

TD ANALYSE DES DONNEES

Exemple d'ACP : Etude olfacto-gustative de cidres

Plusieurs caractéristiques du cidre ont été mesurées sur 10 cidres différents. Les résultats de l'ACP sont présentés page suivante.

cidre	odeur	sucré	acide	amer	astringence	suffocante	piquante	alcool	parfum	fruitée
1	2,14	1,86	3,29	2,29	2	0,14	2,29	1,86	1,29	1,29
2	2,43	0,79	2,71	2,57	2	0,43	2,57	2,86	0,43	0,14
3	2,71	3,14	2,57	2,57	1,43	0,14	2,14	0,86	2,29	1,71
4	3	3,71	2,14	2,07	1,57	0	1,29	1	3,14	3,14
5	3,43	1,29	2,86	3,14	2,17	1	1,86	2,86	1,14	0,29
6	3,14	0,86	2,86	3,79	2,57	0,14	1,71	3,29	0,14	0
7	3,14	1,14	2,86	2,86	2	0,43	1,71	1,86	0,14	0
8	2,43	3,71	3,21	1,57	1,71	0	1	0,57	2,57	2,86
9	5,1	2,86	2,86	3,07	1,79	1,71	0,43	1,43	0,57	2,71
10	3,07	3,14	2,57	3	2	0	0,43	1,29	2,57	3,07

Partie I : Examen des données

a. Justifier l'utilisation d'une ACP.

b. Expliquer les différences obtenues entre une ACP normée et non normée?

Pour réaliser les différents traitements avec R, il faut charger les packages rgl, ade4, ou FactoMineR, et éventuellement Rcmdr (voir la fiche installation).

```
> cidre <- read.table("../echange/cidre.txt")
```

Paramètres statistiques

```

      mean      sd      n
acid  2.793 0.3285676 10
alco  1.788 0.9372869 10
amer  2.693 0.6244473 10
astr  1.924 0.3221525 10
fruit 1.521 1.3484843 10
odeu  3.059 0.8217657 10
parf  1.428 1.1271577 10
piku  1.543 0.7425040 10
sucr  2.250 1.1826994 10
suffo 0.399 0.5538441 10

