

Biostatistiques

Rappels de cours et travaux dirig s

Analyse des donn es M2

auteur : Jean-Marc Labatte
jean-marc.labatte@univ-angers.fr

SOMMAIRE

Table des matières

VIII – ANALYSE EN COMPOSANTES PRINCIPALES.....	4
Fiche 36 – Principe de la méthode ACP.....	6
Fiche 37 – Aides à l'interprétation.....	8
Fiche 38 – Représentation graphique des variables.....	10
Fiche 39 – Représentation graphique des individus.....	11
IX ANALYSE FACTORIELLE DES CORRESPONDANCES.....	14
Fiche 40 – Principe de l'AFC.....	15
Fiche 41 – Aides à l'interprétation.....	17
Fiche 42 – Représentation graphique des profils.....	19
X CLASSIFICATION.....	22
Fiche 43 – Exemple manuel de CAH.....	24
Fiche 44 – Exemple de CAH avec ade4.....	25
ANNEXES.....	28
Annexe A : Lois de probabilités usuelles.....	29
Annexe B : Construction d'un test statistique.....	32
Annexe C : Installation de R.....	33

VIII – ANALYSE EN COMPOSANTES PRINCIPALES

Objectifs et données pour l'ACP

Cette technique s'applique à des tableaux décrivant chaque individu par p variables quantitatives X_k . Les techniques classiques ne permettent que l'étude de la liaison entre deux variables : corrélation, régression et nuage de points par exemple.

L'**objectif** est ici de faire une synthèse de l'ensemble du tableau afin de :

- **synthétiser les liaisons entre variables** (cercle des corrélations), définir les variables qui vont dans le même sens, dans un sens opposé, indépendantes ...
- représenter dans un plan les individus afin de déterminer les individus proches ou éloignés, les regrouper en classe homogène, ... On parle de **topologie des individus**.
- construire de **nouvelles variables**, appelées composantes principales, non corrélées et qui permettent de synthétiser l'information

Ainsi, au lieu d'analyser le tableau à travers p variables, on se limitera à l'étude de quelques variables synthétiques, les composantes principales. La difficulté sera de donner un sens à ces variables et de proposer une analyse des résultats.

Le **tableau** se présente sous la forme :

	X_1	...	X_j	...	X_p
individu 1	x_{11}		x_{1j}		x_{1p}
...					
individu i	x_{i1}		x_{ij}		x_{ip}
...					
individu n	x_{n1}		x_{nj}		x_{np}

Exemple : Nous étudions dans cette partie les masses de différentes parties d'un groupe de 23 bovins constitué de 12 charolais (1 à 12) et 11 zebus (13 à 23).

Les variables représentent: poids vif. poids de la carcasse. poids de la viande de première qualité. poids de la viande totale. poids du gras. poids des os.

Retrouver à l'aide du logiciel R et son interface R-commander les différents résultats ci-dessous et expliquer leur apport respectif dans l'analyse des données.

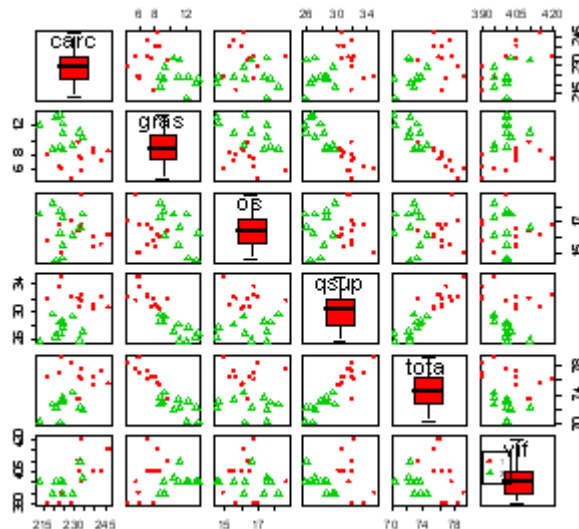
1. Paramètres statistiques: Moyenne et écart-type par race

Variable: carc				Variable: os				Variable: tota			
	mean	sd	n		mean	sd	n		mean	sd	n
1	233.0000	8.790491	12	1	16.30833	0.9949494	12	1	76.60000	1.502120	12
2	224.2727	6.018154	11	2	16.51818	1.2584261	11	2	72.56364	1.297130	11
Variable: gras				Variable: qsup				Variable: vif			
	mean	sd	n		mean	sd	n		mean	sd	n
1	7.258333	1.439986	12	1	31.99167	1.344658	12	1	402.5000	9.885711	12
2	10.845455	1.758615	11	2	27.66364	1.343334	11	2	399.7273	4.221159	11

Matrice des corrélations

	vif	carc	qsup	tota	gras	os
vif	1.00	0.64	-0.09	-0.13	0.16	-0.06
carc	0.64	1.00	0.28	0.39	-0.33	-0.09
qsup	-0.09	0.28	1.00	0.89	-0.86	-0.06
tota	-0.13	0.39	0.89	1.00	-0.91	-0.12
gras	0.16	-0.33	-0.86	-0.91	1.00	-0.27
os	-0.06	-0.09	-0.06	-0.12	-0.27	1.00

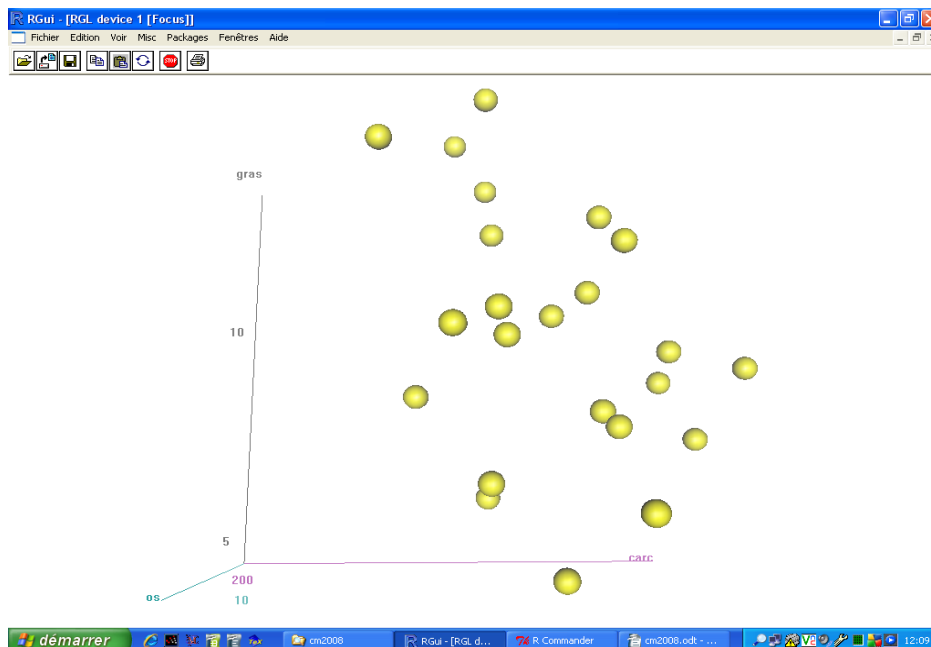
2. Représentation graphique



3. Nuage de points

Dans R, on peut utiliser les commandes suivantes pour construire des nuages 3d. On pourra changer les 3 variables. A défaut, utiliser R commander.

```
>library(rgl)
>attach(zebu)
>col <- ifelse(race==1,"blue","red")
>plot3d(gras,tota,qsup,type="s",col=col)          les échelles sont différentes
suivant les axes
>plot3d(gras,tota,qsup,type="s",col=col,xlim=c(0,450),ylim=c(0,450),zlim=c(0,450))
les échelles sont les mêmes
```



Quelques liens :

<http://pbil.univ-lyon1.fr/R/enseignement.html>
http://www.unilim.fr/pages_perso/vincent.jalby/m1sm/documents/m1sm_S_03.pdf
<http://infolettres.univ-brest.fr/~carpenti/2006-2007/Ana-mult-1-2007.doc>
<http://www.lirmm.fr/~guindon/dess/acp.df>

Fiche 1 – Principe de la méthode ACP

Chaque individu est décrit ici par $p=6$ variables quantitatives. Un individu est représenté par un point dont les coordonnées sont les valeurs prises par les 6 variables (espace à $p=6$ dimensions). On peut ainsi mesurer la distance entre deux individus à l'aide d'une distance classique entre deux points.

Le principe de l'ACP répond simultanément aux deux objectifs suivant :

- **Pour les individus**

L'objectif de la méthode ACP est de projeter les individus sur des axes appelés axes factoriels en conservant le mieux possible les distances entre individus. Cela revient à **déformer le moins possible le nuage de points initial lorsqu'on le projette sur un axe ou un plan**.

Dans la pratique, la projection sur l'axe F_1 permet d'obtenir le maximum de dispersion (=inertie = variance en une dimension) des points projetés sur l'axe.

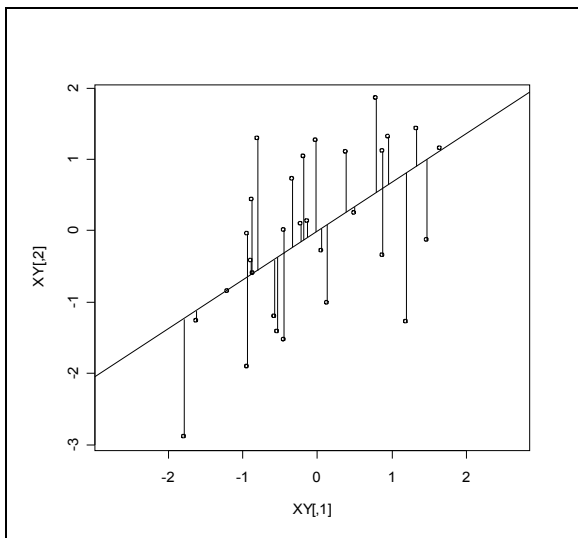
- **Pour les variables**

Cela revient à **construire des variables, appelées composantes principales, par combinaison linéaire des variables initiales et telles que ces nouvelles variables aient la plus grande variance possible**. Les composantes principales sont de plus non corrélées.

