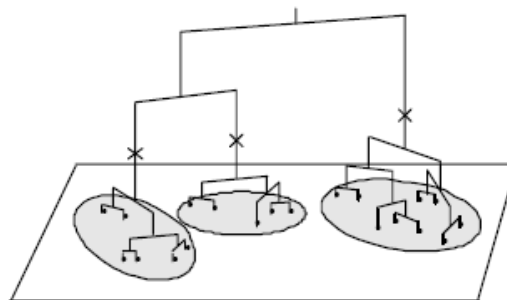
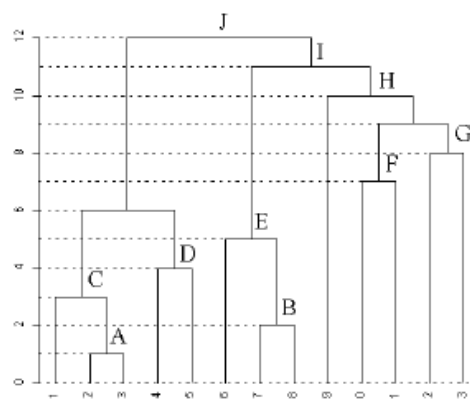


Documents de cours : CAH

A. Introduction



B. Exemples de distance

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

```
01100001010010...
01010001100010...
```

On note n le nombre d'enregistrements qui est la somme de
 a est le nombre de concordances 11
 b le nombre de concordances 10
 c le nombre de concordances 01
 d le nombre de concordances 00.

Ainsi deux espèces sont présentes ensemble dans un même relevé a fois, deux relevés possèdent a espèces en commun. Les deux objets définissent donc la table de contingence 2-2 :

	1	0	Tot
1	a	b	$a+b$
0	c	d	$c+d$
Tot	$a+c$	$b+d$	n

Les quatre nombres de la table définissent une similarité entre les deux objets. On peut utiliser :

$$S_1 = \frac{a}{a+b+c}$$

Indice de communauté de Jaccard

$$S_2 = \frac{a+d}{n}$$

Indice de Sokal & Michener

$$S_3 = \frac{a}{a+2(b+c)}$$

Indice de Sokal & Sneath

$$S_4 = \frac{a+d}{a+2(b+c)+d}$$

Indice de Rogers et Tanimoto

$$S_5 = \frac{2a}{2a+b+c}$$

Indice de Sorensen

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

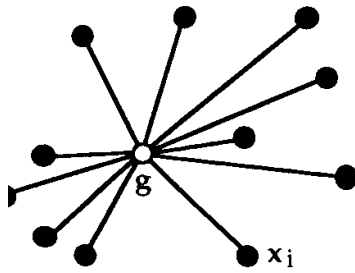
Soit A un tableau de fréquences alléliques avec t lignes (populations) et m colonnes (formes alléliques). Soit v le nombre de loci. Le locus j a $m(j)$ formes alléliques.

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

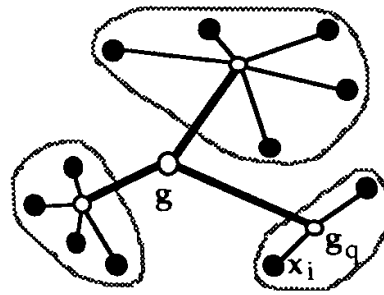
1 — Distance de Nei ²² (Voir ²³) :

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

C. Décomposition de l'inertie



Inertie totale



= Inertie inter-classe + Inertie intra-classe

D. Centres mobiles

Quelques méthodes

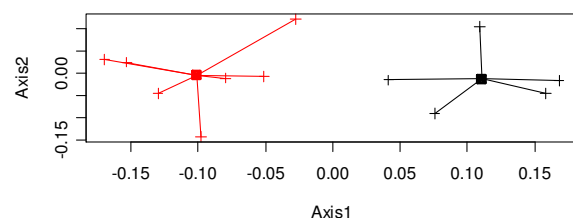
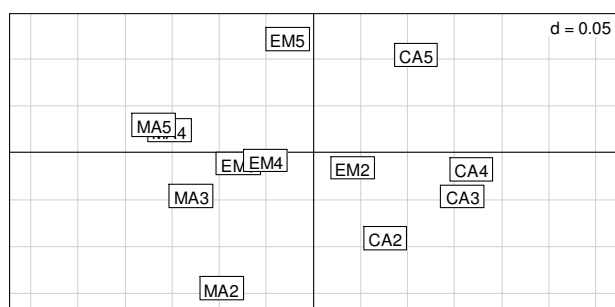
pam partitionning around medoids (more robust)
 clara other partitionning, well-suited for large data sets
 daisy dissimilarity matrix: qualitative or quantitative variables
 dist dissimilarity matrix: quantitative variables only
 fanny fuzzy clustering
 agnes Computes agglomerative hierarchical clustering of the dataset

