1 Objectif

Description de deux alternatives à l'analyse en composantes principales (ACP) : les composants Principal Factor Analysis et Harris Component Analysis [algorithmes non itératifs] de Tanagra. Comparaisons avec SAS, R (package PSYCH) et SPSS.

L'analyse en composantes principales (ACP) est une technique factorielle très populaire. Elle est utilisée, entres autres, pour synthétiser les informations contenues dans un ensemble de données composé de variables actives exclusivement quantitatives. Elle est largement décrite dans la très abondante littérature en langue française consacrée à l'analyse de données.

La situation est un peu différente lorsqu'on s'intéresse aux ouvrages en langue anglaise. L'ACP y est présente certes, mais d'autres approches sont également mises en avant, notamment l'analyse en facteurs principaux [AFP] (Principal Factor Analysis en anglais) que l'on dit même préférable à l'ACP (ah bon ?). Malgré tout, la plupart des auteurs concèdent que cette dernière reste la plus utilisée.

Qu'est ce qui les distingue, qu'est-ce qui les réunit ?[1] Ce sont des techniques factorielles, raison pour laquelle on les confond bien souvent. Mais l'ACP cherche à résumer de manière la plus efficace possible l'information disponible en s'intéressant à la variabilité totale portée par chaque variable de la base. Il s'agit donc d'une technique de compression, intéressante surtout lorsque l'on cherche à exploiter les facteurs dans des études subséquentes (ex. analyse discriminante sur facteurs[2]). En revanche, l'AFP cherche à structurer l'information en s'intéressant à la variabilité commune aux variables. L'idée est de mettre en avant des facteurs sous-jacents (variables latentes) qui associent deux ou plusieurs colonnes des données. L'influence des variables qui font cavalier seul, indépendantes des autres, devrait être écartée.

Elles sont donc différentes de par la nature des informations qu'elles exploitent. Mais la nuance n'est pas évidente. D'autant plus qu'elles sont souvent regroupées dans le même outil dans certains logiciels (ex. « PROC FACTOR » dans SAS, « ANALYZE / DATA REDUCTION / FACTOR » dans SPSS, etc.), que les tableaux des résultats sont identiques, et que les interprétations sont finalement très proches.

Dans ce tutoriel, nous décrirons trois techniques d'analyse factorielle pour variables quantitatives (Principal Component Analysis - ACP, Principal Factor Analysis, Harris Component Analysis). **Nous nous en tiendrons aux algorithmes non itératifs pour les deux dernières.** L'ACP, maintes fois présentée[3], servira surtout de repère pour les deux suivantes. Nous les distinguerons en détaillant la matrice (de corrélation pour l'ACP) qui sera présentée à l'algorithme de diagonalisation. Ce prisme

[1] D. Suhr, « Principal Component Analysis vs. Exploratory Factor Analysis », http://www2.sas.com/proceedings/sugi30/203-30.pdf ; STATSOFT Electronic Statistics Book, « How to Reduce Number of Variables and Detect Relationships, Principal Components and Factor Analysis », http://www.statsoft.com/textbook/principal-components-factor-analysis/

[2] « Analyse discriminante sur axes principaux » - http://tutoriels-data-mining.blogspot.fr/2008/03/analyse-discriminante-sur-axes.html

[3] Ex. « ACP avec Tanagra – Nouveaux outils » - http://tutoriels-data-mining.blogspot.fr/2012/06/acp-avec-tanagra-nouveaux-outils.html

permet de comprendre le type d'information que les méthodes mettent en avant à l'issue des calculs. Pour appuyer l'exposé, nous préciserons chaque étape des opérations sous le logiciel R en mettant en miroir les résultats fournis par SAS. Par la suite, nous décrirons leur mise en œuvre sous les logiciels Tanagra (section 4), R avec le package PSYCH (section 5) et SPSS 12.0.1 (section 6).

2 Données

Nous travaillons sur le fichier « beer_rnd.xls ». Il décrit les préférences (une note allant de 0 à 100, seuls les multiples de 5 sont acceptés) de **n = 99** individus sur des thèmes relatifs à la bière [la boisson] (coût, taille, etc.). Nous avions déjà traité ce fichier précédemment[4]. Nous corsons l'affaire en rajoutant 7 variables générées aléatoirement (rnd1…rnd7). Nous avons donc en tout **p = 14** variables dans la base. Notre objectif est de jauger la capacité des techniques factorielles à discerner les informations « utiles » que portent les données c.-à-d. leur capacité à mettre en évidence les relations existants entre les variables en présence de bruit.

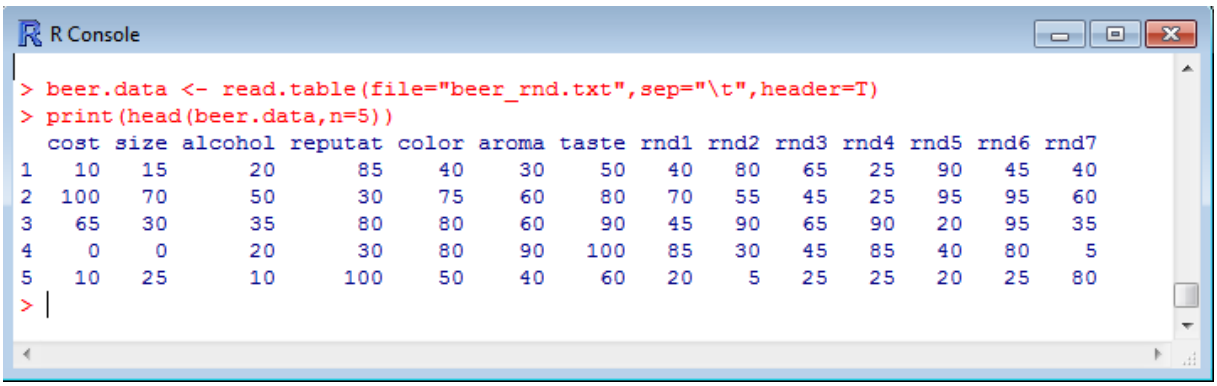
Remarque : Notons que « bruit » n'est peut être pas toujours le terme le plus approprié pour qualifier ces variables additionnelles. Une variable peut être tout à fait légitime mais n'a aucun rapport avec le sujet de l'étude. Elle peut aussi, et ça devient particulièrement compliquée à ce stade, être parfaitement pertinente mais représenter une « dimension » à elle seule. Il est évidemment hors de question de tenter de l'évacuer de l'étude dans ce cas.

Nous montrons dans la copie d'écran suivante les 5 premières lignes du fichier de données.

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 Description des approches, détail des calculs sous R

Nous implémentons les différentes techniques factorielles en détaillant les étapes sous R. Dans un premier temps, nous chargeons le fichier « beer_rnd.txt » (version au format texte) et nous affichons les 5 premières observations.



[4] Voir http://tutoriels-data-mining.blogspot.fr/2012/06/acp-avec-tanagra-nouveaux-outils.html

3.1 Analyse en composantes principales

La matrice des corrélations **C** (de taille p x p) est le point de départ de l'ACP normée. Sous le logiciel R, nous la calculons et l'affichons avec les instructions suivantes :

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

Sa structure est très instructive.

	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 - Matrice des corrélations

La matrice indique les relations entre les variables prises deux à deux. Nous distinguons facilement des « blocs » dans notre fichier :

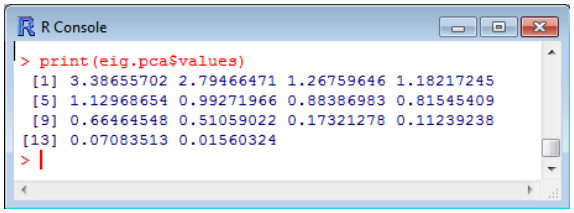
- (cost, size et alcohol) forment un premier groupe avec des corrélations croisées très fortes. Il caractérise les buveurs impénitents qui en veulent beaucoup (size) pour par cher (cost), avec un fort degré d'alcoolémie (alcohol).
- (color, aroma et taste) forme un second groupe. Il caractérise les esthètes de la bière.
- (reputat) est liée négativement avec ce second groupe, mais à un degré moindre (corrélation autour de 0.5 en valeur absolue). Il semble que les esthètes soient peu sensibles à la réputation des marques de bières (*ils se fient à leur propre jugement…*).
- Enfin, nous avons un troisième groupe où les variables sont très peu liées. Elles ne le sont pas non plus avec les autres colonnes de toute manière. Le contraire eut été étonnant, elles ont été générées aléatoirement.

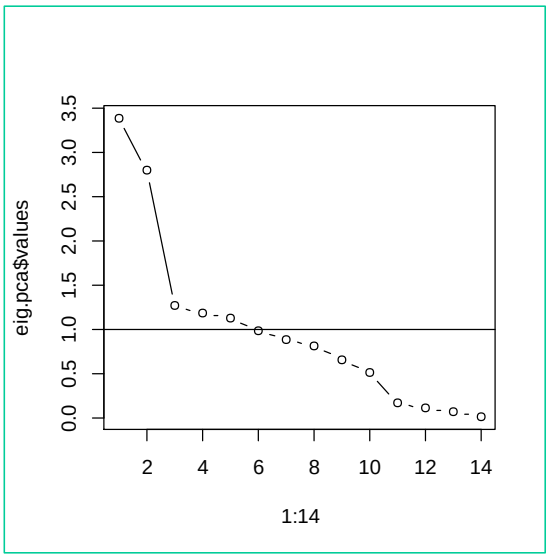
Bien évidemment, la corrélation d'une variable avec elle-même est égale à 1. C'est ce que nous indique la diagonale principale de la matrice. Cette dernière remarque est tout sauf anodine. En effet, l'ACP va en tenir compte lorsqu'elle diagonalisera la matrice de corrélation. Elle traitera la variabilité totale des variables en leur accordant la même importance.

Valeurs propres. Voici les instructions pour diagonaliser la matrice et afficher les valeurs propres (et le scree plot) :

```
#calcul des valeurs et vecteurs propres de la matrice de corrélation
eig.pca <- eigen(beer.cor)
```

```
#affichage
print eigenvalues
print("eigenvalues")
print(eig.pca$values)
#screeplot
plot(1:14,eig.pca$values,type="b")
abline(a=1,b=0)
```





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

(R) (SAS)

Figure 2 - Tableau des valeurs propres - ACP

Les résultats concordent en tout point avec ceux de la PROC FACTOR de SAS[5]. Cette dernière indique clairement qu'elle traite la variabilité totale des variables avec la mention « Prior Communality Estimates : ONE ».

Vient alors un sujet délicat en ACP, combien d'axes faut-il retenir ? Si on s'en tient à la règle de Kaiser (valeur propre ≥ 1), il faudrait sélectionner 5 axes, voire 6 (valeur propre = 0.9927). Ce n'est pas vraiment choquant si l'on se penche sur la nature des données. Les 7 variables ont été générées de manière indépendante, il est difficile de compresser l'information en les ramenant à un nombre

[5] Nous avons utilisé l'instruction suivante :

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

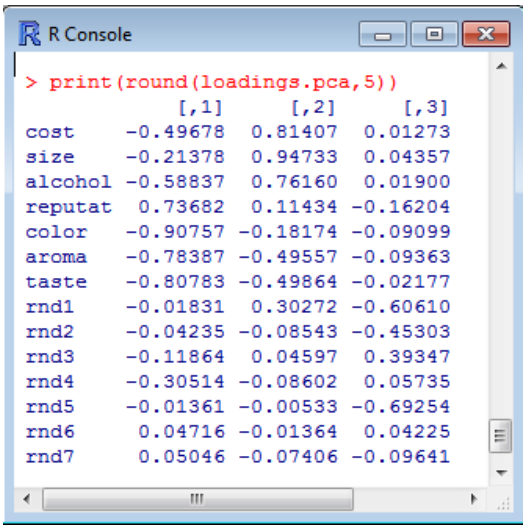
réduit de facteurs. Nous savons en revanche que ce n'est pas du tout adapté si l'on cherche à étudier les relations entre les variables dans notre fichier.

Avec l'éboulis des valeurs propres, on n'en conserverait que 3 si l'on prend en compte le coude, 2 si on décide de l'évacuer (Cattell lui-même a varié sur ce point). On se rapproche de la bonne solution qui est de 2 facteurs au regard de la structure de la matrice des corrélations (Figure 1).

Loadings ou Factor pattern. Ce tableau retrace les corrélations des variables avec les axes c.-à-d. si nous calculons la corrélation des variables avec les colonnes des coordonnées des observations sur les axes, nous obtiendrons ces chiffres. En pratique, nous les obtenons en multipliant les vecteurs propres par la racine carrée des valeurs propres. Pour les trois premiers axes :

```
#corrélation des variables avec les facteurs
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))
```

Nous retrouvons sur les deux premiers facteurs les principales tendances que l'on constatait dans la matrice des corrélations : (color, aroma et taste) s'opposent à (reputat) ; (cost, size et alcohol) sont concordants. Malgré les réticences initialement exprimées, et le fait que l'on ait volontairement perturbé le fichier, l'ACP sait mettre en avant les informations importantes dans le fichier.



