ISFA 2° année
6 février 2002 - 2 heures

Toutes les analyses en composantes principales considérées sont ici du type ACP normée, encore appelée aussi ACP sur matrice de corrélation. Répondre dans la place impartie, brièvement, avec un argument qui vous semble déterminant.

1. Rang d'une matrice de corrélation

Soit un tableau à 3 lignes (individus) et 3 colonnes (variables) :

$$\mathbf{A} = \begin{bmatrix} \sqrt{2}/2 & -1/\sqrt{6} & 0 \\ -\sqrt{2}/2 & 2/\sqrt{6} & -\sqrt{2}/2 \\ 0 & -1/\sqrt{6} & \sqrt{2}/2 \end{bmatrix}$$

1.1. Quelle est la matrice de corrélation associée à **A** ?

1.2. Montrer que pour toute matrice **X** le rang de **X** est celui de $\mathbf{X}^t\mathbf{X}$.

1.3. Quel est le rang de la matrice de corrélation associée à **A** ?

1.4. Montrer que le rang d'une matrice de corrélation calculée entre p variables mesurées sur n individus est inférieur ou égal à n-1 si p > n-1.

2. Essais avec n=p

Un expérimentateur avisé désire se faire une opinion personnelle du comportement de l'analyse en composantes principales sur des tableaux artificiels. Il considère 3 tableaux comportant n = 16 lignes et p = 16 colonnes. Le premier est appelé **talea** car il a été généré par une procédure de tirage aléatoire.

```
> talea
   V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16
1   1  0  0  1  1  0  1  0  0   0   0   0   0   0   1   1
2   1  1  1  1  1  0  1  0  0   1   1   0   1   1   1   1
3   1  1  1  1  0  0  0  1  0   0   1   0   0   1   1   0
4   1  1  1  0  0  1  0  1  0   1   0   0   0   1   1   0
5   1  0  0  0  0  0  1  0  0   0   0   0   1   1   0   0
6   1  0  0  0  0  0  1  0  1   0   1   1   1   0   1   1
7   1  0  1  1  0  0  0  1  1   0   0   0   1   0   1   1
8   0  1  0  0  1  1  1  1  0   0   0   0   0   1   1   1
9   0  0  1  0  0  0  0  1  1   1   0   1   1   0   0   1
10  0  1  0  0  1  1  0  1  0   1   1   1   1   1   1   0
11  0  0  1  0  1  0  0  0  1   0   1   0   0   0   0   1
12  0  0  1  0  1  0  1  0  0   1   0   1   0   1   1   1
13  1  1  1  0  1  1  0  0  1   0   0   0   1   0   0   0
14  0  0  1  1  0  1  0  1  1   1   1   0   1   1   0   1
15  1  1  1  1  0  1  1  1  0   0   0   0   1   1   0   0
16  0  0  1  1  1  0  1  0  0   0   1   0   1   1   0   1
```

Le second est appelé **tgrad** car il représente une structure simple définie par un gradient.

