# Classification des variables qualitatives

## Regroupement de variables, regroupement de modalités

Ricco RAKOTOMALALA

Université Lumière Lyon 2

# PLAN

1. Classification de variables. Quoi ? Pourquoi ?
   a. CAH à partir d'une matrice de dissimilarité
   b. Insuffisances de la classification de variables
2. Classification de modalités. Quoi ? Pourquoi ?
   a. Distance entre modalités – Indice de Dice
   b. CAH sur les modalités
   c. Interprétation des classes
3. Autres pistes pour la classification de modalités
4. Bilan
5. Bibliographie

# Classification de variables qualitatives

Pourquoi ? Quel intérêt ?

# Classification de variables

Démarche : créer des groupes de variables similaires c.-à-d. porteuses de la même dimension d'information

➔ Les variables dans un même groupe sont similaires (liées entre elles)

➔ Les variables dans des groupes différents sont dissemblables (aussi orthogonales que possible)

Quel intérêt ?

1. Comprendre les structures sous-jacentes aux données. Constituer un résumé des informations portées par les données (approche complémentaire à la classification des individus).

2. Détecter les redondances, en vue par exemple d'une réduction de nombre de variables dans un autre processus (ex. analyse supervisée)
   a. En prétraitement, pour organiser ou réduire l'espace de recherche
   b. En post-traitement, pour positionner les variables non sélectionnées dans les modèles

# Un exemple : vote au congrès (1984)

n = 435 individus (député US)

p = 6 variables actives

| Variable | Modalités | Statut |
|---|---|---|
| affiliation | democrat, republican | illustrative |
| budget | yes, no, neither | active |
| physician | yes, no, neither | active |
| salvador | yes, no, neither | active |
| nicaraguan | yes, no, neither | active |
| missile | yes, no, neither | active |
| education | yes, no, neither | active |

Affiliation politique

Variable illustrative

Vote effectué sur différents thèmes, 3 valeurs possibles : yes, no, neither (ni l'un, ni l'autre c.-à-d. absent, a été présent mais n'a pas pris part au vote ex. conflit d'intérêt)

Variables actives

Comprendre les votes qui sont le plus liés entre eux

Etablir les relations avec l'affiliation politique

Remarque : Votes liés ne veut pas dire « vote 'yes' concomitants » -- un vote 'yes' pour un sujet peut être lié à un vote 'no' pour un autre sujet !!!

# CAH à partir d'une matrice de dissimilarités

S'appuyer sur le V de Cramer pour mesurer la liaison entre les variables

# Mesurer la liaison entre 2 variables qualitatives

| $A \setminus B$ | $b_1$ | $b_l$ | $b_L$ | Total |
|---|---|---|---|---|
| $a_1$ | | | | |
| | | $\vdots$ | | |
| $a_k$ | $\cdots$ | $n_{kl}$ | $\cdots$ | $n_{k.}$ |
| | | $\vdots$ | | |
| $a_K$ | | | | |
| Total | | $n_{.l}$ | | $n$ |

## KHI-DEUX d'écart à l'indépendance

$$\chi^2 = \sum_k \sum_l \frac{(n_{kl} - e_{kl})^2}{e_{kl}}$$

$$e_{kl} = \frac{n_{k.} \times n_{.l}}{n}$$

# P(AB) observés

# P(A) x P(B) Sous hyp. d'indépendance

## V de Cramer

$$v = \sqrt{\frac{\chi^2}{n \times \min(K-1, L-1)}}$$

- Symétrique
- $0 \leq v \leq 1$

Ex.

| Nombre de budget | physician | | | |
|---|---|---|---|---|
| budget | n | neither | y | Total général |
| n | 25 | | 146 | 171 |
| neither | 3 | 6 | 2 | 11 |
| y | 219 | 5 | 29 | 253 |
| Total général | 247 | 11 | 177 | 435 |

$\chi^2 = 355.48$

$p.value < 0.0001$

$v = 0.639$

Forte liaison

Liaison significative

# Matrice des similarités – Matrice des dissimilarités

```
#function for calculating Cramer's v
cramer <- function(y,x){
  K <- nlevels(y)
  L <- nlevels(x)
  n <- length(y)
  chi2 <- chisq.test(y,x,correct=F)
  print(chi2$statistic)
  v <- sqrt(chi2$statistic/(n*min(K-1,L-1)))
  return(v)
}
```

## Matrice des similarités (v de Cramer)

| | budget | physician | salvador | nicaraguan | missile | education |
|---|---|---|---|---|---|---|
| budget | 1 | 0.639 | 0.507 | 0.517 | 0.439 | 0.475 |
| physician | 0.639 | 1 | 0.576 | 0.518 | 0.471 | 0.509 |
| salvador | 0.507 | 0.576 | 1 | 0.611 | 0.558 | 0.470 |
| nicaraguan | 0.517 | 0.518 | 0.611 | 1 | 0.545 | 0.469 |
| missile | 0.439 | 0.471 | 0.558 | 0.545 | 1 | 0.427 |
| education | 0.475 | 0.509 | 0.470 | 0.469 | 0.427 | 1 |