Exemple

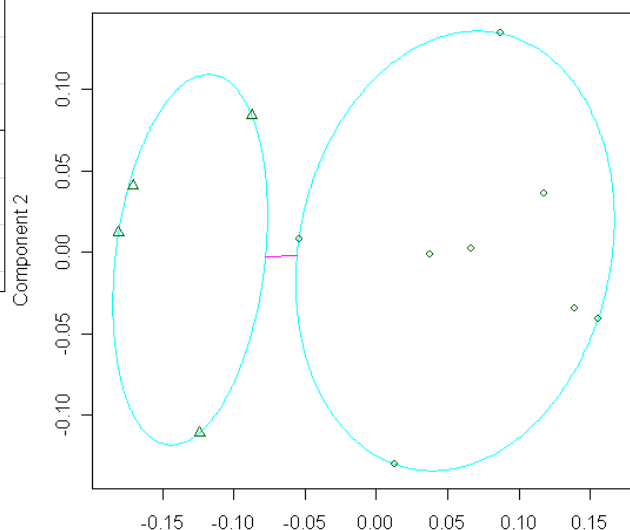
```

### kmeans et pam sur le tableau csp
csp=read.table("csp.txt")
library(ade4)
afc=dudi.coa(csp,scannf=FALSE)
cl=kmeans(afc$li,2)
cl
plot(afc$li, col = cl$cluster, pch=3)
points(cl$centers, col = 1:2, pch = 7, lwd=3)
segments( afc$li[cl$cluster==1,1], afc$li[cl$cluster==1,2],cl$centers[1,1],
cl$centers[1,2])
segments( afc$li[cl$cluster==2,1], afc$li[cl$cluster==2,2],cl$centers[2,1],
cl$centers[2,2],col=2)
library(cluster)
clusplot( afc$li, pam(d,2)$clustering )

```



CLUSPLOT(afc\$li)



Component 1
These two components explain 100 % of the point variability.

E. CAH

Quelques critères d'agrégation

1 - ward

$$D(A \cup B, X) = \frac{(n_A + n_X)D(A, X)}{(n_A + n_B + n_X)} + \frac{(n_B + n_X)D(B, X)}{(n_A + n_B + n_X)} - \frac{n_X D(A, B)}{(n_A + n_B + n_X)}$$

2 - single

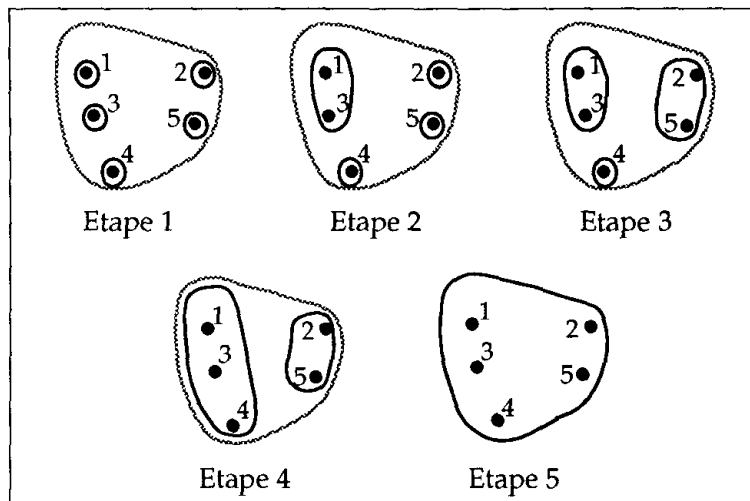
$$D(A \cup B, X) = \min(D(A, X), D(B, X)) = \frac{1}{2}D(A, X) + \frac{1}{2}D(B, X) - \frac{1}{2}|D(A, X) - D(B, X)|$$

3 - complete

$$D(A \cup B, X) = \max(D(A, X), D(B, X)) = \frac{1}{2}D(A, X) + \frac{1}{2}D(B, X) + \frac{1}{2}|D(A, X) - D(B, X)|$$

