Td-TP Ch 2 : Analyse en composantes principales

0. OBJECTIFS, DONNEES ET PRINCIPES DE L'*ACP*

Objectifs et données de 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}

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

Principe de la méthode *ACP*

Chaque individu est décrit ici par p variables quantitatives. Un individu est représenté par un point dont les coordonnées sont les valeurs prises par les p variables (espace à p 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.

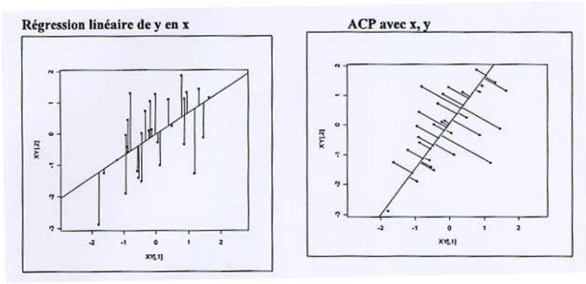
- **Pour les variables**

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.

M1 Ingénierie Statistique, Dpt de Mathématiques, Université de Nantes : Analyse de données (L. Bellanger)

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 $F1\ F2$, ou $F1\ F3$ ...

Interprétation graphique de l'*ACP*

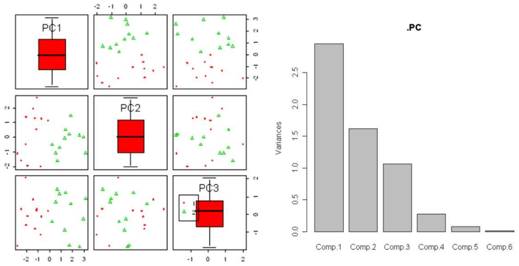


Remarque importante : En général, du fait de l'**hétérogénéité des variables initiales** et de leurs unités, **on réduit ces variables**. On parle alors d'***ACP* normée**.
Une variable est dite réduite quand sa variance vaut 1. De la sorte, chaque variable initiale aura une même importance dans l'analyse 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 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.

Fig gauche : Projection des individus dans les plans définis par les axes $F1\ F2\ F\ 3$
Fig droite : Projection de l'inertie sur les différents axes factoriels



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 ($F1\ F2$ 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)

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.

Sous R, avec ade4 [1]:

Réaliser l'*ACP* : `> library(ade4) ; acp=dudi.pca(tableau)`

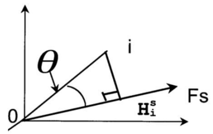
Extraire les valeurs propres : `> acp$eig`

Etude de l'inertie et calcul des % : `> acp$eig/sum(acp$eig)*100`

```
inertierow<-inertia.dudi(acp,row.inertia=TRUE)
inertiecol<-inertia.dudi(acp,col.inertia=TRUE)
names(inertierow) ; inertierow$TOT
```

2. QUALITE DE REPRESENTATION *qtl*

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 $qlt_{F1} + qlt_{F2}$ pour le plan $F1$ $F2$.
- *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 corrélations 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*.

[1] Il existe d'autres fonctions sous R permettant de réaliser une ACP telles que `PCA(FactoMineR)`.

M1 Ingénierie Statistique, Dpt de Mathématiques, Université de Nantes : Analyse de données (L. Bellanger)

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* :

Ctr axe F_1	-	+
variables		
individus		

On réalise ensuite une interprétation.

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

Pour les lignes (individus) `> inertieL<-inertia.dudi(acp, row.inertia=TRUE)`

ctr des lignes en % `> inertieL$row.abs/100`

qlt des lignes en % `> inertieL$row.rel/100`

Pour les colonnes (variables) `> inertieC<-inertia.dudi(acp, col.inertia=TRUE)`

ctr des colonnes en % `> inertieC$col.abs/100`

qlt des colonnes en % `> inertieC$col.rel/100`

Représentations graphique

1. 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 corréles.
- On synthétise chaque axe en précisant les variables qui contribuent le plus en positif ou en négatif (étude des ctr).

Exemple sous R avec ade4 :

```
> inertieV=inertia.dudi(acp,col.inertia=TRUE)
> s.corcircle(zebu.zebu.acp$co,xax=1,yax=2) # cercle des corrélations
```

2. **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** $F1$ $F2$. 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 ces individus.
- ● L'utilisation d'**individus supplémentaires** non utilisés dans l'ajustement mais a posteriori permet également d'éclairer l'analyse.