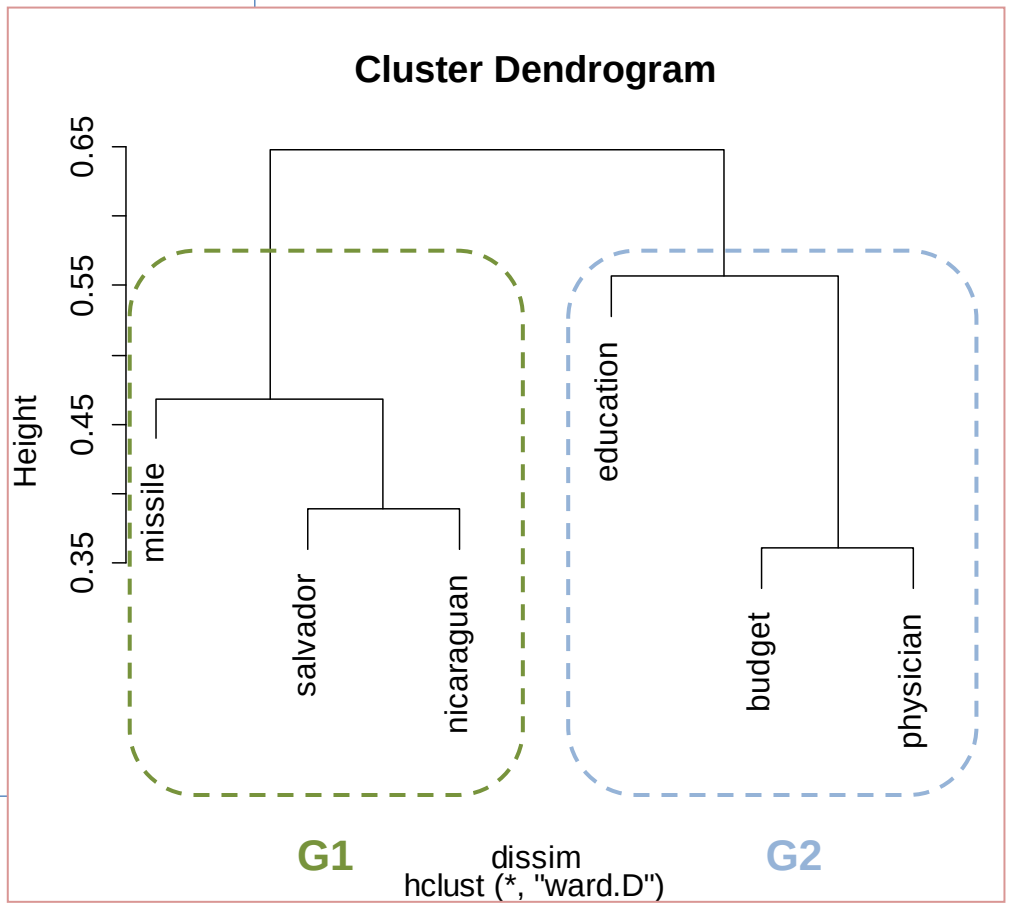
## Matrice des dissimilarités (1-v)

| | budget | physician | salvador | nicaraguan | missile | education |
|---|---|---|---|---|---|---|
| budget | 0 | 0.361 | 0.493 | 0.483 | 0.561 | 0.525 |
| physician | 0.361 | 0 | 0.424 | 0.482 | 0.529 | 0.491 |
| salvador | 0.493 | 0.424 | 0 | 0.389 | 0.442 | 0.530 |
| nicaraguan | 0.483 | 0.482 | 0.389 | 0 | 0.455 | 0.531 |
| missile | 0.561 | 0.529 | 0.442 | 0.455 | 0 | 0.573 |
| education | 0.525 | 0.491 | 0.530 | 0.531 | 0.573 | 0 |

On peut se baser sur cette matrice pour réaliser une CAH

# hclust() sous R – Distance = (1 – v), méthode = Ward

```
#similarity matrix
sim <- matrix(1,nrow=ncol(vote.active),ncol=ncol(vote.active))
rownames(sim) <- colnames(vote.active)
colnames(sim) <- colnames(vote.active)
for (i in 1:(nrow(sim)-1)){
 for (j in (i+1):ncol(sim)){
   y <- vote.active[,i]
   x <- vote.active[,j]
   sim[i,j] <- cramer(y,x)
   sim[j,i] <- sim[i,j]
 }
}
#distance matrix
dissim <- as.dist(1-sim)
#clustering
tree <- hclust(dissim,method="ward.D")
plot(tree)
```




On obtient une vision des structures de liaisons entre les variables. Ex. ‘’budget’’ et ‘’physician’’ sont liées c.-à-d. il y a une forte cohérence des votes (v = 0.639) ; budget et salvador moins (v = 0.507), etc… mais on ne sait pas sur quoi repose ces relations…

# La méthode ClustOfVar (Chavent et al, 2012)

Définir la notion de variable « moyenne » (variable latente), représentative d'un groupe de variables qualitatives.

F = 1er axe de l'ACM (analyse des correspondances multiples)
$\eta(.)$ rapport de corrélation
$\lambda$ dispersion (inertie) liée au groupe

$$\lambda = \sum_{j=1}^{p} \eta^2(X_j, F)$$

Ouvre la porte à différentes stratégies de construction de groupes.

➔ De type ascendant (CAH) : minimiser la perte d'inertie à chaque étape de regroupement

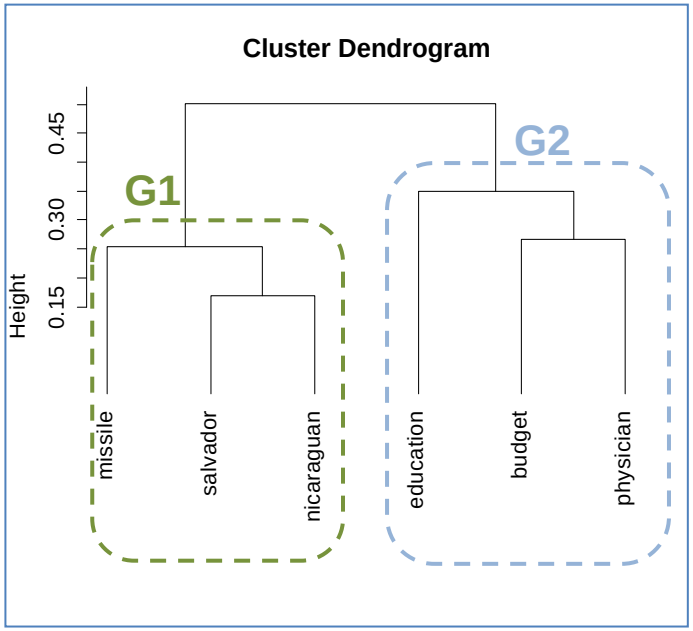
➔ De type K-Means : commencer avec une partition aléatoire initiale, réallouer itérativement les variables aux groupes au sens du max du carré du rapport de corrélation avec la variable latente