```

```

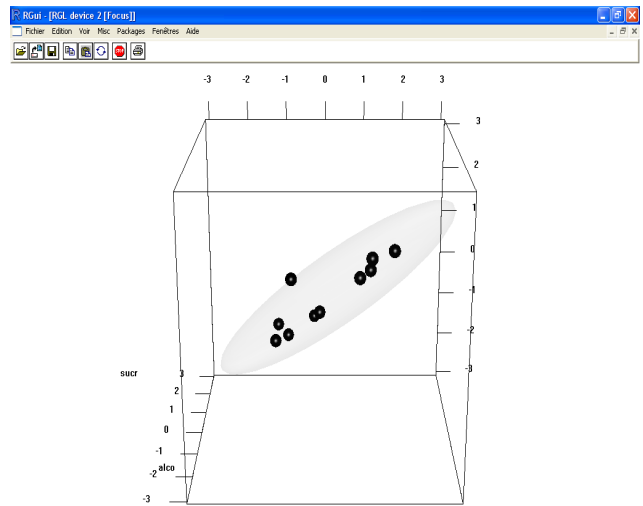
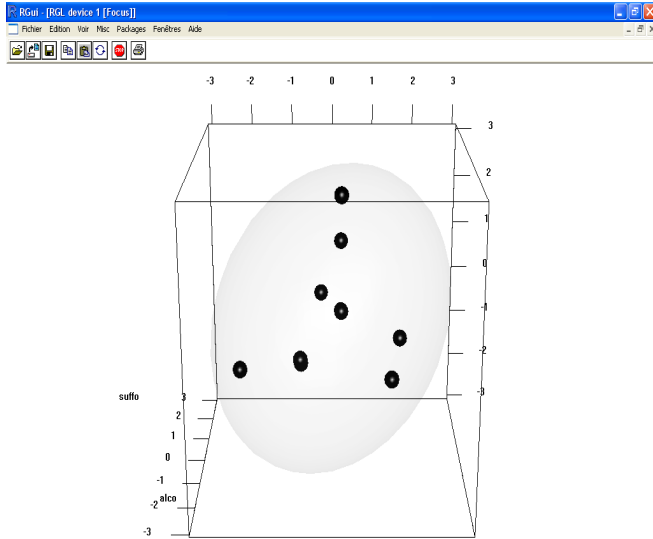
> round(cor(cidre),2)
      odeu  sucr  acid  amer  astr  suffo  piku  alco  parf  fruit
odeu  1.00  0.08 -0.16  0.49  0.04  0.84 -0.61  0.03 -0.29  0.18
sucr  0.08  1.00 -0.29 -0.60 -0.77 -0.19 -0.61 -0.92  0.87  0.95
acid -0.16 -0.29  1.00 -0.08  0.34  0.14  0.14  0.15 -0.40 -0.27
amer  0.49 -0.60 -0.08  1.00  0.71  0.38 -0.03  0.70 -0.63 -0.50
astr  0.04 -0.77  0.34  0.71  1.00  0.07  0.14  0.86 -0.66 -0.64
suffo 0.84 -0.19  0.14  0.38  0.07  1.00 -0.23  0.22 -0.50 -0.10
piku -0.61 -0.61  0.14 -0.03  0.14 -0.23  1.00  0.48 -0.33 -0.73
alco  0.03 -0.92  0.15  0.70  0.86  0.22  0.48  1.00 -0.76 -0.83
parf -0.29  0.87 -0.40 -0.63 -0.66 -0.50 -0.33 -0.76  1.00  0.80
fruit 0.18  0.95 -0.27 -0.50 -0.64 -0.10 -0.73 -0.83  0.80  1.00

```

c. Déterminer trois groupes de variables qui présentent des corrélations entre elles ($r > 0.5$).

Examen graphique :

```
>library("rgl")
>cidreR = as.data.frame(scale(cidre))
>attach(cidreR)
nuage 1 :
>plot3d(acid,alco,suffo,type="s",xlim=c(-3,3),ylim=c(-3,3),zlim=c(-3,3))
>plot3d(ellipse3d(cor(cbind(acid,alco,suffo))),col="grey",alpha=0.05,add=TRUE)
nuage 2 :
>plot3d(parf,alco,sucr,type="s",xlim=c(-3,3),ylim=c(-3,3),zlim=c(-3,3))
>plot3d(ellipse3d(cor(cbind(parf,alco,sucr))),col="grey",alpha=0.05,add=TRUE)
```



- d. Que représentent les ellipses dans la représentation en 3D.
- e. Expliquez les différences entre les ellipses obtenues dans les deux nuages.

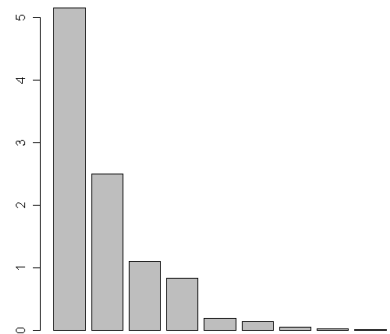
Partie II : ACP normée du tableau**1. Nombre de facteurs retenus**

```
> round(acp$eig,2)
[1] 5.15 2.50 1.10 0.83 0.19 0.14 0.05 0.02 0.01

> round(cumsum(acp$eig*10),2)
[1] 51.54 76.56 87.53 95.87 97.81 99.21 99.70
99.94 100.00
```

a. Les deux premiers facteurs ont été retenus ici. Quel est le pourcentage de variance expliqué par ces deux facteurs ?

b. Que signifie ce pourcentage ?