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 - Corrélation des variables avec les axes - ACP

Finalement, l'enjeu réside surtout dans le choix du bon nombre de facteurs, qui n'était absolument pas évident dans notre configuration. Il est pourtant crucial pour une bonne lecture des résultats. En effet, si nous incluons le 3ème axe dans l'analyse (valeur propre = 1.268 ; 9% de l'information disponible), nous conclurons à une préférence concordante des individus sur les critères RND1 et

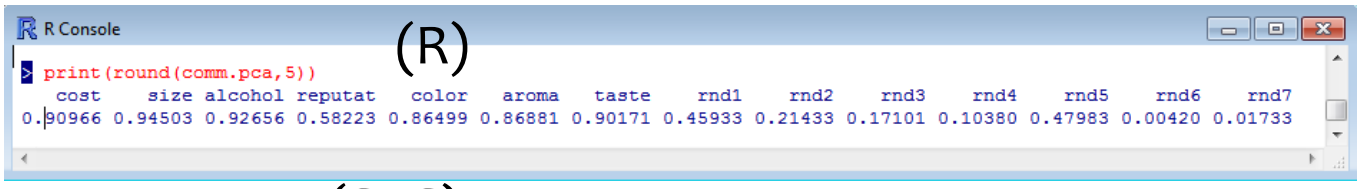
RND5. Et nous savons pertinemment qu'il n'en est rien puisque leurs valeurs ont été générées aléatoirement.

Communalities. Ce tableau indique la qualité de représentation des variables sur les axes retenus. Pour chaque variable, il s'agit simplement de la somme des cosinus carrés, qui lui-même est égal au carré de la corrélation. En intégrant les trois premiers axes, nous faisons sous R :

```
#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))
```

Toutes les variables « vraies » du fichier sont bien représentées sur les 3 premiers axes.

(R)



(SAS)

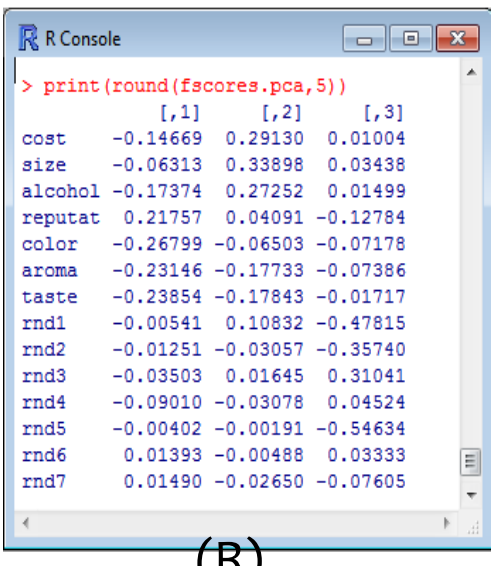
Final Communality Estimates: Total = 7.448818													
cost	size	alcohol	reputat	color	aroma	taste	rnd1	rnd2	rnd3	rnd4	rnd5	rnd6	rnd7
0.90966494	0.94502828	0.92656390	0.58223291	0.86499268	0.86880794	0.90171079	0.45933306	0.21433226	0.17100692	0.10379690	0.47982577	0.00419547	0.01732637

Factor scores – 1. Ce tableau fournit les coefficients des facteurs. Ils permettent de projeter les individus dans le nouveau repère. Pour les obtenir, on multiplie l'inverse de la matrice de corrélation avec les « loadings » :

```
#inversion de la matrice des corrélations
inv.beer.cor <- solve(beer.cor)
# factor scores
fscores.pca <- inv.beer.cor%*%loadings.pca
print(round(fscores.pca,5))
```

Aux signes près, qui sont attribués arbitrairement, nous avons les mêmes valeurs sous SAS.

(R)



(SAS)

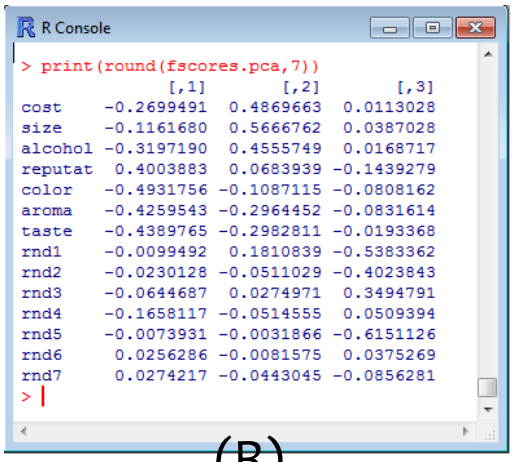
Standardized Scoring Coefficients				
		Factor1	Factor2	Factor3
cost	cost	0.14669	0.29130	-0.01004
size	size	0.06313	0.33898	-0.03438
alcohol	alcohol	0.17374	0.27252	-0.01499
reputat	reputat	-0.21757	0.04091	0.12784
color	color	0.26799	-0.06503	0.07178
aroma	aroma	0.23146	-0.17733	0.07386
taste	taste	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

En appliquant ces coefficients sur les individus de la base d'apprentissage (les variables doivent être centrées et réduites) et en calculant la variance de chaque axe, nous obtiendrons la valeur 1, quel que soit l'axe. C'est ce proposent les logiciels tels que SAS ou SPSS.

Factor scores – 2. L'autre piste serait d'attribuer les scores de manière à ce que les variances des axes coïncident avec leurs valeurs propres. C'est le choix de Tanagra et des principaux packages de R (princomp, prcomp pour la variance [1/(n-1)], PCA de FactoMineR). Pour ce faire, nous multiplions les coefficients calculés précédemment avec la racine carrée des valeurs propres :

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

Les résultats sont identiques à ceux de Tanagra cette fois-ci.



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 – Contribution aux axes. Les coefficients de projection ont un rôle opérationnel. Ils permettent de calculer les coordonnées des individus dans le repère factoriel. Mais ils peuvent aussi servir à l'interprétation. En effet, ils s'appliquent sur des variables standardisées (centrées et réduites), il est par conséquent possible de les comparer pour identifier les variables qui ont le plus fort impact dans la définition de chaque facteur. On parle alors de « contributions aux axes »[6].

A partir du tableau des « Factor Scores » (non corrigée par les valeurs propres, mais tout de manière les résultats auraient été identiques), nous avons calculé le carré des coefficients pour chaque facteur. La contribution est alors égale à cette valeur rapportée à la somme totale.

Par exemple, pour le facteur 1, le coefficient de « cost » est 0.14669, passé au carré il est égal à 0.02152, rapporté à la somme des carrés (0.29528) sa contribution est 0.02152/0.29528 = 7.287%.

[6] **Remarque** : On trouve rarement cette lecture des contributions dans la littérature. On les présente souvent comme le carré des corrélations (des loadings) rapporté à la valeur propre pour chaque axe. Et on ajoute alors que leur valeur ajoutée par rapport aux corrélations est faible dans l'interprétation des résultats en ACP. C'est tout à fait vrai. L'intérêt de notre présentation est que nous disposons maintenant d'une procédure - et d'un critère numérique - pour évaluer la capacité à évacuer les variables « rnd » dans la définition des facteurs

Voici notre feuille de calcul pour l'ensemble des variables :

Standardized Scoring		
	Factor1	Factor2
cost	0.14669	0.2913
size	0.06313	0.33898
alcohol	0.17374	0.27252
reputat	-0.21757	0.04091
color	0.26799	-0.06503
aroma	0.23146	-0.17733
taste	0.23854	-0.17843
rnd1	0.00541	0.10832
rnd2	0.01251	-0.03057
rnd3	0.03503	0.01645
rnd4	0.0901	-0.03078
rnd5	0.00402	-0.00191
rnd6	-0.01393	-0.00488
rnd7	-0.01490	-0.02650

Squared Coefficients	
Factor1	Factor2
0.02152	0.08486
0.00399	0.11491
0.03019	0.07427
0.04734	0.00167
0.07182	0.00423
0.05357	0.03145
0.05690	0.03184
0.00003	0.01173
0.00016	0.00093
0.00123	0.00027
0.00812	0.00095
0.00002	0.00000
0.00019	0.00002
0.00022	0.00070
Total 0.29528	0.35783

Contributions	
Factor1	Factor2
0.07287	0.23714
0.01350	0.32112
0.10223	0.20755
0.16031	0.00468
0.24322	0.01182
0.18143	0.08788
0.19270	0.08897
0.00010	0.03279
0.00053	0.00261
0.00416	0.00076
0.02749	0.00265
0.00005	0.00001
0.00066	0.00007
0.00075	0.00196
CTR(rnd) 3.37%	4.08%

La lecture est cohérente avec celle des « loadings ». Via leurs contributions, nous distinguons les groupes de variables sur chaque axe et, clairement, (rnd1, …, rnd7) pèsent très peu dans la définition du premier (3.37%) comme du second facteur (4.08%).

Conclusion. Tous ces résultats sont bien connus et très largement décrits dans la littérature. Revenir dessus était cependant nécessaire pour comprendre les apports des techniques qui seront présentées dans ce qui suit.

3.2 Analyse en facteurs principaux (AFP)

L'analyse en facteurs principaux (*principal factor analysis* ou *common factor analysis* ou *principal axis factoring* en anglais[7]) cherche à identifier les facteurs (les variables latentes) qui structurent les données. Ils sont associés à deux ou plusieurs variables de la base. L'analyse s'intéresse donc exclusivement à la variabilité partagée entre les variables[8].

Le point de départ est toujours la matrice de corrélation. Mais, pour chaque variable X_j, nous remplaçons les valeurs de la diagonale principale (qui est égale à 1 lorsque l'on tient compte de toute la variabilité disponible) par la part de variance expliquée par les autres colonnes. Concrètement, il s'agit du coefficient de détermination R_j^2 de la régression de Xj sur les (p-1) autres variables. On parle de « prior communalities » (ou « initial estimates of communalities »). Nous utiliserons le terme de « communalités initiales »[9].

Nous travaillerons donc à partir de la matrice **F** suivante (Figure 4) pour notre nouvelle analyse :

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

[88] M. Chavent, V. Kuentz, J. Saracco, « Analyse en Facteurs : présentation et comparaison des logiciels SAS, SPAD et SPSS », Revue Modulad, n°37, 2007.

[9] Certains outils utilisent le terme « communautés »…

	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 – Matrice F à diagonaliser pour l'analyse en facteurs principaux

Les blocs de variables sont les mêmes. Mais nous remarquons que (cost,…, taste) peuvent être expliquées à partir des autres, à l'inverse des variables (rnd1,.., rnd7). En effet, nous lisons dans cette matrice : le coefficient de détermination de la régression de « cost » sur (size, alcohol, …, rnd7) est égal à 0.96 ; R^2(size / cost,alcohol, …, rnd7) = 0.94 ; … ; R^2(rnd1/cost, alcohol, …) = 0.14 ; etc.

A priori, nous sommes obligés de mener « p » régressions pour calculer les communalités. Ce qui peut s'avérer très lourd surtout si la base est volumineuse. En pratique, nous les extrayons à partir de l'inverse C^{-1} de la matrice des corrélations C. Nous avons :

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

Où $\left(c_{jj}^{-1}\right)$ est le j^{ème} élément diagonal de la matrice C^{-1}.

La quantité $u_j = 1 - R_j^2 = \frac{1}{c_{jj}^{-1}}$ est appelée « uniqueness » dans les logiciels anglo-saxons. Elle correspond à la part de variance (résiduelle) de X_j non expliquée par les (p-1) autres variables.

Voyons le détail des calculs pour nos données. Nous inversons la matrice des corrélations :

Inverse of the correlation matrix

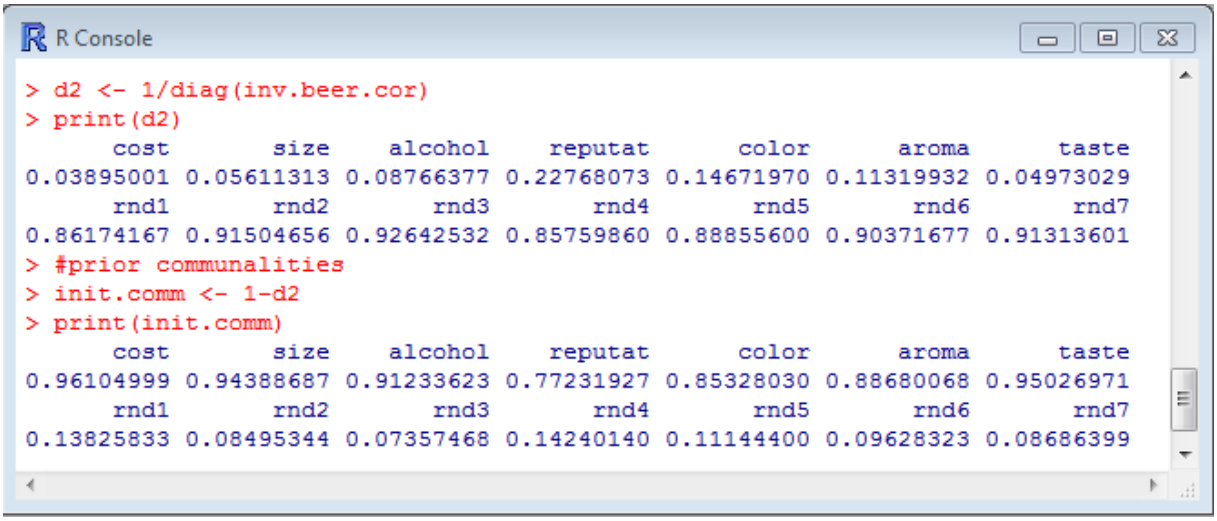
	cost	size	alcohol	reputat	color	aroma	taste	rnd1	rnd2	rnd3	rnd4	rnd5	rnd6	rnd7
cost	25.67	-17.54	-10.39	-7.39	-0.55	9.10	-18.10	0.62	0.91	0.06	-0.74	-1.00	0.10	0.38
size	-17.54	17.82	1.77	4.62	0.86	-5.00	12.72	-0.54	-0.66	-0.01	0.57	0.87	-0.40	-0.51
alcohol	-10.39	1.77	11.41	4.48	-2.84	-4.04	8.98	-0.48	-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 de la matrice des corrélations

Via la diagonale, pour « cost » nous avons $u_{cost} = \frac{1}{25.67} = 0.04$ et $R^2_{cost} = 1 - 0.04 = 0.96$. Voici les instructions idoines sous R :

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

Nous obtenons les vecteurs des « uniqueness » et des communalités.