1. ClustOfVar s'applique au cas de variables mixtes (qualitatives, quantitatives), elle s'appuie sur l'AFDM (analyse factorielle des données mixtes) pour calculer la variable latente
2. C'est une généralisation de la méthode CLV (Vigneau et Qannari, 2003) qui ne traite que des variables quantitatives et s'appuye sur l'ACP (analyse en composantes principales)

# ClustOfVar sur le fichier « vote »

```
library(ClustOfVar)
arbre <- hclustvar(X.quali=vote.active)
plot(arbre)

mgroups <- kmeansvar(X.quali=vote.active,init=2,nstart=10)
print(summary(mgroups))
```




```
Data:
   number of observations:  435
   number of variables:  6
   number of clusters:  2

Cluster  1 :
          squared loading
budget                0.79
physician             0.83
education             0.76


Cluster  2 :
           squared loading
salvador              0.89
nicaraguan            0.86
missile               0.83


Gain in cohesion (in %):  32.5
```

On retrouve les mêmes résultats qu'avec la CAH basée sur la matrice des dissimilarités (1 – v de Cramer)

# Problème d'interprétation des résultats

Le regroupement de variables qualitatives donne une vision parcellaire de la structure des relations entre les variables…

# Analyser les groupes – Ex. G2

| Nombre de budget | physician | | | |
|---|---|---|---|---|
| budget | n | neither | y | Total général |
| n | 25 | | 146 | 171 |
| neither | 3 | 6 | 2 | 11 |
| y | 219 | 5 | 29 | 253 |
| Total général | 247 | 11 | 177 | 435 |

**v = 0.639**

| Nombre de budget | education | | | |
|---|---|---|---|---|
| budget | n | neither | y | Total général |
| n | 28 | 10 | 133 | 171 |
| neither | 4 | 4 | 3 | 11 |
| y | 201 | 17 | 35 | 253 |
| Total général | 233 | 31 | 171 | 435 |

**v = 0.475**

| Nombre de budget | education | | | |
|---|---|---|---|---|
| physician | n | neither | y | Total général |
| n | 202 | 16 | 29 | 247 |
| neither | 6 | 4 | 1 | 11 |
| y | 25 | 11 | 141 | 177 |
| Total général | 233 | 31 | 171 | 435 |

**v = 0.509**

**2 phénomènes sous-jacents**

Budget = y
Physician = n
Education = n

Budget = n
Physician = y
Education = y

Imaginez le boulot s'il y a un grand nombre de variables !

# Positionner les variables illustratives

```
#2 subgroups
groups <- cutree(tree,k=2)
print(groups)
#Cramer's v : affiliation vs. attributes
cv <- sapply(vote.active,cramer,x=vote.data$affiliation)
print(cv)
#mean of v for each group
m <- tapply(X=cv,INDEX=groups,FUN=mean)
print(m)
```

G1

G2

| Variable | Affiliation (v de Cramer) | **Moyenne (v)** |
|---|---|---|
| nicaraguan | 0.660 | **0.667** |
| missile | 0.629 | |
| education | 0.688 | |
| budget | 0.740 | **0.781** |
| physician | 0.914 | |
| salvador | 0.712 | |

- L'appartenance politique pèse (un peu) plus sur les votes en G2 qu'en G1 (consignes de votes, sujets « sensibles », ?)
- Mais on ne sait pas quel est le sens de la relation (republican → ?, democrat → ?)

# Classification des modalités (1)

Comprendre les liaisons entre les variables

En identifiant les associations entre les modalités

# Distance entre modalités – Indice de Dice

Indice de Dice, écart au carré entre les indicatrices des modalités →

**Carré** d'une distance euclidienne

$$\delta^2_{jj'} = \frac{1}{2}\sum_{i=1}^{n}\left(m_{ij} - m_{ij'}\right)^2$$

*i* est l'individu n°i

*j* est la j$^{ème}$ modalité de la base

$m_{ij}$ est une indicatrice de j$^{ème}$ modalité

Transformation du tableau de données en tableau d'indicatrices

```
#dummy coding
library(ade4)
disj <- acm.disjonctif(vote.active)
print(head(vote.active))
print(head(disj))
```

```
> print(head(vote.active))
  budget physician salvador nicaraguan missile education
1      n         y        y          n       n         y
2      n         y        y          n       n         y
3      y   neither        y          n       n         n
4      y         n  neither          n       n         n
5      y         n        y          n       n   neither
6      y         n        y          n       n         n
> print(head(disj))
  budget.n budget.neither budget.y physician.n physician.neither physician.y salvador.n salvador.neither salvador.y nicaraguan.n
1        1              0        0           0                 0           1          0                 0          1            1
2        1              0        0           0                 0           1          0                 0          1            1
3        0              0        1           0                 1           0          0                 0          1            1
4        0              0        1           1                 0           0          0                 1          0            1
5        0              0        1           1                 0           0          0                 0          1            1
6        0              0        1           1                 0           0          0                 0          1            1
  nicaraguan.neither nicaraguan.y missile.n missile.neither missile.y education.n education.neither education.y
1                  0           0         1                 0         0           0                 0           1
2                  0           0         1                 0         0           0                 0           1
3                  0           0         1                 0         0           1                 0           0
4                  0           0         1                 0         0           1                 0           0
5                  0           0         1                 0         0           0                 1           0
6                  0           0         1                 0         0           1                 0           0
```

Codage disjonctif complet

```
#Dice index
dice <- function(m1,m2){
 return(0.5*sum((m1-m2)^2))
}
#Dice index matrix
d2 <- matrix(0,ncol(disj),ncol(disj))
for (j in 1:ncol(disj)){
 for (jprim in 1:ncol(disj)){
  d2[j,jprim] <- dice(disj[,j],disj[,jprim])
 }
}
colnames(d2) <- colnames(disj)
rownames(d2) <- colnames(disj)
#transform the matrix in a R 'dist' class
d <- as.dist(sqrt(d2))
```