2/ Analyse des variables

```
> inertie <- inertia.dudi(acp, col.inertia=TRUE)
```

[coordonnées des variables]

```
> round(acp$co, 2)
```

	Comp1	Comp2
odeu	-0.08	-0.98
sucr	0.97	-0.16
acid	-0.33	0.15
amer	-0.72	-0.47
astr	-0.83	-0.03
suffo	-0.31	-0.79
piqu	-0.49	0.72
alco	-0.94	0.04
parf	0.91	0.20
fruit	0.91	-0.29

[ctr en %]

```
> inertia$col.abs/100
```

	Comp1	Comp2
odeu	0.13	38.70
sucr	18.40	1.05
acid	2.07	0.94
amer	9.97	8.68
astr	13.49	0.04
suffo	1.84	24.96
piqu	4.65	20.59
alco	17.28	0.06
parf	15.95	1.54
fruit	16.21	3.44

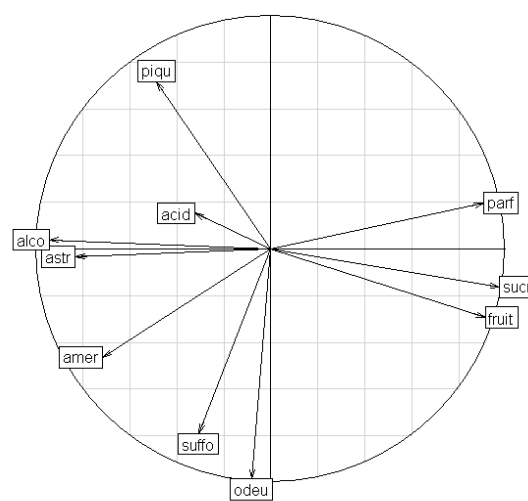
[qlt en %]

```
> inertia$col.re/100
```

	Comp1	Comp2	con.tra
odeu	-0.69	-96.83	10
sucr	94.84	-2.63	10
acid	-10.65	2.35	10
amer	-51.38	-21.71	10
astr	-69.54	-0.11	10
suffo	-9.48	-62.44	10
piqu	-23.97	51.52	10
alco	-89.09	0.16	10
parf	82.23	3.85	10
fruit	83.56	-8.60	10

```
> s.corcircle(acp$co, xax=1, yax=2)
```

- Comment reconnaît-on sur la figure des variables qu'une variable est bien représentée ?
- Quelles sont les variables mal représentées dans le plan F1-F2 ? Justifier votre réponse.
- A l'aide de la figure sur les variables, préciser la variable la plus corrélée positivement à alcool, la plus corrélée négativement à alcool, la moins corrélée à alcool.
- Quelles sont les variables qui ont contribué à l'axe F1 ? Justifier votre réponse.
-
- Donner une signification à cet axe.
- Quelles sont les variables qui ont contribué à l'axe F2 ? Justifier votre réponse.
- Donner une signification à cet axe.



3/ Analyse des individus

```
> inertie <- inertia.dudi(acp, row.inertia=TRUE)
```

Composantes principales

```
> round(acp$li, 2)
```

	Axis1	Axis2
1	-0.53	1.87
2	-2.15	1.41
3	1.82	0.90
4	3.32	0.20
5	-2.20	-0.62
6	-3.57	-0.04
7	-1.69	0.12
8	2.94	1.01
9	0.29	-4.09
10	1.78	-0.76

