Cluster Analysis with R

HAC and K-Means

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

Data importation, descriptive statistics

DATASET

Goal of the study

This tutorial describes a cluster analysis process. We deal with a set of cheeses (29 instances) characterized by their nutritional properties (9 variables). The aim is to determine groups of homogeneous cheeses in view of their properties.

We inspect and test two approaches using two procedures of the R software: the Hierarchical Agglomerative Clustering algorithm (hclust) ; and the K-Means algorithm (kmeans).

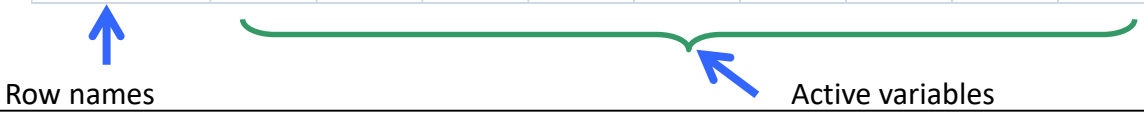
The data file "fromage.txt" comes from the teaching page of Marie Chavent from the University of Bordeaux. The excellent course materials and corrected exercises (commented R code) available on its website will complete this tutorial, which is intended firstly as a simple guide for the introduction of the R software in the context of the cluster analysis.

Processing tasks

- Importing the dataset. Descriptive statistics.
- Cluster analysis with hclust() and kmeans()
- Potential solutions for determining the number of clusters
- Description and interpretation of the clusters

Cheese dataset

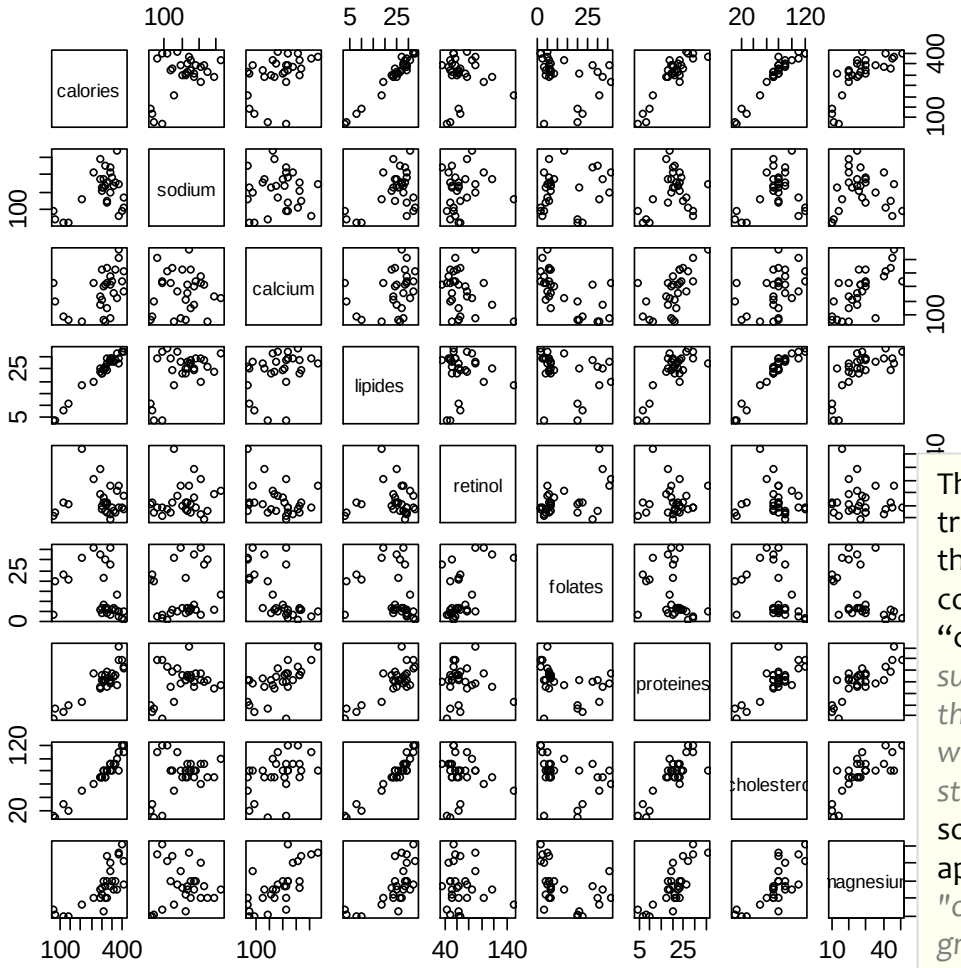
Fromages	calories	sodium	calcium	lipides	retinol	folates	proteines	cholesterol	magnesium
CarredelEst	314	353.5	72.6	26.3	51.6	30.3	21	70	20
Babybel	314	238	209.8	25.1	63.7	6.4	22.6	70	27
Beaufort	401	112	259.4	33.3	54.9	1.2	26.6	120	41
Bleu	342	336	211.1	28.9	37.1	27.5	20.2	90	27
Camembert	264	314	215.9	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