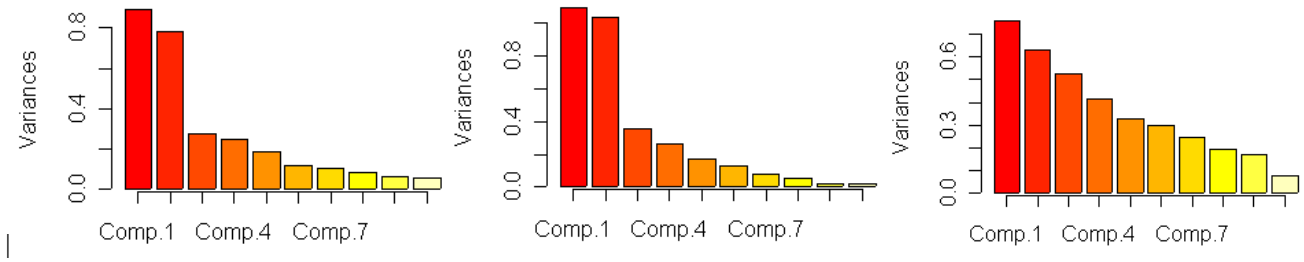
```
> tgrad
   V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16
1   1  1  0  0  0  0  0  0  0   0   0   0   0   0   0   0
2   1  1  1  0  0  0  0  0  0   0   0   0   0   0   0   0
3   1  1  1  1  0  0  0  0  0   0   0   0   0   0   0   0
4   0  1  1  1  1  1  0  0  0   0   0   0   0   0   0   0
5   0  0  1  1  1  1  1  1  1   0   0   0   0   0   0   0
6   0  0  1  1  0  1  1  1  1   0   0   0   0   0   0   0
7   0  0  0  1  1  1  1  1  1   1   0   0   0   0   0   0
8   0  0  0  0  1  1  1  1  1   1   1   1   0   0   0   0
9   0  0  1  0  0  0  1  1  1   0   0   0   0   0   0   0
10  0  0  0  0  0  0  1  1  1   1   1   1   1   0   0   0
11  0  0  0  0  0  0  1  1  1   0   1   1   0   1   0   0
12  0  0  0  0  0  0  0  0  1   1   1   1   1   1   0   0
13  0  0  0  0  0  0  0  0  1   0   1   1   0   1   0   0
14  0  0  0  0  0  0  0  0  0   0   1   1   1   1   1   0
15  0  0  0  0  0  0  0  0  0   0   0   0   1   1   1   1
16  0  0  0  0  0  0  0  0  0   0   0   0   0   1   1   1
```

Le troisième est appelé **tparti** car il représente une structure simple définie par une partition.

```
> tparti
   V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16
1   0  1  0  1  0  0  0  0  0   0   0   0   0   0   0   0
2   1  1  1  1  1  0  0  0  0   0   0   0   0   0   0   0
3   0  1  1  0  0  0  0  1  0   0   0   0   0   0   1   0
4   1  1  0  1  0  0  0  0  0   0   0   0   0   0   0   0
5   0  1  1  0  1  0  0  0  0   0   0   0   0   0   0   0
6   0  1  0  1  0  0  0  0  0   0   0   0   0   0   0   0
7   0  0  0  0  0  1  0  1  1   1   0   0   0   0   0   0
8   0  0  1  0  0  1  1  1  1   1   1   0   0   1   0   0
9   0  0  0  0  0  1  0  1  1   0   1   0   0   0   0   0
10  0  0  0  0  0  0  1  1  0   1   1   0   0   0   0   0
11  0  0  0  0  0  1  0  1  0   1   0   0   0   0   0   0
12  0  0  0  0  0  0  0  0  0   0   0   1   1   1   1   1
13  0  0  0  0  0  0  0  0  0   0   0   0   1   1   0   1
14  0  0  0  0  0  0  1  0  0   0   0   1   0   1   0   1
15  0  1  0  0  0  0  0  0  0   1   0   1   1   0   1   1
16  0  0  0  0  0  0  0  0  0   0   0   0   1   1   0   1
```

On pourra penser que les tableaux simulent des résultats obtenus par 16 élèves (lignes) sur un test contenant 16 questions (colonnes) ou la présence-absence de 16 caractères (colonnes) sur un échantillon de 16 individus (lignes).

Les trois tableaux sont envoyés dans une simple analyse en composantes principales normée dite encore sur matrice de corrélation.