# Matrice des distances

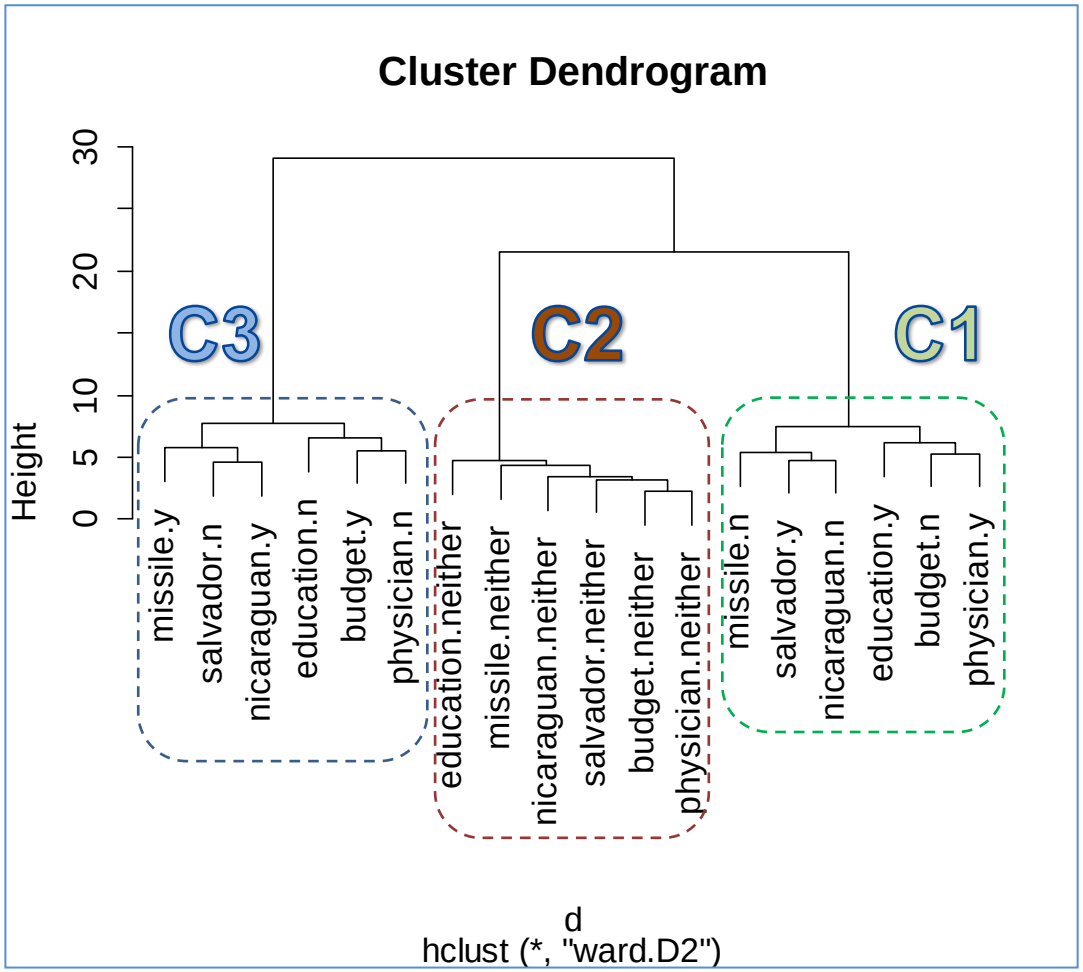
Les cooccurrences sont naturellement inexistantes pour les indicatrices issues d'une même variable (distance est naturellement élevée)

Une valeur faible indique une forte cooccurrence entre les modalités (ex. budget = n et physician = y, …)