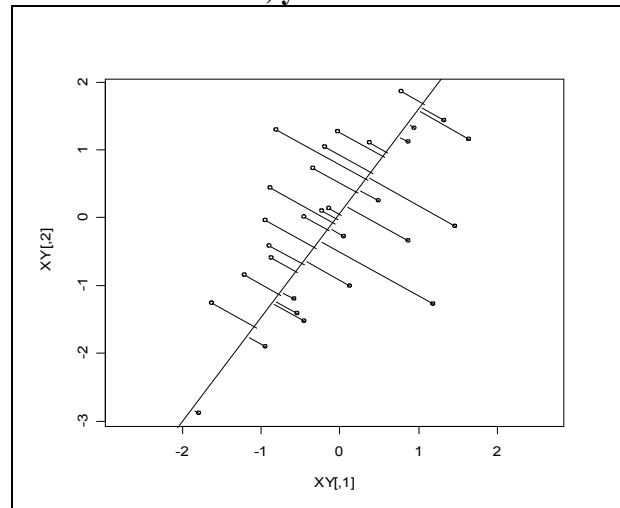
On ne s'intéresse alors qu'aux composantes principales qui ont la plus forte **variance (=valeur propre de l'axe)**. On construit ensuite des nuages de points des individus en fonction de ces composantes principales dans les plans factoriels $F_1 F_2$, ou $F_1 F_3$...

Interprétation graphique de l'ACP

Régression linéaire de y en x



ACP avec x, y



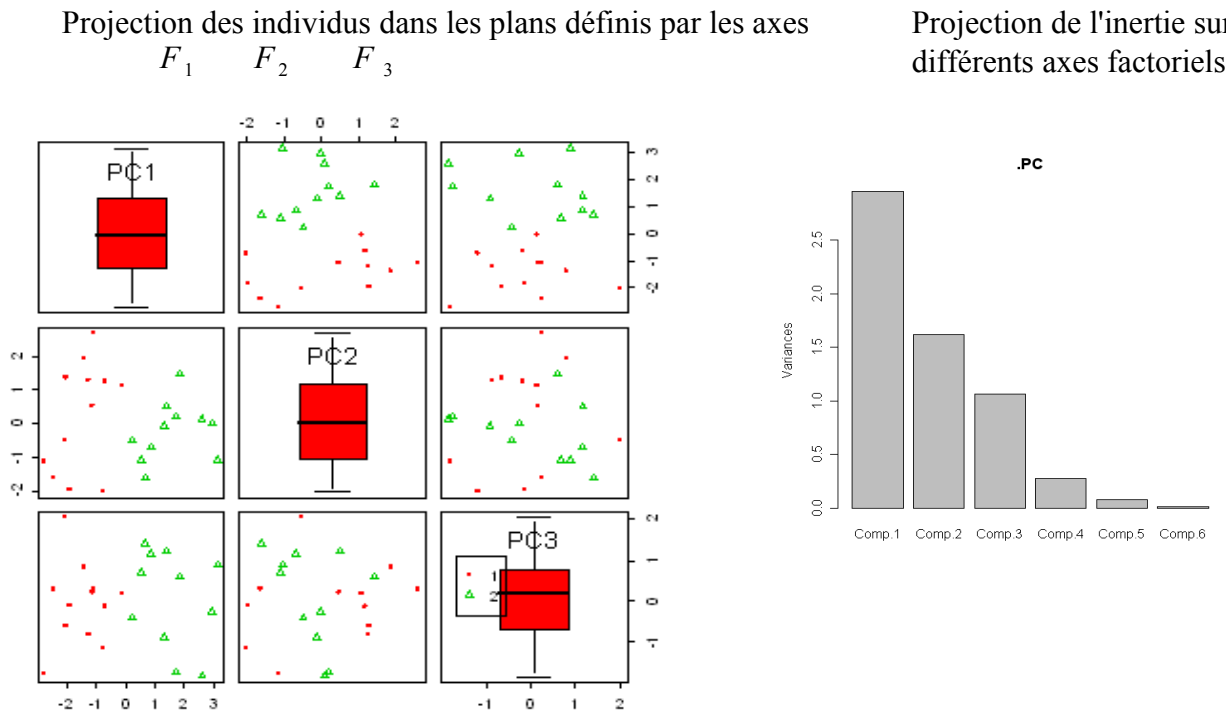
Remarque importante : En général, du fait de l'**hétérogénéité des variables initiales** et de leurs unités, **on réduit ces variables**. On parle alors d'**ACP normée**.

Une variable est dite réduite quand sa variance vaut 1. De la sorte, chaque variable initiale aura une même importance dans l'analyse car sa contribution est proportionnelle à sa variance.

Dans la pratique, on normalise presque toujours et surtout lorsque les variables sont exprimées dans des unités différentes.

L'objectif est ainsi de construire des variables qui synthétisent la dispersion du nuage. Si plusieurs variables initiales sont ainsi fortement corrélées entre elles, celles-ci sont alors représentées par une

composante principale qui les résume. Au final, au lieu de travailler avec $p=6$ variables, on peut espérer travailler sur 2 ou 3 variables synthétiques qui résument l'essentiel de l'information. On retrouve une partie des résultats de l'acp dans Rcmdr, statistiques, ajustement multivarié.



Mais il est fortement conseillé sous R d'utiliser la librairie ade4.

```
>library(ade4)
>zebu.acp <- dudi.pca(zebu[,1:6])
>zebu.acp$eig (valeurs propres)
>zebu.acp$li (composantes principales)
>zebu.acp$co (coordonnées des variables) .... voir l'aide en ligne
```

Guide pratique de l'analyse ACP

- **Etape 1** : Sélection des axes et des plans retenus principalement par rapport aux valeurs propres.
- **Etape 2** : Projection des variables et individus dans un plan donné (F_1 F_2 en premier)
 - Examen des *qlt* dans le plan pour éliminer les individus mal représentés
 - Bilan des *ctr* pour un axe afin de donner un sens à cet axe (opposition, tendance ...)
 - Topographie des variables et individus afin d'identifier des groupes, des oppositions, des tendances notamment à l'aide de la fonction *s.class*
 - Utiliser ses connaissances sur le sujet pour proposer des explications sur les résultats de l'analyse
 - Utiliser des individus ou variables supplémentaires ou des profils type (moyenne des H et des F par exemple)

Fiche 2 – Aides à l'interprétation

1. Valeurs propres λ et choix des axes

Pour définir le nombre d'axes étudiés, on étudie les valeurs propres obtenues. Chaque valeur propre correspond à la part d'inertie projetée sur un axe donné.

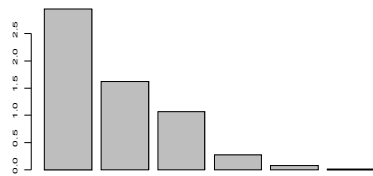
Remarque importante: La somme des valeurs propres est égale à l'inertie totale du nuage (= nombre de variables en ACP normé). On caractérise ainsi chaque axe par le pourcentage d'inertie qu'il permet d'expliquer.

On ne retient donc que les axes avec les plus fortes valeurs propres. Le choix des axes retenus est un peu délicat. On peut donner quelques règles :

- **Règle de Kaiser en ACP normée:** on ne s'intéresse qu'aux axes avec une valeur propre supérieure à 1 (= inertie d'une variable initiale).
- **Règle de l'inertie minimale :** On sélectionne les premiers axes afin d'atteindre un % donné d'inertie expliquée (70% par exemple).
- **Règle du coude :** On observe souvent de fortes valeurs propres au départ puis ensuite de faibles valeurs avec un décrochage dans le diagramme. On retient les axes avant le décrochage.
- **Règle de bon sens :** On analyse les plans et axes et on ne retient que ceux interprétables.

Exemple zebu

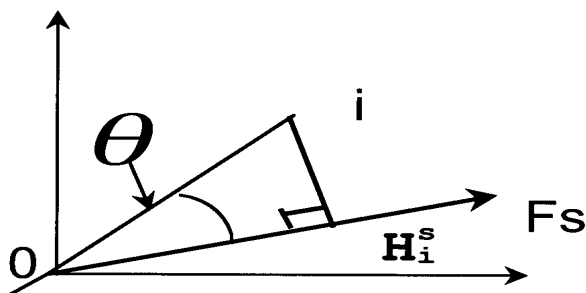
```
> round(acp$eig, 2)
[1] 2.95 1.62 1.07 0.27 0.08 0.01
> round(acp$eig/sum(acp$eig)*100)
[1] 49 27 18 5 1 0
```



Avec ade4:	réaliser l'acp	<code>>acp=dudi.pca(tableau)</code>
	Extraire les valeurs propres	<code>>acp\$eig</code>
	Calcul des %	<code>>acp\$eig/sum(acp\$eig)*100</code>

2. Qualité de représentation $q_l t$

Les individus représentés dans un plan factoriel ne sont pas forcément correctement représentés.



Qualité de représentation *qlt*:

Si l'angle θ est grand, le point initial est éloigné de sa projection. On utilise le paramètre $\cos^2 \theta$ pour caractériser la qualité de représentation (*qlt*) sur un axe.

- Plus qlt_i est proche de 1 plus il est bien représenté.
- Plus qlt_i est proche de 0 plus il est mal représenté.
- Dans un plan, on calcule la somme des deux *qlt*, par exemple $qlt_{F_1} + qlt_{F_2}$ pour le plan $F_1 F_2$.
- *qlt* correspond en fait au rapport de l'inertie du projeté sur l'inertie du point initial.

Qualité globale : Dans un plan donné, on définit également la qualité globale comme le pourcentage d'inertie qu'explique le plan. C'est par rapport à cette qualité globale que l'on évalue la *qlt* d'un individu ou d'une variable.

Remarque : La *qlt* des variables peut s'analyser de la même façon mais l'utilisation du cercle des corrélation est plus intuitive.

Bilan : On commencera donc toujours l'analyse d'un plan factoriel en précisant l'existence (ou non) d'individus ou variables mal représentés et en justifiant par les *qlt*.

3. Contribution *ctr*

Lors de la construction d'un axe factoriel, certaines variables et certains individus ont des rôles plus importants. On calcule un paramètre appelé contribution, *ctr*, qui permet de calculer cette influence.