Sous R avec ade4 :

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

Composantes principales	[CTR en 10000ième]	[Qlt en 10000ième]
`> round(acp$li,3)`	`> inertie$row.abs`	`> inertie$row.re`

Plan *F1* *F2*	Plan ***F1*** ***F3***
`> s.label(acp$li,xax=1,yax=2)`	`> s.label(acp$li,xax=1,yax=3)`

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

I. EXERCICE « A LA MAIN »

$$T = \begin{bmatrix} -2 & 3 & -1 \\ -1 & 1 & 0 \\ 2 & -1 & -1 \\ 1 & -3 & 2 \end{bmatrix}$$

1. Calculer les moyennes
2. Matrice de variances-covariances S
3. Matrice de corrélations
4. *ACP* (T,Q=I_3,D=1/4 I_4) : déterminer les valeurs propres, les axes, les composantes principales.

M1 Ingénierie Statistique, Dpt de Mathématiques, Université de Nantes : Analyse de données (L. Bellanger)

II ETUDE D'UN NUAGE DE POINTS

1. Construire le nuage centré de 50 individus caractérisés par un couple de variables suivant une loi normale d'espérance (1,2) et de matrice de covariance $\Sigma = \begin{pmatrix} 1 & 1.5 \\ 1.5 & 4 \end{pmatrix}$ (fonction **`mvrnorm`** de la library `MASS`**)**.

```
> library(MASS) ; library(car)
> XY<-mvrnorm(50,mu=c(1,2),Sig=matrix(c(1,1.5,1.5,4),2,2))
> XY<-scale(XY,scale=FALSE)
> plot(XY,asp=1)
> XY<-as.data.frame(XY)
> colnames(XY)<-c("X","Y")
> points(mean(XY)[1],mean(XY)[2],col="green")
> ellipse(c(0,0),matrix(c(1,1.5,1.5,4),2,2),radius=2)
```

2. Construire une fonction qui pour un vecteur unitaire $u = (u_1, u_2)^T$

- calcule l'inertie projetée sur l'axe,
- calcule l'affixe de la projection de chaque individu sur l'axe,
- dans la même fenêtre
  - représenter l'histogramme des affixes et l'ajustement d'une loi normale (fonction **`dnorm`**)
  - représenter le nuage de points et le vecteur

```
> fct=function(X,u)
{
S=var(X)
vp=eigen(S)$vectors
In=t(u)%*%S%*%u # inertie projetée sur axe u
F=as.matrix(X)%*%u # affixes de la projection
par(mfrow=c(1,2))
histo=hist(F,proba=T,col=7,main="Histogramme des affixes",xlim=c(-10,10))
lines(seq(-10,10,0.1),dnorm(seq(-
10,10,0.1),mean(F,na.rm=TRUE),sd(F,na.rm=TRUE)),col="blue")
plot(XY,asp=1,main="Nuage de points")
arrows(0,0,u[1],u[2],length=0.05,angle=30,code=2,col="blue")
points(mean(X)[1],mean(X)[2],col="red")
liste=list(inertie=In,affixes=F)
return(liste)
}
```

D'où

```
> covXY<-var(XY)
> S=covXY ; S
> u1=(matrix(c(1/sqrt(2),1/sqrt(2)),byrow=F)) ; u1
> u2=(matrix(c(1/sqrt(2),-1/sqrt(2)),byrow=F)) ; u2
```

les vecteurs précédents sont unitaires et orthogonaux.

```
fct(XY, u1)
fct(XY, u2)
```

III. ANALYSE EN COMPOSANTES PRINCIPALES (*ACP*) PAR ALGEBRE MATRICIELLE

Pour cet exemple, nous prenons les données pp392 du livre de P. Legendre.
Lire les données et les transférer dans une matrice :

Ind	Var1	Var2
Obj1	2	1
Obj2	3	4
Obj3	5	0
Obj4	7	6
Obj5	9	2

```
> Y <- read.table("Ex_ACP_p_392.txt",row.names, h=T)
> Y.mat <- as.matrix(Y)
```

Centrer la matrice Y par colonne en appliquant aux colonnes la fonction `.scale`:

```
> Y.cent <- apply(Y.mat,2,scale,center=TRUE,scale=FALSE)
```

Note : le paramètre « 2 ».indique de calculer le centrage par colonne (et non par ligne)

Calculer la matrice de covariance :

```
> Y.cov <- cov(Y.mat)
# ou encore (opération équivalente) :
> Y.cov <- cov(Y.cent)
```

Calculer les valeurs propres et les vecteurs propres :

```
> Y.eig <- eigen(Y.cov) #
```

Qu'obtenez-vous de `Ycov.svd <- svd(Y.cov)` ?
Qu'obtenez-vous de `Y.svd <- svd(Y.cent)` ?

