Classification automatique sous R

CAH et K-Means

Ricco.Rakotomalala
http://eric.univ-lyon2.fr/~ricco/cours

Importation des données, description

DONNÉES

Objectif de l'étude
Classification automatique de fromages

Objectifs de l'étude

Ce document retranscrit une démarche de classification automatique d'un ensemble de fromages (29 observations) décrits par leurs propriétés nutritives (ex. protéines, lipides, etc. ; 9 variables). L'objectif est d'identifier des groupes de fromages homogènes, partageant des caractéristiques similaires.

Nous utiliserons essentiellement deux approches en nous appuyant sur deux procédures du logiciel R : la classification ascendante hiérarchique (CAH) avec hclust() ; la méthode des centres mobiles (k-means) avec kmeans().

Le fichier « fromage.txt » provient de la page de cours de Marie Chavent de l'Université de Bordeaux. Les excellents supports et exercices corrigés (commentaires + code programme R) que l'on peut y trouver compléteront à profit ce tutoriel qui se veut avant tout un guide simple pour une première prise en main du logiciel R dans le contexte de la classification automatique.

Traitements réalisés

- Chargement et description des données
- Classification automatique avec hclust() et kmeans()
- Pistes pour la détection du nombre adéquat de classes
- Description – interprétation des groupes

Données disponibles

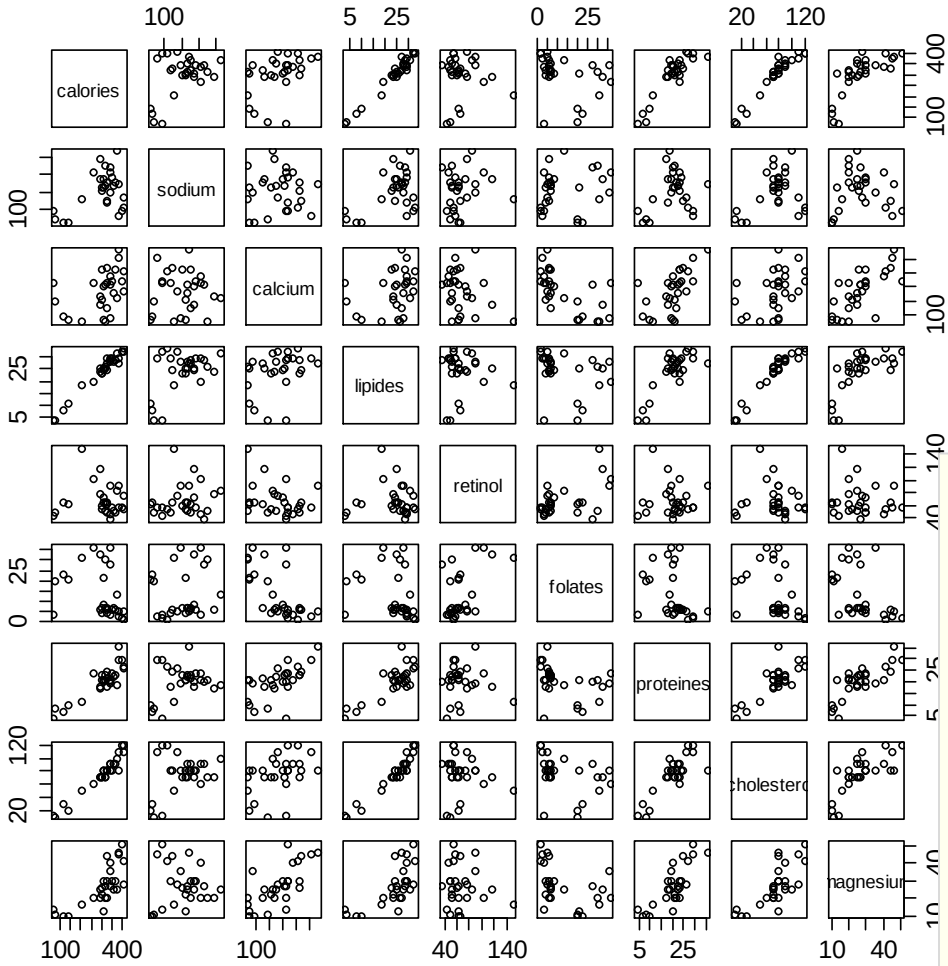
Fromages	calories	sodium	calcium	lipides	retinol	folates	proteines	cholesterol	magnesium
CarredelEst	314	353.5	72.6	26.3	51.6	30.3	21	70	20
Babybel	314	238	209.8	25.1	63.7	6.4	22.6	70	27
Beaufort	401	112	259.4	33.3	54.9	1.2	26.6	120	41
Bleu	342	336	211.1	28.9	37.1	27.5	20.2	90	27
Camembert	264	314	215.9	19.5	103	36.4	23.4	60	20
Cantal	367	256	264	28.8	48.8	5.7	23	90	30
Chabichou	344	192	87.2	27.9	90.1	36.3	19.5	80	36
Chaource	292	276	132.9	25.4	116.4	32.5	17.8	70	25
Cheddar	406	172	182.3	32.5	76.4	4.9	26	110	28
Comte	399	92	220.5	32.4	55.9	1.3	29.2	120	51
Coulomniers	308	222	79.2	25.6	63.6	21.1	20.5	80	13
Edam	327	148	272.2	24.7	65.7	5.5	24.7	80	44
Emmental	378	60	308.2	29.4	56.3	2.4	29.4	110	45
Fr.chevrepatemolle	206	160	72.8	18.5	150.5	31	11.1	50	16
Fr.fondu.45	292	390	168.5	24	77.4	5.5	16.8	70	20
Fr.frais20nat.	80	41	146.3	3.5	50	20	8.3	10	11
Fr.frais40nat.	115	25	94.8	7.8	64.3	22.6	7	30	10
Maroilles	338	311	236.7	29.1	46.7	3.6	20.4	90	40
Morbier	347	285	219	29.5	57.6	5.8	23.6	80	30
Parmesan	381	240	334.6	27.5	90	5.2	35.7	80	46
Petitsuisse40	142	22	78.2	10.4	63.4	20.4	9.4	20	10
PontlEveque	300	223	156.7	23.4	53	4	21.1	70	22
Pyrenees	355	232	178.9	28	51.5	6.8	22.4	90	25
Reblochon	309	272	202.3	24.6	73.1	8.1	19.7	80	30
Rocquefort	370	432	162	31.2	83.5	13.3	18.7	100	25
SaintPaulin	298	205	261	23.3	60.4	6.7	23.3	70	26
Tome	321	252	125.5	27.3	62.3	6.2	21.8	80	20
Vacherin	321	140	218	29.3	49.2	3.7	17.6	80	30
Yaourtlaitent.nat.	70	91	215.7	3.4	42.9	2.9	4.1	13	14