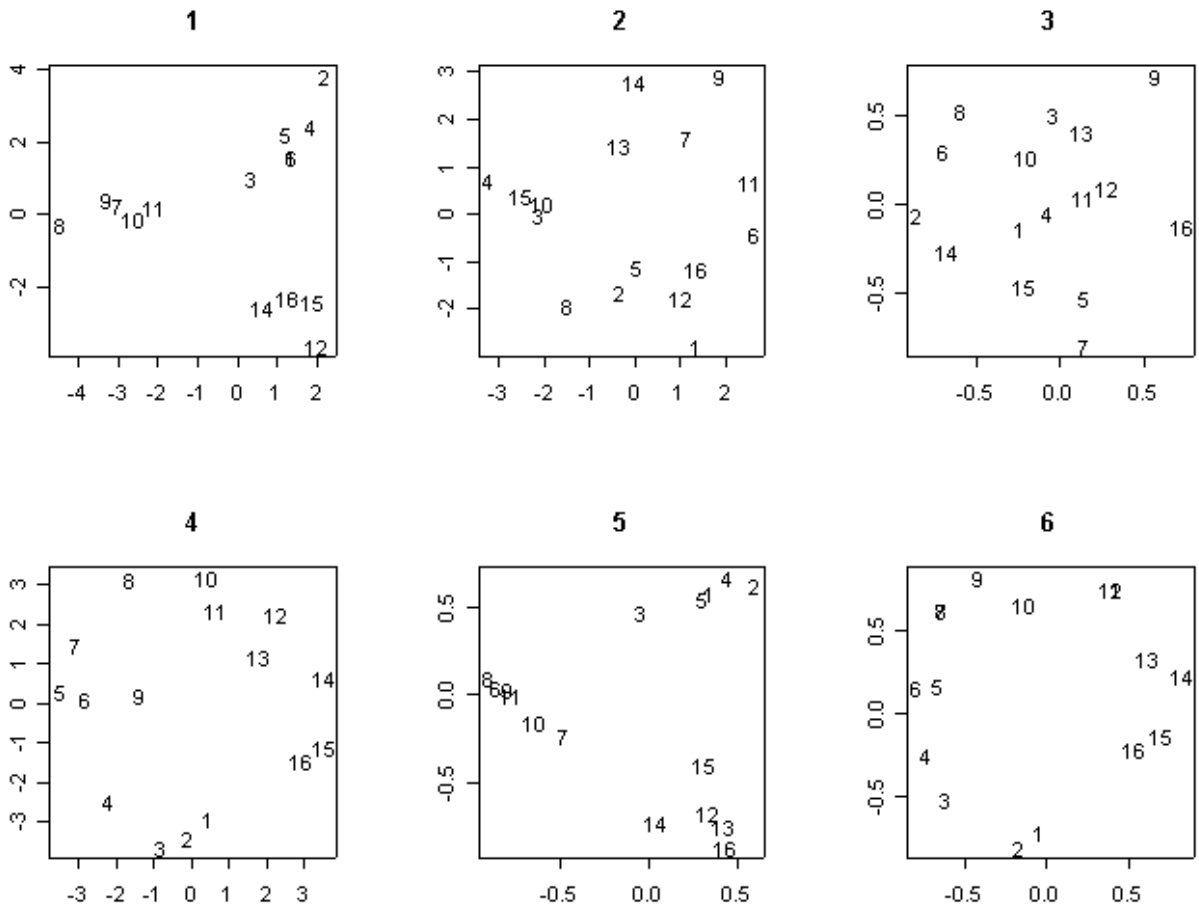


2.1. Peut-on attribuer à chaque analyse le graphe qui lui revient ?

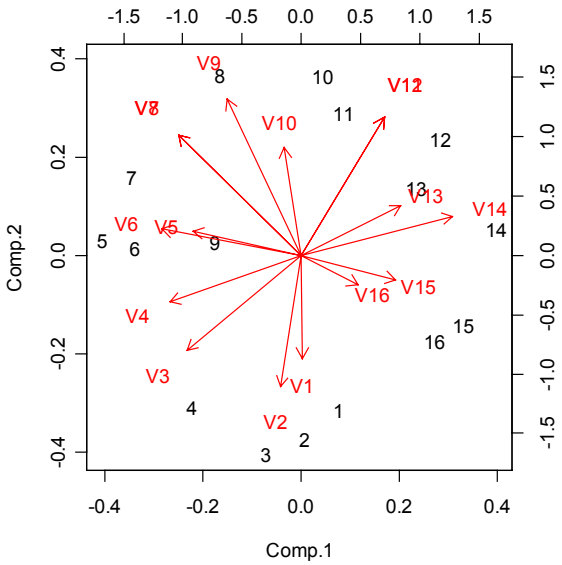
On trace les trois cartes factorielles des lignes (plan 1-2) et les trois cartes factorielles des colonnes (plan 1-2) dans le désordre (ci-après).

