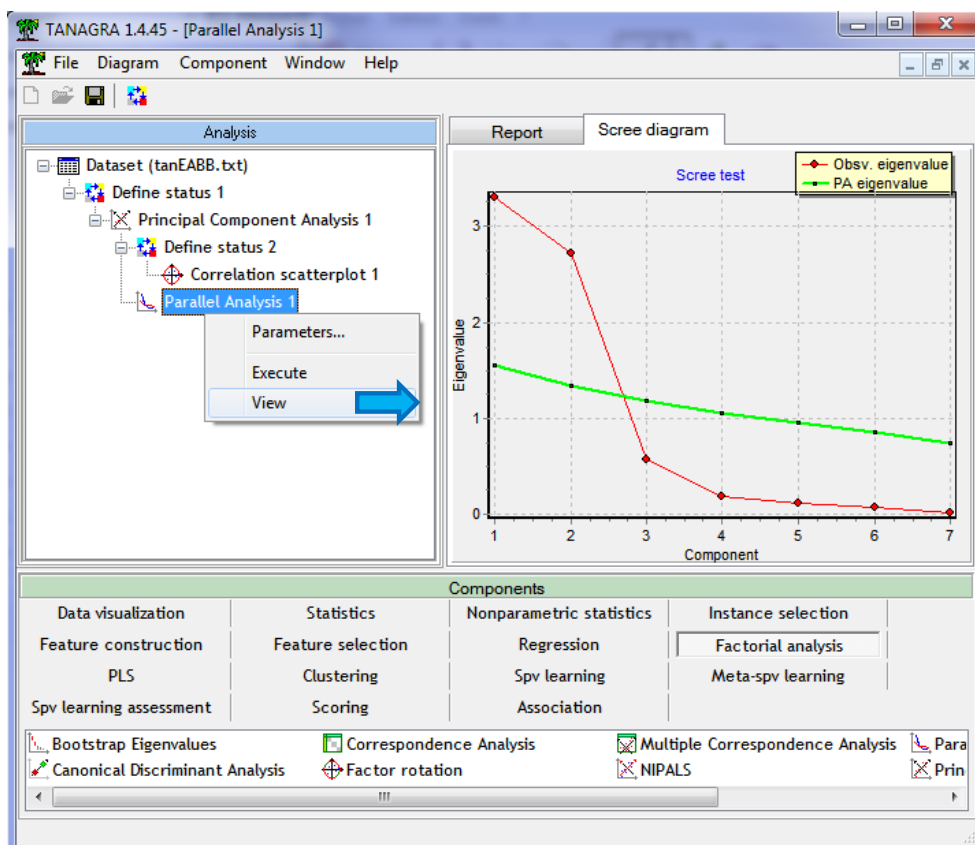
Nous insérons le composant PARALLEL ANALYSIS (onglet FACTORIAL ANALYSIS) à la suite de l'analyse en composantes principales, nous actionnons le menu PARAMETERS.

¹⁰ Cf. <http://blogs.sas.com/content/iml/2011/07/05/jittering-to-prevent-overplotting-in-statistical-graphics/>

¹¹ Mais pas à l'AFC (analyse factorielle des correspondances) qui prend comme point de départ un tableau de contingence. S'il fallait générer les valeurs aléatoirement, il faudrait respecter le nombre de lignes et de colonnes des données, mais aussi les profils marginaux lignes et colonnes. L'affaire devient vite très compliquée.



Le processus est réitéré 200 fois (REPLICATIONS), le seuil critique est définie au risque 5% (CONFIDENCE = 0.95), RANDOM SEED sert à l'initialisation du générateur de nombres aléatoires. Nous validons et nous cliquons sur VIEW.



L'éboullis des valeurs propres est en rouge, les seuils critiques en vert. Il semble effectivement que 2 facteurs soient pertinents dans notre ACP. Le détail des valeurs sont fournies dans l'onglet REPORT.