Définition: La contribution *ctr* est définie comme la proportion de l'inertie de l'axe expliquée par la variable ou l'individu.

Règles d'interprétation :

- L'analyse se fait axe par axe, en parallèle sur les variables et les individus.
- Plus *ctr* est grande, plus l'influence de l'individu est grande. On ne retient donc que les plus fortes valeurs (il y a souvent un décrochage après quelques valeurs).
- *ctr* est considéré comme positif si l'individu est dans la partie positive de l'axe.
- *ctr* est considéré comme négatif si l'individu est dans la partie négative de l'axe.
- Le bilan des *ctr* peut être présenté pour un axe donné sous forme d'un tableau avec les principales *ctr* + et – des individus et des variables, en précisant la valeur de *ctr* :

<i>ctr</i> axe F_1 :	-	+
Variables :		

Individus :

On réalise ensuite une interprétation.

Sous R, ces paramètres sont obtenus avec les commandes suivantes :

Pour les lignes (individus)	<code>>inertieL<-inertia.dudi(acp, row.inertia=TRUE)</code>
<i>ctr</i> des lignes en %	<code>>inertieL\$row.abs/100</code>
<i>qlt</i> des lignes en %	<code>>inertieL\$row.rel/100</code>
Pour les colonnes (variables)	<code>>inertieC<-inertia.dudi(acp, col.inertia=TRUE)</code>
<i>ctr</i> des colonnes en %	<code>>inertieC\$col.abs/100</code>
<i>qlt</i> des colonnes en %	<code>>inertieC\$col.rel/100</code>

Fiche 3 – Représentation graphique des variables

Les composantes principales sont construites comme des combinaisons linéaires des variables initiales. Pour visualiser les liaisons entre la composante principale et les variables initiales, on représente en ACP normée les variables dans les plans factoriels. Les coordonnées des variables sont les coefficients de corrélation de ces variables avec les composantes principales.

Les règles de lecture du cercle des corrélations sont :

- On ne prend en compte **que les variables proches du cercle des corrélations**. Dans le cas contraire, la variable est non corrélée à la composante principale et est donc mal représentée.
- **La liaison entre variables bien représentées s'analyse à travers la direction et le sens de leur vecteur :**
 - si les vecteurs ont même direction et même sens, les variables sont corrélées positivement,
 - si les vecteurs ont même direction mais de sens contraire, les variables sont corrélées négativement,
 - si les vecteurs sont perpendiculaires, les variables sont non corrélées.
- On synthétise chaque axe en précisant les variables qui contribuent le plus en positif ou en négatif (étude des ctr)

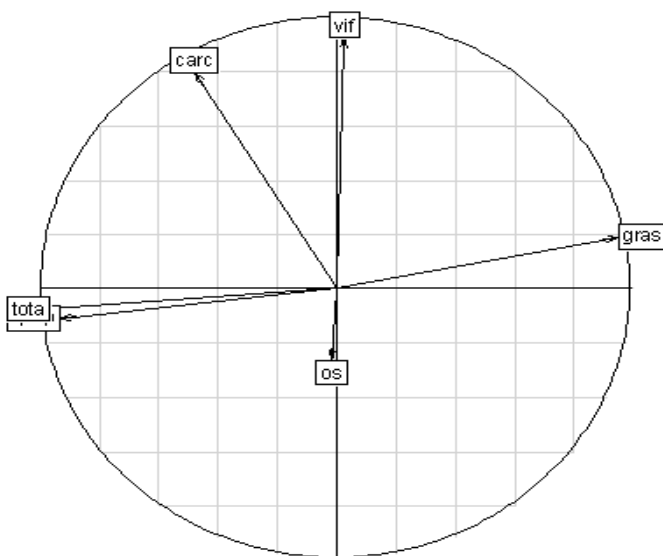
Exemple :

```
inertieV=inertia.dudi(acp,col.inertia=TRUE)
```

Composante principale		ctr / 10 000	
Comp1	Comp2	inertieV\$col.abs	
FL	-0.99	-0.05	FL 2043 189
RW	-0.94	0.35	RW 1833 8066
CL	-0.99	-0.10	CL 2054 720
CW	-0.99	-0.07	CW 2035 326
BD	-0.99	-0.10	BD 2035 699

qlt / 10 000		
Comp1	Comp2	con.tra
FL	-9786	-29 2000
RW	-8776	1224 2000
CL	-9835	-109 2000
CW	-9745	-49 2000
BD	-9746	-106 2000

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



Fiche 4 – Représentation graphique des individus

Les individus sont associés à des points de l'espace dont les coordonnées sont les variables. On peut mesurer la **distance entre ces individus** en utilisant simplement la distance euclidienne classique entre ces deux points (comme au collège...).

La construction des composantes principales conduit à **rendre minimale la déformation des distances entre individus lorsque l'on projette les individus dans le plan factoriel** $F_1 F_2$. Ainsi les distances que l'on observe entre les individus dans le plan factoriel sont globalement les plus proches possible des distances réelles entre ces individus.

L'analyse des plans factoriels permet ainsi d'observer les individus proches entre eux ou au contraire éloignés. Il est ainsi possible de construire des groupes, d'observer des tendances ...

Les règles de lecture des plans factoriels sont :

- **Seuls les individus bien représentés sont pris en compte** dans l'interprétation.
 - On calcule la somme des *qlt* dans le plan et on vérifie que cette somme n'est pas trop faible par rapport à la qualité moyenne du plan.
- On réalise le **bilan en positif et en négatif des individus qui ont la plus forte contribution** pour un axe donné.
 - On donne ainsi en parallèle avec l'analyse des variables une signification concrète à ces axes en terme d'opposition entre individus et variables ou tendance particulière.
- **On réalise des groupes**, à l'aide éventuelle de la fonction `s.class`, dans le cas de groupes préexistants (homme-femme par exemple) ou on construit arbitrairement ces groupes en raison des proximités entre individus.
- En présence de trop nombreux individus, on peut utiliser des **individus type** et réaliser une analyse sur ce individus.
- L'utilisation d'**individus supplémentaires** non utilisés dans l'ajustement mais a posteriori permet également d'éclairer l'analyse.

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

Composantes principales

```
> round(acp$li, 3)
```

	Axis1	Axis2	Axis3
1	-2.691	-1.137	-1.786
2	-0.050	1.102	0.180
3	-1.072	0.499	0.202
4	-0.671	1.228	-0.149
5	-0.756	-2.009	-1.152
6	-1.999	1.337	-0.616
7	-2.402	-1.625	0.276
8	-1.213	1.278	-0.819
9	-1.401	1.906	0.823
10	-1.869	-1.954	-0.105
11	-1.065	2.663	0.276
12	-2.032	-0.507	2.028
13	1.287	-0.082	-0.898
14	0.214	-0.485	-0.409
15	0.713	-1.635	1.418
16	0.586	-1.076	0.701
17	0.880	-0.699	1.164
18	1.735	0.187	-1.741
19	2.561	0.113	-1.835
20	1.801	1.457	0.593
21	1.365	0.502	1.204
22	2.949	-0.008	-0.253
23	3.130	-1.055	0.897

[CTR en 10000ième]

```
> inertie$row.abs
```

	Axis1	Axis2	Axis3
1	1067	347	1302
2	0	326	13
3	169	67	17
4	66	405	9
5	84	1083	541
6	589	480	155
7	850	708	31
8	217	438	274
9	289	975	276
10	515	1024	4
11	167	1904	31
12	609	69	1677
13	244	2	329
14	7	63	68
15	75	718	820
16	51	311	200
17	114	131	553
18	443	9	1236
19	966	3	1374
20	478	570	143
21	275	68	591
22	1281	0	26
23	1444	298	329

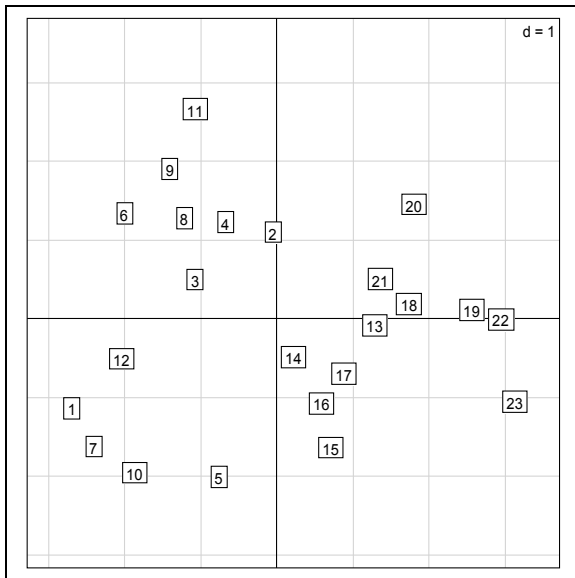
[Qlt en 10000ième]

```
> inertie$row.re
```

	Axis1	Axis2	Axis3	con.tra
1	-5975	-1068	-2634	878
2	-10	4836	129	182
3	-7332	1591	260	114
4	-1994	6672	-98	164
5	-948	-6695	-2202	437
6	-6224	2785	-591	465
7	-6356	-2907	84	658
8	-3728	4140	-1701	286
9	-2274	4212	785	625
10	-4653	-5085	-15	544
11	-1359	8497	91	605
12	-4712	-294	4690	635
13	6643	-27	-3231	181
14	556	-2848	-2028	60
15	945	-4975	3742	389
16	1612	-5446	2310	154
17	2909	-1836	5094	193
18	4939	57	-4972	442
19	6551	13	-3365	725
20	5421	3547	588	434
21	4490	606	3494	301
22	9644	0	-71	654
23	8109	-920	667	876