Verifier les valeurs propres et les vecteurs propres:

```
> Y.eig$values ; Y.eig$vectors
```

Transférer les vecteurs propres dans U (représentation préserve les distances euclidiennes entre les objets) :

```
> U <- Y.eig$vectors
```

Calculer la matrice F des composantes principales et vérifier le contenu de la matrice F :

```
> F <- Y.cent %*% U ; F
```

Diagramme de dispersion des 2 premières colonnes de F :

```
> plot(F[,1], F[,2], xlim=c(-4,4), ylim=c(-3,3), asp=1, xlab="Axe 1", ylab="Axe 2")
```

Notes : (-4,4) = bornes de 'abscisse, (-3,3) = bornes de 'ordonnée.
`asp=1` : le rapport des dimensions abscisse/ordonnée est .fixé à 1
pour obtenir un graphique qui représente correctement les distances entre les objets.
Voici une autre manière de faire. La fonction "`range`" permet de connaître la plage de variation des valeurs sur les axes 1 et 2.
On l'applique aux colonnes de F par la commande "`apply`" :

```
> F.range <- apply(F, 2, range) ; F.range
```

Créer les vecteurs "`xlim`" et "`ylim`" qui fourniront les valeurs limites des axes du graphique :

```
> xlim <- c(F.range[1,1], F.range[2,1])
> ylim <- c(F.range[1,2], F.range[2,2])
> plot(F[,1], F[,2], xlim=xlim, ylim=ylim, asp=1, xlab="Axe 1", ylab="Axe 2")
```

Ajouter au diagramme des .flêches représentant les 2 premières colonnes de la matrice U :

```
> arrows(x0=0,y0=0,U[,1]*3,U[,2]*3)
```

Note : pour cet exemple, les coordonnées des variables (tirées de U) sont multipliées par 3.

IV DVS

1. Soit X_C un tableau centré et X_{CR} sa forme réduite. Etudier la DVS de $\left(\mathbf{X_C}, diag\left(\frac{1}{\sigma_j^2}\right) \;, \frac{1}{n}\mathbf{I_n}\right)$ et $\left(\mathbf{X_{CR}}, \mathbf{I_p}, \frac{1}{n}\mathbf{I_n}\right)$.

2. Effectuer la DVS $(\mathbf{X}, \mathbf{I_p}, \mathbf{I_n})$ de $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 0 \\ -1 & 0 \end{pmatrix}$ et de $\mathbf{B} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$..

V ETUDE D'UN TABLEAU A L'AIDE D'UNE *ACP*

Partie A : calcul à la main

Le tableau de données ci-dessous est constitué de trois variables *x*, *y* et *z*, et de quatre individus A à D. On utilisera les valeurs exactes.

$$\begin{array}{c|ccc} & x & y & z \\ A & 1 & 0 & 0 \\ B & 1 & 2 & 0 \\ C & 2 & 2 & 2 \\ D & 0 & 0 & 2 \end{array}$$

EFFECTUER L'*ACP* NORMEE DU TABLEAU

1) Calculer le centre de gravité g_I du nuage.

2) Calculer le tableau centré réduit.

3) **a.** Calculer la matrice d'inertie S_I du nuage *N(I)*.
b. Que représente cette matrice ?
c. Quelle est l'inertie du nuage ?

4) Recherche des axes principaux d'inertie :
a. Déterminer les valeurs propres de S_I.
b. Vérifier votre résultat à l'aide de la question **3) c.**
c. Déterminer les deux premiers vecteurs propres.

5) **a.** Quelle est la contribution absolue de l'axe 1 à l'inertie du nuage ?
b. Quel est le taux d'inertie extrait par l'axe 1 ?
c. Quelle est la meilleure représentation plane ?