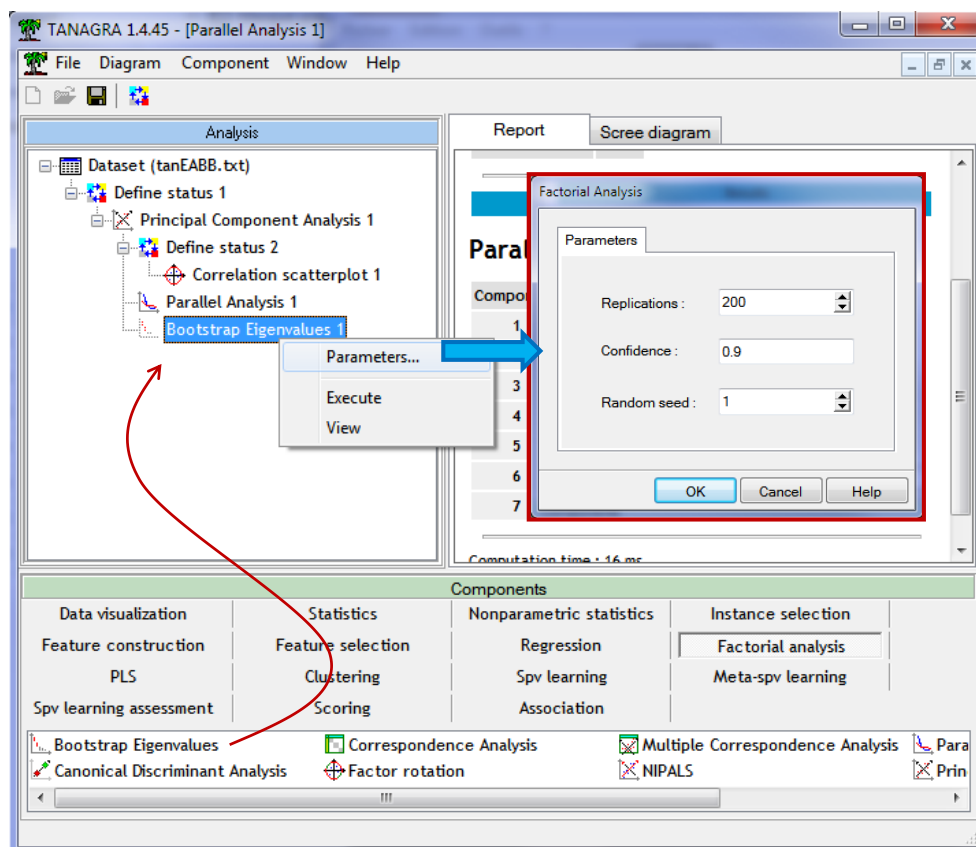
Component	Eigenvalue	(0.95) Critical value
1	3.306082	1.552704
2	2.719206	1.339758
3	0.567371	1.185519
4	0.193431	1.050049
5	0.119954	0.957981
6	0.076735	0.851874
7	0.017222	0.738887