4 - average

$$D(A \cup B, X) = \frac{n_A D(A, X) + n_B D(B, X)}{n_A + n_B}$$



Exemple :

```
### arbre sur CSP ###
csp=read.table("csp.txt")
library(ade4)
afc=dudi.coa(csp,scannf=FALSE)
Distance entre lignes
d=dist.dudi(afc); d
cah=hclust(d^2,"ward")
par(mfrow=c(2,1))
barplot(cah$height[order(cah$height,decreasing=TRUE)])
plot(cah,hang=-1)
cutree(cah,h=0.05)
rect.hclust(cah,h=0.05)
par(mfrow=c(2,1))
s.label(afc$li)
s.class(dfxy=afc$li,fac=as.factor(cutree(cah,h=0.05)))
```

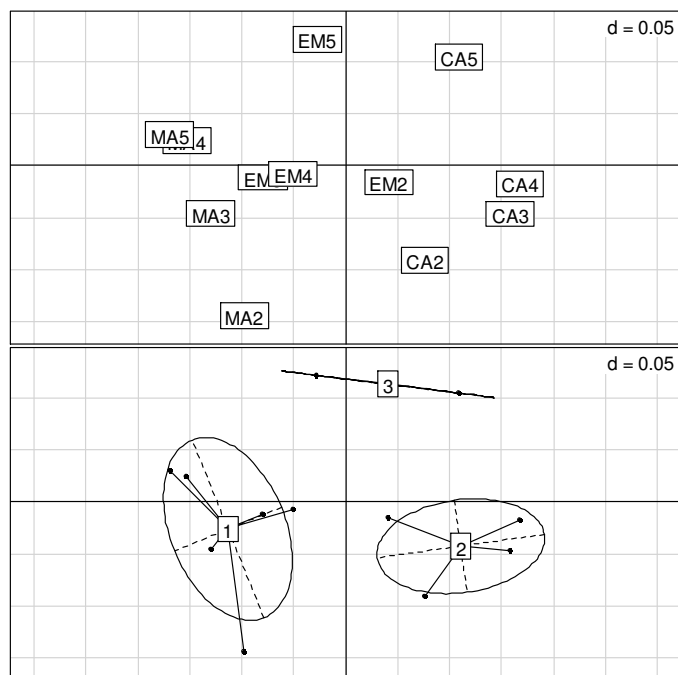
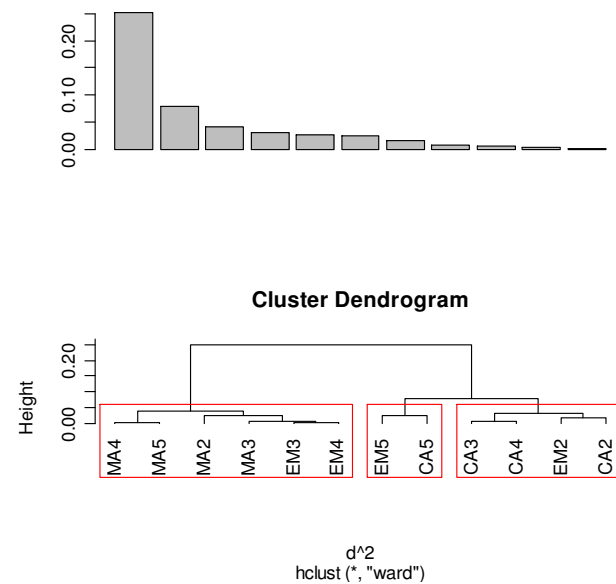
Distance entre colonnes

```
par(mfrow=c(2,1))
barplot(cah$height[order(cah$height,decreasing=TRUE)])
d=dist.dudi(afc,amongrow=FALSE)
cah=hclust(d^2,"ward")
names(cah)
cah$height
plot(cah,hang=-1)
cutree(cah,h=0.05)
rect.hclust(cah,h=0.05)
par(mfrow=c(2,1))
s.label(afc$co)
s.class(dfxy=afc$co,fac=as.factor(cutree(cah,h=0.05)))
```

Entre lignes

```
> cah$height
```

```
[1] 0.0006867328 0.0026901566 0.0050581140 0.0078069586 0.0155507255
[6] 0.0248064759 0.0257161929 0.0304247390 0.0402000210 0.0799327403
[11] 0.2509852077
```

**Entre colonnes**

```
> cah$height
```

```
[1] 0.01169817 0.01346008 0.02225055 0.02547943 0.11309608 0.26209568
```