Label des observations

Variables actives

Fichier de données
Importation, statistiques descriptives et graphiques

```
#modifier le répertoire par défaut
setwd(" … mon dossier …")

#charger les données - attention aux options
fromage <- read.table(file="fromage.txt",header=T,row.names=1,sep="\t",dec=".")

#afficher les 6 premières lignes
print(head(fromage))

#stat. descriptives
print(summary(fromage))

#graphique - croisement deux à deux
pairs(fromage)
```



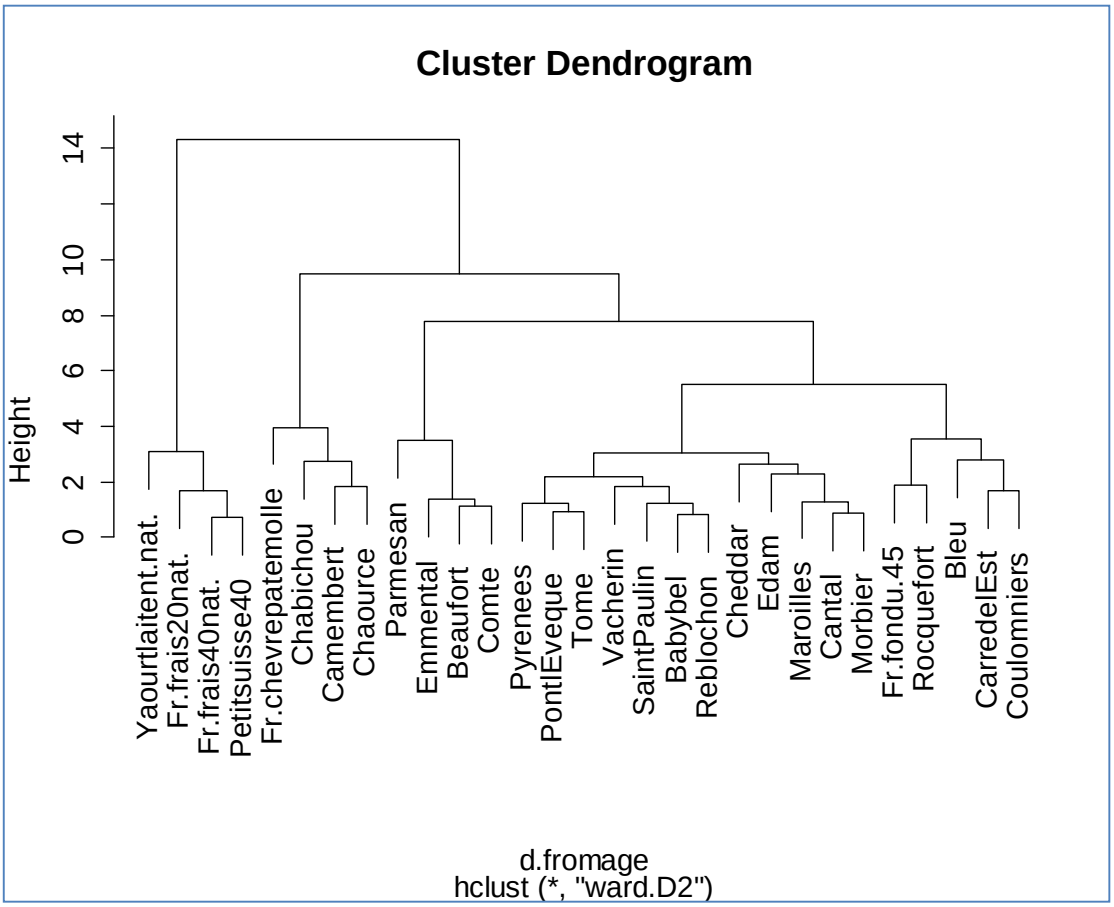
Ce type de graphique n'est jamais anodin. Nous constatons par exemple que (1) « lipides » est fortement corrélé avec « calories » et « cholestérol » (sans trop de surprises) (remarque : la même information va peser 3 fois dans l'analyse) ; (2) dans certaines configurations, des groupes semblent apparaître naturellement (ex. croisement de « protéines » et « cholestérol », avec une corrélation inter-groupes assez marquée).