5.2 Intervalle de variation des valeurs propres – Bootstrap

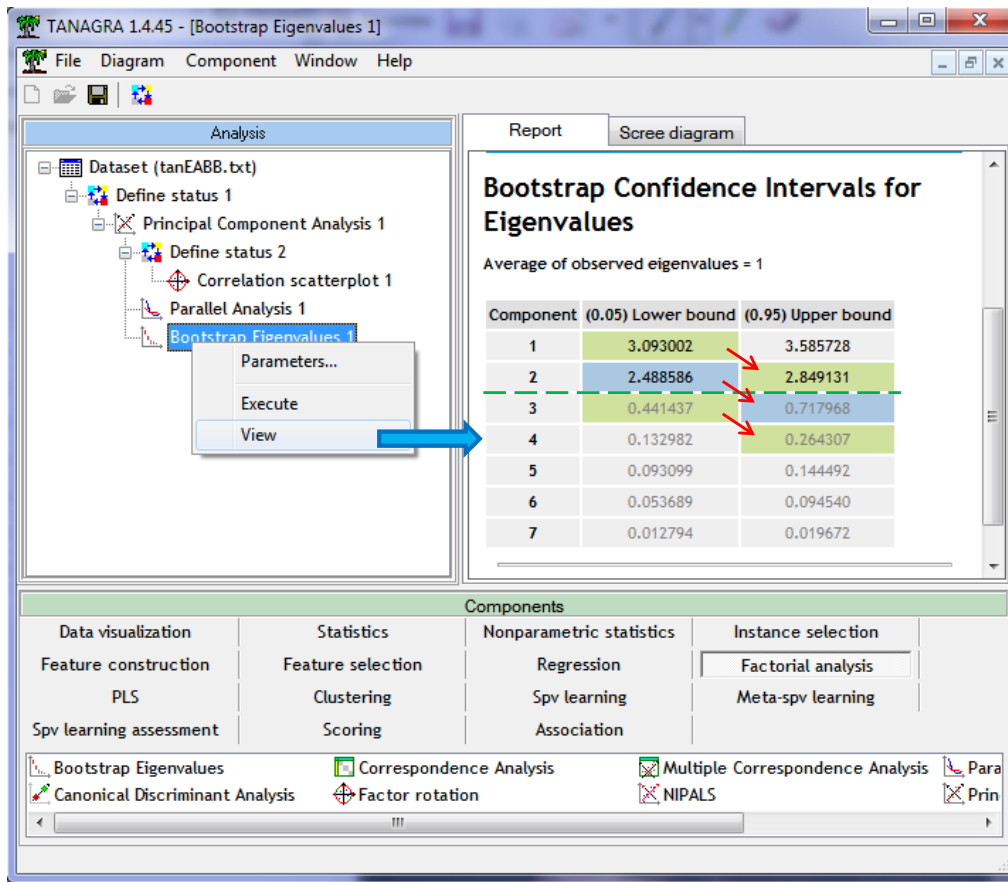
Nous pouvons calculer l'intervalle de confiance des valeurs propres λ_k par bootstrap. Deux règles de détection des facteurs significatifs peuvent s'appliquer : la borne basse de l'intervalle de confiance est supérieure à 1 pour une ACP normée (ou, plus généralement, supérieure à la moyenne des valeurs propres) ; la borne basse de λ_k est supérieure à la borne haute de λ_{k+1} , ce qui indiquerait un décalage significatif entre deux valeurs propres successives.

Le procédé s'applique autant à l'ACP qu'à l'ACM.

Nous insérons le composant BOOTSTRAP EIGENVALUES (onglet FACTORIAL ANALYSIS) à la suite de l'analyse en composantes principales. Nous le paramétrons : REPLICATIONS = 200, CONFIDENCE = 0.90 est le niveau de confiance pour le calcul de l'intervalle, RANDOM SEED = 1.

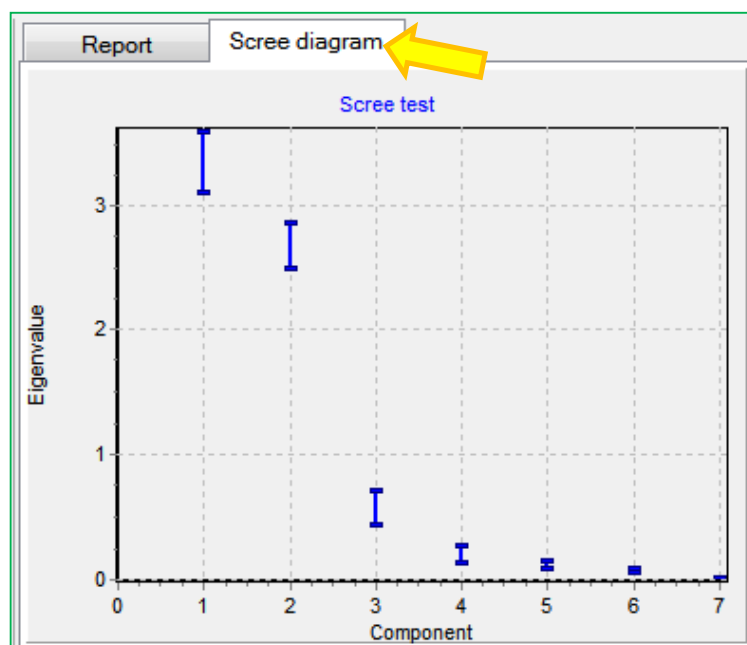


Nous validons et nous lançons les calculs avec VIEW.



Premier résultat : les bornes basses de deux premiers facteurs sont supérieures à 1, avec respectivement 3.09 et 2.49. Les autres valeurs sont grisées.

