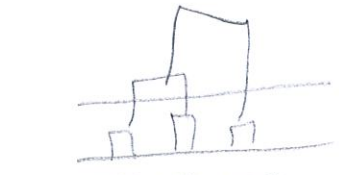
TP classification: méthodologie

Exercice 1

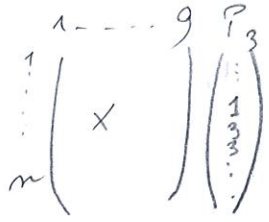
Matrice de données quantitatives $X : (29, 9)$

1) Calories, sodium et calcium ont de ⊕ grandes variances
⇒ ⊕ grande variabilité

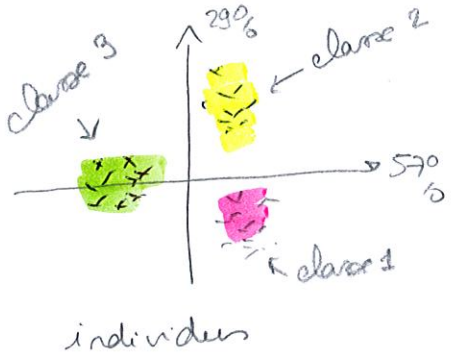
2) On effectue une classificat° hiérarchique de Ward sur X et on récupère la partition en trois classes

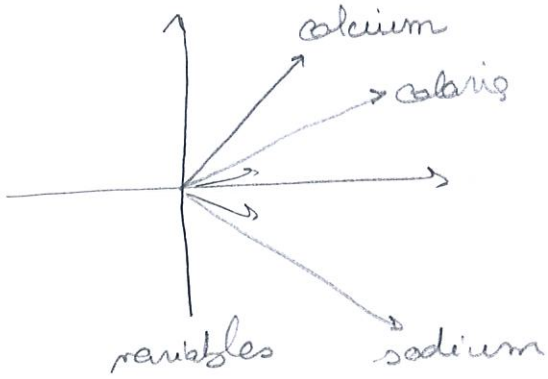


$P_3 = (C_1\ C_2\ C_3)$

$\Rightarrow X_1 =$ 

3) Interprétation de la partition en 3 classes via l'ACP non normée (i.e. ACP sur données centrées sur matrice de covariance)





Interprétation :

- seules les variables de fortes variances contribuent aux axes (car ACP non normée).
- Les fromages projetés à droite (classes 1 et 2) ont des valeurs fortes en calories-calcium-sodium (pâtes dures et fromage gras), à gauche ⊕ faibles (fromage frais)

⇒ On ne peut interpréter les classes qu'en fct des variables de forte variance

Explication: Lorsqu'on calcul la distance entre deux fromages sur les données brutes, on donne implicitement plus d'importance aux variables de forte variance.

Exple:	Calories	Sodium	Calcium	lipides	retinole	
Emmental	378	60	308,2	29,4	56,3	.....
Fr. frais	80	41	146,3	3,5	50	
Ecarts au carré	$(378-80)^2$	$(60-41)^2$	$(308-146)^2$	$(29,4-3,5)^2$	$(56,3-50)^2$	....
	88804	361	26211	670	39,69	....

$distance^2 = 88804 + 361 + 26211 + 670 + \dots$

⊕ "lourd"

⇒ distance ne retrace que la ressemblance sur ces variables de forte variance.

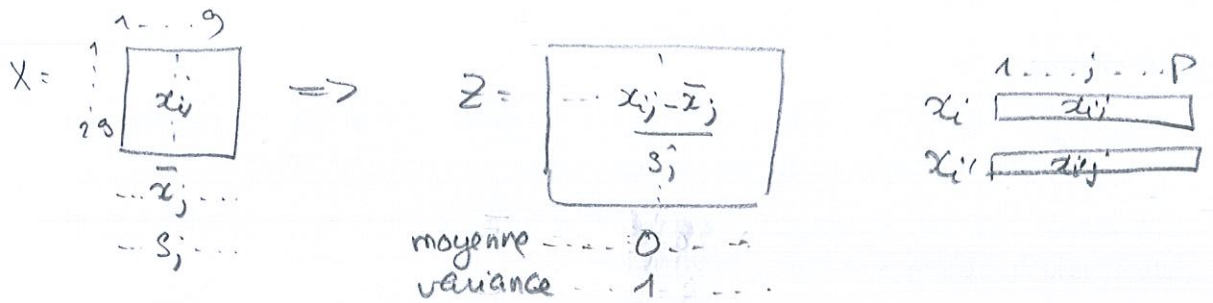
⇒ On va regrouper les fromages dans des classes en fonction de ces variables uniquement

⇒ Centrer-réduire les données pour donner le même poid à toutes les variables dans le calcul des distances.