REPRESENTATION DES INDIVIDUS

1) Compléter dans le tableau ci-dessous les composantes principales (coordonnées des individus).

	Composantes Principales			**qlt** = cos² (/100)			ctr (/100)		
	F^1	F^2	F^3	F^1	F^2		F^1	F^2	
A									
B									
C									
D									

2) Définir la qualité de représentation de *i* sur l'axe $\mathbf{a}_1$ et compléter le tableau ci-dessus.
3) Compléter les contributions absolues des individus à l'inertie de l'axe $\mathbf{a}_1$?
4) Effectuer la représentation graphique du plan (1)-(2).

REPRESENTATION DES VARIABLES

1) Calculer les coordonnées des variables sur les différents axes et compléter le tableau.

2)

	coordonnées			**ctr**		(/100)
	G¹	**G²**		**G¹**	**G²**	
V1						
V2						
V3						

2) Définir la qualité de la représentation de la variable j sur les axes et compléter le tableau.

3) Effectuer la représentation graphique dans les différents plans.

INDIVIDUS ET VARIABLES SUPPLEMENTAIRES

Construire la représentation graphique de l'individus de coordonnées (0,2,0).
Construire la représentation graphique de la variable de coordonnées

(1,-1,1,-1,0).

Partie B : Calculs à l'aide du logiciel R.

Construire une fonction R permettant de déterminer pour un tableau T les valeurs propres ainsi que les composantes principales et qui représente le plan factoriel F^1F^2 pour les individus et les variables.

Partie C : Un second exemple

Reprendre les étapes du I (calcul manuel + vérification sous R) avec le tableau de données :

$$\mathbf{T}=\begin{pmatrix}2&0&0\\0&0&2\\1&1&0\\0&1&1\\1&0&1\\0&2&0\end{pmatrix}\ \text{ou}\ \mathbf{T}=\begin{array}{c}\\A\\B\\C\\D\\E\\F\\ \\\end{array}\begin{pmatrix}1&0&0\\1&1&0\\1&2&0\\0&2&1\\0&1&1\\0&0&1\\v1&v2&v3\end{pmatrix}$$

VI ETUDE D'EXEMPLES SOUS R

Exemple 1 : 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**	**sucre**	**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

Utiliser les résultats ci-dessous pour justifier vos réponses.

a. Justifier l'utilisation d'une *ACP*.
b. Expliquer les différences obtenues entre une *ACP* normée et non normée?
c. Déterminer trois groupes de variables qui présentent des corrélations entre elles (r>0.5).
d. Que représentent les ellipses dans la représentation en 3D.
e. Expliquez les différences entre les ellipses obtenues dans les deux nuages.

Pour réaliser les différents traitements avec `R`, il faut charger les packages **`rgl`**, **`ade4`** et éventuellement **`Rcmdr`** (interface concivial).

```
> 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
piqu   1.543 0.7425040 10
sucr   2.250 1.1826994 10
suffo  0.399 0.5538441 10
```

```
> round(cov(cidre),2)
      odeu  sucr  acid  amer  astr  suffo piqu  alco parf fruit
odeu  0.68  0.07 -0.04  0.25  0.01  0.38 -0.37  0.02 -0.27  0.20
sucr  0.07  1.40 -0.11 -0.44 -0.29 -0.13 -0.53 -1.02  1.16  1.52
acid -0.04 -0.11  0.11 -0.02  0.04  0.02  0.03  0.05 -0.15 -0.12
amer  0.25 -0.44 -0.02  0.39  0.14  0.13 -0.02  0.41 -0.45 -0.42
astr  0.01 -0.29  0.04  0.14  0.10  0.01  0.03  0.26 -0.24 -0.28
suffo 0.38 -0.13  0.02  0.13  0.01  0.31 -0.10  0.12 -0.31 -0.07
piqu -0.37 -0.53  0.03 -0.02  0.03 -0.10  0.55  0.34 -0.28 -0.73
alco  0.02 -1.02  0.05  0.41  0.26  0.12  0.34  0.88 -0.80 -1.05
parf -0.27  1.16 -0.15 -0.45 -0.24 -0.31 -0.28 -0.80  1.27  1.21
fruit 0.20  1.52 -0.12 -0.42 -0.28 -0.07 -0.73 -1.05  1.21  1.82
```