Second résultat : il n’y a pas d’empiètement entre le premier intervalle et le second, ni entre le second et le troisième, ni entre le troisième et le quatrième. Les décalages sont représentés graphiquement dans l’onglet SCREE DIAGRAM.

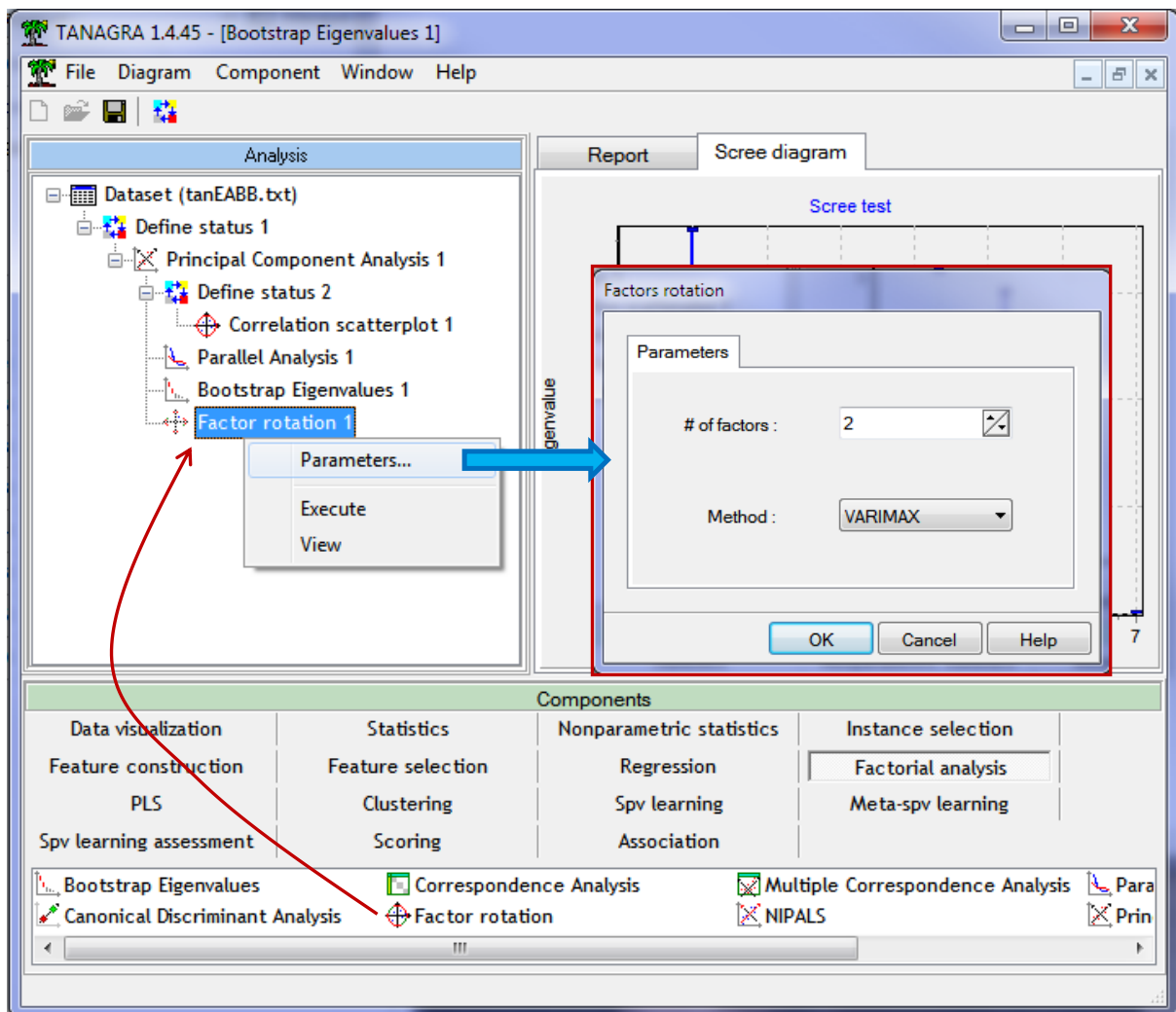


Ce dernier résultat semble accréditer la présence de 3 facteurs pertinents. Mais compte tenu de la faible valeur de la 3^{ème} valeur propre, dont la borne haute de l'intervalle (0.717968) n'est même pas supérieure à 1, on s'en tiendra sagement à 2 facteurs dans cette analyse.

