Clustering of categorical variables

Grouping categorical variables

Grouping categories of nominal variables

Ricco RAKOTOMALALA

Université Lumière Lyon 2

Outline

1. Clustering of categorical variables. Why?
   a. HCA from a dissimilarity matrix
   b. Deficiency of the clustering of categorical variables
2. Clustering categories of nominal variables
   a. Distance between categories – Dice's coefficient
   b. HAC on the categories
   c. Interpretation of the obtained clusters
3. Other approaches for the clustering of categories
4. Conclusion
5. References

Clustering of categorical variables

Why? For what purpose?

Clustering of variables

Goal: grouping related variables

➔ The variables in the same group are highly associated together.

➔ The variables in different groups are not related (in the sense of association measure)

With what objective?

1. Indentify the underlying structure of the dataset. Make a summary of the relevant information (the approach is complementary to the clustering of individuals).

2. Detect redundancies, for instance in order to selecting the variables intended for a subsequent analysis (e.g. supervised learning task)
   a. In a pretreatment phase, in order to organize the search space
   b. In a post-treatment phase, in order to understand the role of the removed variables in the selection process.

An example: Vote dataset (1984)

n = 435 individuals (US Congressmen)

p = 6 active variables

Variable	Categories	Role
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

Political affiliation

Illustrative variable i.e. used for understanding the nature of the groups

Vote on each subject, 3 categories: yes (yea), no (nay), neither (not “yea” or “nay”)

Active variables

Identify the vote which are highly related

Establish their association with the political affiliation

We observe that a vote "yea" to a subject may be highly related to vote "nay" to another subject.

HAC from a dissimilarity matrix

Hierarchical agglomerative clustering

Using the Cramer's V to measure the
association between the nominal variables

Measure of association between 2 nominal variables

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

Pearson's chi-squared statistic

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

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

P(AB) observed

P(A) x P(B) Under the independence assumption

Cramer's v

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

- Symmetrical
- $0 \leq v \leq 1$

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

$\chi^2 = 355.48$

$p.value < 0.0001$

$v = 0.639$

High association

Significant at the 5% level

Similarity matrix – Dissimilarity matrix

```
#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)
}
```

Similarity matrix (Cramer's v)

	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

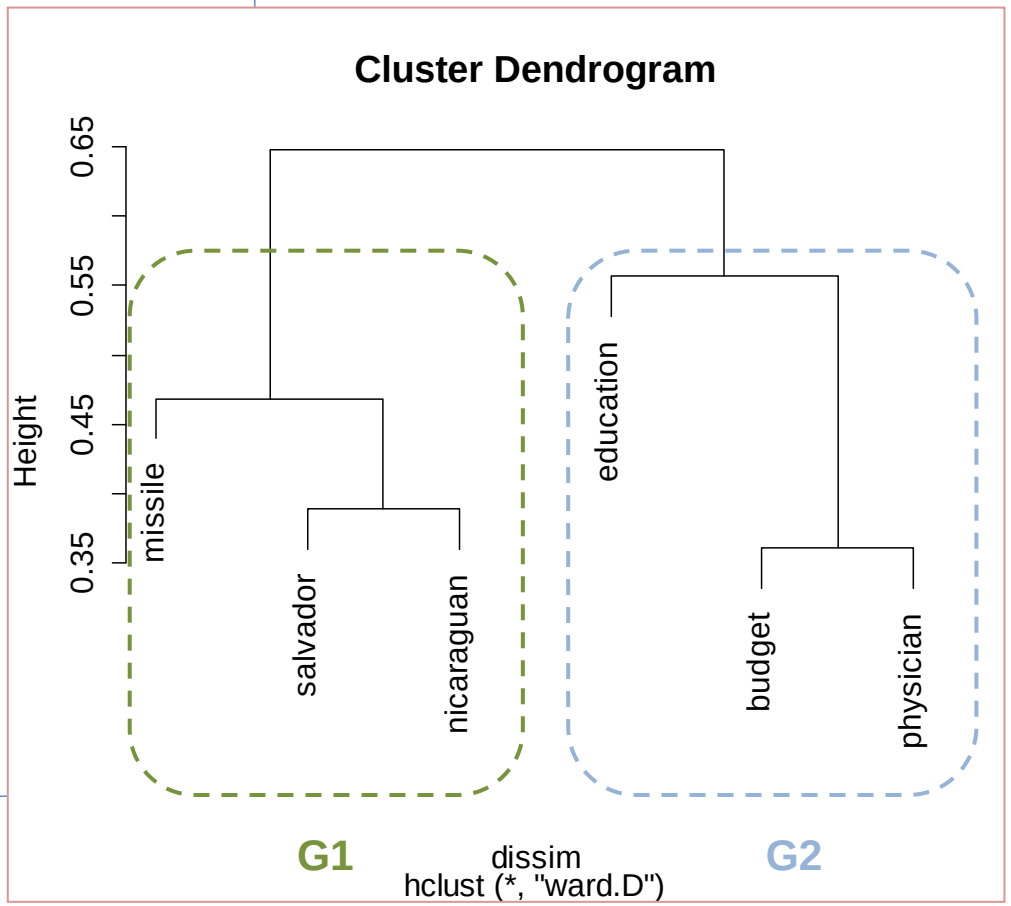
Dissimilarity matrix (1-v)