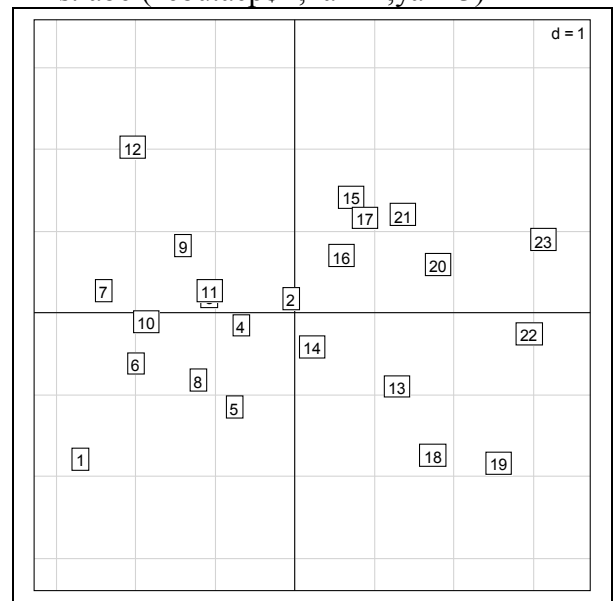
Plan $F_1 F_2$

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

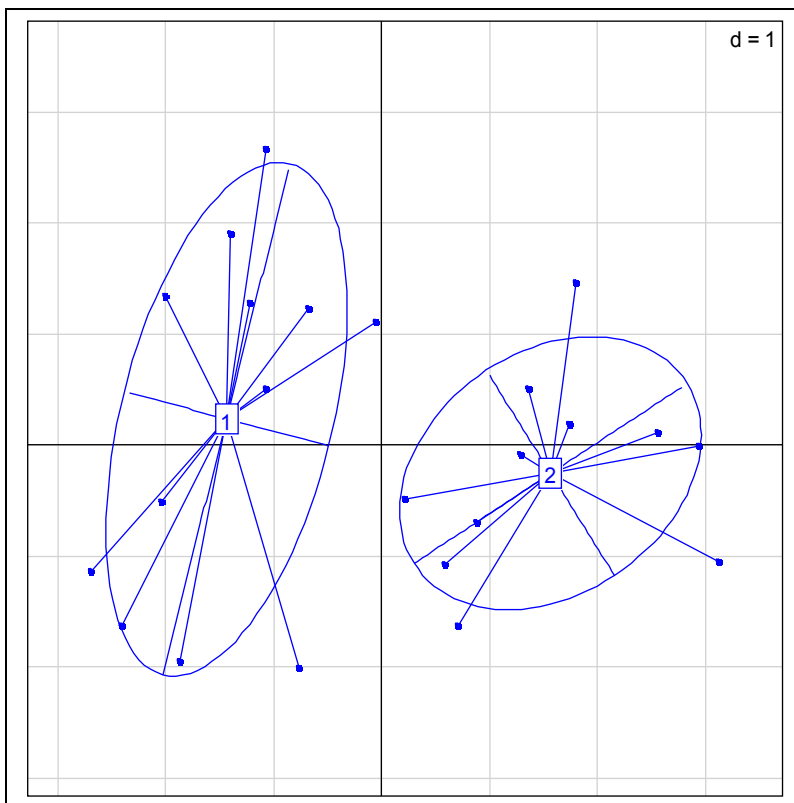


Plan $F_1 F_3$

> s.label(zebu.acp\$li,xax=1,yax=3)



> s.class(dfxy=zebu.acp\$li,fac=race,col=col,xax=1,yax=2)



IX ANALYSE FACTORIELLE DES CORRESPONDANCES

Objectifs et données pour l'AFC

Cette technique s'applique à des tableaux de contingence croisant deux variables qualitatives avec de nombreuses modalités chacune. Les données sont donc les effectifs des individus croisant deux modalités données. Pour de tels tableaux nous disposons du test d'indépendance du χ^2 .

L'**objectif** est ici de faire une synthèse de l'ensemble du tableau afin de répondre aux questions :

- **Pour une variable donnée**, certaines modalités sont-elles proches ou éloignées.
 - La proximité de deux modalités se mesure en comparant leur distribution par rapport à l'autre variable.
Par exemple, yeux bleus et verts sont proches si les deux groupes ont les mêmes distribution de couleurs de cheveux.
- **Entre les deux variables**, certaines modalités « s'attirent-elles » davantage ou au contraire « se repoussent ».
 - On compare la fréquence observée par rapport à la fréquence attendue sous l'hypothèse d'indépendance, si la fréquence observée est plus forte il y a une plus forte association entre les deux et inversement.
Par exemple, les yeux bleus et les cheveux blond « s'attirent », au contraire des yeux noirs et des cheveux blond.

Remarque : L'AFC n'a d'intérêt que si il y a dépendance entre les deux variables, en cas contraire elle n'apporte pas d'information.

Le **tableau** se présente sous la forme :

		variable qualitative 2			
		modalité 1 ...	modalité j	modalité J	
variable qualitative 1	modalité 1	n_{11}	n_{1j}	n_{1J}	
	...				
	modalité i	n_{i1}	n_{ij}	n_{iJ}	
	...				
	modalité I	n_{I1}	n_{Ij}	n_{IJ}	

On constate la symétrie du tableau contrairement au tableau utilisé en ACP.

Exemple 1 (trivial)

On examine la répartition des couleurs de cheveux et d'yeux.

Tableau de contingence

\cheveux	blond	roux	brun	total
yeux				
bleu	20	5	5	
vert	0	15	5	
marron	5	5	40	
total				

Tableau théorique sous H_0 (Indépendance)

\cheveux	blond	roux	brun	total
yeux				
bleu				
vert				
marron				
total				

Test du χ^2 d'indépendance

Fiche 5 – Principe de l'AFC

Pour mesurer les distances entre modalités, il est nécessaire de calculer au préalable la distribution de chaque modalité d'une variable en fonction de l'autre variable.

On définit ainsi **les profils colonnes et les profils ligne qui sont les distributions respectives des modalités des deux variables.**

Un profil ligne (colonne) est déterminé en divisant la ligne (colonne) par le total de la ligne (colonne).

Profils lignes

\cheveux yeux	blond	roux	brun	total
bleu				
vert				
marron				
total				

Profils colonnes

\cheveux yeux	blond	roux	brun	total
bleu				
vert				
marron				
total				

Chaque profil est alors assimilé à un point de coordonnées les proportions par rapport aux modalités de l'autre variable. Le nuage des profils ligne est alors projeté sur des axes factoriels en conservant le maximum d'inertie et il en est de même pour le nuage des profils colonne.

Les principales différences avec l'ACP sont :

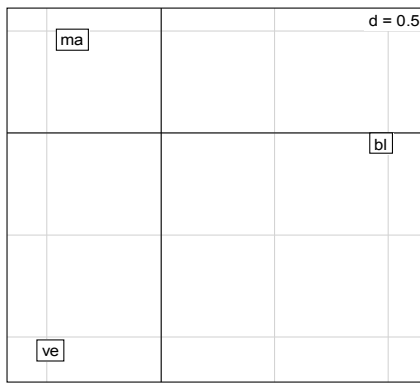
- **Lignes et colonnes sont transformées au préalable en profil et jouent un rôle symétrique.**
- La distance entre deux profils est calculée à l'aide de la **distance du χ^2** (distance entre deux distributions) :

$$d^2_{\chi^2}(i, i') = \sum_{j=1}^J \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

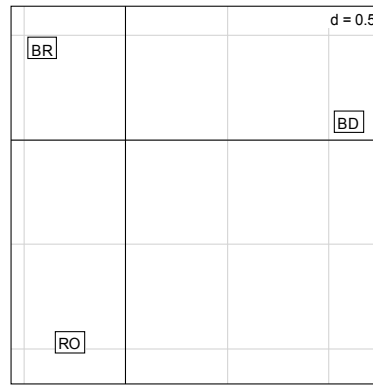
- Les profils sont tous dans un hyperplan car la somme de leurs coordonnées est 1. On a donc un axe de moins qu'en ACP et une valeur propre en moins.
- **Les valeurs propres (inertie projetée sur l'axe) sont inférieures à 1.**
- Les deux nuages représentent des profils et il est d'**usage de représenter les deux nuages dans un même plan** (le profil d'une modalité est « quasi » le barycentre des profils des modalités de l'autre variable).
- **La proximité entre modalités des deux variables indique une attirance entre ces modalités** (l'effectif observé est supérieur à celui attendu sous H_0).
- **La proximité entre modalités d'une même variable indique que les distributions sont voisines** pour ces deux modalités au regard de l'autre variable.

Exemple 1 : Les résultats de l'ajustement donnent :

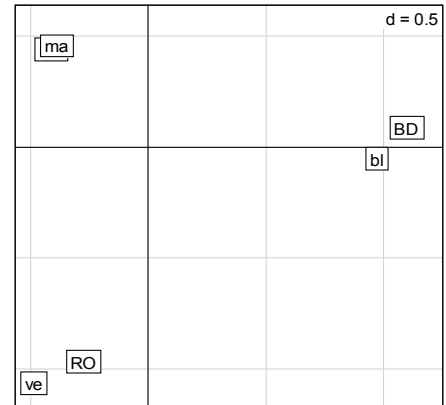
```
> afc$eig
[1] 0.4050000 0.3333333
> (afc$eig/sum(afc$eig))
[1] 0.5485327 0.4514673
```



Profils ligne dans le plan $F_1 F_2$



Profils colonne



Projection simultanée

Exemple 2 : CSP

Le tableau décrit la consommation annuelle en francs d'un ménage pour différentes denrées alimentaires en 1972. MA, EM, CA indiquent la catégorie socio-professionnelle et 2,3,4,5 la taille du foyer.

```
> csp=read.table("csp.txt")
      pain legu frui vian vola lait vin
MA2   332  428  354 1437  526  247 427
EM2   293  559  388 1527  567  239 258
...
```

Statistiques élémentaires

```
chisq.test(csp)
      Pearson's Chi-squared test
X-squared = 1290.386, df = 66, p-value < 2.2e-16
```

Calcul des profils :