5.3 Rotation VARIMAX

La rotation VARIMAX consiste à faire pivoter les axes factoriels de manière à rendre plus tranchées leurs corrélations avec les variables¹². L'interprétation des résultats en est facilitée. Notons que : (1) la part d'inertie expliquée n'est pas modifiée, (2) les composantes restent deux à deux orthogonales. La procédure s'applique uniquement à l'analyse en composantes principales.

Nous insérons le composant FACTOR ROTATION, nous ne travaillons que sur les deux premiers axes.



Nous avons le tableau des corrélations et \cos^2 après et avant rotation. Le premier plan factoriel exprime toujours 86% de l'inertie disponible. Définitivement, le premier axe définit l'esthétisme des consommateurs, le second leur aptitude à la débauche. « Réputation » reste toujours un peu à part, même si l'on constate qu'elle est (relativement) opposée à COLOR, AROMA et TASTE sur le premier axe (*hé oui, quand on sait apprécier, on est moins docile au matraquage publicitaire*).

¹² <http://tutoriels-data-mining.blogspot.fr/2008/04/rotation-varimax-en-acp.html>

Rotated Factor Loadings

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

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

vs. Unrotated Factor Loadings

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

Variance Explained by Each Factor	
Factor1	Factor2
3.2000307	2.8252567

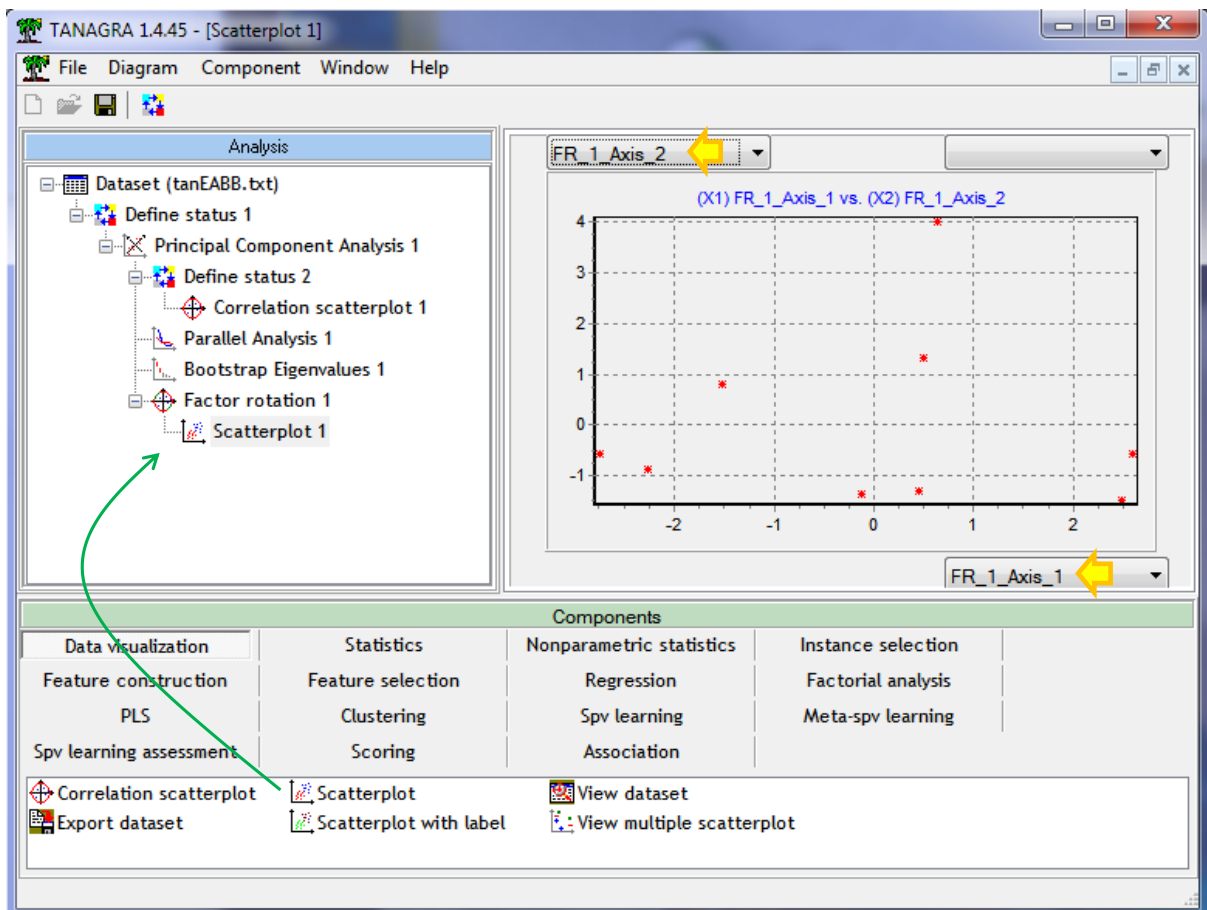
Final Communality Estimates: Total = 6.025287						
cost	size	alcohol	reputat	color	aroma	taste
0.92165047	0.95488671	0.93004894	0.56454880	0.85624989	0.88173669	0.91616599