	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

We can use this matrix as input for the HAC algorithm

hclust() under R – Distance = (1 – v), Ward's method

```
#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)
```



We get a vision of the structures of association between variables. e.g. "budget" and "physician" are related i.e. there is a strong coherence of votes (v = 0.639); budget and salvador are less related (v = 0.507), etc. but we do not know on what association of votes (yes or no) these relationships are based…

ClustOfVar (Chavent and al., 2012)

"Centroid" (representative variable) of a group of variables = latent variable i.e. the group is scored as a single variable

F = 1st factor from the MCA (multiple correspondence analysis)
η(.) correlation ratio
λ Variation within the group

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

Various strategies for grouping are possible.

➔ HAC approaches: minimizing the loss of variation at each step

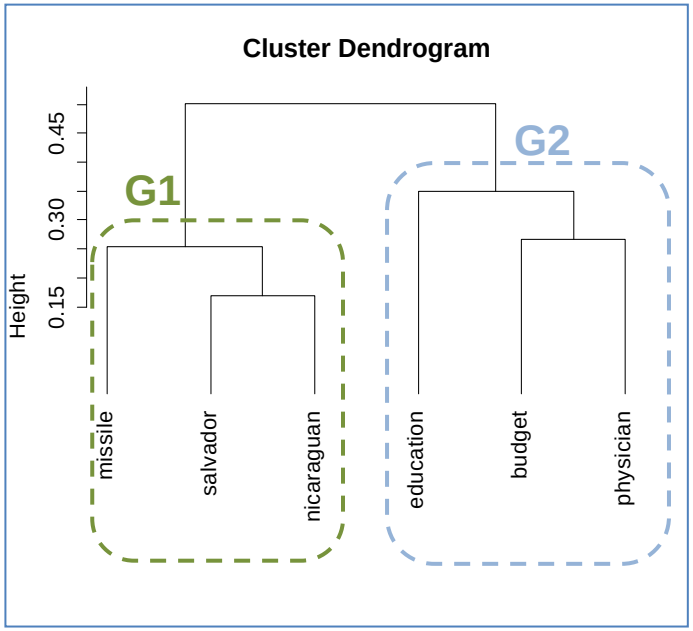
➔ K-Means approach: assign the variables to the closest "centroid" (in the sense of the correlation ratio) during the learning process

1. "ClustOfVar" can handle dataset with mixed numeric and categorical variables. The centroid is defined with first component of the factor analysis for mixed data
2. This is a generalization of the CLV approach (Vigneau and Qannari, 2003) which can handle numeric variables only and is based on PCA (principal component analysis)

ClustOfVar on the « vote » dataset

```
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
```

We obtain the same results as for the HAC on the (1-v) dissimilarity matrix

Issues for the interpretation of the results

The clustering of categorical variables gives a partial vision of the structure of the relationships among variables...