| | budget.n | budget.neither | budget.y | physician.n | physician.neither | physician.y | salvador.n | salvador.neither | salvador.y | nicaraguan.n | nicaraguan.neither | nicaraguan.y | missile.n | missile.neither | missile.y | education.n | education.neither | education.y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| budget.n | 0 | 9.54 | **14.56** | 13.56 | 9.54 | **5.29** | 13.17 | 9.54 | 6.20 | 5.87 | 9.22 | 13.55 | 6.52 | 9.62 | 12.96 | 13.19 | 9.54 | 6.16 |
| budget.neither | 9.54 | 0 | 11.49 | 11.22 | 2.24 | 9.59 | 10.27 | 3.00 | 10.42 | 9.51 | 3.16 | 11.07 | 10.27 | 3.81 | 10.15 | 10.86 | 4.12 | 9.38 |
| budget.y | 14.56 | 11.49 | 0 | 5.57 | 11.27 | 13.64 | 6.52 | 11.18 | 13.29 | 13.47 | 11.40 | 5.70 | 13.13 | 11.02 | 7.07 | 6.48 | 11.18 | 13.30 |
| physician.n | 13.56 | 11.22 | 5.57 | 0 | 11.36 | 14.56 | 5.70 | 11.00 | 13.69 | 13.40 | 11.31 | 5.79 | 13.36 | 10.75 | 6.86 | 6.16 | 11.09 | 13.42 |
| physician.neither | 9.54 | 2.24 | 11.27 | 11.36 | 0 | 9.70 | 10.22 | 3.00 | 10.46 | 9.62 | 3.16 | 10.98 | 10.22 | 3.81 | 10.20 | 10.77 | 4.12 | 9.49 |
| physician.y | 5.29 | 9.59 | 13.64 | 14.56 | 9.70 | 0 | 13.58 | 9.75 | 5.15 | 5.87 | 9.33 | 13.58 | 6.12 | 9.92 | 13.04 | 13.42 | 9.64 | 5.74 |
| salvador.n | 13.17 | 10.27 | 6.52 | 5.70 | 10.22 | 13.58 | 0 | 10.56 | 14.49 | 13.82 | 10.46 | 4.58 | 13.82 | 10.10 | 5.34 | 6.60 | 10.37 | 13.06 |
| salvador.neither | 9.54 | 3.00 | 11.18 | 11.00 | 3.00 | 9.75 | 10.56 | 0 | 10.65 | 9.62 | 3.32 | 11.02 | 10.22 | 4.06 | 10.20 | 10.82 | 4.24 | 9.49 |
| salvador.y | 6.20 | 10.42 | 13.29 | 13.69 | 10.46 | 5.15 | 14.49 | 10.65 | 0 | 4.80 | 10.22 | 14.00 | 5.00 | 10.49 | 13.73 | 13.17 | 10.37 | 6.52 |
| nicaraguan.n | 5.87 | 9.51 | 13.47 | 13.40 | 9.62 | 5.87 | 13.82 | 9.62 | 4.80 | 0 | 9.82 | 14.49 | 5.48 | 9.80 | 13.44 | 13.02 | 9.72 | 6.52 |
| nicaraguan.neither | 9.22 | 3.16 | 11.40 | 11.31 | 3.16 | 9.33 | 10.46 | 3.32 | 10.22 | 9.82 | 0 | 11.34 | 10.12 | 3.81 | 10.39 | 11.00 | 4.12 | 9.33 |
| nicaraguan.y | 13.55 | 11.07 | 5.70 | 5.79 | 10.98 | 13.58 | 4.58 | 11.02 | 14.00 | 14.49 | 11.34 | 0 | 13.71 | 10.86 | 5.70 | 6.60 | 11.02 | 13.17 |
| missile.n | 6.52 | 10.27 | 13.13 | 13.36 | 10.22 | 6.12 | 13.82 | 10.22 | 5.00 | 5.48 | 10.12 | 13.71 | 0 | 10.68 | 14.37 | 12.98 | 10.22 | 6.89 |
| missile.neither | 9.62 | 3.81 | 11.02 | 10.75 | 3.81 | 9.92 | 10.10 | 4.06 | 10.49 | 9.80 | 3.81 | 10.86 | 10.68 | 0 | 10.70 | 10.79 | 4.64 | 9.51 |
| missile.y | 12.96 | 10.15 | 7.07 | 6.86 | 10.20 | 13.04 | 5.34 | 10.20 | 13.73 | 13.44 | 10.39 | 5.70 | 14.37 | 10.70 | 0 | 7.00 | 10.34 | 12.85 |
| education.n | 13.19 | 10.86 | 6.48 | 6.16 | 10.77 | 13.42 | 6.60 | 10.82 | 13.17 | 13.02 | 11.00 | 6.60 | 12.98 | 10.79 | 7.00 | 0 | 11.49 | 14.21 |
| education.neither | 9.54 | 4.12 | 11.18 | 11.09 | 4.12 | 9.64 | 10.37 | 4.24 | 10.37 | 9.72 | 4.12 | 11.02 | 10.22 | 4.64 | 10.34 | 11.49 | 0 | 10.05 |
| education.y | 6.16 | 9.38 | 13.30 | 13.42 | 9.49 | 5.74 | 13.06 | 9.49 | 6.52 | 6.52 | 9.33 | 13.17 | 6.89 | 9.51 | 12.85 | 14.21 | 10.05 | 0 |

# CAH sur les modalités, basée sur l'indice de Dice

```
#cluster analysis on indicator variables
arbre.moda <- hclust(d,method="ward.D2")
plot(arbre.moda)
```

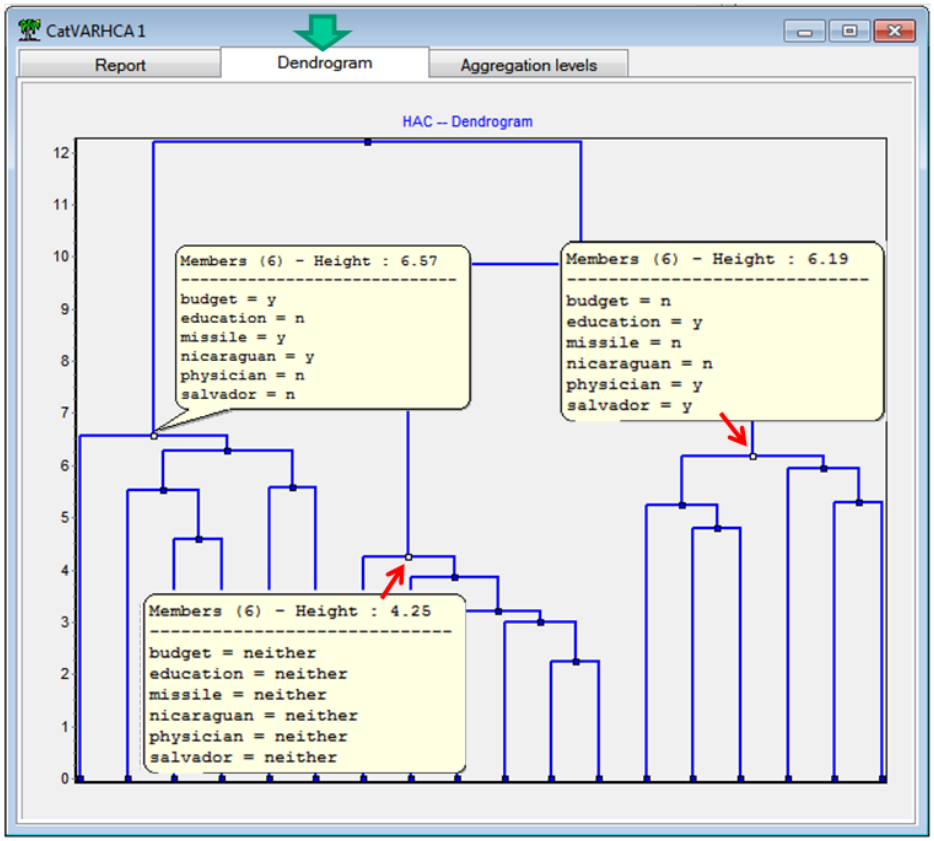



3 groupes maintenant se démarquent. Nous distinguons clairement les relations entre les modalités c.-à-d. les votes concomitants.