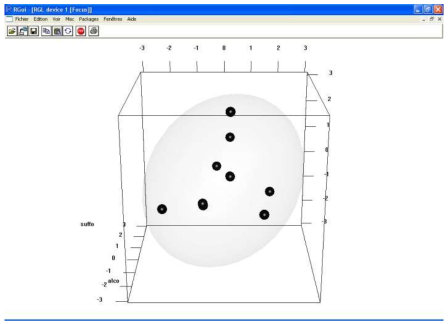
```
> round(cor(cidre),2)
       odeu  sucr  acid  amer  astr  suffo piqu  alco  parf  fruit
odeu  1.00  0.08 -0.16  0.49  0.04  0.84 -0.61  0.03 -0.29  0.18
sucr  0.08  1.00 -0.29 -0.60 -0.77 -0.19 -0.61 -0.92  0.87  0.95
acid -0.16 -0.29  1.00 -0.08  0.34  0.14  0.14  0.15 -0.40 -0.27
amer  0.49 -0.60 -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
piqu -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
```

Examen graphique :

```
> library("rgl")
> cidreR = as.data.frame(scale(cidre)*sqrt(10/9))
> attach(cidreR)
```

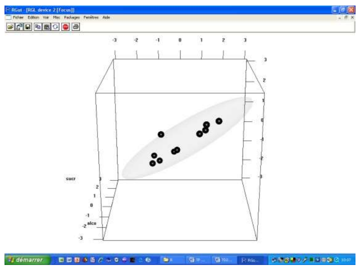
nuage 1 :

```
> 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:

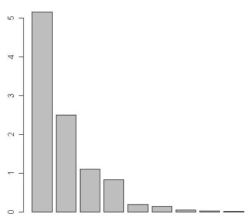
```
> plot3d(parf,alco,sucr,type="s",xlim=c(-3,3),ylim=c(-3,3),zlim=c(-3,3))
> plot3d(ellipse3d(cor(cbind(parf,alco,sucr))),col="grey",alpha=0.05,add=TRUE)
```



Partie II : *ACP* normée du tableau

1. Nombre de facteurs retenus

```
> library(ade4)
> acp<-dudi.pca(cidre,center=T,scale=T,scannf=F)
> 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$co,2)	Comp1	Comp2	[ctr en %] > inertie$col.abs/100 Comp1	Comp2	[qlt en %] > inertie$col.re/100 Comp1	Comp2	con.tra
odeu	-0.08	-0.98	0.13	38.70	-0.69	-96.83	10
sucr	0.97	-0.16	18.40	1.05	94.84	-2.63	10
acid	-0.33	0.15	2.07	0.94	-10.65	2.35	10
amer	-0.72	-0.47	9.97	8.68	-51.38	-21.71	10
astr	-0.83	-0.03	13.49	0.04	-69.54	-0.11	10
suffo	-0.31	-0.79	1.84	24.96	-9.48	-62.44	10
piqu	-0.49	0.72	4.65	20.59	-23.97	51.52	10
alco	-0.94	0.04	17.28	0.06	-89.09	0.16	10
parf	0.91	0.20	15.95	1.54	82.23	3.85	10
fruit	0.91	-0.29	16.21	3.44	83.56	-8.60	10

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

a. Comment reconnaît-on sur la figure des variables qu'une variable est bien représentée ?

b. Quelles sont les variables mal représentées dans le plan F1-F2 ? Justifier votre réponse.

c. 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.

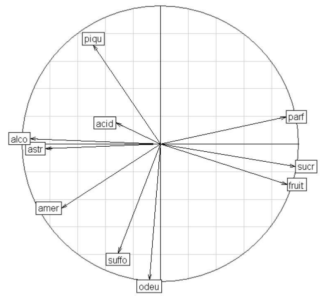
d. Quelles sont les variables qui ont contribuées à l'axe F1 ? Justifier votre réponse.

e.

f. Donner une signification à cet axe

g. Quelles sont les variables qui ont contribuées à l'axe F2 ? Justifier votre réponse.

h. Donner une signification à cet axe.