Data file

Data importation, descriptive statistics and plotting

```
#modifying the default working directory
setwd(" … my directory …")

#loading the dataset - options are essential
fromage <- read.table(file="fromage.txt",header=T,row.names=1,sep="\t",dec=".")

#displaying the first data rows
print(head(fromage))

#summary - descriptive statistics
print(summary(fromage))

#pairwise scatterplots
pairs(fromage)
```



This kind of graph is never trivial. For instance, we note that (1) "lipides" is highly correlated to "calories" and "cholesterol"(*this is not really surprising, but it means also that the same phenomenon will weigh 3 times more in the study*) ; (2) in some situations, some groups seem naturally appeared (*e.g. "proteines" vs. "cholesterol", we identify a group in the southwest of the scatterplot, with high inter-groups correlation*).

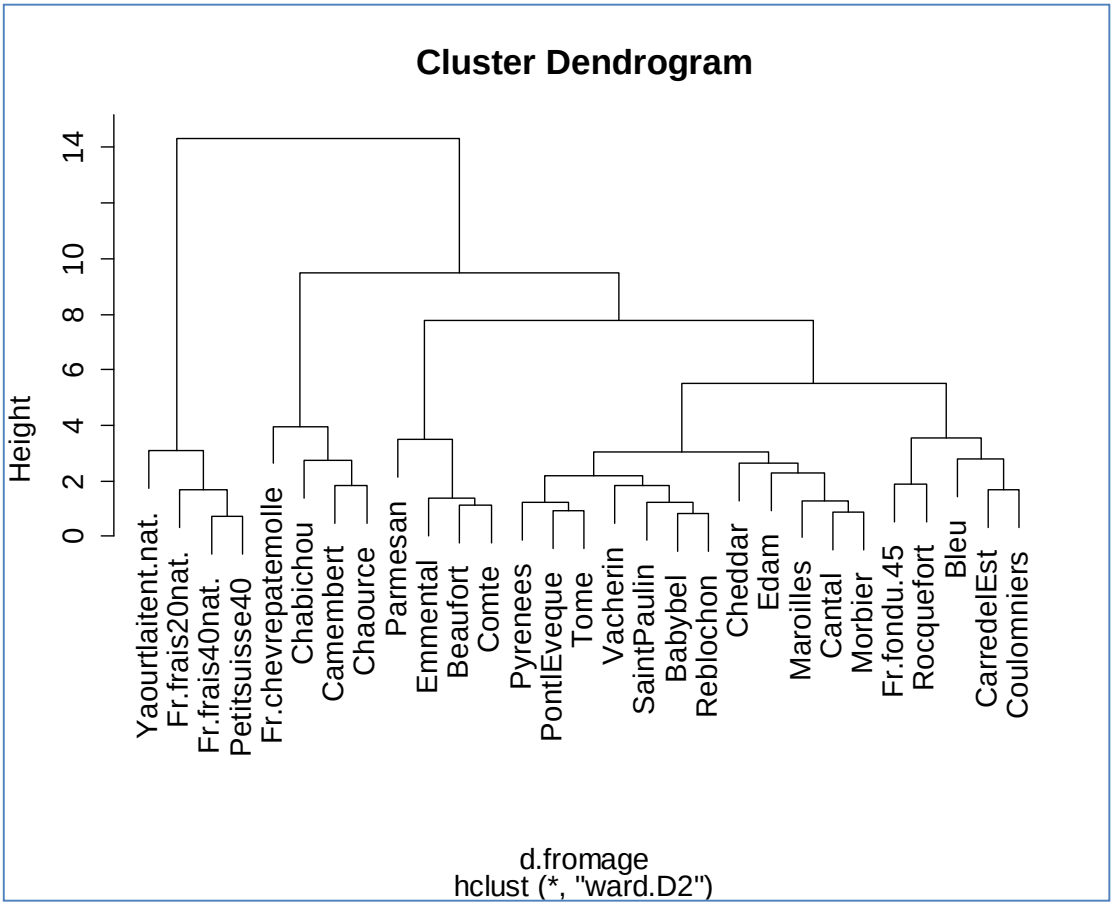
Hierarchical Agglomerative Clustering

HAC (HCLUST)