# CAH sur les modalités avec le logiciel Tanagra

http://tutoriels-data-mining.blogspot.fr/2013/12/classification-de-variables-qualitatives_21.html

**Stratégie d'agrégation : « average linkage »**



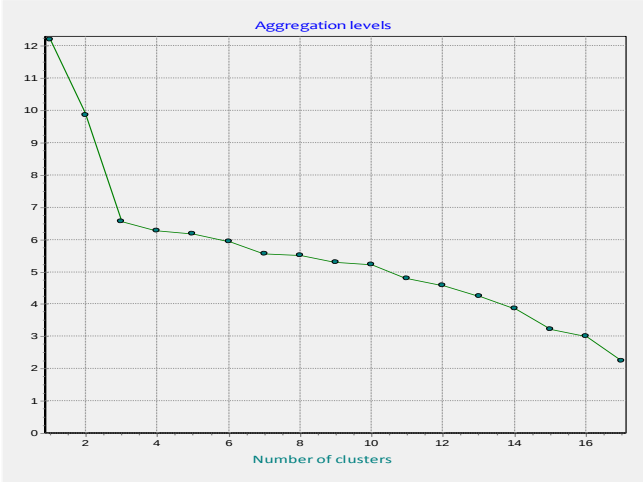

Dendrogramme

(height : distance d'agrégation)

**Clusters' members**

| Cluster | Members | Distance Own Cluster | Distance Next Closest | Ratio (Own / Next) |
|---|---|---|---|---|
| 1 (Size = 6) | budget = y | 5.22 | 11.26 | 0.4640 |
| | education = n | 5.47 | 10.96 | 0.4995 |
| | missile = y | 5.33 | 10.33 | 0.5157 |
| | nicaraguan = y | 4.73 | 11.05 | 0.4279 |
| | physician = n | 5.01 | 11.12 | 0.4507 |
| | salvador = n | 4.79 | 10.33 | 0.4636 |
| 2 (Size = 6) | budget = neither | 2.72 | 9.79 | 0.2781 |
| | education = neither | 3.54 | 9.92 | 0.3569 |
| | missile = neither | 3.35 | 10.00 | 0.3353 |
| | nicaraguan = neither | 2.93 | 9.67 | 0.3027 |
| | physician = neither | 2.72 | 9.84 | 0.2766 |
| | salvador = neither | 2.94 | 9.88 | 0.2973 |
| 3 (Size = 6) | budget = n | 5.01 | 9.50 | 0.5273 |
| | education = y | 5.31 | 9.54 | 0.5562 |
| | missile = n | 5.00 | 10.29 | 0.4861 |
| | nicaraguan = n | 4.76 | 9.68 | 0.4913 |
| | physician = y | 4.70 | 9.65 | 0.4865 |
| | salvador = y | 4.61 | 10.44 | 0.4419 |

Rattachement des modalités aux classes




Courbe des hauteurs d'agrégation

(donne une indication sur le « bon » nombre de clusters)

# CAH sur les modalités, traitement des variables supplémentaires

```
#create 3 groups
dgroups <- cutree(arbre.moda,k=3)

#illustrative variable
illus <- acm.disjonctif(as.data.frame(vote.data$affiliation))
colnames(illus) <- c("democrat","republican")

#distance to illustrative levels
dice.democrat <- sapply(disj,dice,m2=illus$democrat)
tapply(dice.democrat,dgroups,mean)

dice.republican <- sapply(disj,dice,m2=illus$republican)
tapply(dice.republican,dgroups,mean)
```

**Republican**

Budget = n
Physician = y
Salvador = y
Nicaraguan = n
Missile = n
Education = y

**Democrat**

Budget = y
Physician = n
Salvador = n
Nicaraguan = y
Missile = y
Education = n

On comprend mieux le mécanisme des votes des députés.

$\delta^2$

Moyenne des distances (au carré) aux membres des groupes

**Distance to clusters - Supplementary variables**

| Variable = level | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|
| affiliation = republican | 30.9 | 86.6 | 184.0 |
| affiliation = democrat | 186.6 | 130.9 | 33.5 |

# Classification des modalités (2)

S’appuyer sur d’autres mesures de similarités / dissimilarités

# Varclus du package « Hmisc » de R

Mesure de similarité $s_{jj'} = \frac{1}{n}\sum_{i=1}^{n} m_{ij} \times m_{ij'}$

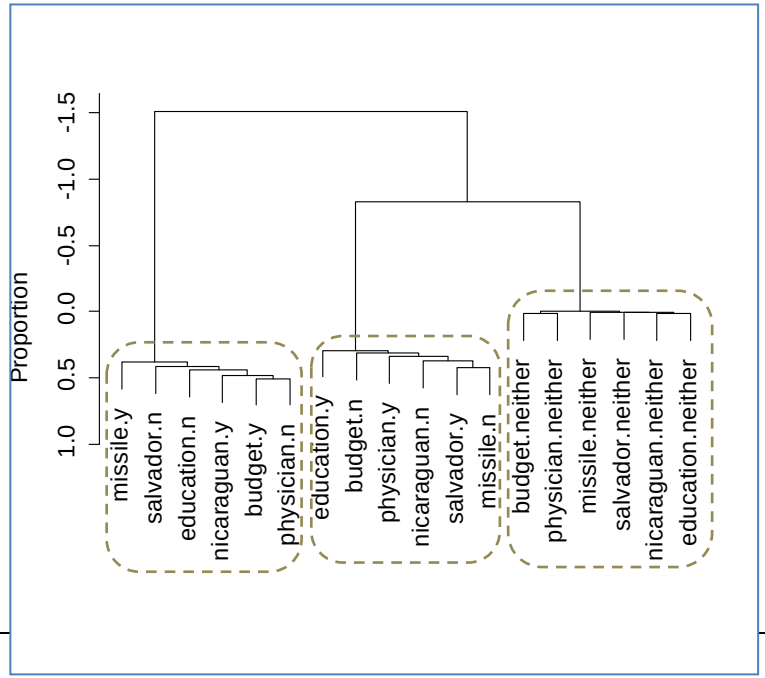
Fréquence conjointe c.-à-d. proportion des individus qui possèdent simultanément les 2 caractères (0 : aucun individu n'a les 2 modalités en commun ; 1 : tous les individus possèdent ces deux caractères)

Mesure de dissimilarité $d_{jj'} = 1 - s_{jj'}$

• Attention, ce n'est pas une distance ($d_{jj} \neq 0$), mais cela ne gêne pas la méthode hclust() appelée en interne.