2.2. Quelles sont les cartes des lignes ?

2.3. Quelles sont les cartes de l'analyse de tparti ?



2.4. Comment obtient-on ce graphique ?



2.5. Laquelle de ces trois analyses invalide l'assertion "quand sont légitimement conservés deux axes, il existe deux faits marquants dans les données" ?

2.6. Commenter l'assertion "si on avait utilisé les tableaux transposés au lieu des tableaux initiaux on aurait obtenu strictement les mêmes résultats numériques".

2.7. Vrai ou faux ? "Ces trois analyses ont donné une valeur propre nulle."

3. Jury de dégustation

25 juges, notés de a à y, ont noté leur préférence pour 8 échantillons de vin notés de A à H. Le juge donne la note 1 au produit qu'il préfère et la note 8 au produit qu'il estime le moins bon. Les résultats forment un tableau macon à 8 lignes (produits) et 25 colonnes (juges) :

```
> macon
  a b c d e f g h i j k l m n o p q r s t u v w x y
A 5 5 4 3 3 4 7 2 1 3 5 4 4 5 4 8 5 7 8 5 4 6 7 2 8
B 4 8 2 4 1 5 2 7 8 8 1 6 3 7 8 5 7 8 1 4 1 5 4 4 6
C 2 6 1 1 6 2 1 5 5 4 3 7 2 2 6 2 1 6 2 1 2 1 2 5 1
D 6 7 5 8 2 6 8 8 6 6 6 5 6 6 3 6 8 1 7 6 7 4 1 6 7
E 1 4 3 2 7 1 6 4 3 1 2 8 1 1 1 3 2 2 6 2 8 2 8 1 2
F 3 2 8 6 5 8 3 3 4 7 8 1 5 8 7 4 4 3 3 8 6 8 6 7 3
G 7 1 6 5 4 7 4 1 7 5 7 3 8 3 2 7 3 5 4 7 3 7 3 8 5
H 8 3 7 7 8 3 5 6 2 2 4 2 7 4 5 1 6 4 5 3 5 3 5 3 4
```

Les dégustateurs sont de 5 catégories professionnelles. Les cinq premiers sont œnologues (a-e), les cinq suivants sont restaurateurs (f-j), les cinq suivants sont négociants (k-o), les cinq suivants sont viticulteurs (p-t) et les cinq derniers sont les organisateurs du concours (u-y) .

3.1. Quelles sont les moyennes et les variances de ce tableau ?

```
> eigen(cor(macon))$values
 [1]  8.774e+00  5.635e+00  3.741e+00  2.672e+00  1.953e+00  1.721e+00
 [7]  5.031e-01  6.413e-16  3.955e-16  3.418e-16  2.001e-16  1.584e-16
[13]  1.263e-16  6.291e-17  2.793e-17 -6.118e-17 -2.028e-16 -2.258e-16
[19] -2.641e-16 -3.113e-16 -4.317e-16 -4.638e-16 -5.280e-16 -6.871e-16
[25] -1.645e-15
```

3.2. Commenter ce résultat.

Dans un concours de ce type, est déclaré premier le produit ayant obtenu le plus grand nombre de places de premier, puis en cas d'ex æquo le plus grand nombre de places de second, puis en cas d'ex æquo le plus grand nombre de places de troisième, puis... Est déclaré second celui qui a la même propriété quand on a enlevé le premier. Est déclaré troisième celui qui a la même propriété quand on a enlevé le premier et le second …

3.3. Quel est le classement du concours ?

```
> sort(apply(macon,1,sum))
  C   E   H   A   B   G   F   D
 76  81 112 119 119 122 130 141
```