Hierarchical Agglomerative Clustering

hclust() function – "stats" package – Always available

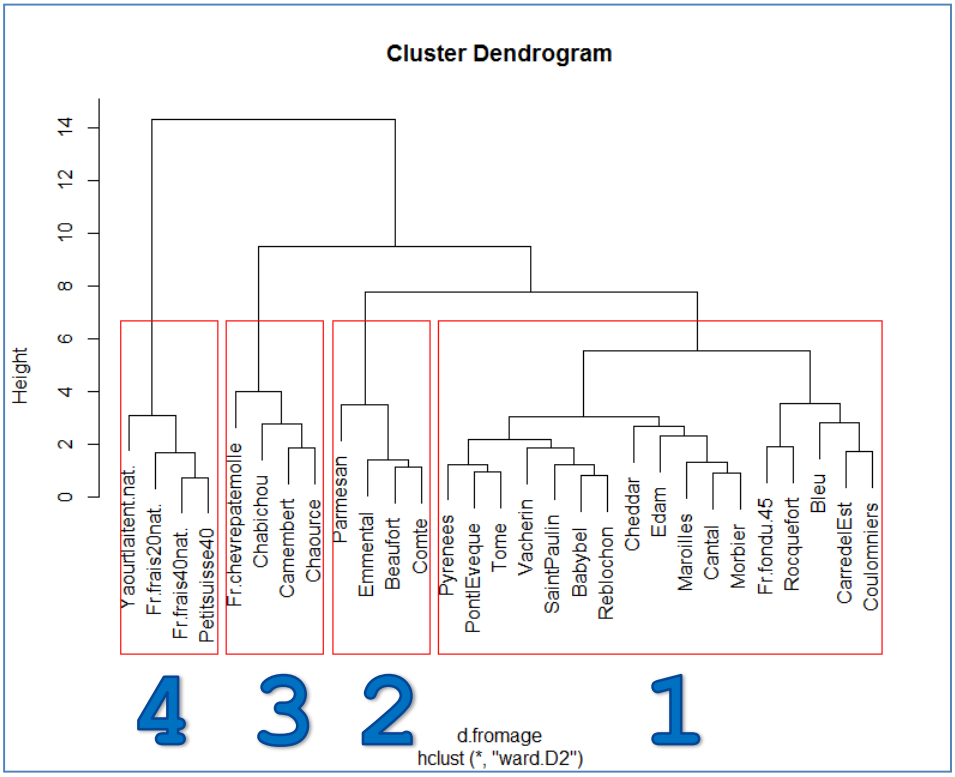
```
# standardizing the variables
# which allows to control the over influence of variables with high variance
fromage.cr <- scale(fromage,center=T,scale=T)

# pairwise distance matrix
d.fromage <- dist(fromage.cr)

# HAC – Ward approach - https://en.wikipedia.org/wiki/Ward's_method
# method = « ward.D2 » corresponds to the true Ward's method
# using the squared distance
cah.ward <- hclust(d.fromage,method="ward.D2")

# plotting the dendrogram
plot(cah.ward)
```



The Dendrogram suggests a partitioning in 4 groups. It is noted that a group of cheeses, the "fresh Cheeses" (far left), seems very different to the others, to the point that we could have considered also a partitioning in 2 groups only. We will discuss this dimension longer when we combine the study with a principal component analysis (PCA).

```
# dendrogram with highlighting of the groups
rect.hclust(cah.ward,k=4)

# partition in 4 groups
groupes.cah <- cutree(cah.ward,k=4)

# assignment of the instances to clusters
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



The 4th group corresponds to the "fresh cheeses".
The 3rh to the "soft cheeses".
The 2nd to the "hard cheeses".
The 1st is a bit the "catch all" group.

My skills about cheese stop there (thanks to Wikipedia). For characterization using the variables, it is necessary to go through univariate (easy to read and interpret) or multivariate statistical techniques (which take into account the relationships between variables).

K-Means Clustering – Relocation method

K-MEANS

K-Means clustering

The R's kmeans() function ("stats" package also, *such as hclust*)

```
# k-means from the standardized variables
# center = 4 – number of clusters
# nstart = 5 – number of trials with different starting centroids
# indeed, the final results depends on the initialization for kmeans
groupes.kmeans <- kmeans(fromage.cr,centers=4,nstart=5)