• $d_{jj'} = 1$ forcément pour 2 modalités provenant d'une même variable. Leur réunion ne peut intervenir qu'à la fin du processus d'agrégation (CAH).

```
# chargement du package
library(Hmisc)
# appel de la fonction  : cf. aide pour les options
v <-
varclus(as.matrix(disj),type="data.matrix",
similarity="bothpos",method="ward.D")
plot(v)
```




Partition en 3 groupes toujours aussi « évidente »

# Classification des modalités (3)

Tandem clustering

# Tandem clustering

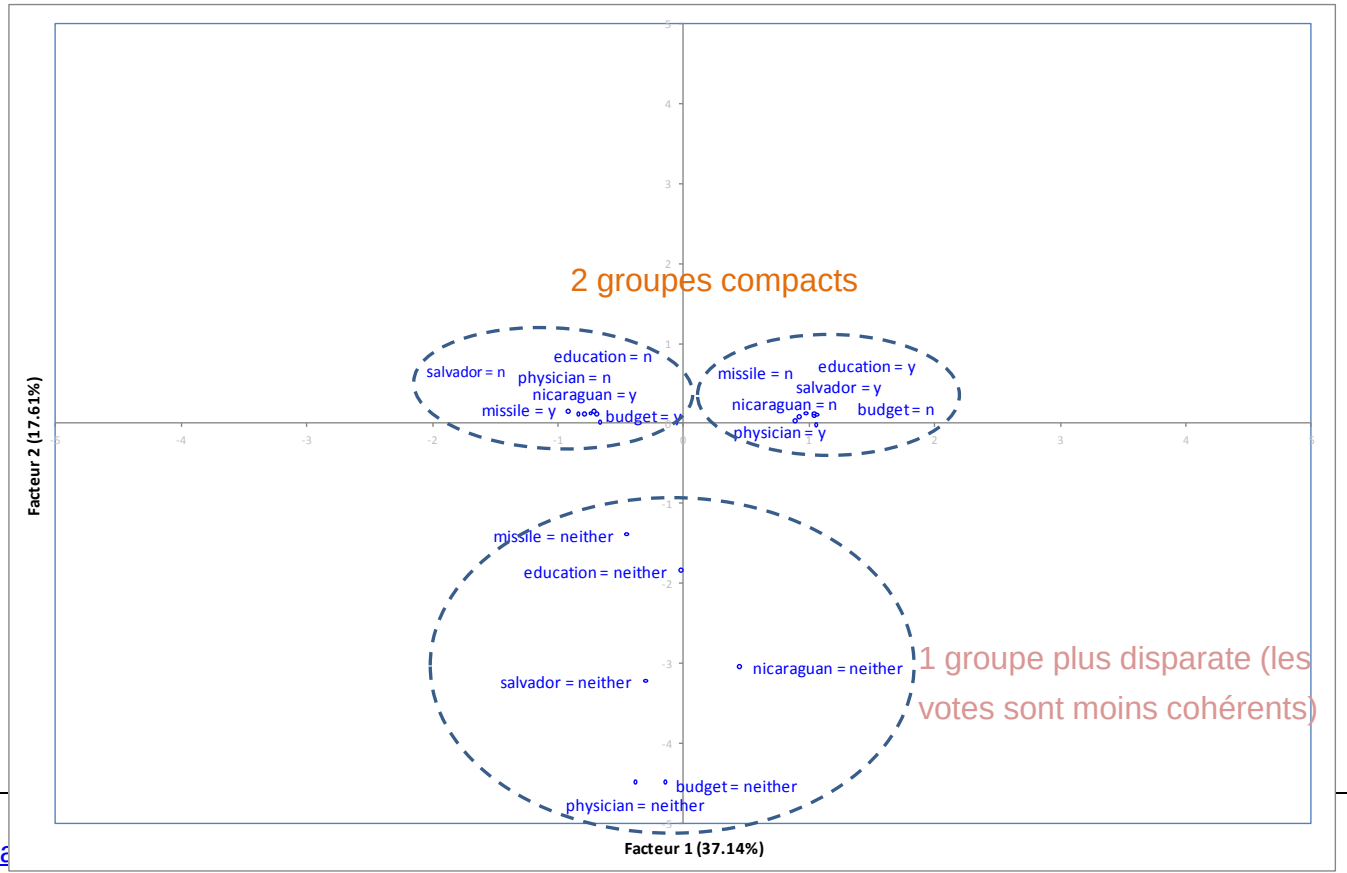
Coordonnées factorielles des modalités via une ACM (analyse des correspondances multiples).

Analyse en 2 temps :

1. Projeter les modalités dans un nouvel espace de représentation
2. Réaliser une classification avec la distance euclidienne

Individus = modalités. Réaliser une CAH (ou tout autre méthode de classification) dans le nouveau repère. Note : on peut n'utiliser qu'un sous-ensemble des axes, c'est une forme de régularisation (nettoyage des données). Problème : choix du nombre de facteurs.