Nous affectons les communalités estimées aux éléments diagonaux de C pour obtenir la matrice F :

```
#nouvelle matrice à diagonaliser
beer.cor.pfa <- beer.cor
#remplacer la diagonale par les communalités
diag(beer.cor.pfa) <- init.comm
#trace de la matrice
print(sum(diag(beer.cor.pfa)))
```

Nous pouvons définir formellement les termes de la matrice F comme suit :

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

La trace de la matrice est égale à $\sum_{j=1}^{p} R_j^2$ =7.0137 ; c'est la quantité totale d'information que l'on essaiera de décomposer.

Valeurs propres. Nous diagonalisons la matrice F pour obtenir les valeurs propres[10].

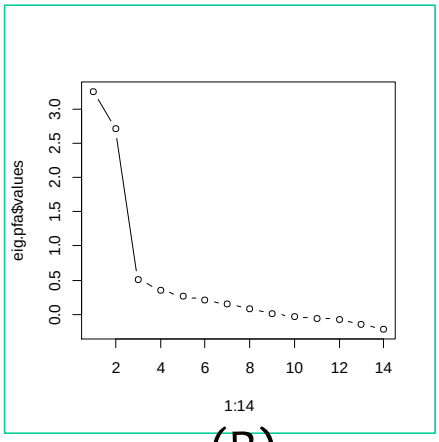
[10] Sous SAS, nous exécutons les instructions suivantes, l'option « priors = smc » est primordiale :

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

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

SAS et R fournissent les mêmes valeurs.

```
> 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)

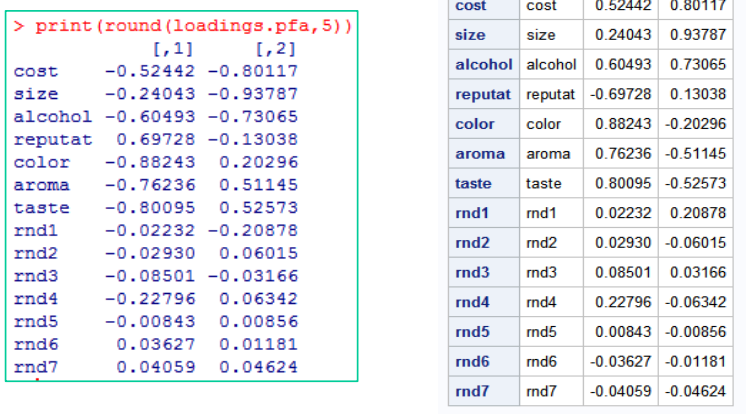
Première surprise, certaines valeurs propres sont négatives. Ca n'en est pas vraiment une, la matrice **F**, contrairement à **C**, n'est pas semi-définie positive[11]. Concrètement, jusqu'au 4ème axe (la somme des valeurs propres n'excède pas la trace de la matrice F), nous traitons bien la variance partagée. A partir du 5ème, la variance intrinsèque aux variables pèse sur les facteurs, au point qu'il faut retrancher de l'information (à partir du 10ème axe) pour que la somme totale des valeurs propres soit bien égale à la trace de la matrice que nous avons diagonalisée.

Une modélisation en 2 facteurs est la plus appropriée au regard de l'éboulis des valeurs propres. Le décalage est très accentué entre la 2nde et la 3ème valeur propre. Ces deux premiers axes restituent 84.92% de l'information à traiter. Cette issue était moins évidente s'agissant de l'ACP (voir Figure 2).

Loadings ou Factor pattern. De nouveau, nous calculons le tableau des « loadings » des variables avec les facteurs, en nous restreignant aux deux premiers.

```
#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))
```

[11] http://en.wikipedia.org/wiki/Positive-definite_matrix#Positive-semidefinite



```
> 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" - Analyse en facteurs principaux

Loadings ≠ corrélation. Attention, les « loadings » ne correspondent plus exactement aux corrélations des variables avec les axes dans le cas de l'AFP. Il s'agirait plutôt des coefficients standardisés de la régression de chaque variable avec les facteurs[12]. Dans les faits, l'usage du tableau des « loadings » reste le même. Il sert à interpréter les axes.

Communalités. Ce tableau est encore plus intéressant dans cette analyse puisqu'il s'agit de confronter l'information restituée avec l'information initialement exploitable pour chaque 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))
```

L'information partagée par les « vraies » variables de la base sont parfaitement restituées sur les 2 premiers axes, à un degré légèrement moindre pour « reputat » cependant. Ces facteurs permettent d'appréhender les relations existantes entre les variables.

```
> 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 - Confrontation des communalités initiales et estimées – AFP

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

Pour la cohérence, nous noterons que la somme des valeurs propres des deux facteurs sélectionnés est égale à la somme des communalités estimées.

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

Factor scores. De nouveau ici, nous obtenons les coefficients de fonctions de projection en multipliant l'inverse la matrice des corrélations avec les « loadings ».

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

Il faut les appliquer sur les variables centrées et réduites pour obtenir les coordonnées des individus dans le plan factoriel.

```
> 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 des variables « rnd ». Nous calculons les contributions à partir du tableau des « Factor Scores », nous constatons que le rôle des « rnd » est considérablement amoindri par rapport à l'ACP (0.30% vs. 3.37 pour l'ACP sur le premier axe ; 0.13% vs. 4.08% pour le second). C'est le principal intérêt de cette approche dans l'étude des relations entres les variables : celles qui ne sont pas liées avec les autres voient leur impact très largement amoindri sur les premiers facteurs.

Standardized Scoring Coefficients		
	Factor1	Factor2
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

Squared Coefficients	
Factor1	Factor2
0.00596	0.41914
0.04505	0.02619
0.14652	0.00227
0.00194	0.00771
0.01854	0.00292
0.01469	0.00006
0.36252	0.27831
0.00036	0.00029
0.00000	0.00007
0.00049	0.00007
0.00040	0.00032
0.00041	0.00003
0.00003	0.00011
0.00014	0.00005
Total 0.59705	0.73753

Contributions	
Factor1	Factor2
0.00998	0.56830
0.07546	0.03551
0.24541	0.00308
0.00324	0.01045
0.03106	0.00396
0.02461	0.00008
0.60719	0.37735
0.00060	0.00039
0.00000	0.00010
0.00082	0.00009
0.00068	0.00044
0.00068	0.00004
0.00005	0.00015
0.00023	0.00006
CTR(rnd) 0.30%	0.13%

« Fidélité » des facteurs. En théorie, la variance des facteurs vaut 1. Sauf que nous ne disposons pas des « vrais » coefficients définis sur la population, mais des estimations obtenues à partir d'un échantillon. La variance observée des coordonnées des individus sur les axes donne une indication sur leur fiabilité (des axes).

Nous obtenons ces variances en effectuant la somme du produit des coefficients de projection avec les « loadings ». Sous R, nous utilisons les instructions suivantes pour les deux axes :

```
#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))
```

Nous avons :

(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 utilise l'appellation « carré du coefficient de corrélation multiple des variables avec chaque facteur ». En effet, cette variance correspond également au carré de la corrélation entre la variable latente théorique (définie sur la population) et son estimation par le facteur (obtenue sur l'échantillon). L'utiliser comme indicateur de crédibilité des facteurs est dès lors tout à fait approprié. Une valeur élevée (≥ 0.7 selon certaines références) indique une bonne stabilité. Le facteur n'est pas le fruit d'un artefact statistique, il correspond à une réalité qui existe dans la population.

Dans notre étude, si nous tentons une modélisation en 5 axes, nous obtenons les carrés des corrélations suivantes :

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 des facteurs observés - Analyse en facteurs principaux

Manifestement, une modélisation en 2 facteurs est vraiment la bonne solution dans cette analyse.

3.3 Une stratégie itérative pour l'analyse en facteurs principaux

Il existe une variante itérative de l'analyse en facteurs principaux censée fournir une meilleure qualité de projection. La procédure précédente sert de première étape (section 3.2). Ensuite, nous remplaçons la diagonale de la matrice F par les communalités estimées. Nous relançons alors

l'analyse. On procède ainsi jusqu'à convergence du système c.-à-d. jusqu'à les communalités soient stables (ex. SAS). On peut aussi fixer arbitrairement le nombre maximum d'itérations (Ex. SPSS).

Attention, il se peut que les communalités d'une des variables soit supérieure à 1 dans certaines situations. On parle de « problème de Heywood ». Cela indique clairement une incohérence dans le processus. Les causes sont diverses, entres autres parce qu'il est possible que l'on ait sélectionné un nombre inapproprié de facteurs[13].

3.4 Analyse de Harris en composantes principales (Harris)

L'analyse de Harris et l'analyse en facteur principaux procèdent de la même logique. Ils cherchent à décomposer l'information partagée par les variables en travaillant à partir d'une version modifiée de la matrice de corrélations. Ici, l'idée consiste à exacerber les corrélations entre deux variables si elles (l'une des deux ou les deux simultanément) sont fortement liées aux autres.

Concrètement, nous partons de la matrice F (Figure 1), et nous en déduisons la nouvelle matrice H à diagonaliser avec la formule suivante :

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

Sur nos données, nous obtenons la matrice H (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 - Matrice H à diagonaliser pour l'analyse de Harris

Prenons un exemple pour expliciter les calculs. La corrélation entre cost et size est relativement élevée : 0.88. Par ailleurs, la part de variance de cost (resp. size) expliquée par les autres est R^2_{cost} = 0.961 (R^2_{size}=0.944). Tous deux sont fortement liés aux autres variables (au moins une partie). Nous en déduisons les « uniqueness » : u_{cost} = 0.039 et u_{size} = 0.056.

Ainsi, leur liaison prend plus d'intensité dans la nouvelle matrice :

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

Nous observons dans la matrice les mêmes groupes que précédemment (Figure 1). Mais, par rapport aux matrices C (ACP) et F (AFP), le différentiel des valeurs est accentué. Les blocs ressortent plus. L'analyse devrait exploiter à bon escient cette particularité.

[13] Voir http://v8doc.sas.com/sashtml/stat/chap26/sect21.htm

Sous R, nous utilisons une formulation matricielle que l'on retrouve sur les sites de SPSS (*qui désigne la méthode par l'appellation «Image (Kaiser, 1963) »*) et SAS (Harris, 1962)[14] :

```
#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)))
```

La trace de la matrice est bien plus élevée [**Tr(H) = 88.87841**] dans cette analyse.

```
[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
```

Valeurs propres. Nous diagonalisons H :

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

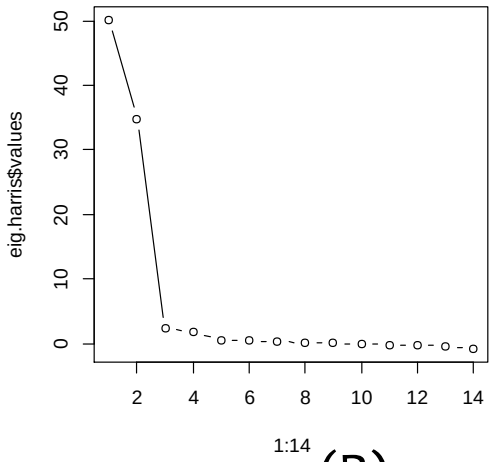
Ici également, nous pouvons obtenir des valeurs propres négatives. Nous savons pourquoi maintenant (voir l'analyse en facteurs principaux, section 8). L'information la plus intéressante est que le « saut » entre la seconde et la troisième valeur propre est vraiment marquée pour Harris.

[14] Le code sous SAS pour lancer l'analyse est :

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

Sans aucun doute possible, une modélisation en 2 facteurs est la plus appropriée. Nous disposons alors de 95.52% de l'information disponible [(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. Ce tableau sert toujours à l'interprétation. La formule est légèrement modifiée cependant, nous devons tenir compte des « uniqueness » u_i des 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))
```

Les deux blocs de variables se démarquent d'avantage avec la méthode de Harris.

(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

L'association des variables (cost, …, taste) aux facteurs est plus nette sans qu'il soit nécessaire de procéder à une rotation quelconque (*nous reparlerons de la rotation des axes dans la section* 4).

Variance non pondérée. SAS fournit une autre information, la somme des carrés des loadings par facteur. Il l'interprète comme la variance expliquée non pondérée qui leur est associée. Nous avons **2.81752** et **3.09661** pour respectivement le premier et le second axe. Elle est différente de la variance pondérée qui correspond à la valeur propre.

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

Le code pour les obtenir sous R est simplissime.

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

Communalités. Pour obtenir les communalités restituées par les axes, nous calculons la somme des carré des loadings, mais par variable cette fois-ci.

```
#communalités
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))
```

L'intérêt toujours est de pouvoir les confronter avec les communalités initiales.