# displaying the results
print(groupes.kmeans)

# crosstabs with the clusters coming from HAC
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"
```

Size of each group

Mean for each variable (standardized) conditionally to the group membership

Cluster membership for each case

Variance explained: 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

Correspondences between HAC and k-Means

The 4th group of the HAC is equivalent to the 1st group of the K-Means. After that, there are some connections, but they are not exact.

Note: You may not have exactly the same results with the K-means.

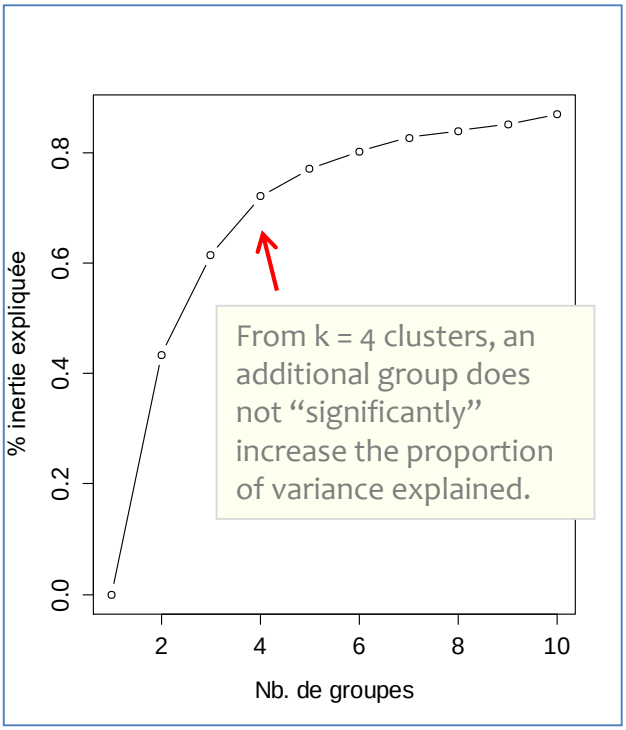
K-Means Algorithm

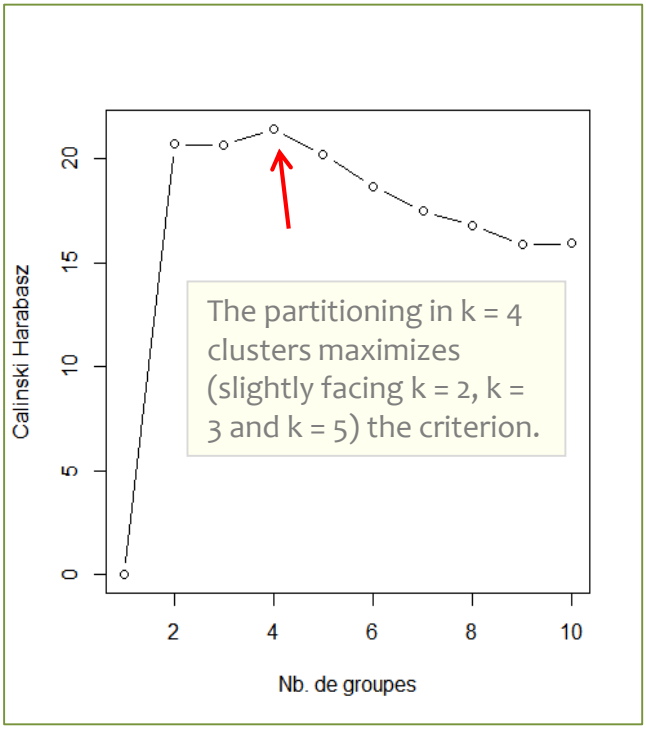
Determining the number of clusters

K-Means, unlike the CAH, does not provide a tool to help us to detect the number of clusters. We have to program them under R or use procedures provided by dedicated packages. The approach is often the same: we vary the number of groups, and we observe the evolution of an indicator of quality of the partition.

Two approaches here: (1) the elbow method, we monitor the percentage of variance explained when we increase the number of clusters, we detect the elbow indicating that an additional group does not increase significantly this proportion ; (2) Calinski Harabasz criterion from the "fpc" package (*the aim is to maximize this criterion*).
See: https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set

```
# (1) elbow method
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
}
# plotting
plot(1:10,inertie.expl,type="b",xlab="Nb. de groupes",ylab="% inertie expliquée")

# (2) Calinski Harabasz index - fpc package
library(fpc)
# values of the criterion according to the number of clusters
sol.kmeans <- kmeansruns(fromage.cr,krange=2:10,criterion="ch")
# plotting
plot(1:10,sol.kmeans$crit,type="b",xlab="Nb. of groups",ylab="Calinski-Harabasz")
```





Conditional descriptive statistics and visualization

INTERPRETING THE CLUSTERS

Interpreting the clusters
Conditional descriptive statistics

The idea is to compare the means of the active variables conditionally to the groups. It is possible to quantify the overall amplitude of the differences with the proportion of explained variance. The process can be extended to auxiliary variables that was not included in the clustering process, but used for the interpretation of the results. For the categorical variables, we will compare the conditional frequencies. The approach is straightforward and the results easy to read. We should remember, however, that we do not take into account the relationship between the variables in this case (some variables may be highly correlated).