[ctr en %]

```
> inertie$row.abs/100
```

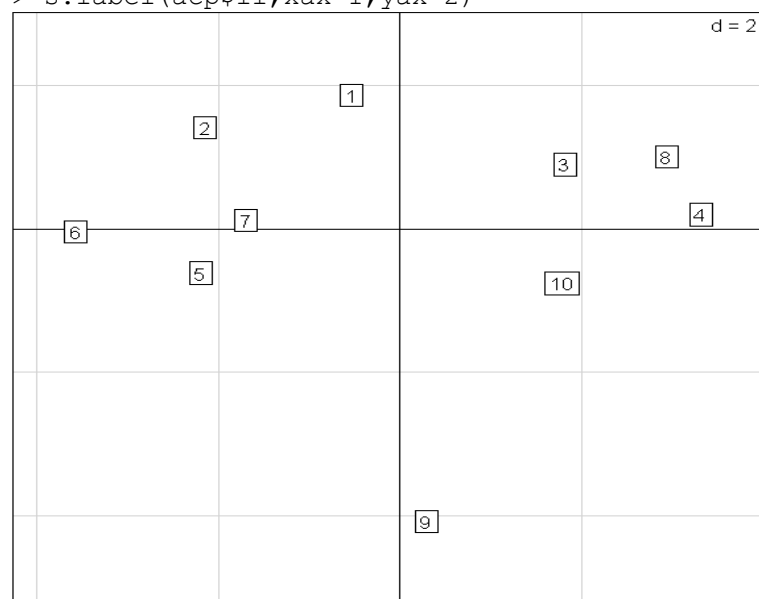
	Axis1	Axis2
1	0.55	13.91
2	8.95	7.97
3	6.41	3.24
4	21.35	0.17
5	9.37	1.53
6	24.78	0.01
7	5.55	0.06
8	16.74	4.04
9	0.16	66.77
10	6.14	2.31

[qlt en %]

```
> inertie$row.re/100
```

	Axis1	Axis2	con.tra
1	-4.71	58.00	6.00
2	-56.65	24.50	8.14
3	49.20	12.07	6.72
4	77.55	0.29	14.19
5	-78.09	-6.19	6.18
6	-82.36	-0.01	15.51
7	-69.70	0.35	4.11
8	63.92	7.48	13.50
9	0.46	-91.86	18.19
10	42.36	-7.76	7.4

```
> s.label(acp$li, xax=1, yax=2)
```



- Comment évalue-t-on si un individu est bien représenté dans un plan ?
- Quel est l'individu le mieux représenté dans le plan F1-F2 ? Justifier votre réponse.
- Quels sont les 3 individus les moins bien représentés dans le plan F1-F2 ? Justifier votre réponse.
- Quels sont les individus qui ont contribué à l'axe F1 ? À l'axe F2 ? Justifier votre réponse.
- Proposer 4 groupes de cidres en précisant clairement les principales caractéristiques de ces groupes.

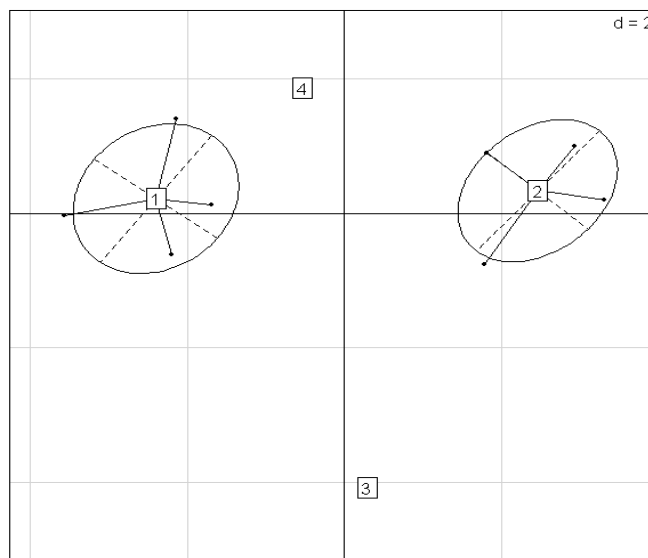
5. Vers la classification.

Les individus semblent se répartir en quatre groupes :

groupe 1 : 2 5 6 7 groupe 2 : 3 4 8 10 groupe 3 : 9 groupe 4 : 1

Créons un facteur indiquant le groupe :

```
> fac <- as.factor(c(4,1,2,2,1,1,1,2,3,2))
> s.class(dfxy=acp$li, fac=fac, xax=1, yax=2)
```



6) Autre Représentations complémentaires



scatter(acp)

Autre exemple d'ACP :

Traiter l'exemple decathlon décrivant les score obtenu aux 10 épreuves du décathlon par différents athlètes ainsi que le score final.