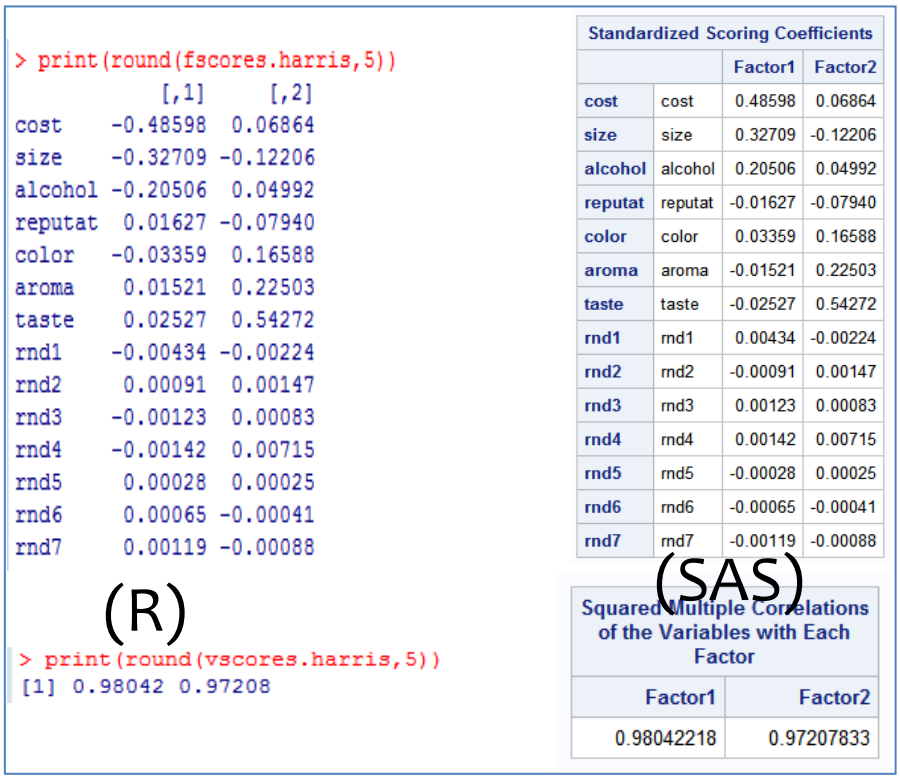
```
> print(round(cbind(init.comm,comm.harris),5))
        init.comm comm.harris
cost      0.96105     0.94399
size      0.94389     0.93907
alcohol   0.91234     0.86767
reputat   0.77232     0.45497
color     0.85328     0.82313
aroma     0.88680     0.84004
taste     0.95027     0.93847
rnd1      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. La démarche pour obtenir les coefficients de fonctions de projection sur les facteurs et les variances de ces derniers est identique à celle de l'AFP.

```
#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 et SAS fournissent les mêmes valeurs.



```
> 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
```

(R)

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

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

(SAS)

Squared Multiple Correlations of the Variables with Each Factor

Factor1	Factor2
0.98042218	0.97207833

Contribution des variables aux facteurs. Nous calculons les contributions des variables aux facteurs à partir du tableau des « factor scores ». L'impact des variables « rnd » sur les deux premiers facteurs est quasi nul ! Elles sont complètement évacuées. C'est un peux ce que l'on cherchait à obtenir depuis le début de ce tutoriel.

Standardized Scoring Coefficients

	Factor1	Factor2
cost	0.48598	0.06864
size	0.32709	-0.12206
alcohol	0.20506	0.04992
reputat	-0.01627	-0.0794
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

Squared Coefficients

Factor1	Factor2
0.23618	0.00471
0.10699	0.01490
0.04205	0.00249
0.00026	0.00630
0.00113	0.02752
0.00023	0.05064
0.00064	0.29454
0.00002	0.00001
0.00000	0.00000
0.00000	0.00000
0.00000	0.00005
0.00000	0.00000
0.00000	0.00000
0.00000	0.00000

Contributions

Factor1	Factor2
0.60948	0.01174
0.27610	0.03714
0.10851	0.00621
0.00068	0.01572
0.00291	0.06859
0.00060	0.12623
0.00165	0.73422
0.00005	0.00001
0.00000	0.00001
0.00000	0.00000
0.00001	0.00013
0.00000	0.00000
0.00000	0.00000
0.00000	0.00000

Total	0.38750	0.40117	**CTR(rnd)**	**0.01%**	**0.01%**

3.5 Bilan

Les tableaux de loadings et de contributions sont les deux éléments comparaison des méthodes dans notre étude. Ils nous permettent de juger leur capacité à identifier les groupes, en associant au mieux les variables « utiles » aux premiers axes, et en évacuant les variables « bruitées ». Pour ce qui nous concerne, force est de constater que la méthode Harris semble la plus performante. Elle identifie les blocs, avec des valeurs plus tranchées (en valeur absolue) des loadings. Les groupes de variables sont d'emblée bien identifiés.

Factor Pattern - PCA

	Factor1	Factor2
cost	0.49678	0.81407
size	0.21378	0.94733
alcohol	0.58837	0.7616
reputat	-0.73682	0.11434
color	0.90757	-0.18174
aroma	0.78387	-0.49557
taste	0.80783	-0.49864
rnd1	0.01831	0.30272
rnd2	0.04235	-0.08543
rnd3	0.11864	0.04597
rnd4	0.30514	-0.08602
rnd5	0.01361	-0.00533
rnd6	-0.04716	-0.01364
rnd7	-0.05046	-0.07406

Factor Pattern - PFA

	Factor1	Factor2
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	0.0293	-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 - Harris

	Factor1	Factor2
cost	0.96686	0.09576
size	0.93749	-0.2453
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	0.1909	-0.069
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

Figure 10 - Comparaison des méthodes - Les "loadings"

Néanmoins, nous le verrons plus loin dans ce document, les approches fournissent des résultats très similaires après rotation des axes.

4 Analyse sous Tanagra

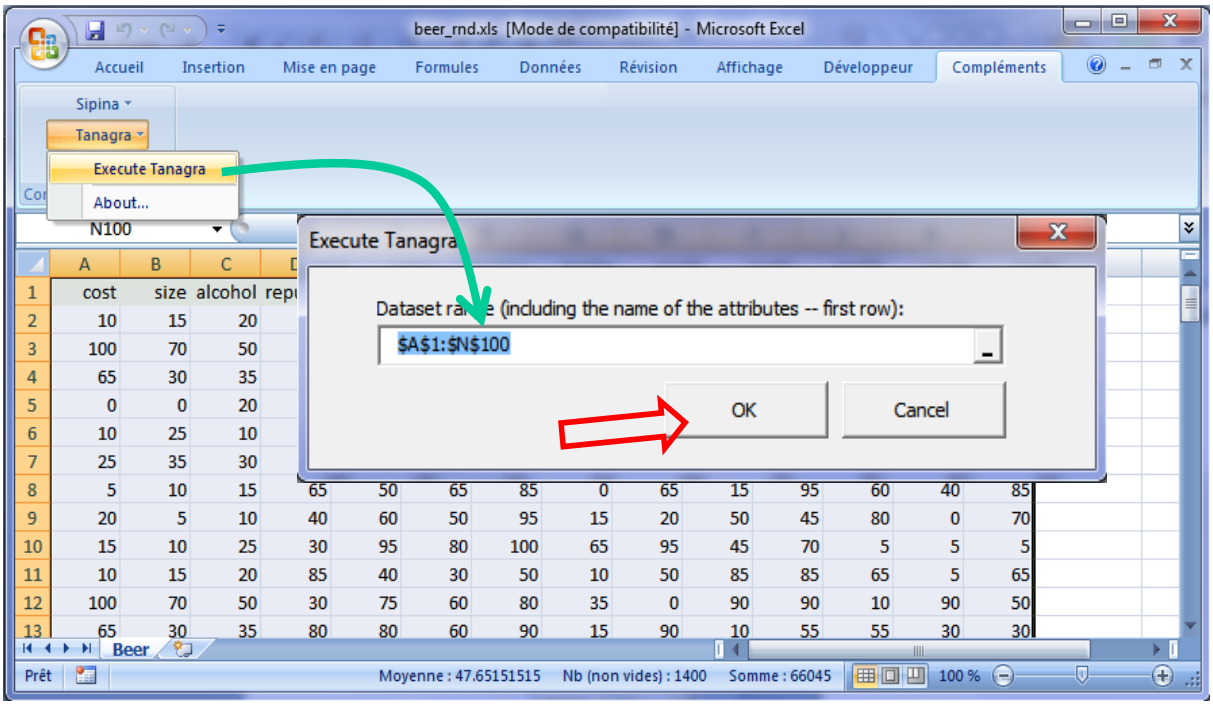
L'analyse de facteurs principaux et l'analyse de Harris ont été implémentées dans la version 1.4.47 de Tanagra. Dans cette section, nous décrivons leur mise en œuvre, toujours sur le fichier « beer_rnd ». Les résultats sont strictement identiques à ceux de R et SAS. Tanagra se distingue surtout par une mise en forme des sorties qui cherche à attirer l'œil de l'analyste sur les résultats les plus marquants. Nous utiliserons également la rotation des facteurs (VARIMAX[15]) dans cette section.

4.1 Importation des données

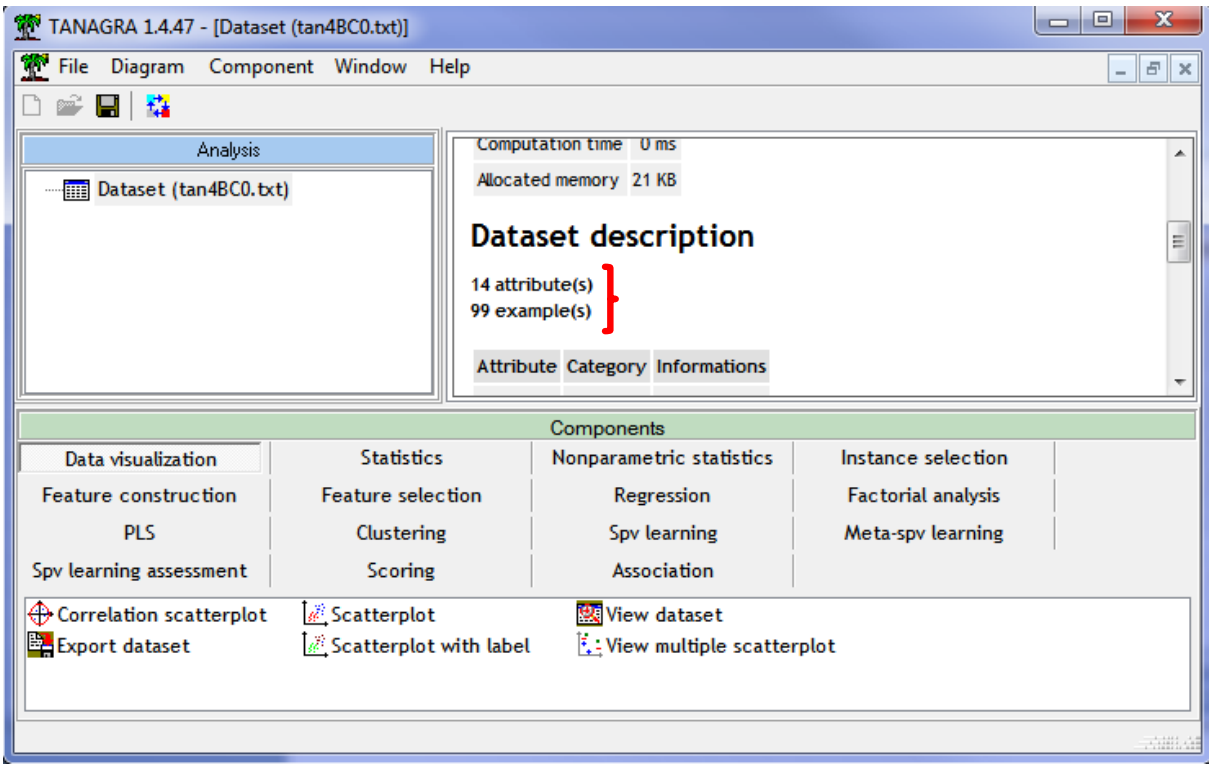
Nous utilisons la macro complémentaire « tanagra.xla » pour envoyer le fichier « beer_rnd.xls » d'Excel vers Tanagra[16].

[15] « Rotation VARIMAX en ACP » - http://tutoriels-data-mining.blogspot.fr/2008/04/rotation-varimax-en-acp.html

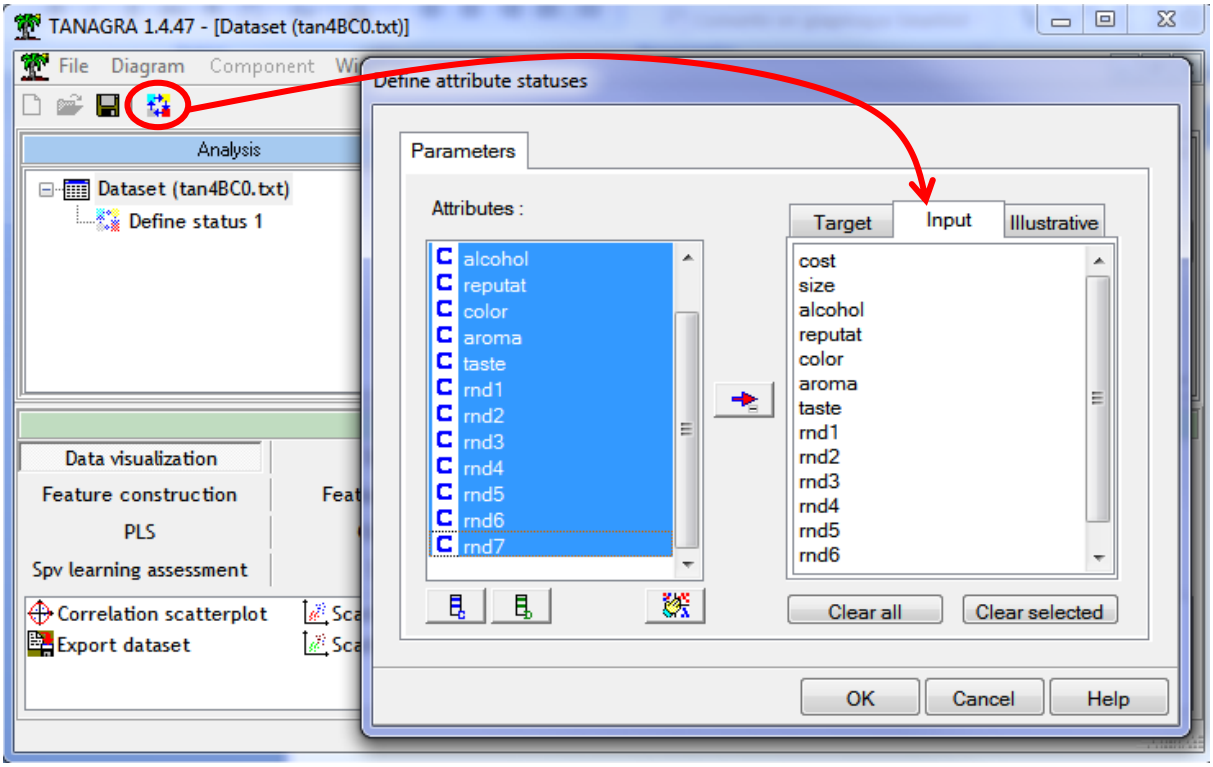
[16] Voir « L'add-in Tanagra pour Excel 2007 et 2010 » - http://tutoriels-data-mining.blogspot.fr/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html



Tanagra est démarré, les données chargées. Nous disposons de n = 99 observations décrites par p = 14 variables.