```
#Function for calculating summary statistics – y cluster membership variable
stat.comp <- function(x,y){
  #number of clusters
  K <- length(unique(y))
  #nb. Of instances
  n <- length(x)
  #overall mean
  m <- mean(x)
  #total sum of squares
  TSS <- sum((x-m)^2)
  #size of clusters
  nk <- table(y)
  #conditional mean
  mk <- tapply(x,y,mean)
  #between (explained) sum of squares
  BSS <- sum(nk * (mk - m)^2)
  #collect in a vector the means and the proportion of variance explained
  result <- c(mk,100.0*BSS/TSS)
  #set a name to the values
  names(result) <- c(paste("G",1:K),"% epl.")
  #return the results
  return(result)
}

#applying the function to the original variables of the dataset
#and not to the standardized variables
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

The definition of the groups is – above all – dominated by fat content (lipids, cholesterol and calories convey the same idea) and protein.

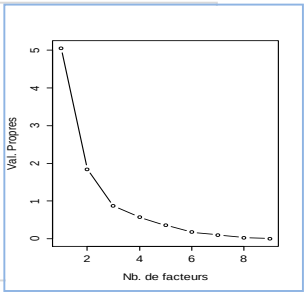
Group 4 is strongly determined by these variables, the conditional means are very different.