Exemple d'AFC :

Les données proviennent d'une société d'assurance automobile. Les deux variables retenues pour l'analyse sont le mode de règlement et la situation matrimoniale (<http://pbil.univ-lyon1.fr/R/enseignement.html>, tdr620).

	celibataire	concubin	divorce	marie	veuf
annuel	209	1483	41	320	60
mensuel	34	151	1	70	10
semestriel	535	2448	33	897	135
trimestriel	77	245	4	139	9

1. Comment s'appelle ce tableau? Quelle est la première question qu'il pose ?

2. Tester l'indépendance entre les deux variables.

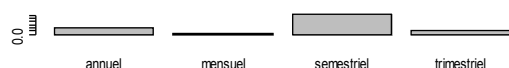
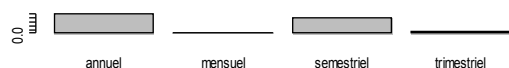
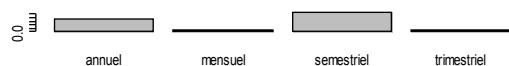
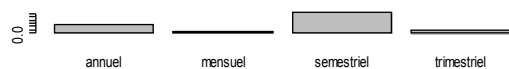
3. Est-il préférable de comparer les profils colonnes ou les profils ligne? Interpréter les résultats

X-squared = 129.2212, df = 12, p-value < 2.2e-16

4. Est-il préférable de comparer les profils colonnes ou les profils ligne? Interpréter les résultats.

```
profilcol <- prop.table(sitpay, 2)
par(mfrow=c(4,1))
barplot(profilcol[,1])
```

	celibataire	concubin	divorce	marie	veuf
annuel	24	34	52	22	28
mensuel	4	3	1	5	5
semestriel	63	57	42	63	63
trimestriel	9	6	5	10	4