Profils ligne

```
> round(csp/apply(csp,1,sum),2)
      pain legu frui vian vola lait  vin
MA2   0.08 0.15 0.10 0.40 0.15 0.06 0.07
EM2   0.07 0.15 0.11 0.37 0.18 0.04 0.08
CA2   0.10 0.14 0.08 0.37 0.13 0.08 0.10
MA3   0.09 0.15 0.10 0.36 0.14 0.08 0.09
EM3   0.07 0.14 0.11 0.39 0.19 0.04 0.06
MA4   0.12 0.14 0.08 0.35 0.14 0.09 0.09
EM4   0.09 0.14 0.10 0.37 0.15 0.08 0.08
CA4   0.07 0.13 0.11 0.40 0.19 0.05 0.05
MA5   0.12 0.14 0.08 0.34 0.14 0.09 0.09
EM5   0.10 0.17 0.09 0.35 0.15 0.09 0.05
CA5   0.07 0.15 0.12 0.37 0.16 0.08 0.04
```

Profils colonne

```
> round(t(t(csp)/apply(t(csp),1,sum)),2)
      pain legu frui vian vola lait  vin
MA2  0.06 0.05 0.06 0.06 0.05 0.06
EM2  0.05 0.06 0.06 0.07 0.06 0.06
CA2  0.07 0.09 0.09 0.09 0.10 0.05
MA3  0.08 0.06 0.06 0.07 0.06 0.08
EM3  0.07 0.07 0.07 0.07 0.06 0.07
CA3  0.08 0.10 0.11 0.10 0.12 0.06
MA4  0.10 0.08 0.06 0.07 0.07 0.10 0.09
EM4  0.09 0.08 0.08 0.08 0.08 0.09 0.09
CA4  0.07 0.09 0.10 0.10 0.12 0.07 0.06
MA5  0.12 0.09 0.07 0.08 0.08 0.12 0.11
EM5  0.11 0.11 0.09 0.09 0.09 0.12 0.07
CA5  0.10 0.12 0.15 0.12 0.12 0.13 0.06
```

Poids des lignes et colonnes

```
> round(apply(csp,1,sum)/sum(csp),2)
      MA2 EM2 CA2 MA3 EM3 CA3 MA4 EM4 CA4 MA5 EM5 CA5
0. 0.06 0.09 0.07 0.07 0.10 0.08 0.08 0.10 0.09 0.10 0.12

> round(apply(csp,2,sum)/sum(csp),2)
      pain legu frui vian vola lait  vin
0.  0.14 0.10 0.37 0.16 0.07 0.07
```

Comment décrire ce tableau ? Quelles sont les relations entre variables et modalités de chaque variables ? Comment présenter ce tableau à un public non averti ?

Fiche 6 – Aides à l'interprétation

1. Valeurs propres λ et choix des axes

Pour définir le nombre d'axes étudiés, on étudie les valeurs propres obtenues. Chaque valeur propre correspond à la part d'inertie projeté sur un axe donné.

Remarques importantes:

- La somme des valeurs propres est toujours égale à l'inertie totale du nuage. On caractérise ainsi chaque axe par le % d'inertie qu'il permet d'expliquer.
- En AFC, les **valeurs propres sont toutes inférieures à 1**.

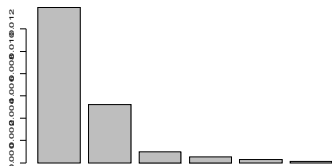
On ne retient donc que les axes avec les plus fortes valeurs propres. Le choix des axes retenus est un peu délicat. On peut donner quelques règles :

- **Règle du coude** : On observe souvent de fortes valeurs propres au départ puis ensuite de faibles valeurs avec un décrochage dans le diagramme. On retient les axes avant le décrochage.
- **Règle de l'inertie minimale** : On sélectionne les premiers axes afin d'atteindre un % donné d'inertie expliquée (70% par exemple).
- **Règle du bon sens** : On analyse les plans et axes et on ne retient que ceux interprétables.

Exemple 2 CSP

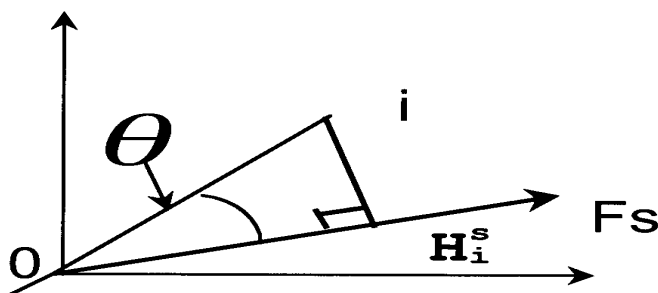
Avec ade4: réaliser l'afc `>afc=dudi.coa(tableau)`
Extraire les valeurs propres `>afc$eig`
Calcul des % `>afc$eig/sum(afc$eig)`

```
> round(afc$eig,3)
[1] 0.014 0.005 0.001 0.001 0.000 0.000
> round(afc$eig/sum(afc$eig)*100)
[1] 66 25 5 2 1 1
```



2. Qualité de représentation *qlt*

Les profils projetés dans un plan factoriel ne sont pas forcément correctement représentés. Dans ce cas, les interprétations sont erronées. Il est indispensable de vérifier la bonne représentation au préalable.



Qualité de représentation qlt :

Si l'angle θ est grand, le point initial est éloigné de sa projection. On utilise le paramètre $\cos^2 \theta$ pour caractériser la qualité de représentation (qlt) sur un axe.

- Plus qlt_i est proche de 1 plus il est bien représenté.
- Plus qlt_i est proche de 0 plus il est mal représenté.
- Dans un plan, on calcule la somme des deux qlt , par exemple $qlt_{F_1} + qlt_{F_2}$ pour le plan $F_1 F_2$.
- qlt correspond en fait au rapport de l'inertie du projeté sur l'inertie du point initial.

Qualité globale : Dans un plan donné, on définit également la qualité globale comme le pourcentage d'inertie qu'explique le plan. C'est par rapport à cette qualité globale que l'on évalue la qlt d'un profil.

Bilan : On commencera donc toujours l'analyse d'un plan factoriel en précisant l'existence (ou non) de profils mal représentés et en justifiant par les qlt .

3. Contribution ctr

Lors de la construction d'un axe factoriel, certains profils ont des rôles plus importants. On calcule un paramètre appelé contribution, ctr , qui permet de calculer cette influence.

Définition: La contribution ctr est définie comme la proportion de l'inertie de l'axe expliquée par le profil pour un axe donné.

Règles d'interprétation :

- L'analyse se fait axe par axe, en parallèle sur les profils ligne et colonne.
- Plus ctr est grande, plus l'influence du profil est grande. On ne retient donc que les plus fortes valeurs (il y a souvent un décrochage après quelques valeurs).
- ctr est considéré comme positif si le profil est dans la partie positive de l'axe.
- ctr est considéré comme négatif si le profil est dans la partie négative de l'axe.
- Le bilan des ctr peut être présenté pour un axe donné sous forme d'un tableau avec les principes $ctr +$ et $-$ des profils, en précisant la valeur de ctr :

ctr axe F_1 :	-	+
Profils colonnes :		

Profils lignes :

On réalise ensuite une interprétation.

Sous R, ces paramètres sont obtenus avec les commandes suivantes :

Pour les lignes	<code>>inertieL<-inertia.dudi(acp, row.inertia=TRUE)</code>
ctr des lignes en %	<code>>inertieL\$row.abs/100</code>
qlt des lignes en %	<code>>inertieL\$row.rel/100</code>

Pour les colonnes	<code>>inertieC<-inertia.dudi(acp, col.inertia=TRUE)</code>
ctr des colonnes en %	<code>>inertieC\$col.abs/100</code>
qlt des colonnes en %	<code>>inertieC\$col.rel/100</code>

Fiche 7 – Représentation graphique des profils

Les profils ligne (respectivement colonne) sont associés à des points de l'espace dont les coordonnées sont les distributions conditionnelles. On peut mesurer la **distance entre ces profils** en utilisant la distance du χ^2 entre ces deux points.

La construction des composantes principales conduit à **rendre minimale la déformation des distances entre profils lorsque l'on projette les profils dans le plan factoriel $F_1 F_2$** . Ainsi les distances que l'on observe entre les profils dans le plan factoriel sont globalement les plus proches possible des distances réelles entre ces profils.

L'analyse des plans factoriels permet ainsi d'observer les profils proches entre eux (distribution similaire) ou au contraire éloignés. Il est ainsi possible de construire des groupes, d'observer des tendances ...

Les règles de lecture des plans factoriels sont :

- **Les profils ligne et colonne sont projetés simultanément** afin notamment d'observer les modalités des deux variables présentant de fortes (attraction) ou faibles (répulsion) associations.
- **Seuls les profils bien représentés sont pris en compte** dans l'interprétation.
 - On calcule la somme des qtl dans le plan et on vérifie que cette somme n'est pas trop faible par rapport à la qualité moyenne du plan.
- On réalise le **bilan en positif et en négatif des profils qui ont la plus forte contribution** pour un axe donné.
 - On donne ainsi en parallèle sur les lignes et colonnes une signification concrète à ces axes en terme d'attraction, de similarité, entre modalités ou en tendance particulière.
- **On réalise des groupes**, à l'aide éventuelle de la fonction `s.class`, dans le cas de groupes préexistants (homme-femme par exemple) ou on construit arbitrairement ces groupes en raison des proximités entre profils.
- L'utilisation de **profils supplémentaires** non utilisés dans l'ajustement mais a posteriori permet également d'éclairer l'analyse.