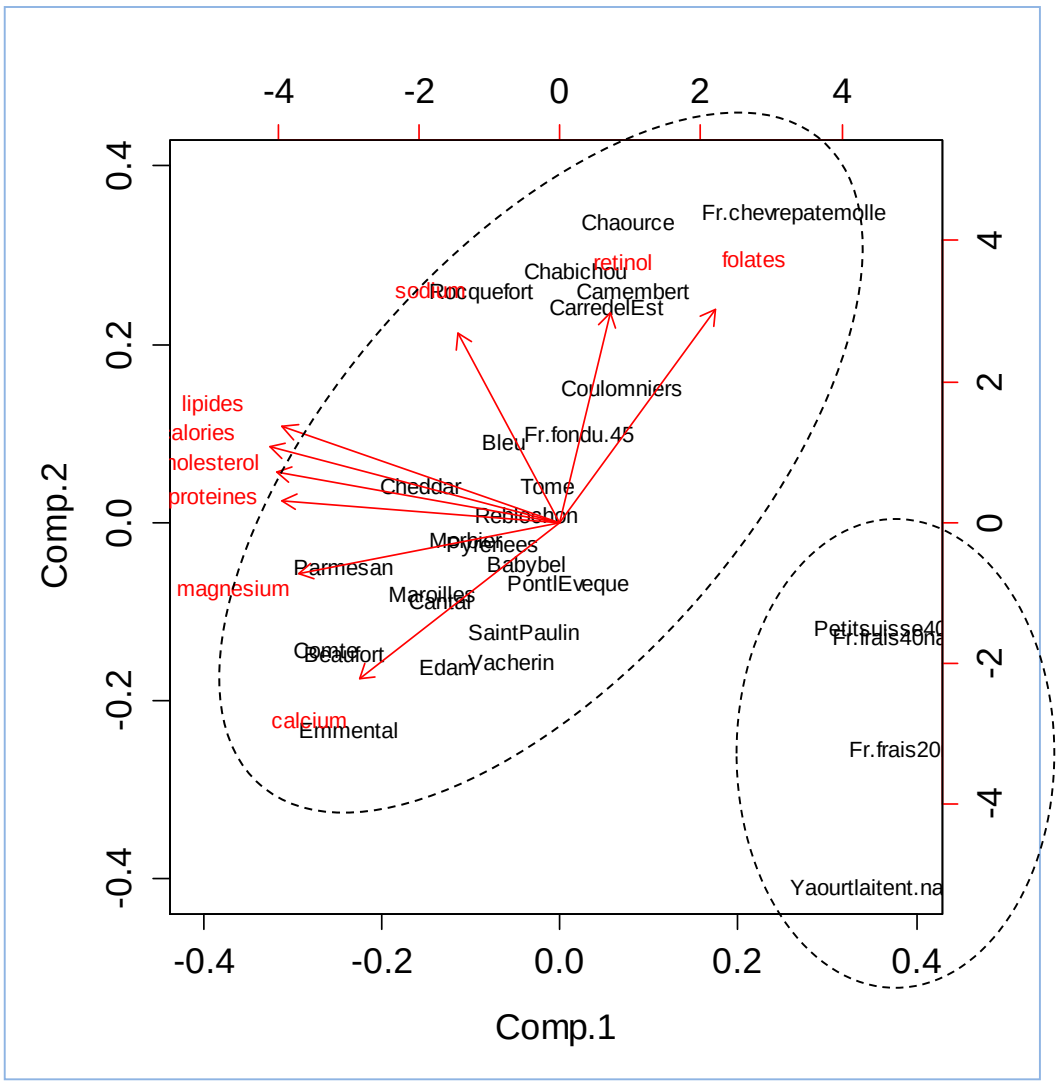
Interpreting the clusters

Principal component analysis (PCA) **(1/2)**

When we combine the cluster analysis with factor analysis, we benefit from the data visualization to enhance the analysis. The main advantage is that we can take the relationship between the variables into account. But, on the other hand, we must also be able to read the outputs of the factor analysis correctly.

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

#scree plot – Retain the two first factors
plot(1:9,acp$sdev^2,type="b",xlab="Nb. de factors",ylab="Eigen Val.")

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



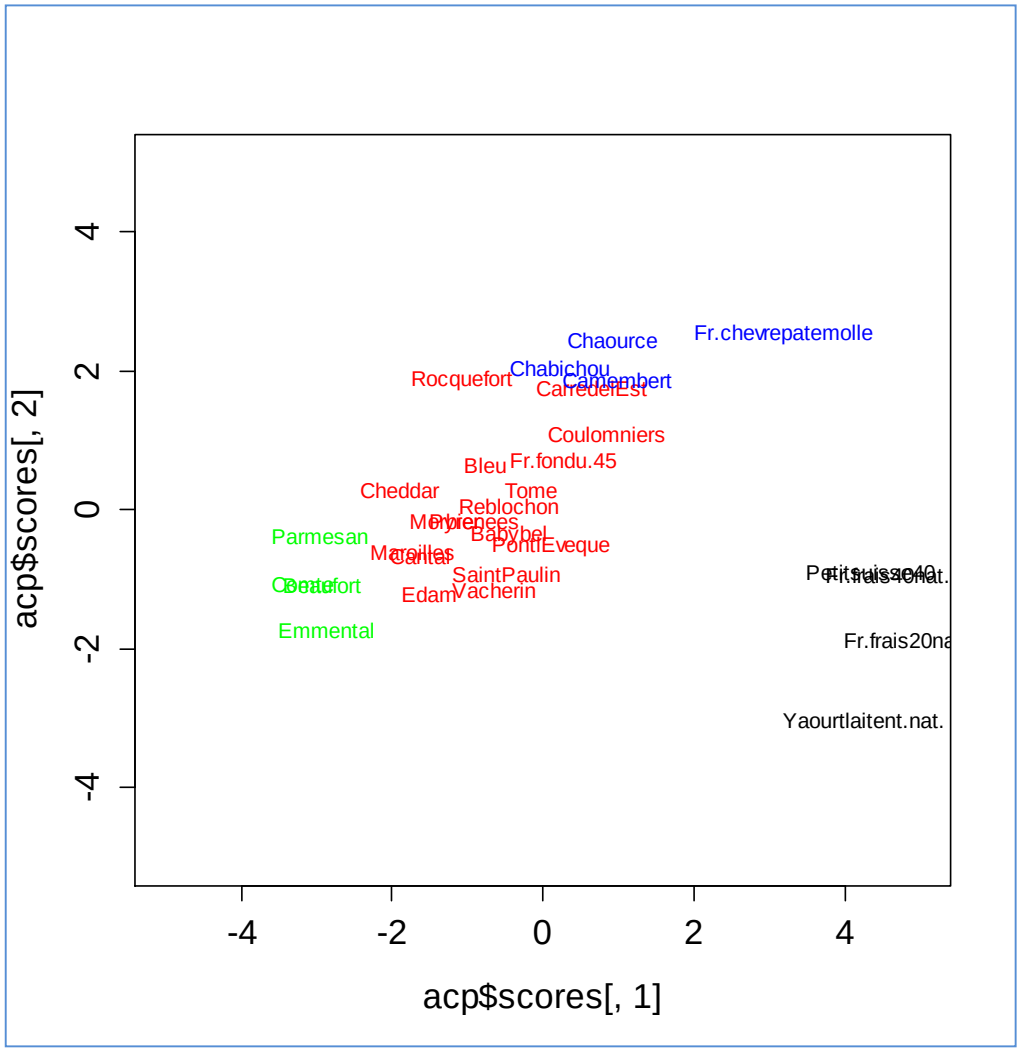


We note that there is a problem. The "fresh cheeses" group dominates the available information. The other cheeses are compressed into the left part of the scatter plot, making difficult to distinguish the other groups.