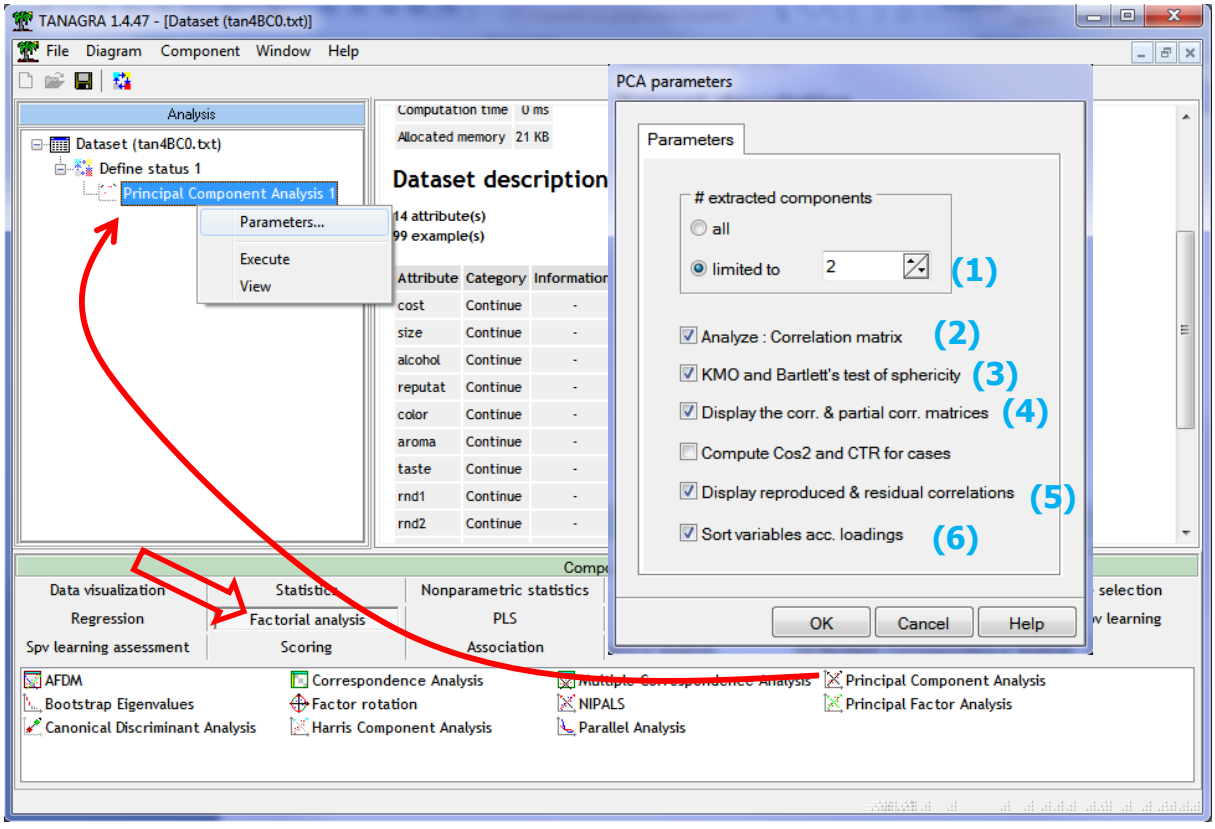
Pour initier une analyse, nous devons désigner les variables actives. Nous utilisons le composant DEFINE STATUS. Nous plaçons toutes les variables en INPUT.



4.2 Analyse en composantes principales et rotation varimax

4.2.1 Analyse en composantes principales

Nous ajoutons le composant PRINCIPAL COMPONENT ANALYSIS (onglet Factorial Analysis) pour exécuter une ACP. Nous actionnons au menu contextuel PARAMETERS pour accéder aux paramètres.

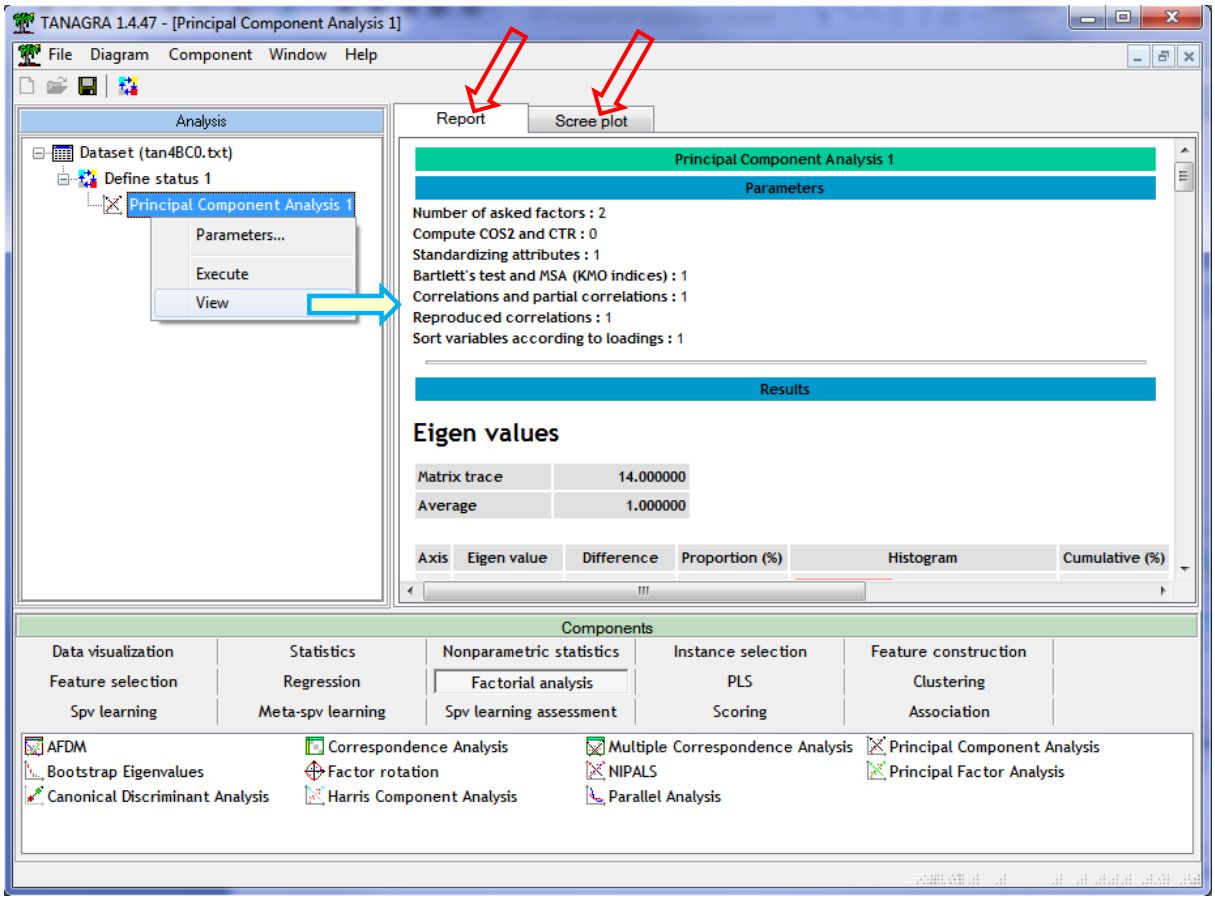


Voici la liste des options sélectionnées pour cette étude :

1. Nous choisissons 2 facteurs.
2. Il s'agit d'une ACP normée, travaillant à partir de la matrice des corrélations.
3. La statistique MSA de Kaiser-Mayer-Olkin et le test de Bartlett seront calculées.
4. Les matrices des corrélations brutes et partielles seront affichées.
5. Les matrices des corrélations reproduites (par l'ACP) et résiduelles seront affichées.
6. Les tableaux de résultats sont triés selon les valeurs des loadings sur les axes, permettant de mieux appréhender les relations entre les variables.

Nous validons en cliquant sur OK.

Nous accédons aux résultats en cliquant sur le menu contextuel VIEW.



Voyons-en le détail.

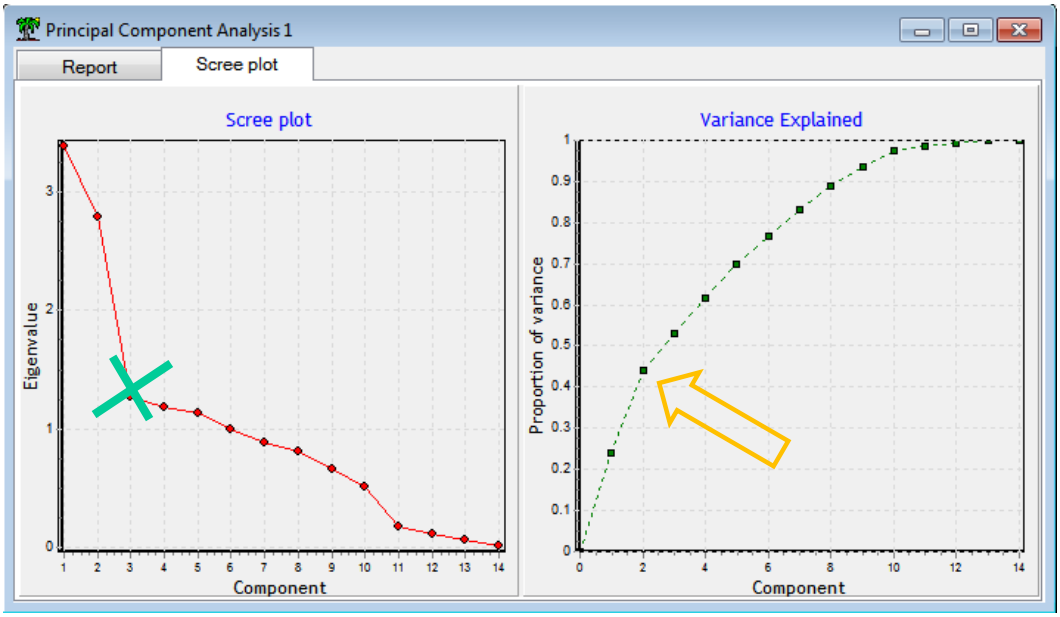
Valeurs propres. Le tableau des valeurs propres indique la proportion de variance reproduite sur les facteurs, individuellement (Proportion) et cumulativement (Cumulative).

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. L'éboulis des valeurs propres et la courbe de la proportion de variance cumulée aident au choix du nombre de facteurs. En ce qui nous concerne, nous savons déjà que deux suffisent. A posteriori, on se rend compte donc qu'il ne fallait pas inclure le « coude » du scree plot dans la sélection, la « cassure » dans la courbe de la variance expliquée nous aurait éclairé sur ce point.



Autres indicateurs pour le choix des axes. Tanagra intègre une panoplie d'outils pour la détection du bon nombre d'axes. Clairement la règle de Kaiser est inadaptée (valeur propre supérieure à 1). Il nous conduirait à retenir 5 voire 6 axes. Le test de Karlis-Saporta-Spinaki (A) est meilleur, il prône bien une décomposition en deux facteurs, tout comme le test des bâtons brisés de Legendre[17] (B).

[17] Voir « ACP avec Tanagra – Nouveaux outils », « ACP avec R – Détection du nombre d'axes » et « ACP sous R – Indice KMO et test de Bartlett » pour une description détaillée de ces outils.

Significance of Principal Components

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

(A)

Eigenvalue table - Test for significance

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

(B)

Test de Bartlett. Il vise à tester l'existence d'au moins un facteur. Sa pertinence est très discutable. Il est quasiment tout le temps significatif. C'est le cas ici.

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

Indice KMO. Il indique le degré de redondance entre les variables, autorisant ou non une factorisation efficace. Notre indice global est 0.491, ce qui ne semble pas très bon[18]. Ce critère est à prendre avec beaucoup de prudence. J'en parle uniquement parce qu'il est présent dans certains logiciels qui ont pignon sur rue.

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

Factor loadings. Vient par la suite le tableau des loadings. Les variables sont triées selon la valeur absolue de la corrélation avec les axes successifs c.-à-d. elles sont triées sur le premier facteur pour

[18] Voir http://en.wikiversity.org/wiki/Exploratory_factor_analysis

les variables présentant une corrélation supérieure à 0.5 en valeur absolue ; le reste des variables sont triées sur le second axe, avec la même contrainte ; etc. L'objectif est de mettre en évidence les groupes. Ici, nous constatons que (color, taste, aroma, reputat) sont associées au premier facteur ; (alcohol, size, cost) au second[19].