3.4. Obtient-on le même résultat en classant par la somme des rangs ?

```
> w_cor(macon)
> max(w[row(w)>col(w)])
[1] 0.9524
> min(w[row(w)>col(w)])
[1] -0.9048
```

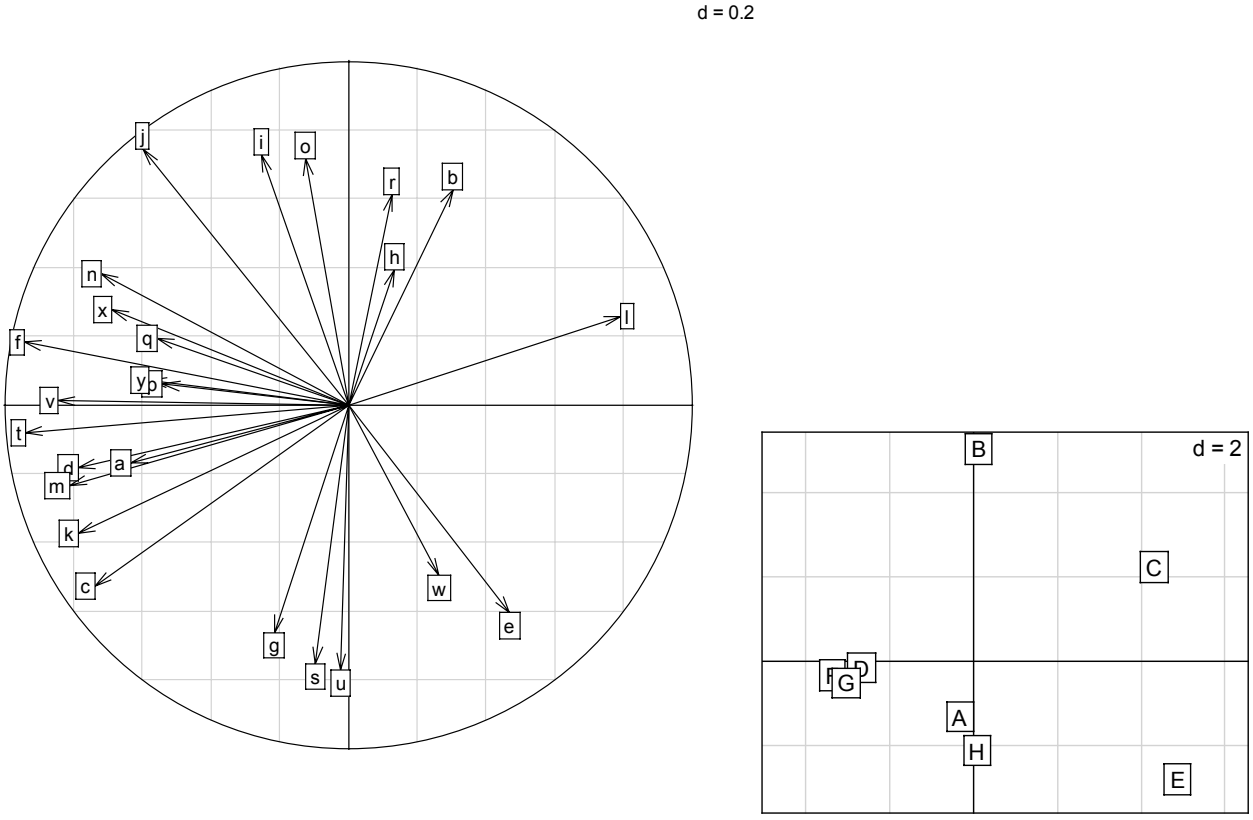
3.5. Existe-t-il deux juges ayant fait exactement le même rangement ?

```
> princomp(macon,cor=T)
Call:
princomp(x = macon, cor = T)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8
2.962e+00 2.374e+00 1.934e+00 1.635e+00 1.398e+00 1.312e+00 7.093e-01 4.059e-08
```

```
   Comp.9   Comp.10   Comp.11   Comp.12   Comp.13   Comp.14   Comp.15   Comp.16
3.953e-08 2.836e-08 2.432e-08 2.187e-08 1.838e-08 1.335e-08 1.157e-08 9.352e-09
  Comp.17   Comp.18   Comp.19   Comp.20   Comp.21   Comp.22   Comp.23   Comp.24
0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
  Comp.25
0.000e+00

 25  variables and  8 observations.
```