Classification ascendante hiérarchique

CAH (HCLUST)

Classification ascendante hiérarchique
La procédure hclust() de R (package « stats » - toujours chargée)

```
#centrage réduction des données
#pour éviter que variables à forte variance pèsent indûment sur les résultats
fromage.cr <- scale(fromage,center=T,scale=T)

#matrice des distances entre individus
d.fromage <- dist(fromage.cr)

#CAH - critère de Ward
#method = « ward.D2 » correspond au vrai critère de Ward
#utilisant le carré de la distance
cah.ward <- hclust(d.fromage,method="ward.D2")

#affichage dendrogramme
plot(cah.ward)
```

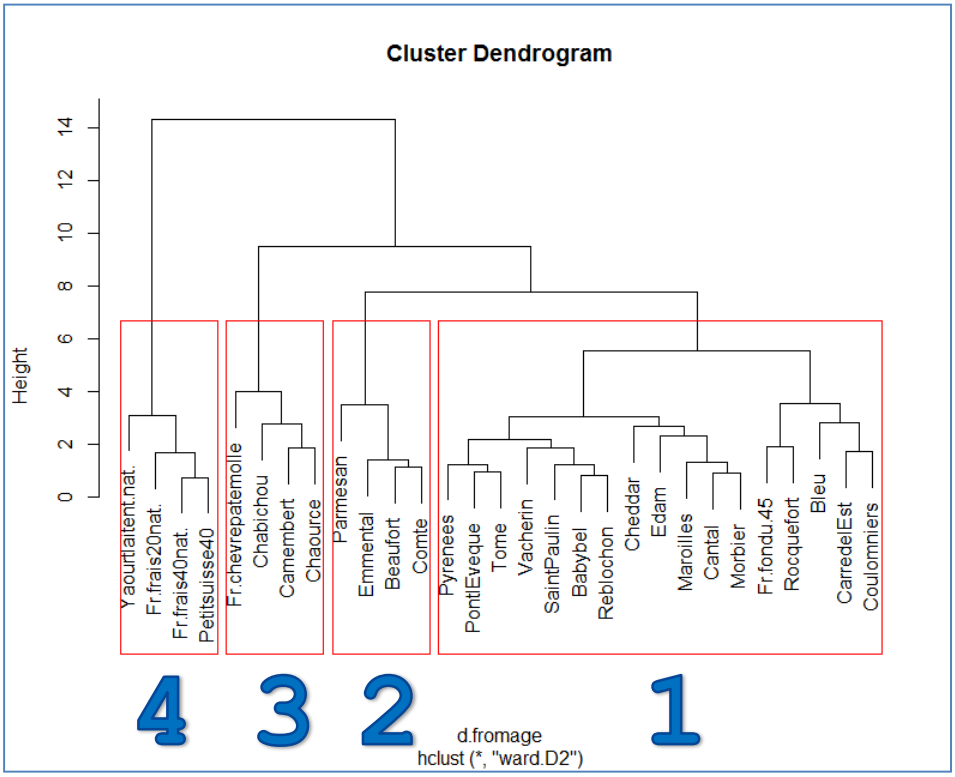


Le dendrogramme « suggère » un découpage en 4 groupes. On note qu'une classe de fromages, les « fromages frais » (tout à gauche), se démarque fortement des autres au point qu'on aurait pu envisager aussi un découpage en 2 groupes seulement. Nous y reviendrons plus longuement lorsque nous mixerons l'analyse avec une analyse en composantes principales (ACP).

Classification ascendante hiérarchique
Découpage en classes – Matérialisation des groupes

```
#dendrogramme avec matérialisation des groupes
rect.hclust(cah.ward,k=4)

#découpage en 4 groupes
groupes.cah <- cutree(cah.ward,k=4)

#liste des groupes
print(sort(groupes.cah))
```

Fromage	Groupe
CarredelEst	1
Babybel	1
Bleu	1
Cantal	1
Cheddar	1
Coulomniers	1
Edam	1
Fr.fondu.45	1
Maroilles	1
Morbier	1
PontlEveque	1
Pyrenees	1
Reblochon	1
Rocquefort	1
SaintPaulin	1
Tome	1
Vacherin	1
Beaufort	2
Comte	2
Emmental	2
Parmesan	2
Camembert	3
Chabichou	3
Chaource	3
Fr.chevrepatemolle	3
Fr.frais20nat.	4
Fr.frais40nat.	4
Petitsuisse40	4
Yaourtlaitent.nat.	4



Le 4ème groupe est constitué de fromages frais.
Le 3ème de fromages à pâte molle.
Le 2nd de fromages « durs ».
Le 1er est un peu fourre-tout (de mon point de vue).

Mes compétences en fromage s'arrêtent là (merci à Wikipédia). Pour une caractérisation à l'aide des variables de l'étude, il faut passer par des techniques statistiques univariées (simples à lire) ou multivariées (tenant compte des relations entre les variables).