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. Les coefficients des fonctions de projections sont fournis de manière à ce que la variance des coordonnées factorielles soit égale à la valeur propre de l'axe dans Tanagra (conformément à la tradition des références francophones). Attention, ce tableau pouvant servir au déploiement, les variables conservent l'ordonnancement de la base de données d'origine.

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

[19] Les corrélations supérieures à 0.7 en valeur absolue sont surlignées en rouge foncé, en rouge clair lorsqu'elles sont supérieures à 0.5.

Matrice des corrélations. Les variables sont ordonnées selon les loadings (Tableau « Factor Loadings »). Les blocs se démarquent mieux. Cette option est surtout utile lorsque nous traitons une base avec un nombre élevé de variables.

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

Matrice des corrélations partielles. Il mesure l'association entre chaque couple de variables en annihilant l'impact des (p-2) autres. Par exemple, la relation entre les préférences pour « color » et « taste » semble forte (r = 0.80487). Lorsque l'on tient compte des autres variables, elle est moins marquée (r partiel = 0.26931). Elle est donc avant tout déterminée par une ou des tierces variables. On peut penser que « aroma » joue un rôle ici puisque ces trois variables forment un bloc.

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

Corrélation originales, reconstituées et résiduelles. Ce tableau montre la capacité de l'analyse factorielle à reproduire les relations entre les variables, à nombre de facteurs fixé. 3 valeurs sont proposées : la corrélation initialement mesurée, la corrélation reproduite par l'ACP, le résidu.

Tanagra met en évidence les couples de variables présentant une corrélation supérieure à 0.5 en valeur absolue, avec un résidu inférieur à 0.05 en valeur absolue. Dans notre exemple, nous constatons que les corrélations dans les blocs ont été correctement reconstituées par l'ACP.

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)	-

La corrélation reproduite entre deux variables est calculée à partir du tableau des loadings. Nous détaillons ci-dessous les calculs pour le couple « color » et « 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

La corrélation brute est de **0.82324**. A partir du tableau des loadings, nous calculons :

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

Nous en déduisons la corrélation résiduelle :

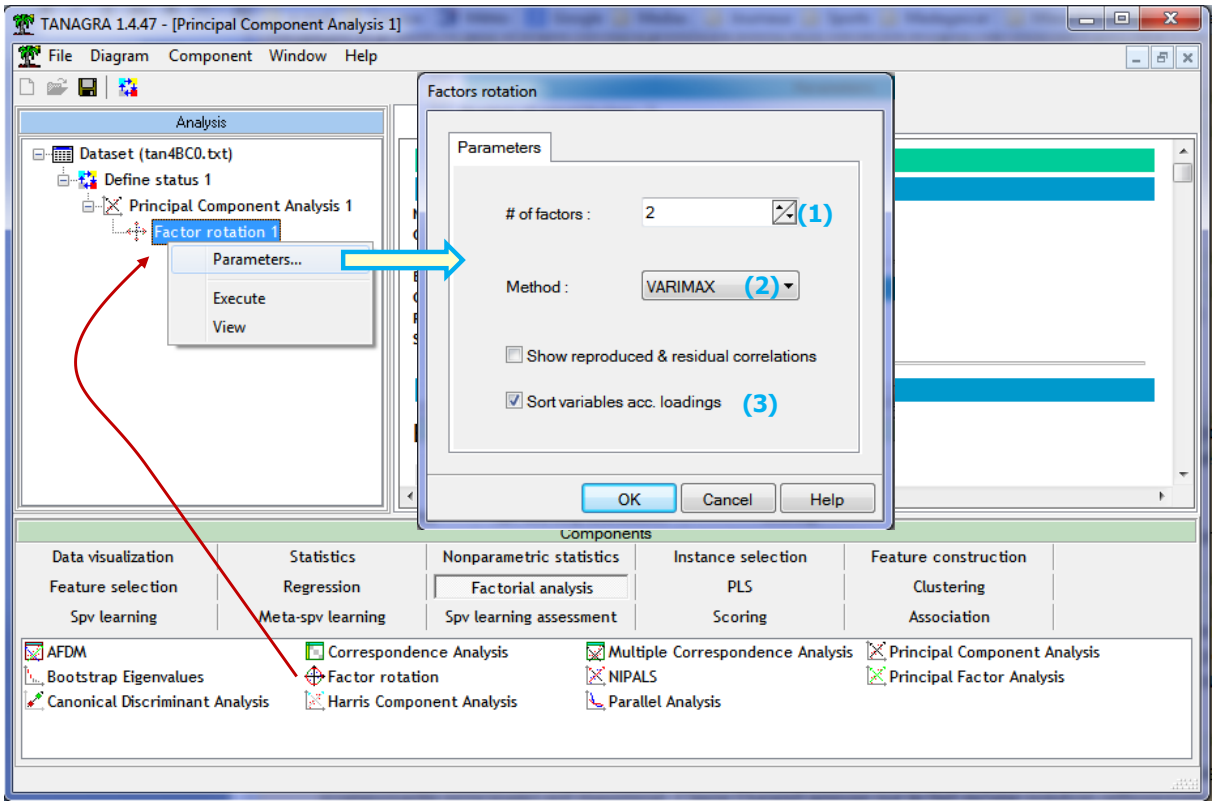
Cor. Résiduelle (color, aroma) = 0.82324 – 0.80148 = **0.02176**

Remarque : Si nous incluons tous les facteurs (14) dans l'analyse, les corrélations reproduites seraient identiques aux corrélations brutes, les corrélations résiduelles seraient par conséquent nulles.

4.2.2 Rotation varimax sur 2 facteurs

La rotation VARIMAX cherche à faire pivoter les axes de manière à maximiser la corrélation de chaque variable avec un des facteurs. L'objectif est de faciliter l'interprétation. La propriété d'orthogonalité des axes est conservée.

Nous introduisons le composant FACTOR ROTATION (onglet FACTORIAL ANALYSIS) dans notre diagramme. Nous le paramétrons comme suit : (1) l'optimisation porte sur deux facteurs ; (2) la méthode VARIMAX est utilisée ; (3) les variables sont triées selon les loadings.



Nous validons, puis nous cliquons sur VIEW.

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

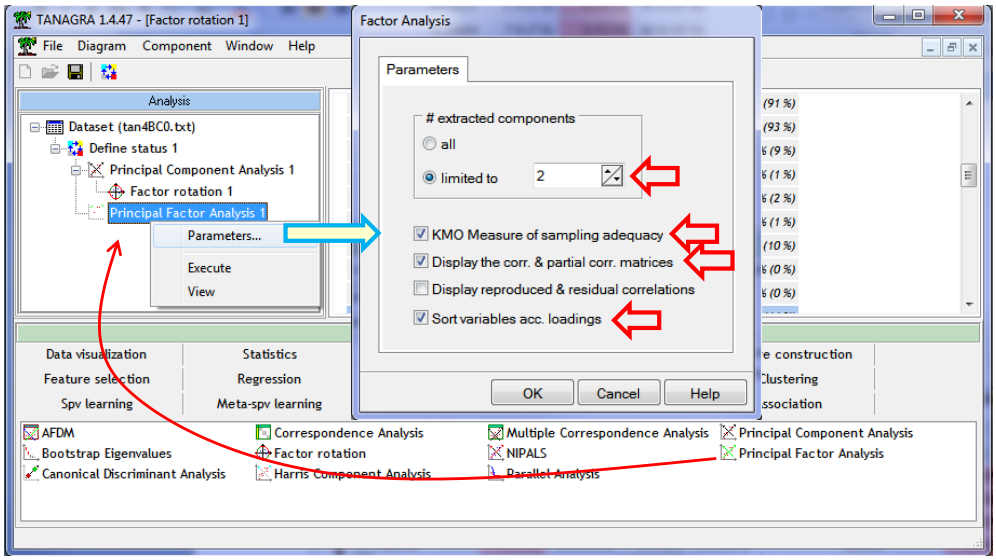
Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
taste	-0.80783	65 % (65 %)	-0.49864	25 % (90 %)
aroma	-0.78387	61 % (61 %)	-0.49557	25 % (86 %)
color	-0.90757	82 % (82 %)	-0.18174	3 % (86 %)
reputat	0.73682	54 % (54 %)	0.11434	1 % (56 %)
size	-0.21378	5 % (5 %)	0.94733	90 % (94 %)
cost	-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 affiche les nouveaux « loadings ». Il met en contrepoint les anciennes valeurs pour que l'on apprécie l'importance de la correction. De même, nous avons les variances rapportées par les axes après (resp. avant) la rotation : 3.30199 sur le 1er axe, 2.87923 sur le 2nd (resp. 3.38656 et 2.79466).

Alors que l'ACP semblait être en retrait par rapport aux autres approches (Figure 10), nous retrouvons après rotation la qualité des résultats – en termes de « loadings » - de l'analyse factorielle de Harris.

4.3 Analyse en facteurs principaux et rotation varimax

Analyse en facteurs principaux. Il nous faut dans un premier temps introduire le composant PRINCIPAL FACTOR ANALYSIS (onglet FACTORIAL ANALYSIS) dans le diagramme, puis le paramétrer.



Nous validons puis nous cliquons sur VIEW pour obtenir les résultats.

Par rapport à l'ACP, quelques particularités sont à signaler. Dans le tableau des « loadings », Tanagra affiche les communalités initiales et estimées.

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 %)*

Les variances des facteurs sont associées aux coefficients de projection (Factor Scores).

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

Rotation VARIMAX. La rotation VARIMAX joue le même rôle pour l'AFP. Nous affichons les résultats fournis par Tanagra (*le tri des variables a été désactivé pour faciliter les comparaisons*) et SAS[20].

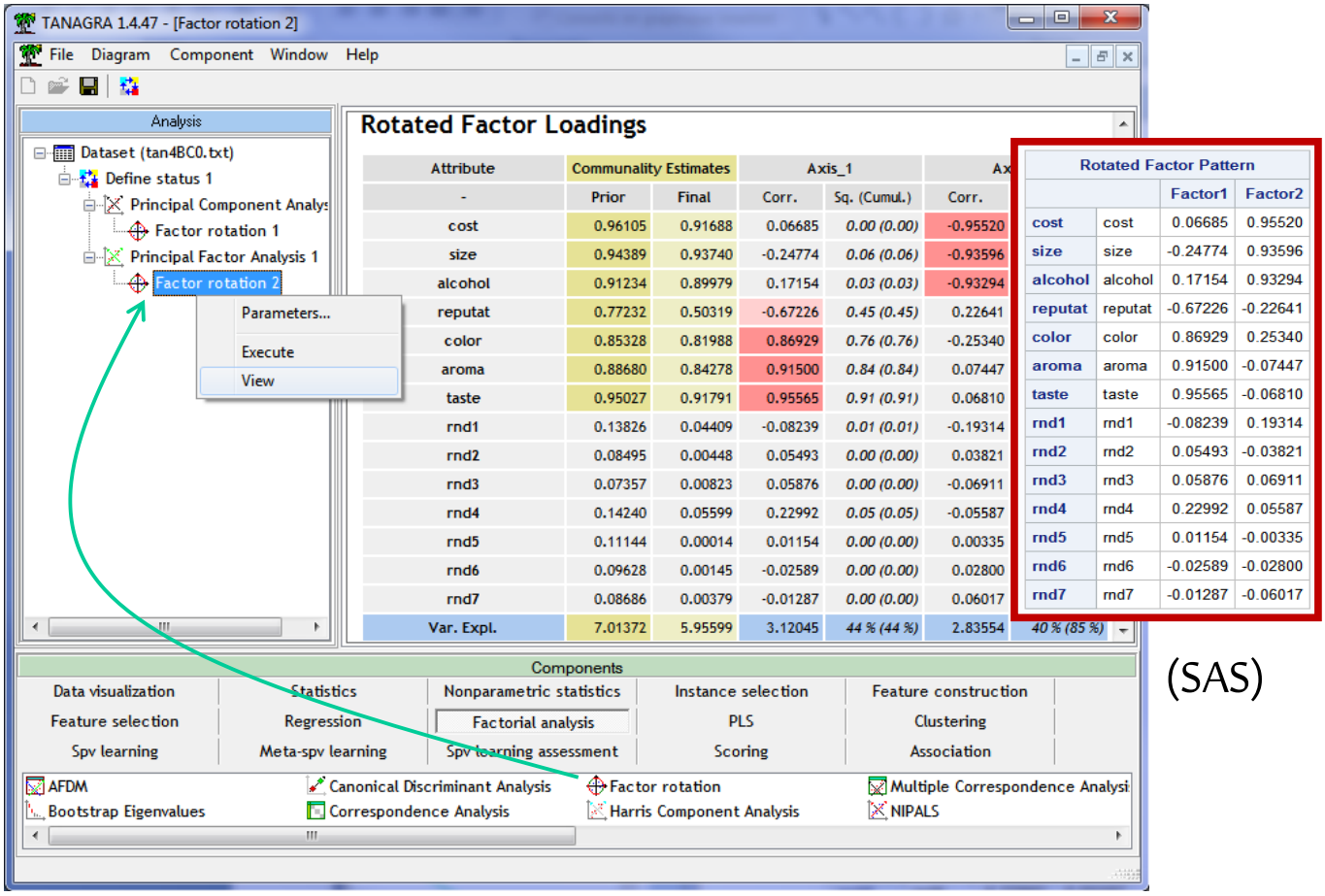
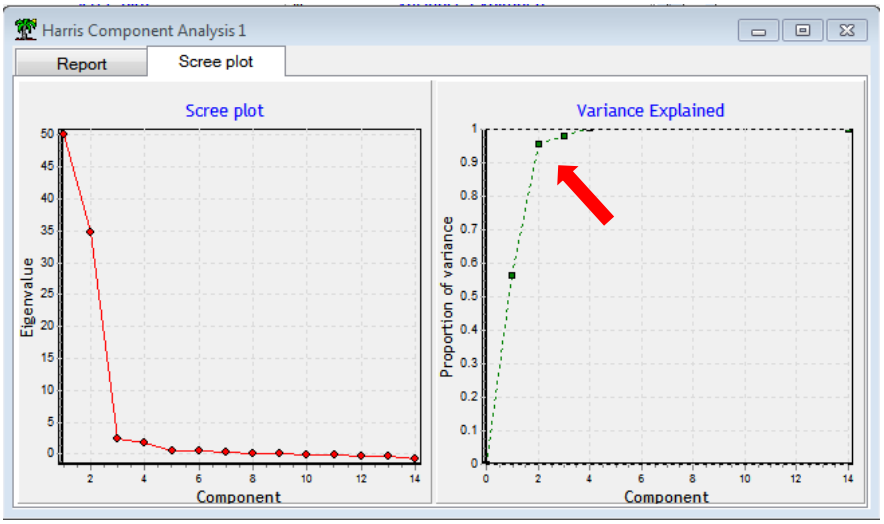