3.6. Que signifient les nombres ci-dessus ?

La figure ci-dessous représente le cercle des corrélations (plan 1-2) et la carte factorielle des lignes de l'ACP normée du tableau macon.



3.7. Que pensez-vous du juge l ?

3.8. Pourquoi la majorité des juges donnent une flèche dirigée vers la gauche alors que les produits préférés (E et C) sont à droite ?

3.9. Illustrer par les données ce qui différencie les juges prenant des valeurs opposées sur le deuxième axe.

3.10. Les jugements portés semblent-ils associés à la catégorie professionnelle des juges ?

Solution

1.

1.1. Quelle est la matrice de corrélation …

A est centrée. Les vecteurs colonnes sont normés. La matrice des cosinus est celle des produits scalaires. Elle vaut :

$$\mathbf{R} = \mathbf{A}^t\mathbf{A} = \begin{bmatrix} 1 & -\sqrt{3}/2 & 1/2 \\ -\sqrt{3}/2 & 1 & -\sqrt{3}/2 \\ 1/2 & -\sqrt{3}/2 & 1 \end{bmatrix}$$

1.2. Montrer que pour toute matrice …

$$\mathbf{Xu} = 0 \Rightarrow \mathbf{X}^t\mathbf{Xu} = 0$$
$$\mathbf{X}^t\mathbf{Xu} = 0 \Rightarrow \mathbf{u}^t\mathbf{X}^t\mathbf{Xu} = 0 \Rightarrow \left\|\mathbf{Xu}\right\|^2 = 0 \Rightarrow \mathbf{Xu} = 0$$

Les deux matrices ont même noyau donc même dimension d'image donc même rang.

1.3. Quelle est le rang de la matrice de corrélation …

C'est le rang de **A**, donc 2 (somme des lignes nulles).

1.4. Montrer que le rang d'une matrice …

Le rang de **R** est celui de $\mathbf{X}^t\mathbf{X}$, donc de **X**. Les colonnes de **X** sont centrées, donc $\mathbf{X}^t\mathbf{1}_n = 0$, donc les lignes de **X** ont une somme nulle, donc le rang de **X** ne peut excéder *n*-1.

2.

2.1. Peut-on attribuer à chaque analyse …

Non, on peut simplement dire que deux des analyses donnent une structure à deux dimensions et qu'une seule des trois indique une absence de structure. On peut juste affirmer que le graphe de droite est celui de l'analyse de talea.

2.2. Quelles sont les cartes des lignes ?

Les figures 1, 2 et 4 sont des cartes de lignes (coordonnées supérieures à 1). Les figures 3, 5 et 6 sont des cartes de colonnes (elles tiennent dans le cercle unité).

2.3. Quelles sont les cartes de l'analyse de tparti ?

Les figures 1 (carte des lignes) et 5 (carte des colonnes) sont celles de l'analyse de tparti (elles expriment la partition en trois groupes du tableau).

2.4. Comment obtient-on ce graphique ?

C'est le biplot de l'analyse de tgrad. On l'obtient par biplot(princomp(tgrad)).

2.5. Laquelle de ces trois analyses invalide…

L'analyse de tgrad permet de dire le contraire. Il n'y a qu'un fait marquant dans les données (ordination régulière des lignes et des colonnes) mais il faut deux axes pour l'exprimer par une carte en forme de cercle.

2.6. Commenter l'assertion "si on avait utilisé les tableaux transposés …

C'est parfaitement idiot. Le centrage et la normalisation des tableaux sont des opérations dissymétriques et on n'aurait pas obtenu des résultats numériques identiques. Par contre, comme les structures sont simples et symétriques sur les lignes et les colonnes, on aurait obtenu des résultats conduisant à une interprétation voisine.