Interpreting the clusters
Principal component analysis (PCA) **(2/2)**

Thus, if we understand easily the nature of the 4th group (fresh cheeses), the others are difficult to understand when they are represented into the individuals factor map (first two principal components).

```
#highlight the clusters into the individuals factor map of PCA
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))
```



For groups 1, 2 and 3 (green, red, blue), we perceive from the biplot graph of the previous page that there is something around the opposition between nutrients (lipids/calories/cholesterol, proteins, magnesium, calcium) and vitamins (retinol, folates). But, in what sense exactly?

Reading is not easy because of the disruptive effect of the 4th group.

In the light of the results of PCA

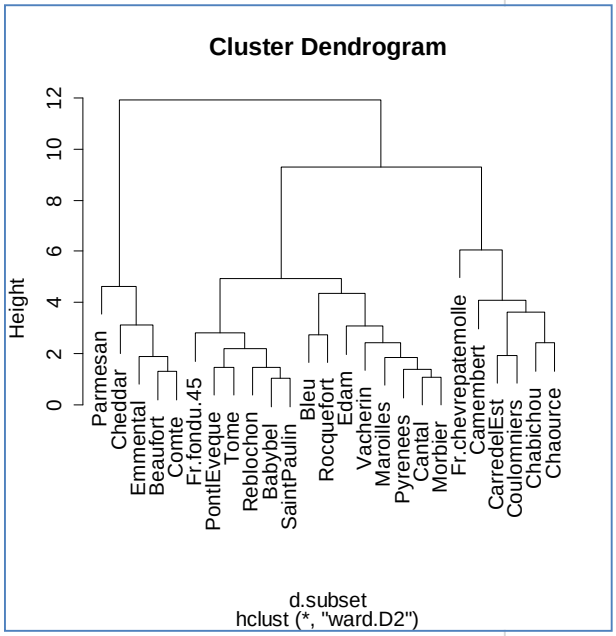
COMPLEMENT THE ANALYSIS

Complement the analysis

Remove the "fresh cheeses" group from the dataset (1/2)

The fresh cheeses are so special – far from all the other observations – that they mask interesting relationships that may exist between the other products. We resume the analysis by excluding them from the treatments.

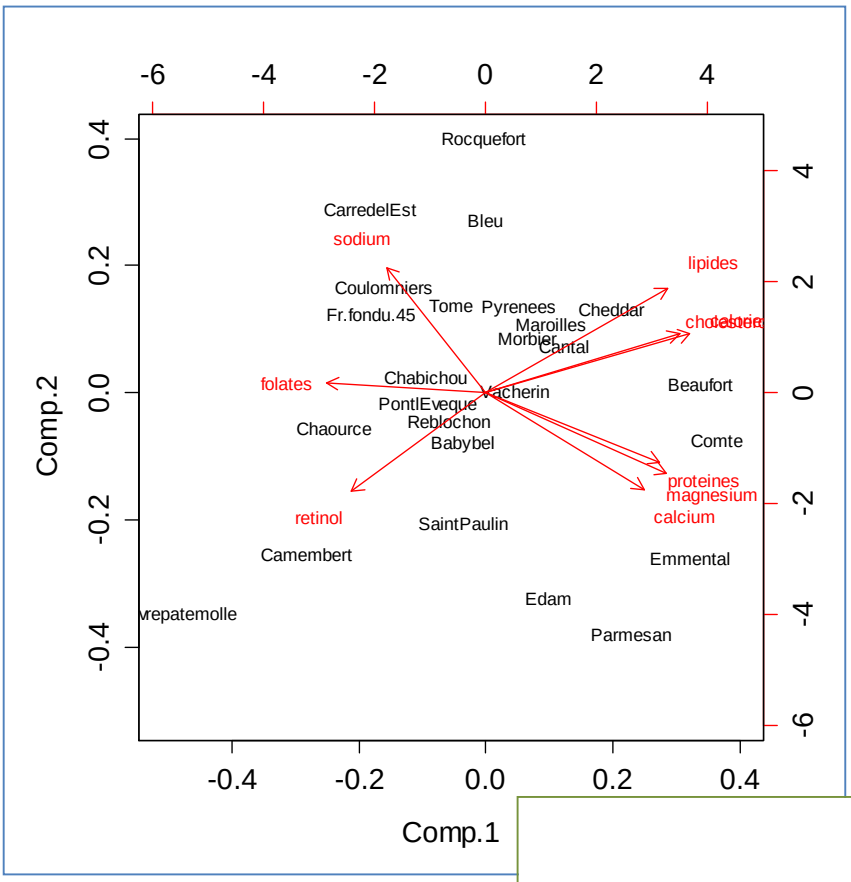
```
#remove the instance corresponding to the 4th group
fromage.subset <- fromage[groupes.cah!=4,]

#standardizing again the dataset
fromage.subset.cr <- scale(fromage.subset,center=T,scale=T)

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

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

#displaying the dendrogram
plot(cah.subset)

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

#displaying the group membership for each case
print(sort(groupes.subset))

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

#scree plot
plot(1:9,acp.subset$sdev^2,type="b")

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

#scatter plot - individuals factor map
plot(acp.subset$scores[,1],acp.subset$scores[,2],type="n",xlim=c(-6,6),ylim=c(-6,6))

#row names + group membership
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))
```