Figure 11 - "Loadings" après rotation varimax - Analyse en facteurs principaux

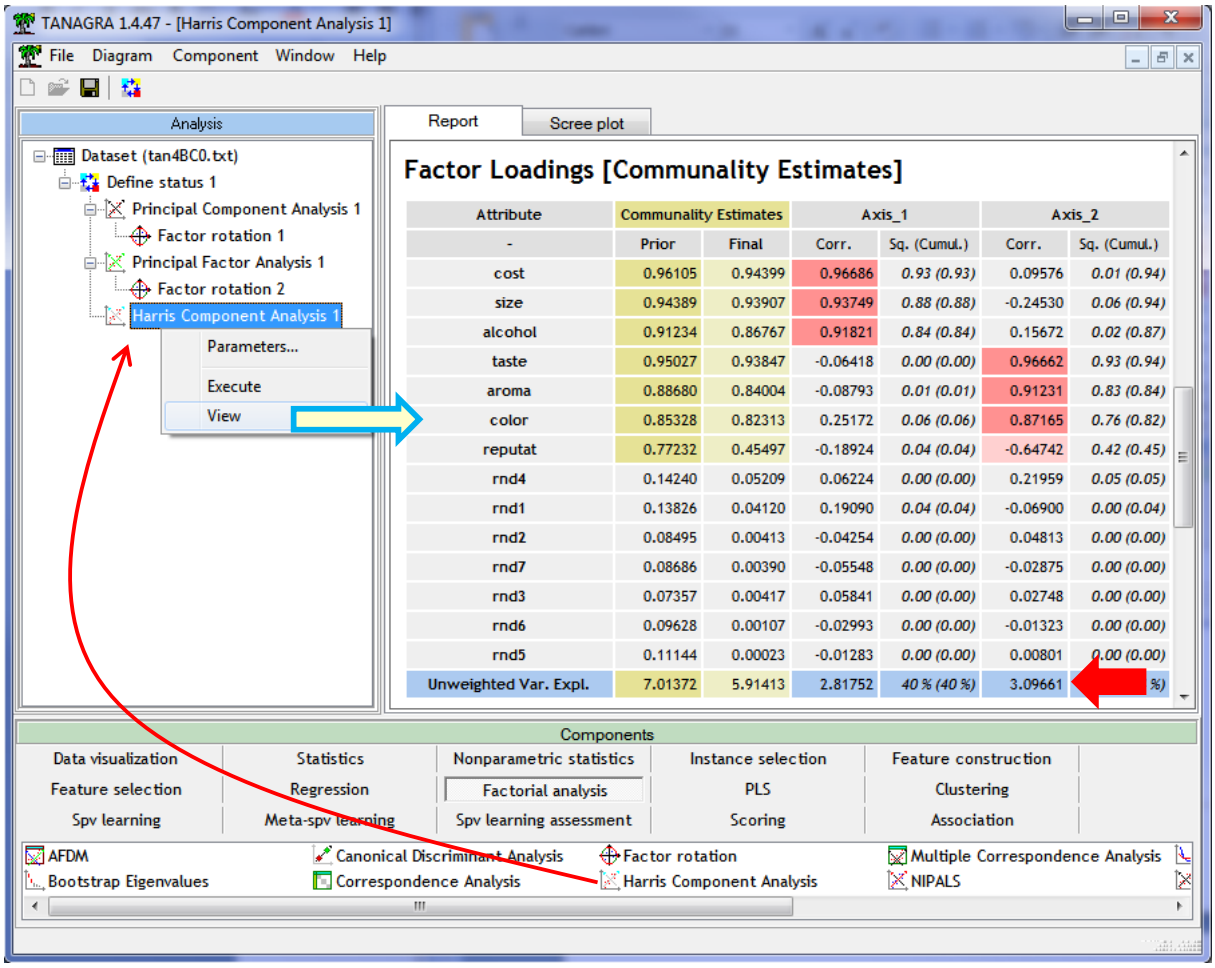
[20] `proc factor data = mesdata.beer_rnd method=principal priors=smc nfactors=2 rotate=varimax;`
`run;`

4.4 Analyse de Harris et rotation varimax

Analyse de Harris. La démarche étant maintenant bien maîtrisée, nous serons moins diserts concernant l'analyse de Harris. Nous ajoutons le composant HARRIS COMPONENT ANALYSIS (onglet Factorial Analysis) dans le diagramme, 2 facteurs sont demandés. Le scree plot et surtout la courbe des variances expliquées cumulées ne laissent aucun doute sur le nombre de facteurs à sélectionner.



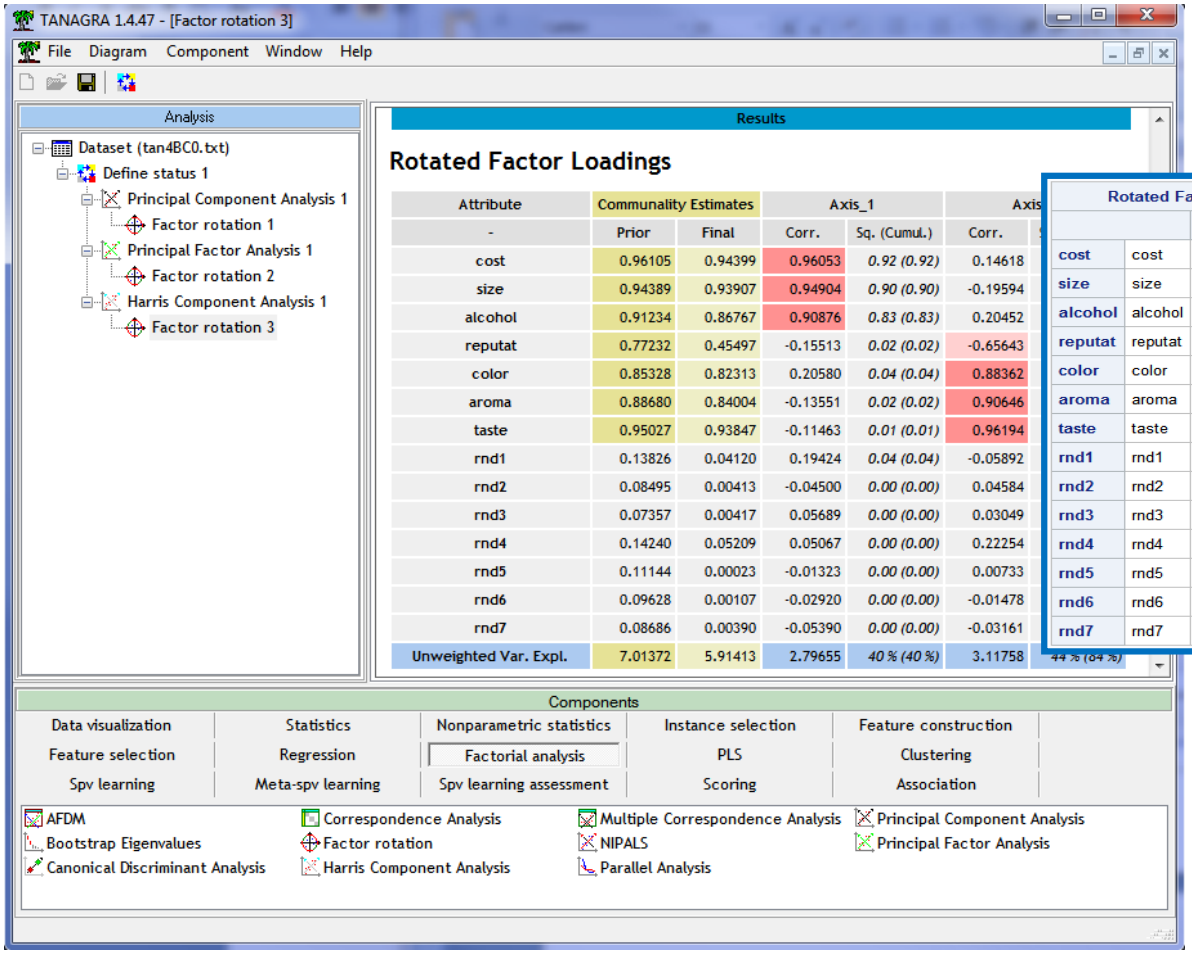
Petite particularité, Tanagra intègre les variances expliquées non pondérées au bas de chaque facteur dans le tableau des « loadings ».



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	*%)*

Rotation varimax. Les associations variables – facteurs étant déjà fortement marquées, la rotation VARIMAX n'amène pas vraiment de valeur ajoutée ici.

Voici les résultats pour Tanagra et SAS[21], notons que ce dernier trie les facteurs en fonction de la variance non pondérée. Ainsi, le 1er facteur de Tanagra correspond au 2nd de SAS et inversement.



4.5 Bilan – Comparaisons après rotation

L'approche de Harris semblait être la plus intéressante précédemment (Figure 10). Après rotation, nous constatons que les trois approches fournissent des résultats de qualité très similaire.

Ainsi, malgré son handicap théorique qui aurait pu le pénaliser dans le contexte dans lequel nous l'avons placé, l'**ACP** se comporte très bien. C'est ce qui explique en grande partie sa très forte popularité**. Bien comprise, bien paramétrée,** elle **est aussi efficace que les autres, même lorsqu'il s'agit d'étudier les relations entre les variables, y compris en présence de variables bruitées**. Le paramétrage consiste surtout à choisir le nombre adéquat de facteurs. A la lecture du tableau des

[21]
```
proc factor data = mesdata.beer_rnd
method=harris
msa
nfactors=2
score
rotate=varimax;
run;
```

valeurs propres (Figure 2), choisir 3 axes n'aurait eu rien de choquant… et nous aurait envoyé sur une fausse piste. Une décomposition en 2 facteurs est manifestement la bonne solution.

Nous retraçons dans le tableau suivant les « loadings » des variables pour chaque méthode, après rotation des axes avec la méthode VARIMAX.

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" des analyses après rotation VARIMAX

5 Analyse avec le package PSYCH de R

« Il y a sûrement un package sous R qui réalise directement ce truc là », voilà une phrase que j'entends souvent. Vive « Google » ! Après quelques recherches, j'ai effectivement (re[22])-découvert le package PSYCH[23]. Finalement, le plus fastidieux aura été de lire attentivement la documentation pour (1) comprendre le type d'analyse menée par les fonctions proposées et (2) faire la correspondance avec nos résultats.

5.1 Analyse en composantes principales

Plusieurs procédures permettent d'instancier l'ACP dans R (entres autres les fonctions **princomp()**[24] et **prcomp()**, accessibles via le le package STAT qui est chargé par défaut). Pour harmoniser la présentation, nous choisissons la procédure **principal()** de PSYCH.