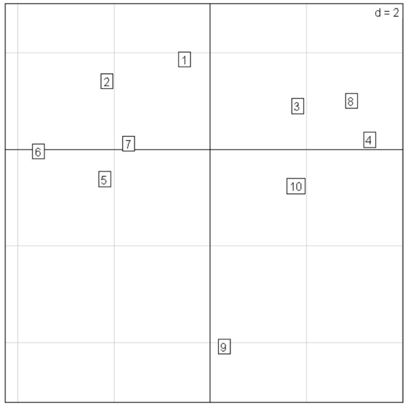
3. Analyse des individus

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

	Composantes principales > round(acp$li,2) Axis1	Axis2	[ctr en %] > inertie$row.abs/100 Axis1	Axis2	[qlt en %] > inertie$row.re/100 Axis1	Axis2	con.tra
1	-0.53	1.87	0.55	13.91	-4.71	58.00	6.00
2	-2.15	1.41	8.95	7.97	-56.65	24.50	8.14
3	1.82	0.90	6.41	3.24	49.20	12.07	6.72
4	3.32	0.20	21.35	0.17	77.55	0.29	14.19
5	-2.20	-0.62	9.37	1.53	-78.09	-6.19	6.18
6	-3.57	-0.04	24.78	0.01	-82.36	-0.01	15.51
7	-1.69	0.12	5.55	0.06	-69.70	0.35	4.11
8	2.94	1.01	16.74	4.04	63.92	7.48	13.50
9	0.29	-4.09	0.16	66.77	0.46	-91.86	18.19
10	1.78	-0.76	6.14	2.31	42.36	-7.76	7.47

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

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

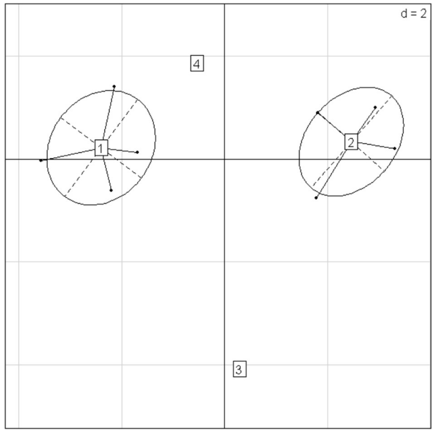


.

4. 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)
```



Exemple II : Charolais – Zebu

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 respectivement : poids vif ; poids de la carcasse ; poids de la viande de première qualité ; poids de la viande totale ; poids du gras ; poids des os.

Analyser les résultats ci-dessous.

```
> zebu<-read.table("zebu.txt",header=T)
> zebu
vif carc qsup tota gras os race
1 395 224 35.1 79.1 6.0 14.9 1
2 410 232 31.9 73.4 9.7 16.4 1
3 405 233 30.7 76.5 7.5 16.5 1
4 405 240 30.4 75.3 8.7 16.0 1
> race <- as.factor(race)
> zebu <- zebu[,1:6]
```

1. Paramètres statistiques:

Moyenne et écart-type par race

```
Variable: carc
      mean       sd     n
1 233.0000 8.790491 12
2 224.2727 6.018154 11
Variable: gras
      mean       sd      n
1  7.258333 1.439986 12
2 10.845455 1.758615 11
```

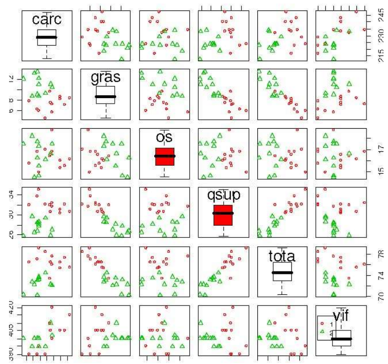
```
Variable: os
      mean         sd     n
1 16.30833 0.9949494 12
2 16.51818 1.2584261 11
Variable: qsup
      mean       sd     n
1 31.99167 1.344658 12
2 27.66364 1.343334 11
```

```
Variable: tota
      mean       sd     n
1 76.60000 1.502120 12
2 72.56364 1.297130 11
Variable: vif
      mean       sd     n
1 402.5000 9.885711 12
2 399.7273 4.221159 11
```

Matrice des corrélations

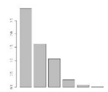
```
       vif  carc  qsup  tota  gras    os
vif   1.00  0.64 -0.09 -0.13  0.16 -0.06
carc  0.64  1.00  0.28  0.39 -0.33 -0.09
qsup -0.09  0.28  1.00  0.89 -0.86 -0.06
tota -0.13  0.39  0.89  1.00 -0.91 -0.12
gras  0.16 -0.33 -0.86 -0.91  1.00 -0.27
os   -0.06 -0.09 -0.06 -0.12 -0.27  1.00
```

2. Représentation graphique



3. Valeurs propres