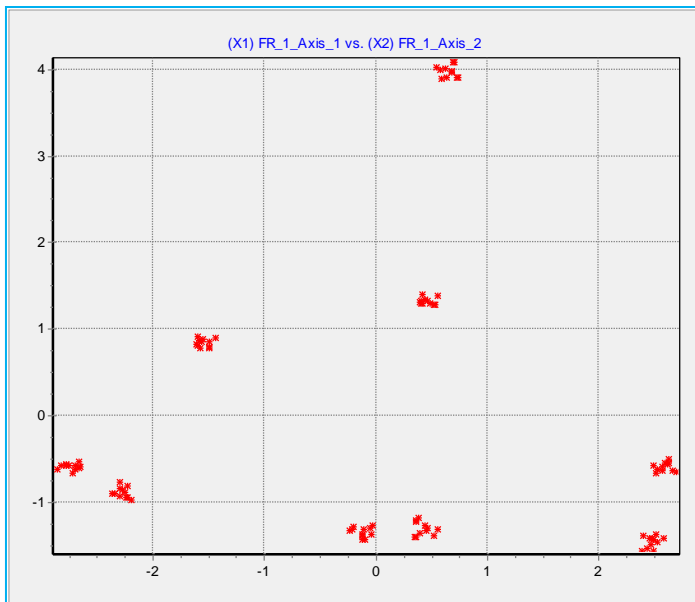
TANAGRA

SAS - PROC FACTOR

5.4 Projection des individus dans le premier plan factoriel

Voyons maintenant comment de situent les consommateurs dans le nouveau repère. Pour ce faire nous insérons le composant SCATTERPLOT (onglet DATA VISUALIZATION).





Les individus sont positionnés selon leurs coordonnées sur les facteurs pivotés (FR_1_AXIS_1 en abscisse, FR_2_AXIS_1 en ordonnée). Manifestement, il y a une forte superposition de nombreux points. Nous actionnons le menu COMPONENT / JITTER. Nous situons mieux les individus.

6 Clustering de variables

Pour confirmer les résultats, nous avons réalisé une classification ascendante hiérarchique sur les variables. Nous insérons le composant VARHCA (onglet CLUSTERING) dans le diagramme. Nous lançons directement les calculs avec le menu VIEW.

Pour confirmer les résultats, nous avons réalisé une classification ascendante

The screenshot shows the TANAGRA 1.4.45 interface with the VARHCA component selected in the Analysis tree. The main window displays the following data:

Cluster members and R-square values

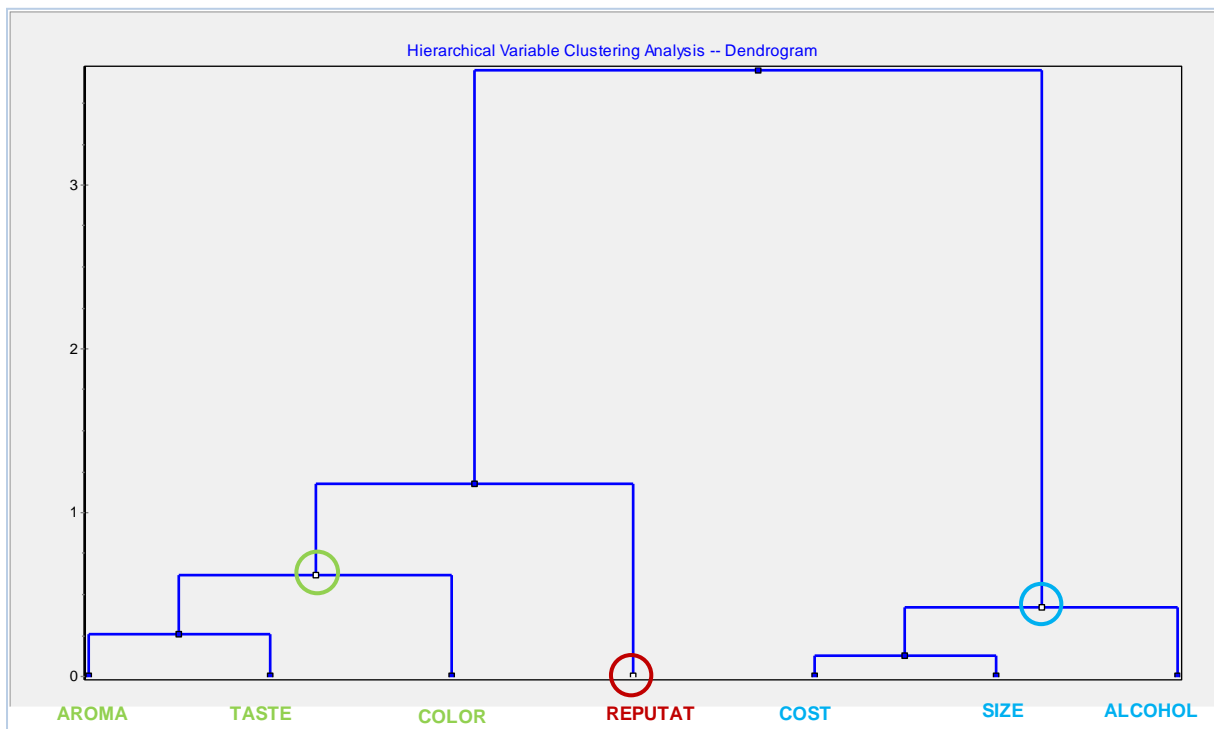
Cluster	Members	Own Cluster	Next Closest	1-R ² ratio
1	cost	0.9316	0.0305	0.0706
	size	0.8945	0.0037	0.1059
	alcohol	0.8935	0.1300	0.1224
2	aroma	0.9065	0.2720	0.1284
	taste	0.8937	0.3925	0.1749
	color	0.8628	0.2744	0.1891
3	reputat	1.0000	0.3499	0.0000

Cluster correlations -- Structure

Attribute	# membership	Cluster 1	Cluster 2	Cluster 3
cost	1	0.9652	0.1212	-0.1748
size	1	0.9458	-0.2065	-0.0612
alcohol	1	0.9453	0.1934	-0.3605
reputat	1	-0.2086	-0.5916	1.0000
color	1	0.2571	0.9289	-0.5238
aroma	1	-0.0754	0.9521	-0.5215
taste	1	-0.0684	0.9454	-0.6265

The Components panel at the bottom shows 'Clustering' selected, and the VARHCA component is highlighted in the component list.

Effectivement, la classification des variables en 3 groupes paraît le plus pertinent¹³. Nous pouvons visualiser le dendrogramme pour situer les proximités.



7 Conclusion

Nous décrivons dans ce tutoriel plusieurs outils relatifs à l'analyse en composantes principales nouvellement (version 1.4.45) implantées dans Tanagra. Bien sûr, ils sont disponibles ici ou là dans différents outils libres (je pense aux packages du logiciel R entre autres¹⁴). Notre seul mérite est de les avoir intégrés de manière unifiée et cohérente dans un seul logiciel.

¹³ Voir <http://tutoriels-data-mining.blogspot.fr/2008/03/classification-de-variables.html>

¹⁴ Ex. le package [PSYCH](#) pour le test de Bartlett.