Exercice 2

Matrice de données quantitatives : $Z \leftarrow X$ centré-réduite.

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad \text{où} \quad s_j = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$



1) Distance euclidienne normalisée par l'inverse de la variance :

$$d_M^2(x_i, x_{i'}) = \sum_{j=1}^{p} \frac{1}{s_j^2} (x_{ij} - x_{i'j})^2 \quad \text{avec} \quad M = \begin{pmatrix} 1/s_1^2 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1/s_p^2 \end{pmatrix}$$

← pondère les écarts par la variance de j.

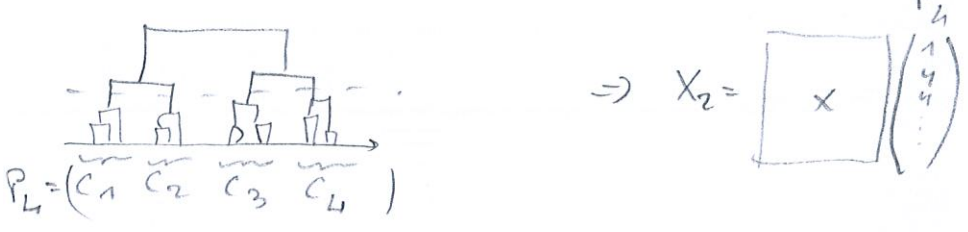
$$= \sum_{j=1}^{n} \left(\frac{x_{ij} - \bar{x}_j - (x_{i'j} - \bar{x}_j)}{s_j} \right)^2$$

$$= \sum_{j=1}^{n} (z_{ij} - z_{i'j})^2 = d^2(z_i, z_{i'})$$

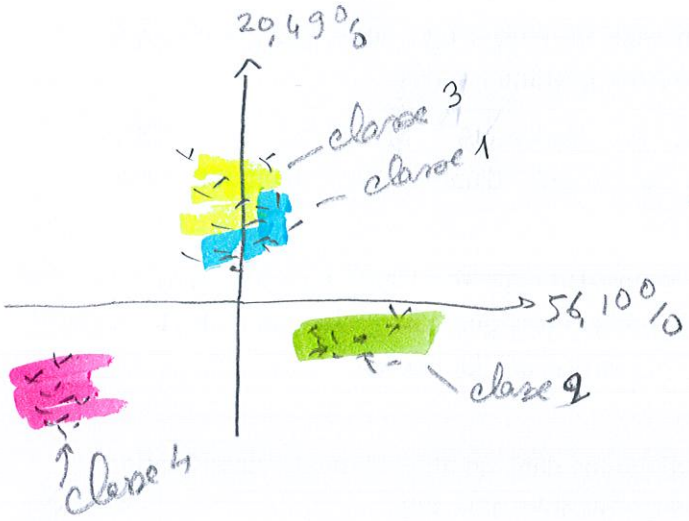
↑ distance euclidienne simple.

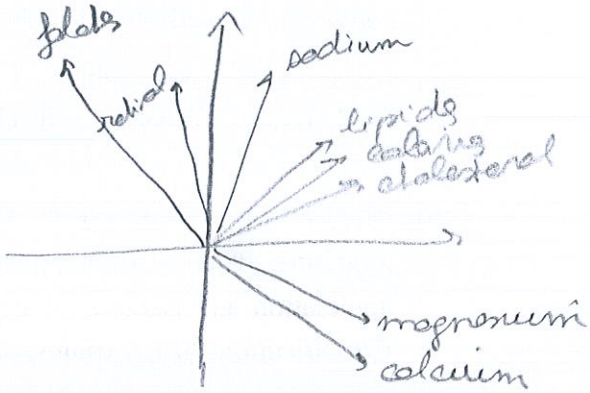
⇒ C'est équivalent de calculer la distance euclidienne simple sur les données centrées-réduites * Exemple p4

2) Classification hiérarchique de Ward sur Z et on récupère la partition en 4 classes



3) Interprétation de la partition en 4 classes via l'ACP normée (i.e. ACP sur données centrées-réduites ou matrice de corrélations)





* 2 dim° ⇒ 76,59 % de l'inertie expliquée + cluster de la valeur propre.

* Toutes les variables jouent un rôle

⇒ classe 4 : peu gras et calorique (fromages frais)

classe 2 : riches en magnesium et calcium (pâtes dures)

classe 3 : riches en fibres

⇒ Aller voir sur classe 3 ?