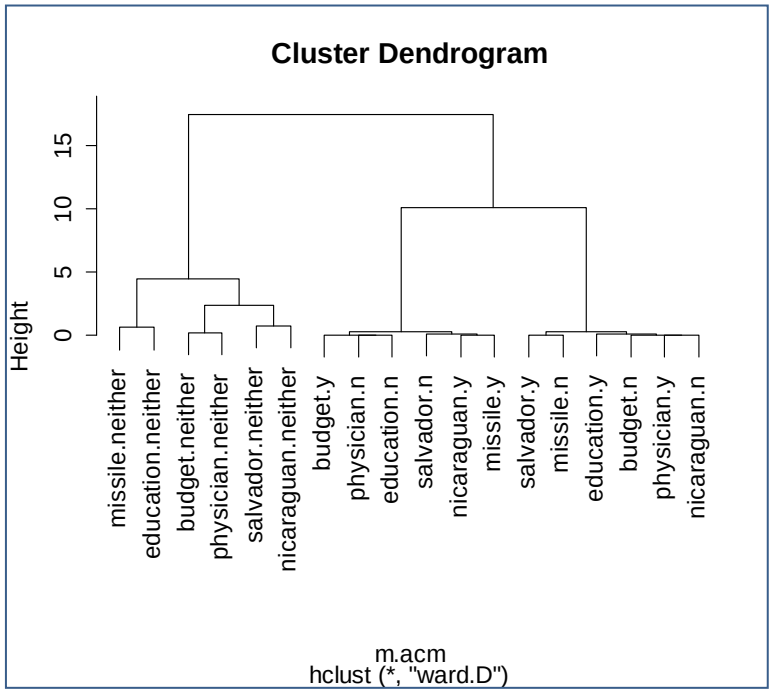
ACM : premier plan factoriel suffit amplement



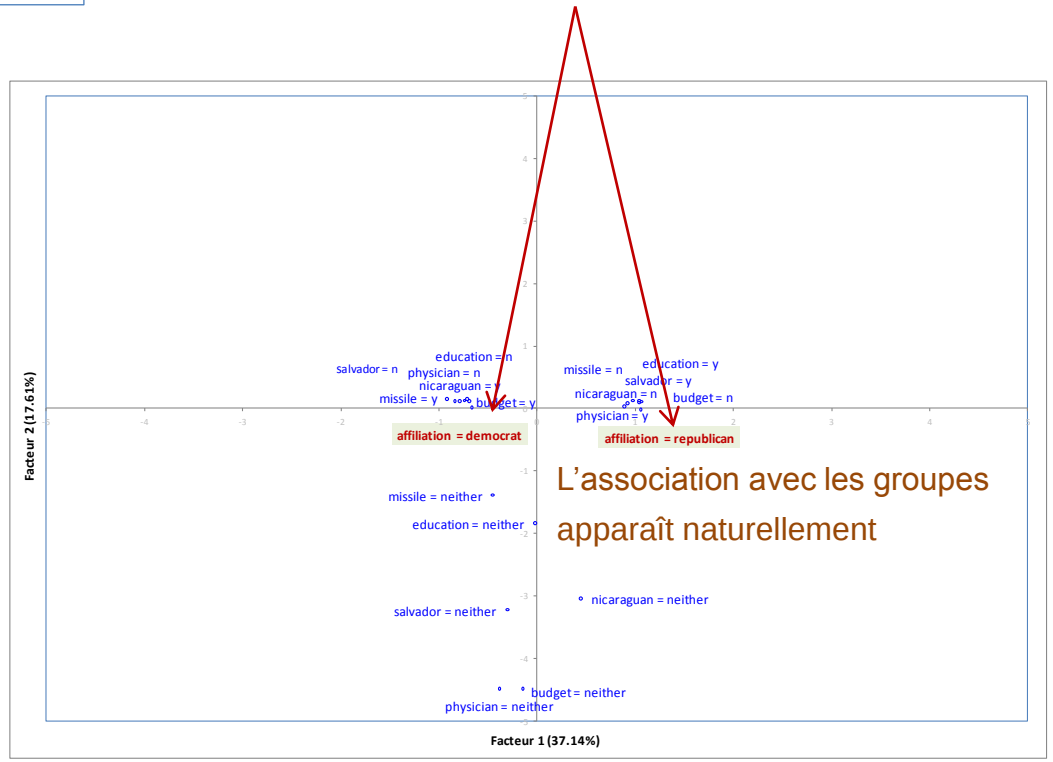
# CAH sur les coordonnées factorielles - Distance euclidienne

```
#MCA with the ade4 package
acm <- dudi.coa(disj,scannf=F,nf=2)
#factorial coordinates of the levels
acm.coord <- data.frame(acm$co)
rownames(acm.coord) <- colnames(disj)
#distance matrix
m.acm <- dist(acm.coord,method="euclidian")
#cluster analysis from the distance matrix m.acm
arbre.acm <- hclust(m.acm,method="ward.D")
plot(arbre.acm)
```

Les individus-modalités n'ont pas la même fréquence (poids). Si elles sont très disparates, il faudrait en tenir compte dans le processus de clustering (cf. l'option « members »).




Positionnement des modalités supplémentaires dans le repère factoriel




L'association avec les groupes apparaît naturellement

# Bilan

# Bilan

La classification de variables qualitatives cherche à regrouper les variables en paquets homogènes : les variables dans un même groupe sont fortement liées entre elles, les variables dans des groupes différents sont faiblement liées.

La méthode apporte une réelle valeur ajoutée quand il s'agit de détecter des redondances, par ex. guider ou aider à interpréter la sélection de variables dans un processus de modélisation prédictive.

Mais elle ne donne pas d'indications sur la nature de l'association entre les variables.

Mieux vaut dans ce cas se tourner vers la classification des modalités des variables qualitatives.

La technique repose essentiellement sur la définition d'un indice de similarité entre modalités.

Mais d'autres pistes existent, par ex. une approche de type « tandem clustering » c.-à-d en 2 temps : une ACM pour situer les modalités dans un repère factoriel, une classification à partir des coordonnées factorielles des modalités.

# Bibliographie

Tutoriel Tanagra, « Classification de variables qualitatives », décembre 2013.

H. Abdallah, G. Saporta, « Classification d'un ensemble de variables qualitatives », in Revue de Statistique Appliquée, Tome 46, N°4, pp. 5-26, 1998.

M. Chavent, V. Kuentz Simonet, B. Liquet, J. Saracco, « ClustOfVar: An R package for the Clustering of Variables », in Journal of Statistical Software, 50(13), september 2012.

F. Harrell Jr, « Hmisc: Harrell Miscellaneous », version 3.14-5.