Guide pratique de l'analyse AFC

- **Etape 1** : Sélection des axes et des plans retenus principalement par rapport aux valeurs propres.
- **Etape 2** : Projection des profils ligne et colonne dans un plan donné ($F_1 F_2$ en premier)
 - Examen des *qtl* dans le plan pour éliminer les profils mal représentés
 - Bilan des *ctr* pour un axe afin de donner un sens à cet axe (opposition, tendance ...)
 - Topographie des profils afin d'identifier des groupes, des oppositions, des tendances notamment à l'aide de la fonction `s.class`
 - Utiliser ses connaissances sur le sujet pour proposer des explications sur les résultats de l'analyse
 - Utiliser des profils supplémentaires ou des profils type (moyenne des H et des F par exemple)

Exemple 2 csp

Profils lignes et colonnes

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

Coordonnées profils ligne

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

	Axis1	Axis2
MA2	-0.10	-0.14
EM2	0.04	-0.02
CA2	0.08	-0.09
MA3	-0.13	-0.05
EM3	-0.08	-0.01
CA3	0.16	-0.05
MA4	-0.15	0.02
EM4	-0.05	-0.01
CA4	0.17	-0.02
MA5	-0.17	0.03
EM5	-0.03	0.12
CA5	0.11	0.10

[CTR en %]

```
> round(inertie$row.abs/100)
```

	Axis1	Axis2
MA2	4	24
EM2	1	0
CA2	4	14
MA3	8	3
EM3	3	0
CA3	18	4
MA4	13	1
EM4	2	0
CA4	20	1
MA5	18	2
EM5	1	27
CA5	10	24

[QLT en %]

```
> round(inertie$row.re/100)
```

	Axis1	Axis2	con.tra
MA2	-29	-64	10
EM2	31	-5	2
CA2	34	-49	7
MA3	-86	-11	6
EM3	-80	-2	3
CA3	87	-7	14
MA4	-94	2	9
EM4	-81	-2	1
CA4	90	-1	14
MA5	-91	3	13
EM5	-5	89	8
CA5	46	43	14

colonne

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

	Comp1	Comp2
pain	-0.19	0.04
legu	0.01	0.07
frui	0.13	0.01
vian	0.04	-0.03
vola	0.12	-0.02
lait	-0.19	0.15
vin	-0.23	-0.19

[CTR en %]

```
> round(inertie$col.abs/100)
```

	Comp1	Comp2
pain	22	3
legu	0	13
frui	11	0
vian	4	5
vola	17	1
lait	18	31
vin	29	48

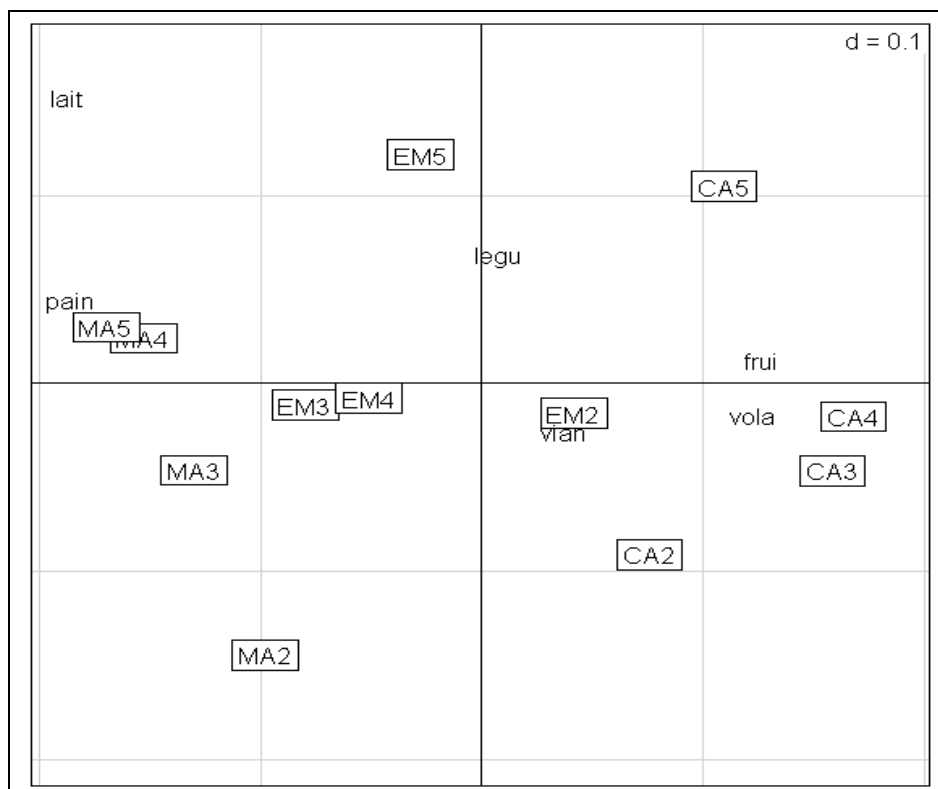
[QLT en %]

```
> round(inertie$col.re/100)
```

	Comp1	Comp2	con.tra
pain	-87	5	17
legu	1	66	5
frui	79	1	10
vian	51	-25	5
vola	84	-2	13
lait	-58	38	20
vin	-61	-38	31

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

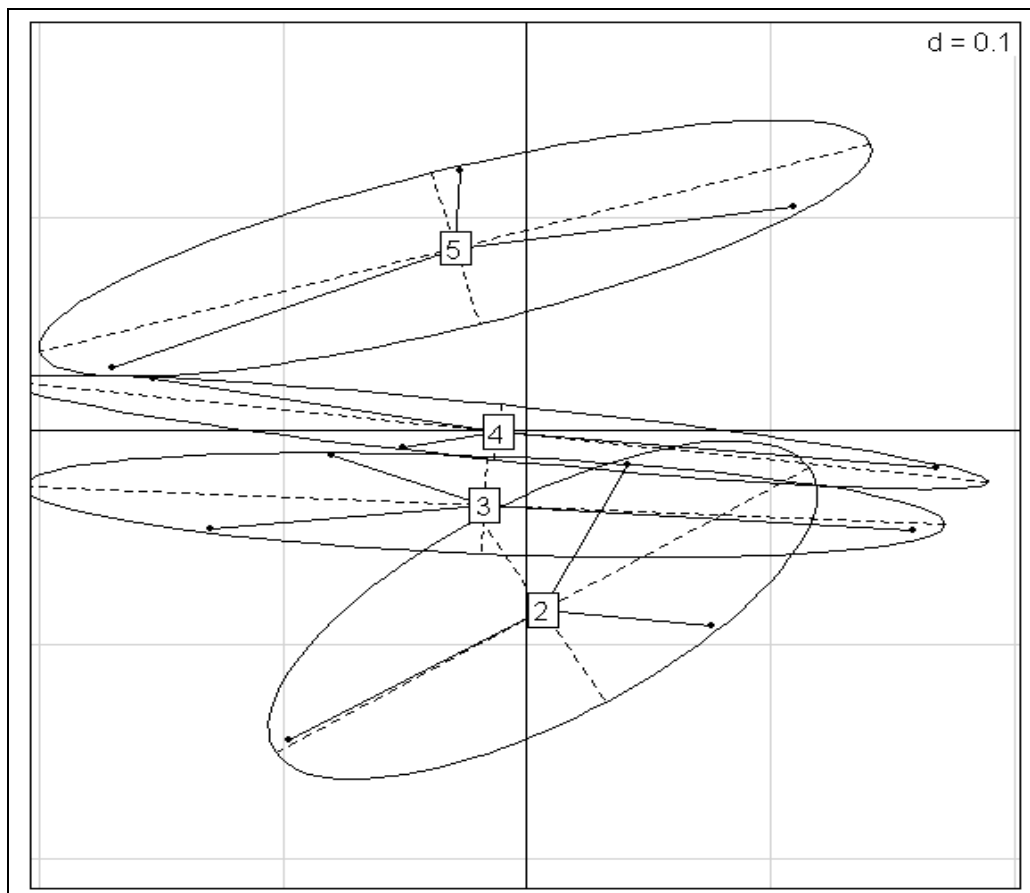
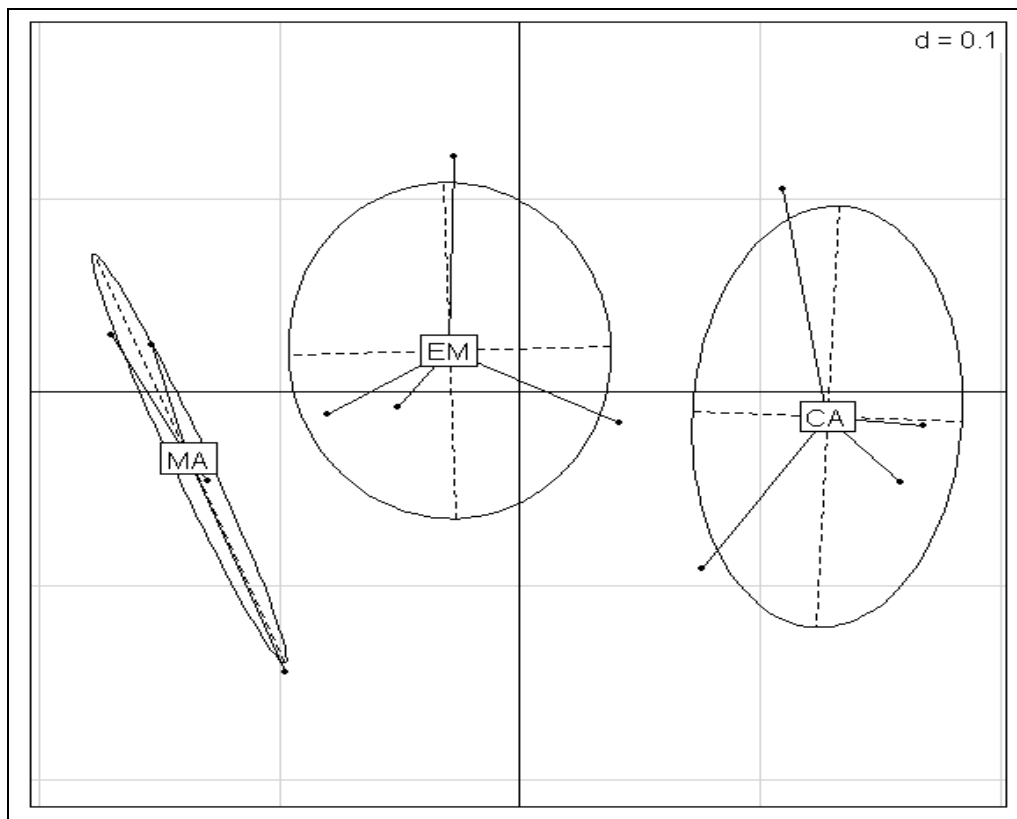
```
> s.label(afc$co,xax=1,yax=2,add.plot=T,boxes=F)
```



```

> pro <- as.factor(c("MA", "EM", "CA", "MA", "EM", "CA", "MA", "EM", "CA", "MA", "EM", "CA"))
> nb <- as.factor(c(2,2,2,3,3,3,4,4,4,5,5,5))
> s.class(afc$li,xax=1,yax=2,fac=pro)
> s.class(afc$li,xax=1,yax=2,fac=nb)

```



X CLASSIFICATION

Objectifs et données :

On dispose de n objets caractérisés par des variables ou par les distances entre ces objets (tableau de distance). L'objectif de la classification est de construire des groupes d'objets homogènes avec comme principe :

- au sein d'un groupe, les objets sont les plus similaires possibles,
- entre les groupes, les différences sont le plus grandes possibles.