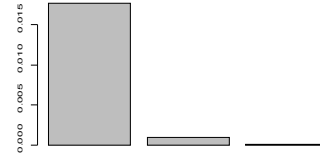


AFC du tableau

1. Qu'apporte l'AFC pour l'étude de ce tableau ?

2. Interpréter les valeurs propres obtenues.

	eigenvalue	% variance	cum variance
dim 1	0.018	94.276	94.276
dim 2	0.001	5.106	99.381
dim 3	0.000	0.619	100.000



```
> res$col
$contrib
      Dim 1      Dim 2      Dim 3
celibataire 14.69729695  2.290090 67.87627123
concubin    26.83118595  0.758794  0.03496332
divorce     12.14464596 37.325609  1.80965172
marie       46.29342775  3.294608 29.74455904
veuf        0.03344339 56.330899  0.53455469

$cos2
      Dim 1      Dim 2      Dim 3
celibataire 0.96270313 0.008124114 2.917276e-02
concubin    0.99846219 0.001529270 8.537074e-06
divorce     0.85658219 0.142580315 8.374980e-04
marie       0.99199432 0.003823509 4.182173e-03
veuf        0.01083095 0.988033114 1.135934e-03

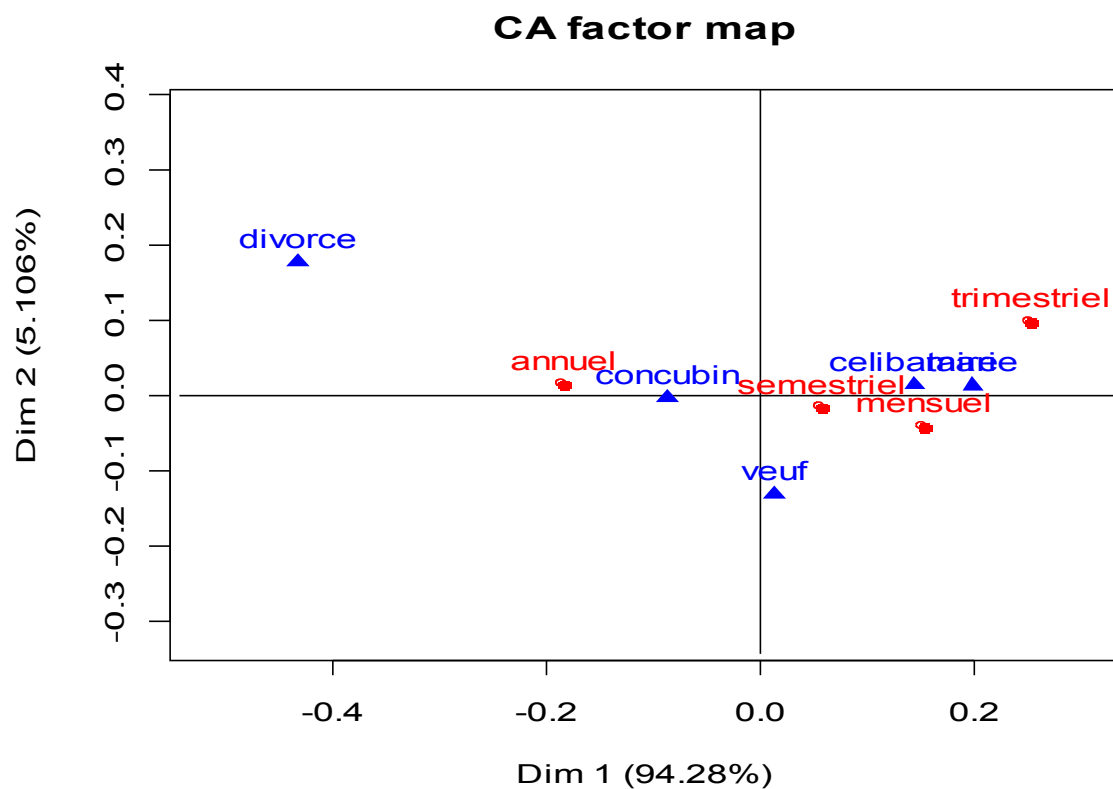
> res$row
$contrib
      Dim 1      Dim 2      Dim 3
annuel    59.338040  7.754347  2.2888620
mensuel    5.139243  6.836175 84.1700687
semestriel 10.770248 17.345222 13.2263652
trimestriel 24.752470 68.064256  0.3147040

$cos2
      Dim 1      Dim 2      Dim 3
annuel    0.9927227 0.007026006 2.512575e-04
mensuel    0.8478129 0.061077684 9.110940e-02
semestriel 0.9130092 0.079633871 7.356896e-03
trimestriel 0.8703153 0.129612128 7.260474e-05
```

3. Interpréter les qualités de représentation ($\cos^2$) des profils colonne et ligne dans le plan $F_1 F_2$.

4. Interpréter les contributions des profils colonne et ligne dans le plan $F_1 F_2$.

5. Interpréter les cartes factorielles (plans de projection)

**Autre exemple d'AFC :**

Traiter l'exemple election2002 décrivant les journaux lus en fonction du vote.

Exemple de CAH

Exemple 1 : On considère la matrice des carrés des distances d'un ensemble de 5 points de même poids.

	A	B	C	D	E
A	0	16	1	9	10
B		0	17	25	2
C			0	4	9
D				0	13
E					0

Réaliser la classification ascendante hiérarchique et le dendrogramme en utilisant le critère du saut minimal.

Exemple 2 : Le tableau décrit pour chaque pays de la CEE la production agricole par unité d'exploitation (x) et le pourcentage de la population active employée dans l'agriculture (y).

pays	x	y
B	16.8	2.7
DK	21.3	5.7
D	18.7	3.5
GR	5.9	22.2
E	11.4	10.9
F	17.8	6.0
IRL	10.9	14.0
I	16.6	8.5
L	21.0	3.5
NL	16.4	4.3
P	7.8	17.4
UK	14.0	2.3