Méthode des centres mobiles

K-MEANS

Méthode des centres mobiles
La procédure kmeans() de R (package « stats » également)

```
#k-means avec les données centrées et réduites
#center = 4 - nombre de groupes demandés
#nstart = 5 - nombre d'essais avec différents individus de départ
#parce que les résultats sont dépendants de l'initialisation
groupes.kmeans <- kmeans(fromage.cr,centers=4,nstart=5)

#affichage des résultats
print(groupes.kmeans)

#correspondance avec les groupes de la CAH
print(table(groupes.cah,groupes.kmeans$cluster))
```

```
K-means clustering with 4 clusters of sizes 4, 14, 6, 5

Cluster means:
    calories     sodium    calcium     lipides    retinol    folates  proteines cholesterol
1 -2.1572744 -1.5213272 -0.7167418 -2.19980413 -0.5136787  0.2955348 -1.8634139  -1.9945017
2  0.3726429  0.5276310  0.1925511  0.41101185 -0.3108901 -0.4505349  0.1522469   0.3181087
3 -0.1309315  0.3941009 -1.0428188 -0.03591228  1.1713977  1.5572630 -0.1847229  -0.2213739
4  0.8395372 -0.7332260  1.2856329  0.65210487 -0.1242419 -0.8436457  1.2861074   0.9705456
   magnesium
1 -1.3884943
2  0.0156683
3 -0.4681630
4  1.6287198

Clustering vector:
       CarredelEst            Babybel           Beaufort               Bleu          Camembert
                 3                  2                  4                  2                  3
            Cantal          Chabichou           Chaource           Cheddar             Comte
                 2                  3                  3                  2                  4
       Coulomniers               Edam          Emmental Fr.chevrepatemolle        Fr.fondu.45
                 3                  4                  4                  3                  2
    Fr.frais20nat.     Fr.frais40nat.          Maroilles            Morbier           Parmesan
                 1                  1                  2                  2                  4
     Petitsuisse40        PontlEveque           Pyrenees          Reblochon         Rocquefort
                 1                  2                  2                  2                  2
       SaintPaulin              Tome           Vacherin Yaourtlaitent.nat.
                 2                  2                  2                  1

Within cluster sum of squares by cluster:
[1]  6.446342 28.737063 25.431001  9.871039
 (between_SS / total_SS =  72.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
```

Effectif des classes

Moyennes des variables actives (centrées et réduites) conditionnellement à l'appartenance aux groupes.

Groupe d'affectation des individus

Proportion d'inertie expliquée par la partition : 72%

groupes.cah	1	2	3	4
1	0	14	2	1
2	0	0	0	4
3	0	0	4	0
4	4	0	0	0

Correspondance CAH – K-Means

Le groupe 4 de la CAH coïncide avec le groupe 1 des K-Means. Après, il y a certes des correspondances, mais elles ne sont pas exactes.

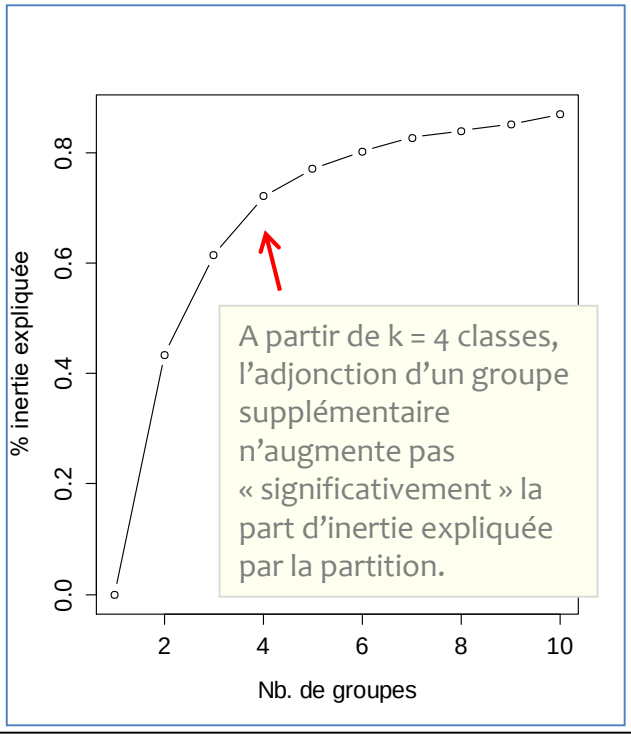
Remarque : Il se peut que vous n'ayez pas exactement les mêmes résultats avec les K-Means.

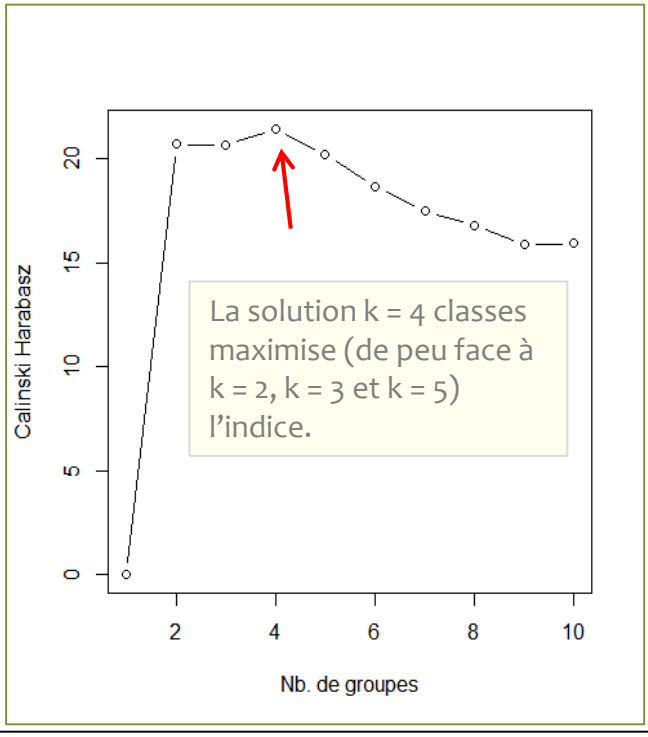
Méthode des centres mobiles
Aide à la détection du nombre adéquat de groupes

K-MEANS, à la différence de la CAH, ne fournit pas d'outil d'aide à la détection du nombre de classes. Nous devons les programmer sous R ou utiliser des procédures proposées par des packages dédiés. Le schéma est souvent le même : on fait varier le nombre de groupes et on surveille l'évolution d'un indicateur de qualité de la solution c.-à-d. l'aptitude des individus à être plus proches de ses congénères du même groupe que des individus des autres groupes.