We can identify three groups. There is less the disrupting phenomenon observed in the previous analysis.

Complement the analysis

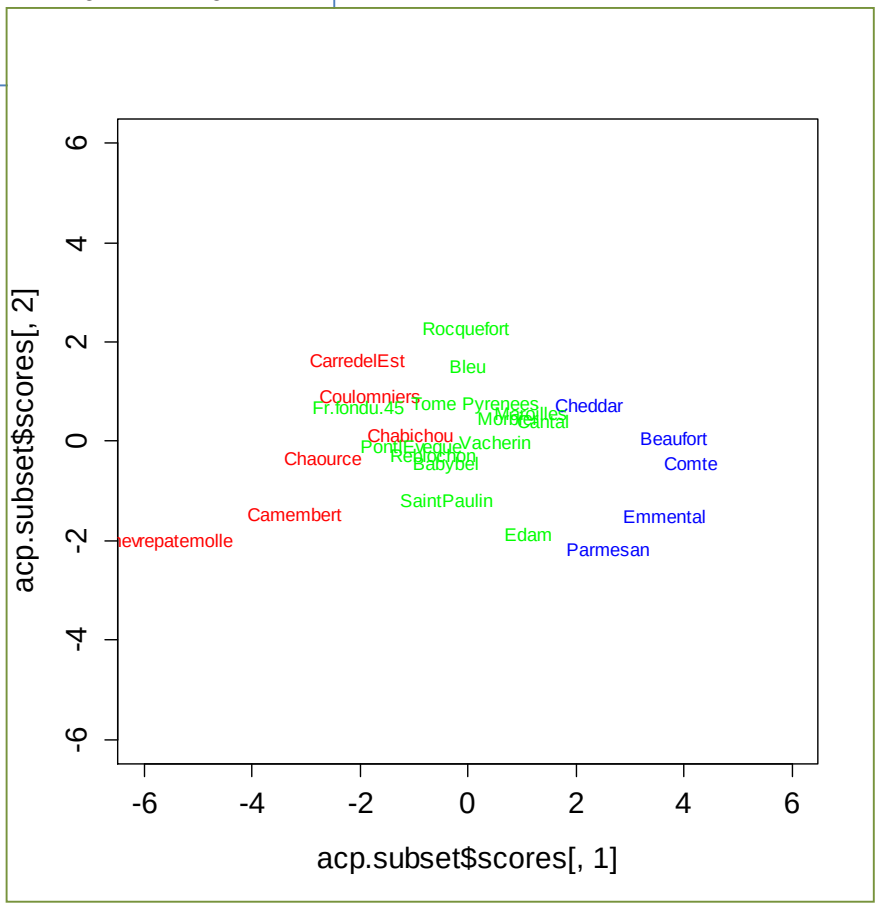
Remove the "fresh cheeses" group from the dataset (2/2)



The results do not contradict the previous analysis. But the associations and oppositions appear more clearly, especially on the first factor.

The location of “folates” is more explicit.

We can also wonder about the interest of keeping 3 variables that convey the same information in the analysis (lipids, cholesterol and calories).



The groups are mainly distinguishable on the first factor.

Some cheeses are assigned to other groups compared to the previous analysis: “Carré de l’est” and Coulommiers on the one hand; Cheddar on the other hand.

And we can do much more amazing things ...

French references:

1. Chavent M. , Teaching page - Source of **"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.