Lien avec l'AFC et l'ACP : L'analyse factorielle et la classification sont souvent conduites en parallèle. On peut ainsi effectuer une classification sur les profils, les variables, les individus. Les résultats de ces classifications peuvent ainsi être mis en parallèle avec l'analyse factorielle et améliorer l'interprétation.

Les méthodes de classification reposent sur la notion de similarité ou de dissimilarité (distance) entre les objets que l'on souhaite regrouper en classes homogènes. On est donc souvent conduit à manipuler des tableaux de distance.

Soit E un ensemble de n objets.

- ❖ Une application d de $E \times E$ dans $\mathbb{R}^+$ est appelée **dissimilarité** si elle vérifie pour tout couple (i, j) de E^2 :
 - $d(i, j) = d(j, i)$
 - $d(i, j) \geq 0$ et $d(i, i) = 0$
 - si de plus $d(i, j) \leq d(i, k) + d(k, j)$, d est appelée **distance**
- ❖ Une application s de $E \times E$ dans $\mathbb{R}^+$ est appelée **similarité** si elle vérifie pour tout couple (i, j) de E^2 :
 - $s(i, j) = s(j, i)$
 - $s(i, j) \geq 0$
 - $s(i, i) \geq s(i, j)$

Exemples de distance et similarité

- **Données numériques (individus et variables quantitatives)**

Pour des objets décrits par des variables quantitatives, on peut utiliser la distance euclidienne classique (ACP)

- **Tableaux de contingence (profils)**

Pour des profils de modalités dans un tableau de contingence, on utilisera la distance du χ^2 (AFC).

- **Pour des données binaires** (présence/absence d'une caractéristique)

Plusieurs indices de similarité existent, comme l'indice de Jaccard :
$$J = \frac{a}{a + b + c}$$

avec a le nombre de caractéristiques communes,
 b celui possédé par i et pas par j
 c celui possédé par j et pas par i .

Exemple : Construire le tableau des indices de Jaccard pour les individus A à E

	C1	C2	C3	C4
A	0	1	1	0
B	1	1	1	1
C	1	0	0	1
D	1	1	1	0

- **Distance génétique** (Chessel et al., <http://pbil.univ-lyon1.fr/R/enseignement.html>)

Soit A un tableau de fréquences alléliques avec t lignes (populations) et m colonnes (formes alléliques). Soit v le nombre de loci. Le locus j a $m(j)$ formes alléliques.

$$m = \sum_{j=1}^v m(j)$$

1 — Distance de Nei ²² (Voir ²³) :

$$D_1(a, b) = -Ln \left(\frac{\sum_{k=1}^v \sum_{j=1}^{m(k)} p_{aj}^k p_{bj}^k}{\sqrt{\sum_{k=1}^v \sum_{j=1}^{m(k)} (p_{aj}^k)^2} \sqrt{\sum_{k=1}^v \sum_{j=1}^{m(k)} (p_{bj}^k)^2}} \right)$$

Principe de la Classification Ascendante Hiérarchique (CAH)

La CAH est une méthode itérative.

1. Au départ chaque objet représente un groupe.
2. A chaque étape est construite une nouvelle partition de moins en moins fine, contenant une classe de moins à chaque étape. A la première étape, on regroupe les deux objets les plus proches et on ne dispose plus alors que de $n - 1$ classes.

Cette méthode aboutit à un emboîtement de partitions visualisé graphiquement par un arbre hiérarchique indicé (dendrogramme).

On regroupe à chaque étape les 2 objets ou les 2 groupes d'objets dont **la ressemblance est la plus forte**. Cette méthode nécessite ensuite de **recalculer la distance des objets restant au groupe ainsi formé**. Ces méthodes de calculs s'appellent le **critère d'agrégation** et définissent la méthode :

- **Critère du saut minimal** : la distance entre h et h' est celle définie par les deux éléments les plus proches.
- **Critère du diamètre ou distance maximale**: la distance entre h et h' est celle définie par les deux éléments les plus éloignées
- **Critère de la moyenne (UPGMA)**: la distance entre h et h' est celle définie par la distance moyenne entre les éléments de h et de h' .
- **Critère de Ward** : Le critère est calculé pour h et h' par $\Delta_{hh'} = \frac{p_h \times p_{h'}}{p_h + p_{h'}} d^2(g_h, g_{h'})$. Le critère de Ward correspond à l'augmentation de l'inertie intra résultant du regroupement de h et h' .

Dans la pratique, le critère de Ward est le plus largement utilisé en CAH.

Liens sur internet :

<http://biom3.univ-lyon1.fr/R/stage/stage7.pdf>

http://zoonek2.free.fr/UNIX/48_R_2004/06.html

<http://www.lsp.ups-tlse.fr/Carlier/Hyper/plaq2html/node130.html>

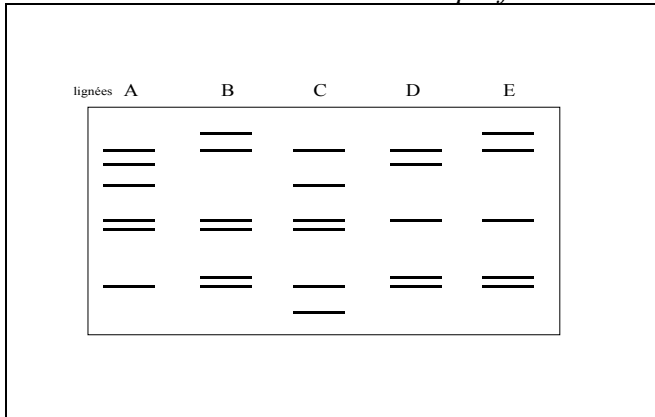
http://www.univ-orleans.fr/SCIENCES/LIFO/Manifestations/FD2004/FDD_Vrain.pdf

Fiche 8 – Exemple manuel de CAH

Un laboratoire veut étudier la ressemblance génétique de cinq lignées de sorgho (A, B, C, D, E). Pour cela, il réalise une analyse par RFLP avec l'enzyme de restriction *Eco* RI et une sonde d'origine inconnue.

Les fragments d'ADN ainsi amplifiés sont ensuite séparés par électrophorèse. Les résultats de l'électrophorèse donnent les profils suivants:

NB : Deux lignées de sorgho présentent autant de fragments d'ADN identiques que de bandes révélées à la même hauteur sur les profils de l'électrophorèse.



- Question 1 : Construisez la matrice de similarité entre les cinq lignées de sorgho à partir du coefficient de similarité de Dice (S_{xy}) défini ci-dessous.

$$S_{xy} = \frac{N_{xy}}{N_x + N_y}$$

N_x = nombre de bandes de la lignée X

N_y = nombre de bandes de la lignée Y

N_{xy} = nombre de bandes identiques entre les lignées X et Y

Similarité	A	B	C	D	E
A					
B					
C					
D					
E					

- Question 2 :

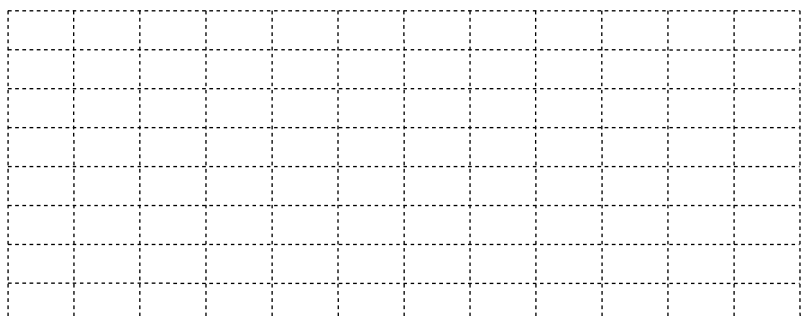
Déterminez la matrice de dissimilarité en utilisant la fonction de similitude linéaire : $D_{xy} = 1 - S_{xy}$

Dissimilarité	A	B	C	D	E
A					
B					
C					
D					
E					

- Question 3 : Effectuez sur la matrice de dissimilarité la CAH en utilisant le critère du saut maximal.