Deux pistes ici : (1) surveiller l'évolution de la proportion d'inertie expliquée par la partition, on cherche le « coude » dans le graphique (nous programmons la procédure) ; (2) utiliser l'indice de Calinski Harabasz, on recherche alors à maximiser ce second critère (nous utilisons la fonction kmeansruns() du package « fpc », on peut aussi choisir l'indice silhouette moyenne). Voir : https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set

```
#(1)évaluer la proportion d'inertie expliquée
inertie.expl <- rep(0,times=10)
for (k in 2:10){
  clus <- kmeans(fromage.cr,centers=k,nstart=5)
  inertie.expl[k] <- clus$betweenss/clus$totss
}
#graphique
plot(1:10,inertie.expl,type="b",xlab="Nb. de groupes",ylab="% inertie expliquée")

#(2) indice de Calinski Harabasz - utilisation du package fpc
library(fpc)
#évaluation des solutions
sol.kmeans <- kmeansruns(fromage.cr,krange=2:10,criterion="ch")
#graphique
plot(1:10,sol.kmeans$crit,type="b",xlab="Nb. de groupes",ylab="Silhouette")
```





Analyses univariées et multivariées

INTERPRÉTATION DES CLASSES

Interprétation des classes
Statistiques comparatives

L'idée est de comparer les moyennes des variables actives conditionnellement aux groupes. Il est possible de quantifier globalement l'amplitude des écarts avec la proportion de variance expliquée. La démarche peut être étendue aux variables illustratives. Pour les catégorielles, nous confronterions les distributions conditionnelles.

L'approche est simple et les résultats faciles à lire. Rappelons cependant que nous ne tenons pas compte des liaisons entre les variables dans ce cas.

```
#fonction de calcul des stats
stat.comp <- function(x,y){
  #nombre de groupes
  K <- length(unique(y))
  #nb. d'observations
  n <- length(x)
  #moyenne globale
  m <- mean(x)
  #variabilité totale
  TSS <- sum((x-m)^2)
  #effectifs conditionnels
  nk <- table(y)
  #moyennes conditionnelles
  mk <- tapply(x,y,mean)
  #variabilité expliquée
  BSS <- sum(nk * (mk - m)^2)
  #moyennes + prop. variance expliquée
  result <- c(mk,100.0*BSS/TSS)
  #nommer les élements du vecteur
  names(result) <- c(paste("G",1:K),"% epl.")
  #renvoyer le vecteur résultat
  return(result)
}

#appliquer stat.comp aux variables de la base originelle fromage
#et non pas aux variables centrées et réduites
print(sapply(fromage,stat.comp,y=groupes.cah))
```

	calories	sodium	calcium	lipides	retinol	folates	proteines	cholesterol	magnesium
G 1	331.11765	262.7941	189.40000	27.15294	60.09412	9.711765	21.37647	82.35294	26.88235
G 2	389.75000	126.0000	280.67500	30.65000	64.27500	2.525000	30.22500	107.50000	45.75000
G 3	276.50000	235.5000	127.20000	22.82500	115.00000	34.050000	17.95000	65.00000	24.25000
G 4	101.75000	44.7500	133.75000	6.27500	55.15000	16.475000	7.20000	18.25000	11.25000
% epl.	87.97373	56.6772	41.27705	86.85973	64.89488	63.494807	82.70802	82.46284	67.71603

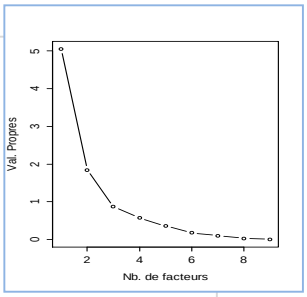
La définition des groupes est – avant tout – dominée par les teneurs en graisses (lipides, cholestérol et calories relèvent de la même idée) et en protéines.