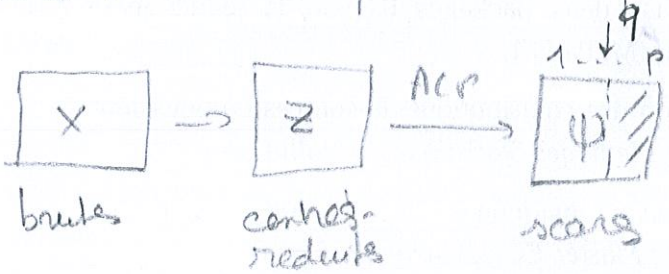
classe 1 : ⊕ salés

(*) Exple :

	calories	sodium	calcium	lipides	retinol ...
Emmental	378	60	308	29	56
Fromage frais, écart au carré variance	80	41	146	3,5	50
	$(378-80)^2$	$(60-41)^2$	$(308-146)^2$	$(29-3,5)^2$	
	8464	1184	8260	66	

Exercice 3

Matrice de données quantitatives : Ψ ← Scores de l'ACP de X.



1) On effectue une classification hiérarchique de Ward sur Ψ (en conservant toutes les composantes principales)

On effectue une classification hiérarchique de Ward sur Z

⇒ Même dendrogramme
⇒ Même indices
⇒ Même hiérarchie

2) On effectue une classification hiérarchique de Ward sur les q premières colonnes de ψ (en conservant les q premières composantes principales)

→ Choix de q : 2 ou 3
→ On retrouve la même partition à 1 individu près
→ Stratégie de SPAD : type faire une ACP et "monter" la hiérarchie sur les q premières composantes.

(+) de SPAD : } Outil d'interprétation des classes

```
#-------TP Classification : méthodologie

#Attention : retirer les accents dans les noms de colonne !!
X<-read.table(file.choose(),sep="",header=TRUE,row.names=1)
dim(X)

#---------Exercice 1-----------------------

#Variance des variables
apply(X,2,var)

#Classification hiérarchique de Ward sur donnees brutes
d <- dist(X)
tree <- hclust(d^2,method="ward")
plot(tree)

P3 <- as.factor(cutree(tree,3))
levels(P3) <- paste("Classe",1:3)
X1 <- data.frame(X,P3=P3)



#Interprétation de la partition

require(FactoMineR)

#ACP non normee
pca1 <- PCA(X1,scale=FALSE,quali.sup=10,graph=FALSE)
pca1$eig

# Nuage des individus et des variables dans le premier plan factoriel
par(mfrow=c(1,2))
plot.PCA(pca1,axes=c(1,2),choix="ind",habillage=10,invisible="quali")
plot.PCA(pca1,axes=c(1,2),choix="var")

#distance entre Emmental et Fromage frais
X[c(13,16),]
(X[13,]-X[16,])^2



#---------Exercice 2-----------------------

#donnees centrees-reduire
n <- nrow(X)
Z<-scale(X,center=TRUE,scale=TRUE)*sqrt((n)/(n-1))

apply(Z,2,mean)
apply(Z,2,sd)*sqrt((n-1)/n)


# Classification sur donnes centrees-reduites:

d <- dist(Z)
tree <- hclust(d^2,method="ward")
plot(tree)

P4 <- as.factor(cutree(tree,4))
levels(P4)<-paste("Classe",1:4)
X2<-data.frame(X,P4=P4)


#ACP normee
pca2 <- PCA(X2,scale=TRUE,quali.sup=10,graph=FALSE)
pca2$eig

#distance normalisee par l'inverse des variances entre Emmental et Fromage frais
X[c(13,16),]
apply(X,2,var)
(X[13,]-X[16,])^2

par(mfrow=c(1,2))
plot.PCA(pca2,axes=c(1,2),choix="ind",habillage=10,invisible="quali")
plot.PCA(pca2,axes=c(1,2),choix="var")

plot.PCA(pca2,axes=c(1,3),choix="ind",habillage=10,invisible="quali")
plot.PCA(pca2,axes=c(1,3),choix="var")

#---------Exercice 3-----------------------

#ACP normee en conservant TOUTES les composantes principales
pca2 <- PCA(X,ncp=9,graph=FALSE)
Psi <- pca2$ind$coord #matrice des scores de l'ACP normee

#Ward sur donnees centrees-reduites
d <- dist(Z)
tree <- hclust(d,method="ward")
plot(tree)
```