Dij					Dij					Dij				

Dendrogramme



Fiche 9 – Exemple de CAH avec ade4

Exemple zebu

```
> library(ade4)
> acp = dudi.pca(zebu[,1:6])
```

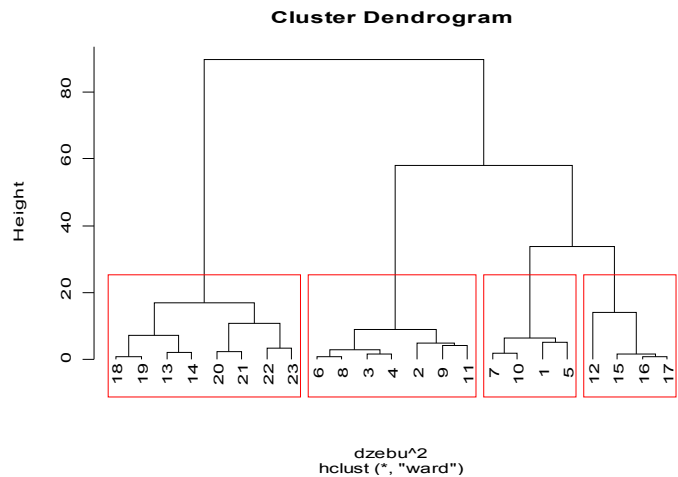
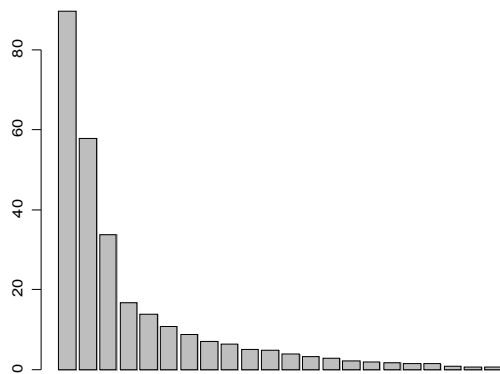
CAH sur les individus

```
> dzebu = dist.dudi(acp, amongrow=TRUE)
> round(dzebu,1)      (extrait pour les 10 premiers individus)
  1  2  3  4  5  6  7  8  9 10
2  4.1
3  3.1 1.8
4  3.7 1.6 1.2
5  2.2 3.6 2.9 3.5
6  3.0 2.6 1.7 1.4 3.7
7  2.5 4.1 2.6 3.4 2.4 3.2
8  3.2 2.1 1.4 0.9 3.4 0.8 3.4
9  4.3 2.1 2.1 2.5 4.6 2.6 4.3 2.6
10 2.1 3.7 2.7 3.5 1.6 3.4 1.3 3.4 4.3

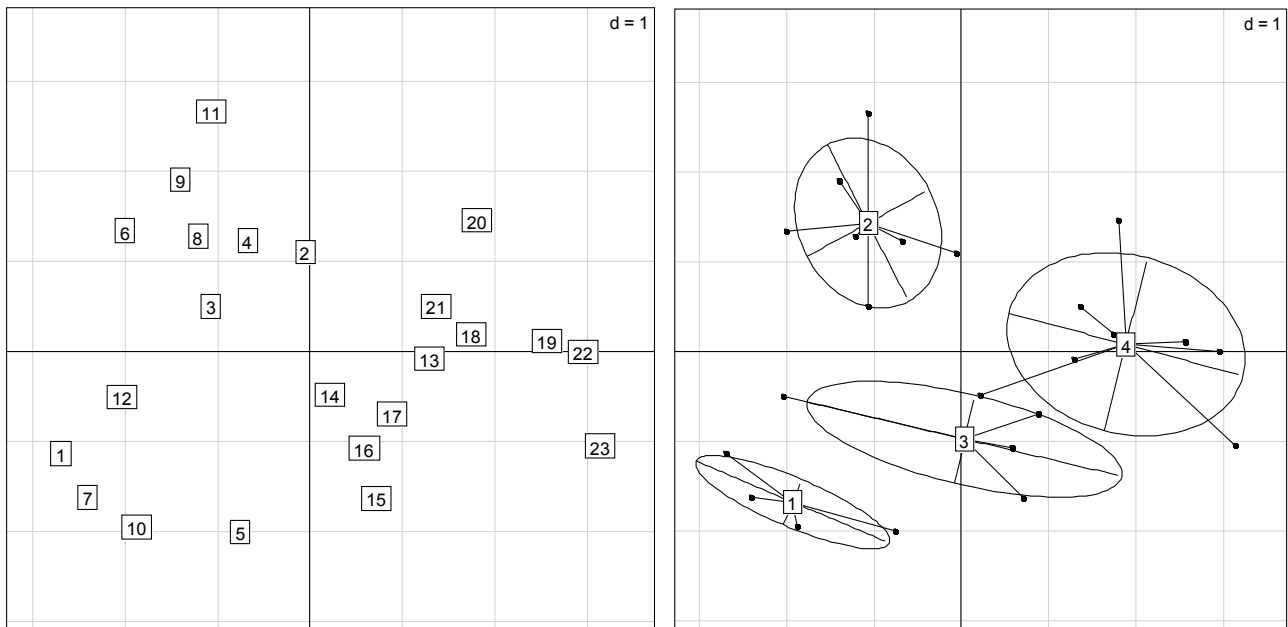
> cah=hclust(dzebu^2,method="ward")

> barplot(cah$height[order(cah$height,decreasing=TRUE)])

> plot(cah,hang=-1)
> rect.hclust(cah,h=20)
```



```
> cutree(cah,h=20)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
 1  2  2  2  1  2  1  2  2  1  2  3  4  4  3  3  3  4  4  4  4  4
> s.label(acp$li)
> s.class(dfxy=acp$li,fac=as.factor(cutree(cah,h=20)))
```



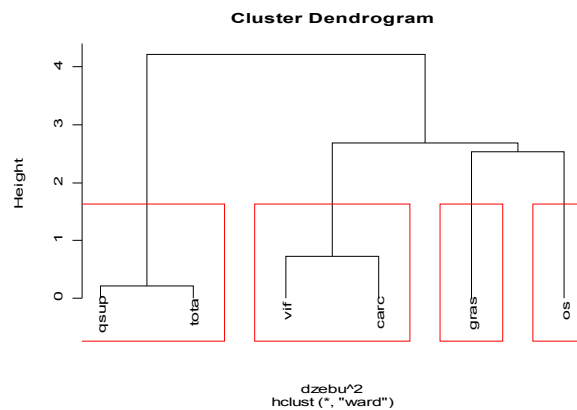
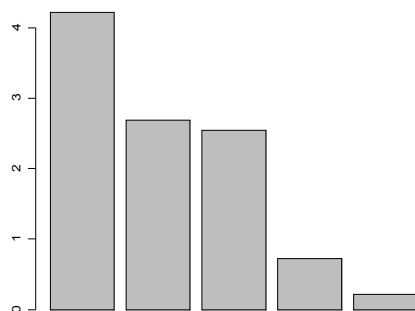
CAH sur les variables

```
> dzebu = dist.dudi(acp, amongrow=FALSE)
> round(dzebu,1)      (extrait pour les 10 premiers individus)
   1   2   3   4   5   6   7   8   9  10
2  4.1
3  3.1 1.8
4  3.7 1.6 1.2
5  2.2 3.6 2.9 3.5
6  3.0 2.6 1.7 1.4 3.7
7  2.5 4.1 2.6 3.4 2.4 3.2
8  3.2 2.1 1.4 0.9 3.4 0.8 3.4
9  4.3 2.1 2.1 2.5 4.6 2.6 4.3 2.6
10 2.1 3.7 2.7 3.5 1.6 3.4 1.3 3.4 4.3

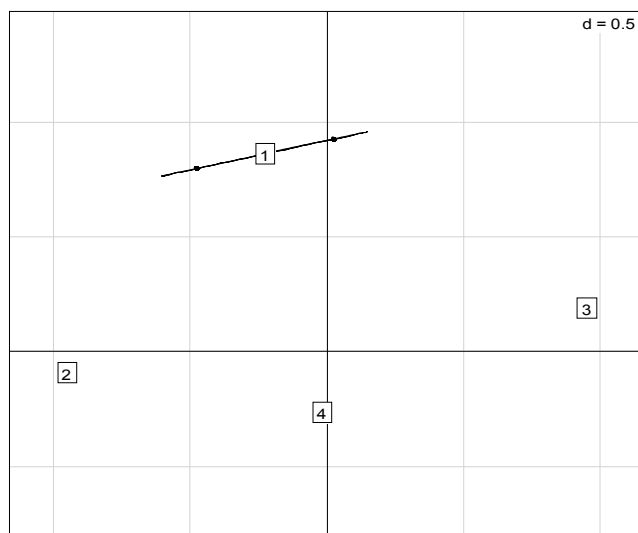
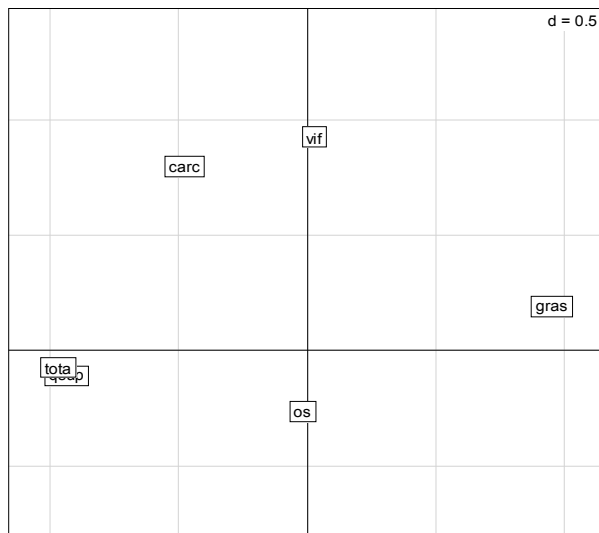
> cah=hclust(dzebu^2,method="ward")

> barplot(cah$height[order(cah$height,decreasing=TRUE)])

> plot(cah,hang=-1)
> rect.hclust(cah,h=1)
```



```
> cutree(cah,h=1)
vif carc qsup tota gras os
1 1 2 2 3 4
> s.label(acp$co)
> s.class(dfxy=acp$co,fac=as.factor(cutree(cah,h=1)))
```



Compléments pour l'analyse de données sous R

FactoMineR

Sous R, taper : `> source(" http://factominer.free.fr/install-facto.r ")`
(sous réserve d'accès à internet)

Utilisation de FactoMineR sous Rcmdr

Sélectionner le tableau sous Rcmdr puis dans le menu FactoMineR choisir la méthode.

- Vous pouvez choisir les variables et les individus en supplémentaires.
- Les résultats sont disponibles sous R dans le fichier de sortie noté par défaut res.
- Vous pouvez également sauver un fichier excel avec outputs.
- Pour afficher des plans factoriels, sélectionner dans FactoMineR les axes à représenter

Aides à l'interprétations à récupérer dans le fichier R res ou le fichier excel:

ACP

- valeurs propres : `res$eig`
- qlt et ctr : Sous R, les qlt sont sous la forme `res$var$cos2` ou `res$ind$cos2`, les ctr `res$var$contrib` ou `res$ind$contrib`.
- Tous sur les variables et individus : `res$var` ou `res$ind`
- variables supplémentaire : `res$quanti.sup` et `res$ind.sup`

AFC

- valeurs propres : `res$eig`
- qlt et ctr : Sous R, les qlt sont sous la forme `res$col$cos2` ou `res$ROW$cos2`, les ctr `res$col$contrib` ou `res$row$contrib`.
- Tous sur les profils ligne et colonne : `res$col` ou `res$var`
- profil supplémentaire : `res$col.sup` `res$row.sup`