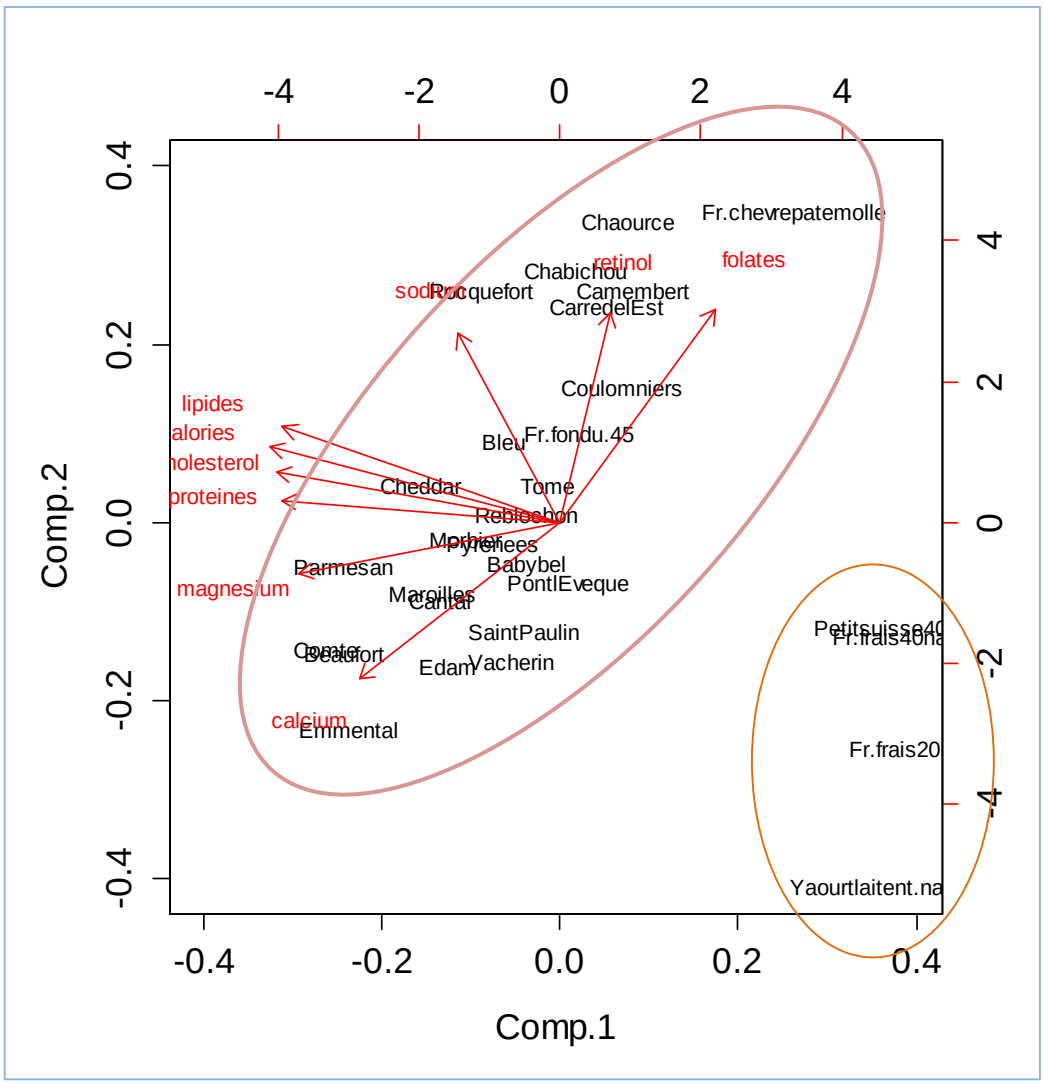
Le groupe 4 est fortement déterminé par ces variables, les moyennes conditionnelles sont très différentes.

Interprétation des classes
Analyse en composantes principales (ACP) **(1/2)**

Avec l'ACP, nous tenons compte des liaisons entre les variables. L'analyse est plus riche. Mais il faut savoir lire correctement les sorties de l'ACP.

```
#ACP normée
acp <- princomp(fromage,cor=T,scores=T)

#screeplot - 2 axes retenus
plot(1:9,acp$sdev^2,type="b",xlab="Nb. de facteurs",ylab="Val. Propres")

#biplot
biplot(acp,cex=0.65)
```