2.7. Vrai ou faux ? "Ces trois analyses ont donné une valeur propre nulle."

VRAI, en vertu de la question 1.4.

3.

3.1. Quelles sont les moyennes et les variances de ce tableau ?

Chaque variable-colonne est une permutation sur {1, 2, …,8}. Les moyennes sont toutes égales à 9/2 et les variances sont toutes égales à 6 si on les calcule avec n -1 et 21/4 si on les calcule avec n.

3.2. Commenter ce résultat.

La matrice des corrélations de macon est de dimension 25 et d'après ce qui précède de rang au plus égal à 7. Le résultat indique avec 15 chiffres significatifs exacts que la matrice a 18 valeurs propres nulles et 7 valeurs propres non nulles.

3.3. Quel est le classement du concours ?

```
apply(macon,1,table)
$C
1 2 3 4 5 6 7
7 8 1 1 3 4 1 premier
$E
1 2 3 4 6 7 8
7 7 3 2 2 1 3 second
$B
1 2 3 4 5 6 7 8
4 2 1 5 3 2 3 5 troisième
$G
1 2 3 4 5 6 7 8
2 1 5 3 4 1 7 2 quatrième
Résultat : C>E>B>G>D>H>A>F
```

```
$D
 1  2  3  4  5  6  7  8
 2  1  1  1  2 10  4  4 cinquième
$H
1 2 3 4 5 6 7 8
1 3 5 4 5 2 3 2 sixième
$A
1 2 3 4 5 6 7 8
1 2 3 6 6 1 3 3 septième
$F
1 2 3 4 5 6 7 8
1 1 6 3 2 3 3 6 huitième
```

3.4. Obtient-on le même résultat en classant par la somme des rangs ?

OUI pour les deux premiers mais pas du tout ensuite.

3.5. Existe-t-il deux juges ayant fait exactement le même rangement ?

NON, car le maximum de la corrélation entre deux juges différents ne vaut pas 1, ce qui est le cas si deux juges ont fait le même classement.

3.6. Que signifient les nombres ci-dessus ?

Ce sont les racines carrées des valeurs propres dites encore valeurs singulières du tableau.

3.7. Que pensez-vous du juge l ?

Il a porté un jugement pratiquement opposé à l'ensemble des autres. Il a mis les deux premiers aux deux dernières places.

3.8. Pourquoi la majorité des juges donnent une flèche …

Parce que le rang est une valeur qui marque la préférence à l'envers (rang faible signifie préférence élevée). Le faisceau de flèches à gauche est un ensemble de variables corrélées négativement avec la première coordonnée factorielle. Elles

prennent des valeurs faibles (négatives, en dessous de la moyenne, donc 1,2 ou 3) pour les composantes négatives, donc C ou E. La majorité des juges à gauche exprime la préférence marquée pour les produits à droite et inversement D, F et G sont les produits rejetés donc avec des rangs élevés.

3.9. Illustrer par les données ce qui différencie les juges …

Ceux d'en haut préfèrent E à B,C alors que ceux d'en bas font l'inverse. Par exemple :

```
> macon[,c("i","o","u","s")]
  i o      u s            i o      u s
A 1 4      4 8          E 3 1 +    8 6 -
B 8 8 -    1 1 +        F 4 7      6 3
C 5 6 -    2 2 +        G 7 2      3 4
D 6 3      7 7          H 2 5      5 5
```

3.10. Les jugements portés semblent-ils associés à la catégorie professionnelle des juges ?

NON. Si on reporte la catégorie professionnelle des juges sur la carte, les cinq groupes sont largement mélangés. C'est une affaire de goût et non de compétence.

```
prof_gl(5,5)
scatter.corcircle(pca1$co,lab=as.character(prof))
scatter.chull(pca1$co,prof,add.p=T,clab=2)
```