```
> library(ade4)
> acp <- dudi.pca(zebu) ; round(acp$eig,2)
[1] 2.95 1.62 1.07 0.27 0.08 0.01
> round(cumsum(acp$eig*10),2) #FAUX!!!
[1] 29.51 45.71 56.37 59.08 59.89 60.00
```



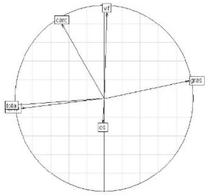
4. Analyse des variables

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

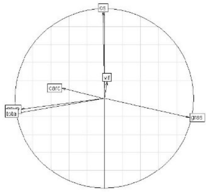
```
     Coordonnées          |    [CTR en 10000ième]      |        [Qlt en 10000ième]
> round(acp$co,2)         | > inertie$col.abs          | > inertie$col.re
    Comp1 Comp2 Comp3     |     Comp1 Comp2 Comp3      |      Comp1 Comp2 Comp3 con.tra
vif  0.03   0.93  0.19    | vif      2 5310   341      | vif       7 8603   364   1667
carc -0.48  0.80  0.12    | carc  779 3930   127       | carc -2299 6366   136   1667
qsup -0.94 -0.11 -0.12    | qsup 2971   73   136       | qsup -8766 -119  -145   1667
tota -0.97 -0.07 -0.16    | tota 3181   33   254       | tota -9387  -53  -270   1667
gras  0.95  0.19 -0.21    | gras 3066  224   429       | gras  9046  363  -458   1667
os   -0.02 -0.26  0.96    | os      1  430 8712        | os       -3 -696 9287   1667
```

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

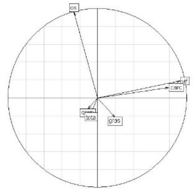
axes1-2



axes1-3



axes 2 - 3



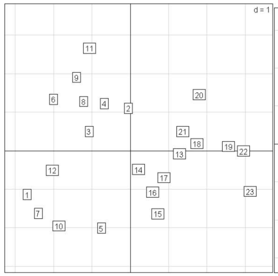
5. Analyse des individus

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

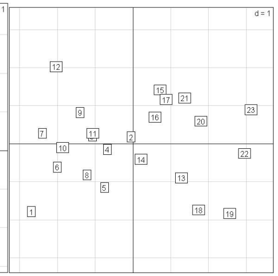
```
 Composantes principales    [CTR en 10000ième]    [Qlt en 10000ième]