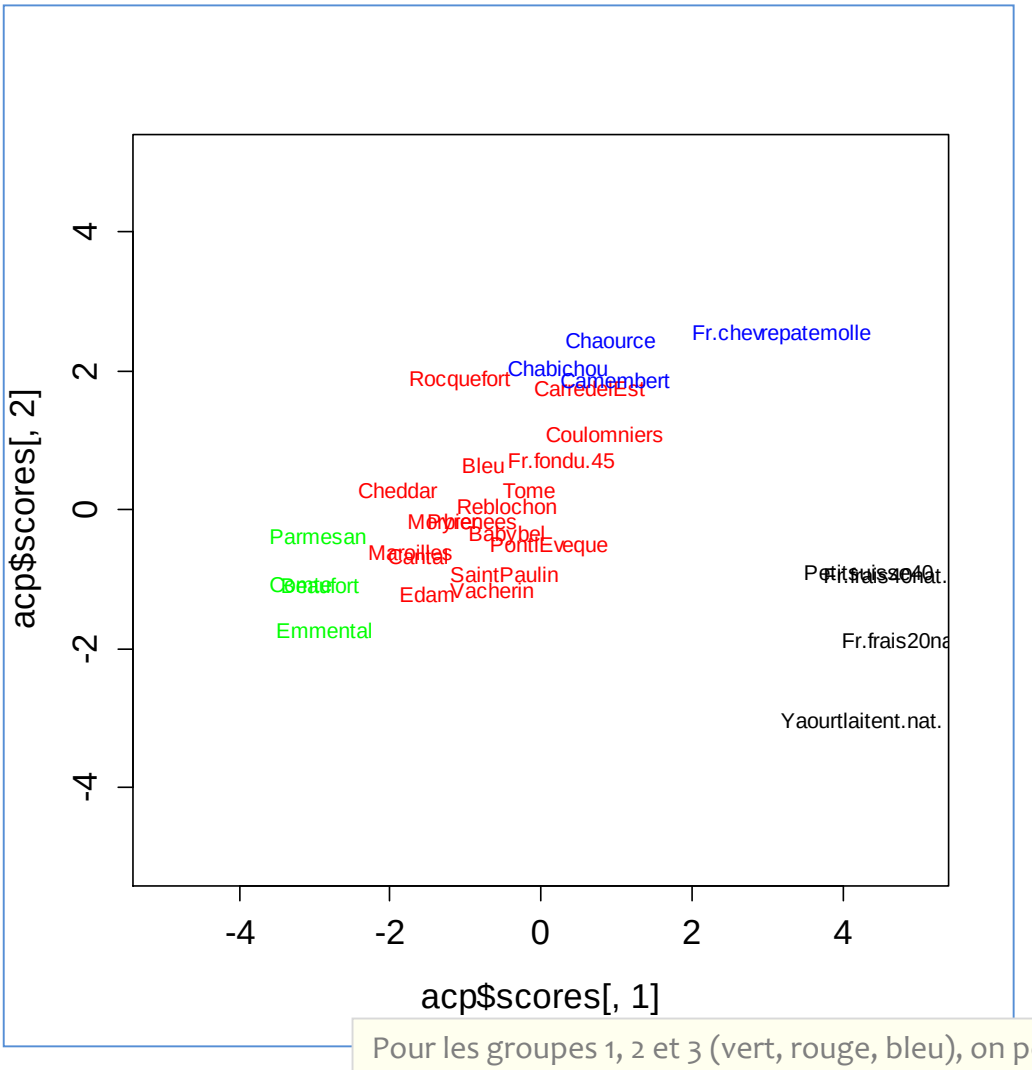


Il y a un problème. Le groupe des fromages frais écrase l'information disponible et tasse les autres fromages dans un bloc qui s'oriente différemment.

Interprétation des classes
Analyse en composantes principales (ACP) **(2/2)**

De fait, si l'on comprend bien la nature du groupe 4 des fromages frais, les autres sont plus compliqués à comprendre lorsqu'ils sont replacés dans le premier plan factoriel.

```
#positionnement des groupes dans le plan factoriel avec étiquettes des points
plot(acp$scores[,1],acp$scores[,2],type="n",xlim=c(-5,5),ylim=c(-5,5))

text(acp$scores[,1],acp$scores[,2],col=c("red","green","blue","black")[groupes.cah],cex
=0.65,labels=rownames(fromage),xlim=c(-5,5),ylim=c(-5,5))
```



Pour les groupes 1, 2 et 3 (vert, rouge, bleu), on perçoit à partir du graphique biplot de la page précédente qu'il y a quelque chose autour de l'opposition entre nutriments (lipides/calories/cholestérol, protéines, magnésium, calcium) et vitamines (rétinol, folates). Mais, dans quel sens exactement ?

La lecture n'est pas facile du fait de l'effet perturbateur du groupe 4.

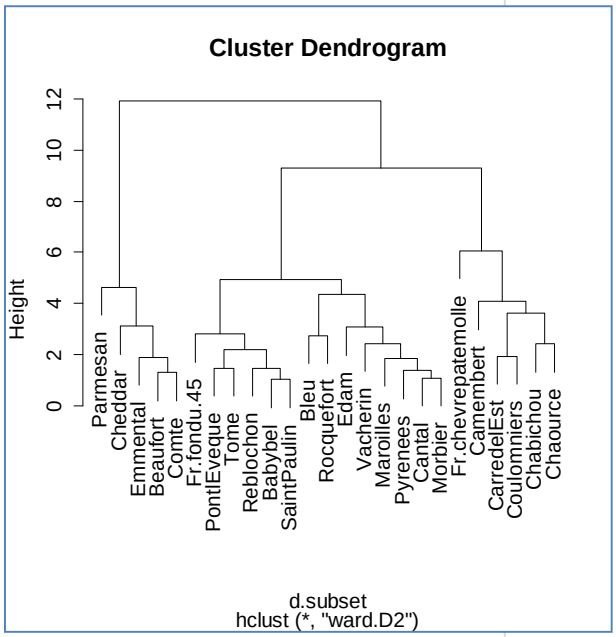
A la lumière des résultats de l'ACP

COMPLÉTER L'ANALYSE

Approfondir l'analyse
Retirer les fromages frais du jeu de données (1/2)

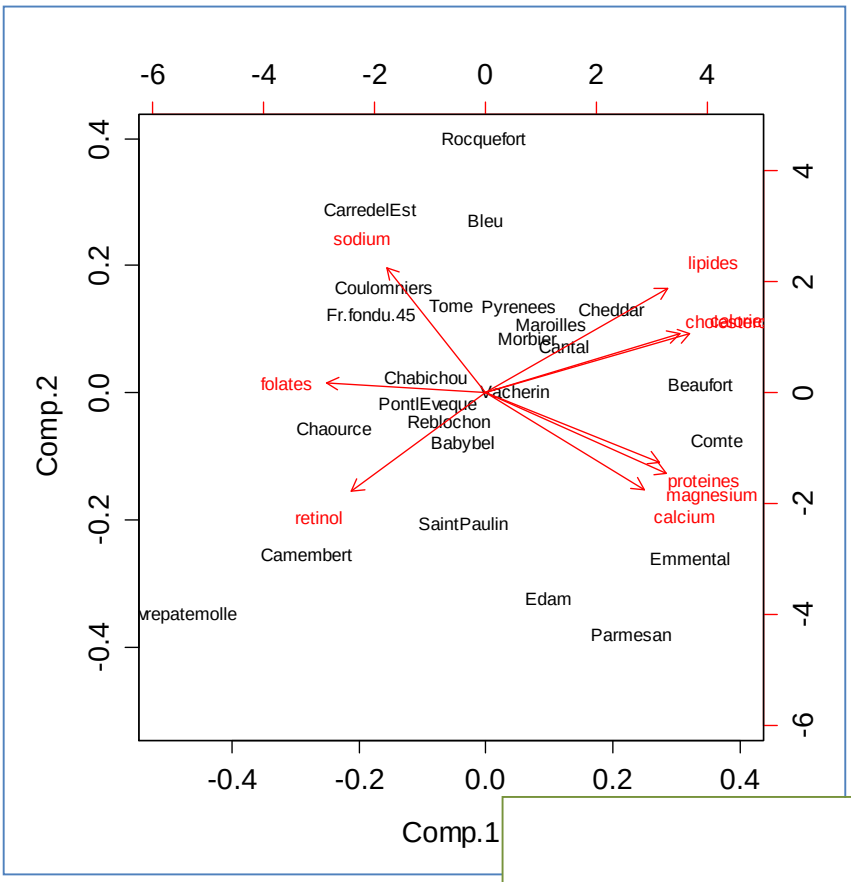
Les fromages frais sont tellement particuliers – éloignés de l'ensemble des autres observations – qu'ils masquent des relations intéressantes qui peuvent exister entre ces produits. Nous reprenons l'analyse en les excluant des traitements.

```
#retirer les 4 obs. du groupe 4
fromage.subset <- fromage[groupes.cah!=4,]

#centrage réduction
fromage.subset.cr <- scale(fromage.subset,center=T,scale=T)

#matrice de distance
d.subset <- dist(fromage.subset.cr)

#cah 2
cah.subset <- hclust(d.subset,method="ward.D2")

#affichage
plot(cah.subset)

#groupes
groupes.subset <- cutree(cah.subset,k=3)

#affichage des groupes
print(sort(groupes.subset))

#acp
acp.subset <- princomp(fromage.subset,cor=T,scores=T)

#screeplot - 2 axes retenus
plot(1:9,acp.subset$sdev^2,type="b")

#biplot
biplot(acp.subset,cex=0.65)

#positionnement des groupes dans le plan factoriel
plot(acp.subset$scores[,1],acp.subset$scores[,2],type="n",xlim=c(-6,6),ylim=c(-6,6))

#etiquettes des points
text(acp.subset$scores[,1],acp.subset$scores[,2],col=c("red","green","blue")[groupes.su
bset],cex=0.65,labels=rownames(fromage.subset),xlim=c(-6,6),ylim=c(-6,6))
```