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

[22] Nous l'avions déjà utilisé… http://tutoriels-data-mining.blogspot.fr/2012/05/acp-sous-r-indice-kmo-et-test-de.html

[23] http://cran.r-project.org/web/packages/psych/index.html

[24] Maintes fois présentée dans nos tutoriels, ex. http://tutoriels-data-mining.blogspot.fr/2009/05/analyse-en-composantes-principales-avec.html

```
print(pca.unrotated$loadings[,])
```

En accord avec Tanagra (*attention, le tableau est trié selon les « loadings » dans Tanagra*) et SAS, nous obtenons :

```
> 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
```

Avec la rotation VARIMAX

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

Idem, nous sommes en accord avec Tanagra et SAS (Figure 12):

```
> 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 Analyse en facteurs principaux

La procédure **fa()** semble convenir. Par défaut, elle réalise l'approche itérative. Mais nous pouvons la paramétrer (demander une seule itération, « **max.iter = 1** ») pour implémenter la méthode non itérative et ainsi retrouver nos résultats.

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

Voyons ce qu'il en est au niveau des « loadings » :

```
> 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
```

Nous modifions l'option « rotate » pour introduire la rotation VARIMAX.

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

Ici également, nous retrouvons les valeurs fournies par Tanagra et SAS (Figure 12) :

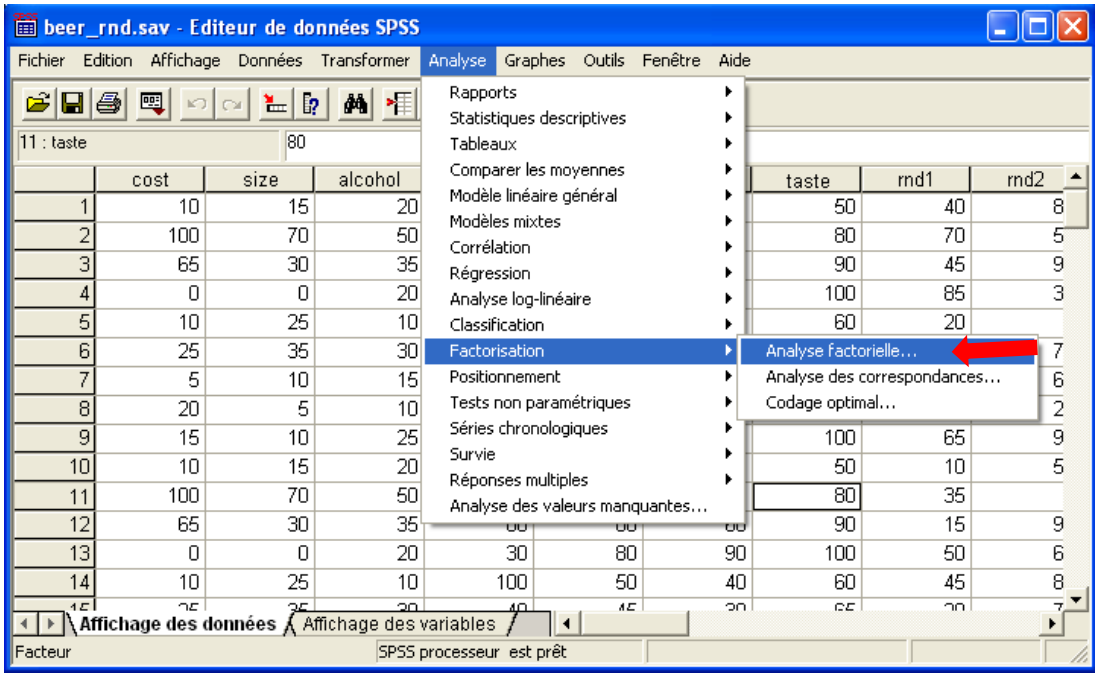
```
> print(pfa.varimax$loadings[,])
                PA1         PA2
cost      0.06686663  0.95520124
size     -0.24772440  0.93596492
alcohol   0.17154705  0.93293433
reputat -0.67226251 -0.22640086
color     0.86929298  0.25338749
aroma     0.91500223 -0.07448421
taste     0.95564970 -0.06811382
rnd1     -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 Analyse de Harris

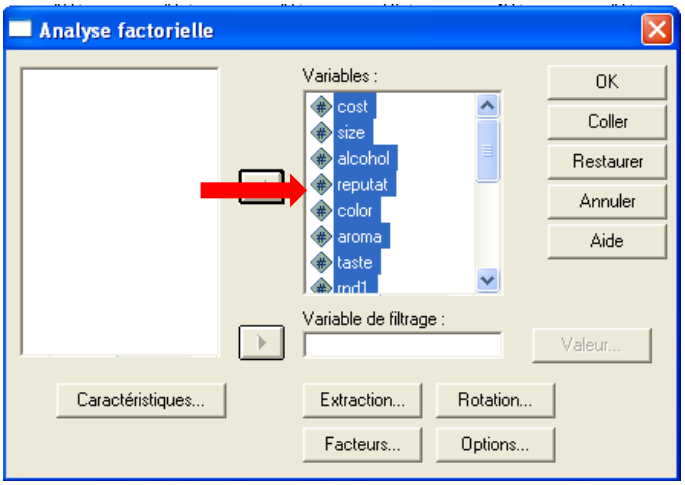
J'ai bien cherché, je n'ai pas trouvé une procédure qui implémente directement l'approche de Harris. Qu'à cela ne tienne, nous avons toujours la possibilité de la programmer comme nous l'avions fait dans la section 3.4. Il faudait la mettre sous forme de fonction pour la rendre plus générique c.-à-d. applicable sur un data.frame quelconque. Le fin du fin serait alors de l'intégrer dans un package que l'on met à la disposition des autres utilisateurs. Cela n'a rien d'insurmontable. Les férus de R pourront le faire très aisément.

6 Analyse en facteurs principaux avec SPSS

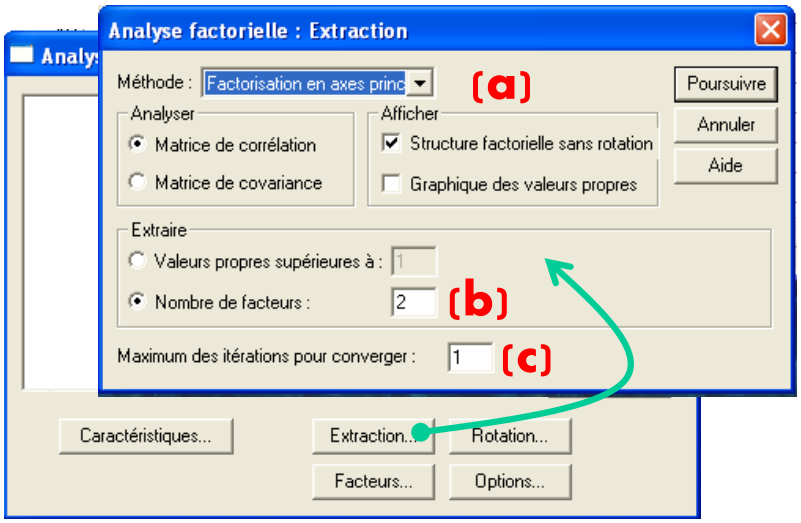
Nous utilisons la version française de SPSS 12.0.1 dans cette section. Le fichier de données a été importé, nous sommes prêts à lancer l'analyse. Nous actionnons le menu ANALYSE / FACTORISATION / ANALYSE FACTORIELLE, la boîte de dialogue de paramétrage apparaît.



Nous sélectionnons les variables de l'étude.

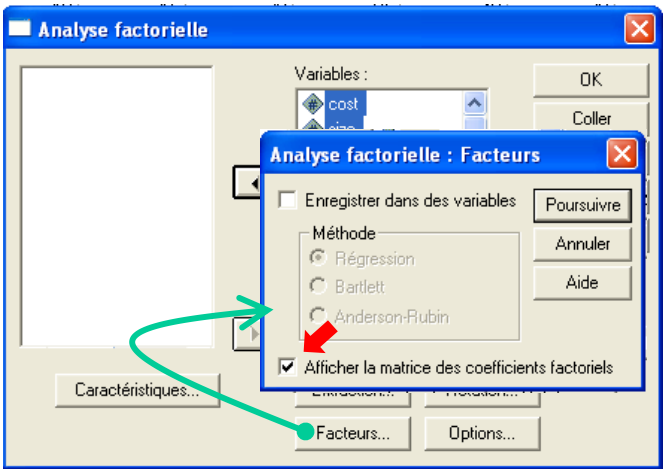


Nous choisissons la technique factorielle en cliquant sur le bouton « Extraction ».

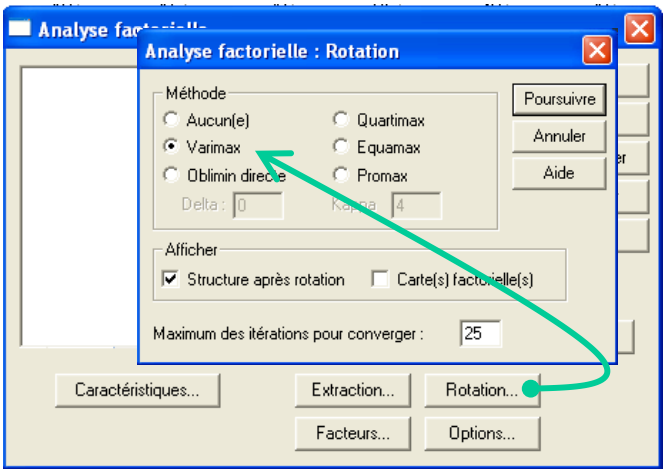


Nous demandons une analyse en facteurs principaux (a), en limitant à 2 le nombre de facteurs à extraire (b) et en limitant à 1 le nombre d'itérations (c). Ce dernier paramètre est important. Par défaut, tout comme la procédure **fa()** du package psych de R, SPSS implémente la procédure itérative. Avec une seule itération, nous retrouvons les résultats de Tanagra et SAS.

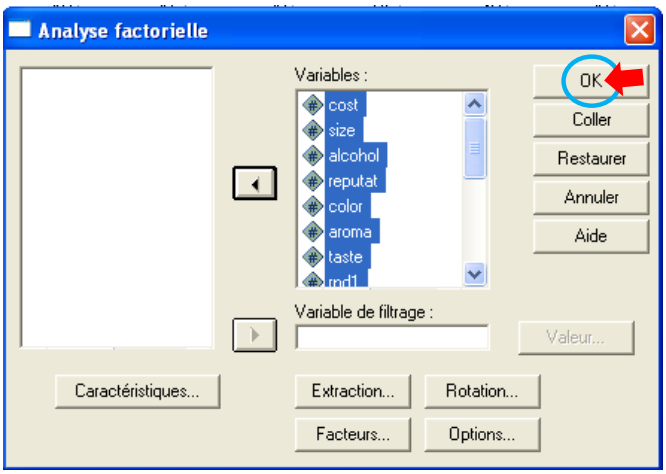
Nous cliquons maintenant sur le bouton « Facteurs ». Nous souhaitons faire afficher les coefficients des fonctions de projection (les « Factor Scores »).



Enfin, avec le bouton Rotation, nous demandons la rotation VARIMAX.



Il ne nous reste plus qu'à valider toutes ces options et lancer les opérations en cliquant sur OK.



SPSS génère un rapport qui retrace les principaux résultats de l'analyse.

Communalités initiales et estimées. La qualité de la représentation est traduite par la confrontation entre les communalités initiales et reproduites pour chaque variable (en comparaison, cf. 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) avant et après rotation. Nous avons ensuite les « loadings » avant [**a**] (cf. Figure 6) et après [**b**] rotation des axes (cf. Figure 11).

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

Coefficients des fonctions de projections (Factor Scores). SPSS fournit directement les coefficients des fonctions de projection après rotation des axes. Nous ne les avons pas montrés précédemment, nous profitons de l'occasion pour les comparer avec ceux fournis par Tanagra.

Sans surprise, nous obtenons exactement les mêmes valeurs. Il en est de même sous 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)

Variances et covariances factorielles. Comme nous l'avions dit précédemment, si la variance des facteurs théoriques est unitaire, celle des facteurs observés peut en différer légèrement (voir page 14). Plus elle s'en écarte, plus le facteur est sujet à caution. Par un processus similaire, les facteurs censés être orthogonaux peuvent être légèrement corrélés entre eux en pratique.

SPSS complète donc le tableau des « Factor Scores » par la « Matrice de covariance factorielle ».

Matrice de covariance factorielle

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

Les covariances (valeurs hors diagonale principale) entre les facteurs est très faible sur notre exemple. Les deux axes sont crédibles.

Variances et covariances factorielles - Modélisation en 5 axes sans rotation. Par curiosité, j'ai essayé une analyse en 5 facteurs sans rotation des axes pour tenter de reproduire les résultats obtenus sous SAS (Figure 8).

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.

Nous retrouvons bien les variances sur la diagonale principale de la matrice. Les axes 3, 4 et 5 sont vraiment sujets à caution avec une variance qui s'éloigne nettement de la valeur 1. Nous constatons également qu'ils sont de surcroît plus ou moins liés entre eux. Ces 3 derniers facteurs sont manifestement instables.

7 Conclusion

Je me suis intéressé aux alternatives à l'analyse en composantes principales lorsque j'ai étudié la PROC FACTOR de SAS décrite un de mes anciens tutoriels[25]. J'avoue avoir été un peu surpris. En effet, j'étais en présence d'approches que je ne connaissais pas malgré une lecture assidue de la littérature francophone dans le domaine. Je me suis rabattu sur les références en langue anglaise. J'ai alors découvert un univers un peu différent – la philosophie « multivariate analysis » n'est pas exactement la même que celle de l'analyse de données à la française – même si les finalités sont très similaires. Heureusement, de très nombreux documents sont accessibles sur web. Il suffit d'introduire les bons mots clés dans les moteurs de recherche.

Après un tour d'horizon, je me suis dit que dans un premier temps, il serait intéressant d'étudier en détail l'analyse en facteurs principaux et l'analyse de Harris (non itératives) qui reposent sur deux façons plus ou moins sophistiquées d'exploiter la matrice de corrélation. Et la meilleure manière de se forger une opinion sur les méthodes est de les programmer. Ce que j'ai fait dans R pour en tracer les grandes lignes et calibrer les calculs, puis dans Tanagra pour disposer d'un outil prêt à l'emploi, facile à mettre en œuvre.

Notons à propos de Tanagra que, tout comme l'ACP, les deux nouvelles méthodes implémentées peuvent être couplées avec le composant de rotation des axes (FACTOR ROTATION) comme nous avions pu le voir dans ce tutoriel, mais également avec les outils de détection du nombre d'axes basés sur des techniques de ré-échantillonnage (BOOTSTRAP EIGENVALUES, PARALLEL ANALYSIS[26]).

Avec un peu de recul, je ne suis pas sûr qu'elles soient décisives au point de supplanter l'ACP, très largement présent dans notre background collectif. En revanche, elles incarnent d'autres points de vue qui, dans certaines circonstances, amènent un éclairage différent sur les informations véhiculées par les données que l'on cherche à décortiquer. C'est surtout en cela qu'elles sont intéressantes.

[25] Tutoriel Tanagra - « ACP sur corrélations partielles (suite) » - http://tutoriels-data-mining.blogspot.fr/2012/06/acp-sur-correlations-partielles-suite.html

[26] Voir http://tutoriels-data-mining.blogspot.fr/2012/06/acp-avec-tanagra-nouveaux-outils.html pour la mise en œuvre de ces outils avec l'ACP, la transposition à l'analyse en facteurs principaux et l'analyse de Harris est directe.