 > round(acp$li,3)           > inertie$row.abs      > inertie$row.re
      Axis1 Axis2 Axis3          Axis1 Axis2 Axis3     Axis1 Axis2 Axis3 con.tra
 1  -2.691 -1.137 -1.786   1    1067   347 1302     1  -5975 -1068 -2634     878
 2  -0.050  1.102  0.180   2       0   326   13     2    -10  4836   129     182
 3  -1.072  0.499  0.202   3     169    67   17     3  -7332  1591   260     114
 4  -0.671  1.228 -0.149   4      66   405    9     4  -1994  6672   -98     164
 5  -0.756 -2.009 -1.152   5      84  1083  541     5   -948 -6695 -2202     437
 6  -1.999  1.337 -0.616   6     589   480  155     6  -6224  2785  -591     465
 7  -2.402 -1.625  0.276   7     850   708   31     7  -6356 -2907    84     658
 8  -1.213  1.278 -0.819   8     217   438  274     8  -3728  4140 -1701     286
 9  -1.401  1.906  0.823   9     289   975  276     9  -2274  4212   785     625
 10 -1.869 -1.954 -0.105  10     515  1024    4    10  -4653 -5085   -15     544
 11 -1.065  2.663  0.276  11     167  1904   31    11  -1359  8497    91     605
 12 -2.032 -0.507  2.028  12     609    69 1677    12  -4712  -294  4690     635
 13  1.287 -0.082 -0.898  13     244     2  329    13   6643   -27 -3231     181
 14  0.214 -0.485 -0.409  14       7    63   68    14    556 -2848 -2028      60
 15  0.713 -1.635  1.418  15      75   718  820    15    945 -4975  3742     389
 16  0.586 -1.076  0.701  16   51 311 200          16 1612 -5446 2310        154
 17  0.880 -0.699  1.164  17   114 131 553         17 2909 -1836 5094        193
 18  1.735  0.187 -1.741  18   443 9 1236          18 4939 57 -4972          442
 19  2.561  0.113 -1.835  19   966 3 1374          19 6551 13 -3365          725
 20  1.801  1.457  0.593  20   478 570 143         20 5421 3547 588          434
 21  1.365  0.502  1.204  21   275 68 591          21 4490 606 3494          301
 22  2.949 -0.008 -0.253  22   1281 0 26           22 9644 0 -71             654
 23  3.130 -1.055  0.897  23   1444 298 329        23 8109 -920 667          876
```


M1 Ingénierie Statistique, Dpt de Mathématiques, Université de Nantes : Analyse de données (L. Bellanger)

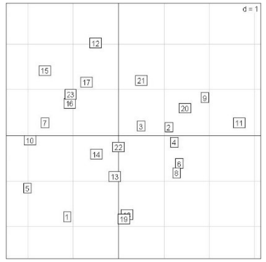
Axe 1- 2



Axe 1 - 3

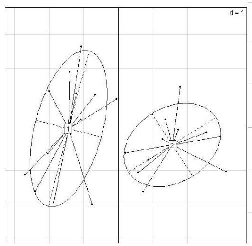


Axe 2-3

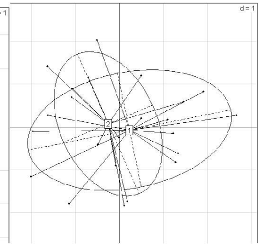


5. Avec les informations sur les races :

Axes 1-2



Axes 2-3



Autres exemples : avec `PCA(FactoMineR)` ou `dudi .pca(ade4)` !

Exercice 1 : `voitur94`

Ce fichier exemple est fourni avec le logiciel winstat. On y trouve les variables suivantes :

Identifian : nom du modèle
Puiss_admi : en chevaux fiscaux
Cylindree : en cm**3**
Moteur : 1=essence, 2=diésel
Trasmissio : 1=traction, 2=propulsion
Longueur : longueur de la voiture
Largeur : largeur de la voiture
Surface : surface de la voiture
Poids_Tota : poids total en Kg
Vit_Maxi : vitesse maximum en km/h
Dep_arret : Temps, en secondes, pour parcourir 1000 m, départ arrêté.
Marque : nom du fabriquant
MarqQL : nom fabriquant codé (avec un chiffre)
Conso_Moye : Consommation moyenne aux 100 Km, en litres (d'essence ou de gazole)

Exercice 2 : `indep` et `depend`

Effectuez une ACP pour chaque fichier. Comparez les éboulis de valeurs propres.

Exercice 3 : `sen`

Le fichier **sen** est issu d'une enquête sur la traction animale. Les variables sont :
EX : numéro d'exploitation
QU : numéro de quartier (village)
AC : nombre d'actifs dans la famille
SP : surface possédée
SU : surface agricole utile
AT : nombre d'ânes de traction
CT : nombre de chevaux de trait
BT : nombre de paires de boeufs de trait
VT : nombre de paires de vaches de trait
BV : nombre de bovins hors exploitation
OV : nombre d'ovins
CP : nombre de caprins

Exercice 4 : `olympic`

Le fichier olympic présente les performances de 33 athlètes. Les variables sont :
dossard : numéro du dossard
m100 : course 100 mètres
long : saut en longueur
poid : lancer du poids
haut : saut en hauteur
m400 : course 400 mètres
m110 : course 110 mètres
disq : lancer du disque
perc : saut à la perche
jave : lancer du javelot
m1500 : course de 1500 mètres

Exercice 5 : `espvie`

Le fichier espvie représente l'espérance de vie de la population de 40 pays en fonction de critères sociaux. Les variables sont :
Pays : nom du pays
EspVie : espérance moyenne de vie de la population
PersTele : nombre d'habitants par téléviseur
PersPhy : nombre d'habitants par physicien
FVie : espérance de vie des femmes
Hvie : espérance de vie des hommes

Exercice 6 : `capitales(ade4)` => Tableau de distances non euclidiennes

Travail à rendre par écrit en binôme lors du dernier TP semaine 43 (M1 IS) ou semaine 45 (M1 Info)

Fonction ACP à construire et application sur les données : `Projet M1 AD 1920.csv`