3 groupes se distinguent. On a moins le phénomène d'écrasement constaté dans l'analyse précédente.

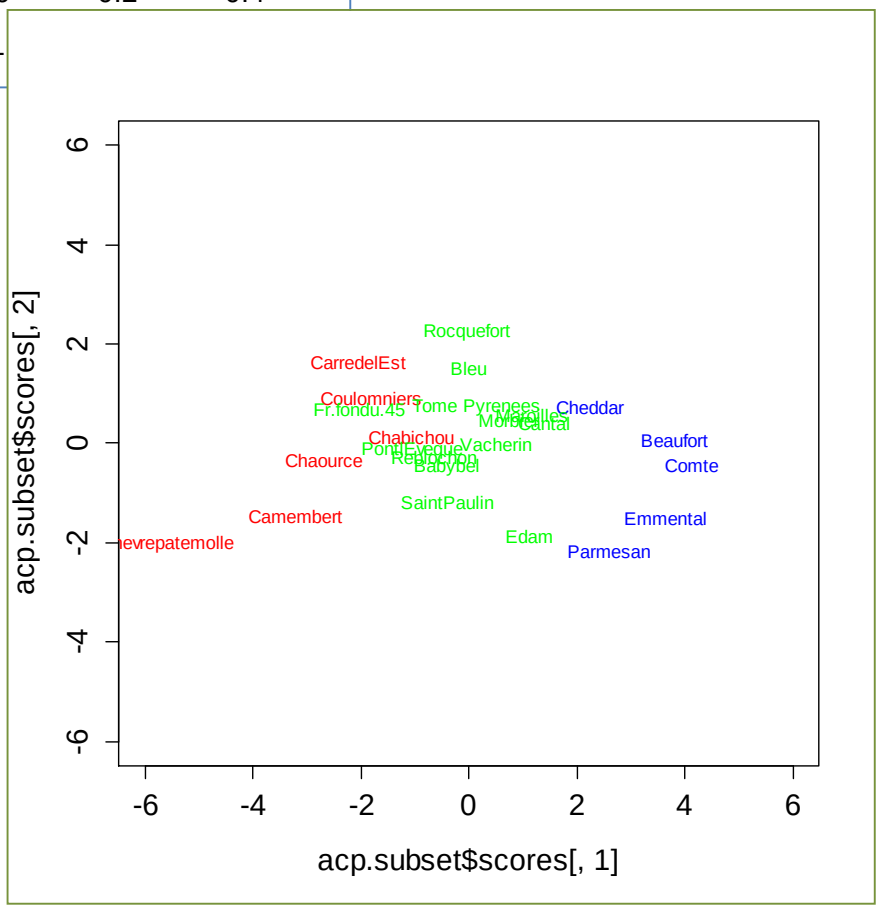
Approfondir l'analyse
Retirer les fromages frais du jeu de données (2/2)



Les résultats ne contredisent pas l'analyse précédente. Mais les concomitances et oppositions apparaissent plus clairement, notamment sur le 1er facteur.

Le positionnement de folates est plus explicite.

On peut aussi s'interroger sur l'intérêt de conserver 3 variables qui portent la même information dans l'analyse (lipides, cholestérol et calories).



Les groupes sont constitués essentiellement sur le 1er facteur.

Quelques fromages ont changé de camp par rapport à l'analyse précédente : carré de l'est et coulommiers d'une part ; cheddar d'autre part.

Et on peut faire bien d'autres choses encore...

Références :

1. Chavent M. , Page de cours - Source des données « fromages.txt »
2. Lebart L., Morineau A., Piron M., « Statistique exploratoire multidimensionnelle », Dunod, 2006.
3. Saporta G., « Probabilités, Analyse de données et Statistique », Dunod, 2006.
4. Tenenhaus M., « Statistique : Méthodes pour décrire, expliquer et prévoir », Dunod, 2007.