Interpreting a cluster – 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

Main associations between the categories

Budget = y
Physician = n
Education = n

Budget = n
Physician = y
Education = y

This kind of analysis cannot be done manually.

Analyzing the illustrative variables

The illustrative variables are used to strengthen the interpretation of the results.

```
#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 (Cramer's v)	**Mean (v)**
nicaraguan	0.660	**0.667**
missile	0.629	
education	0.688	
budget	0.740	**0.781**
physician	0.914	
salvador	0.712	

- The political affiliation has a little more influence for the votes in G2 than in G1 (why? the subjects are more sensitive in G2?)
- We do not know what are the votes of the democrats (republicans)?

Clustering the categories of categorical variables (1)

Identifying the nature of the association between the categorical variables

Distance between categories – Dice's coefficient

Dice coefficient. Squared difference between the dummy coding 0/1 for each category of variables. Square of the Euclidean distance.

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

i is the individual n°i
j is the j[th] category
m_{ij} is an indicator for the j[th] category

Transforming the initial data table into a table of indicator variables.

```
#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
```

Simple coding scheme

Distance matrix

```
#Dice's index
dice <- function(m1,m2){
 return(0.5*sum((m1-m2)^2))
}
#Dice's 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))
```

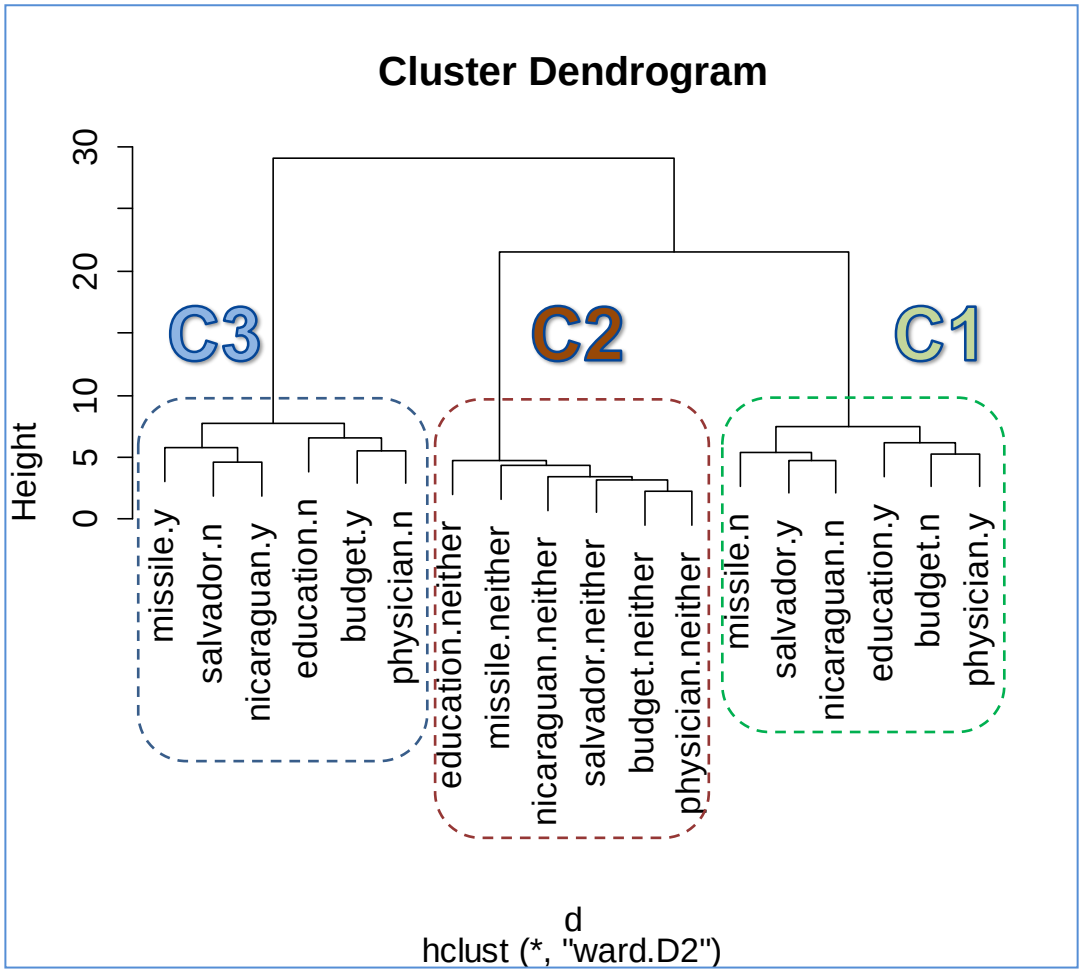
The distance is high for the indicator variables coming from the same categorical variable.

A low value points out a high association between the categories (e.g. budget = n and 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

HAC of the categories, based on the Dice's index

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

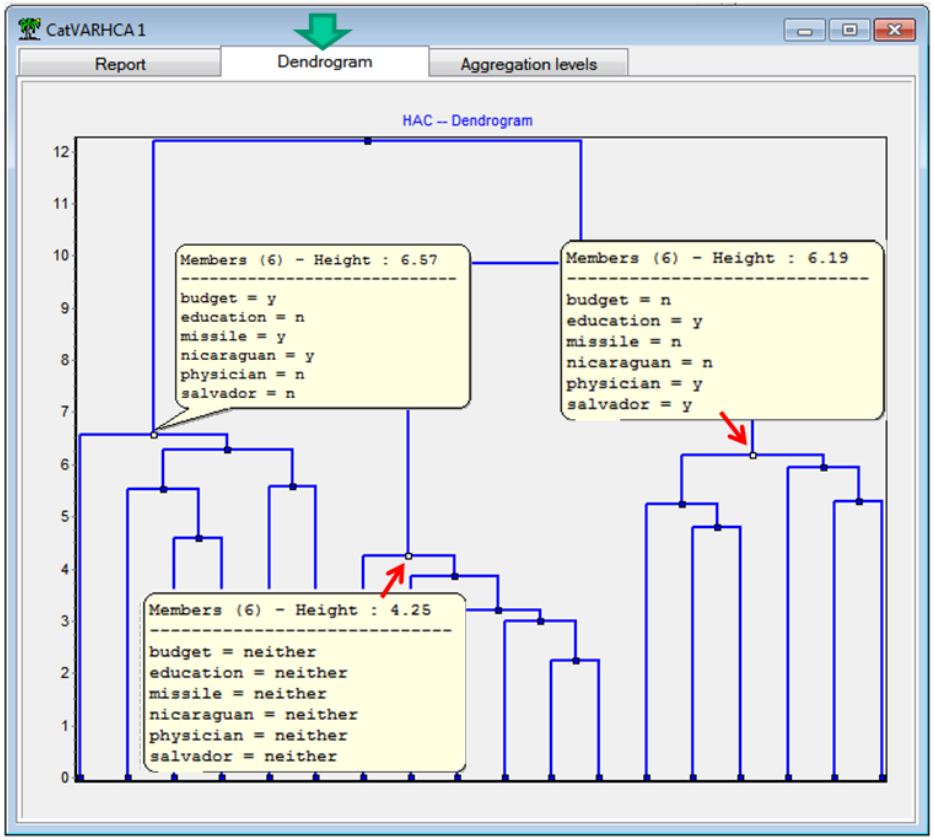


3 groups now are highlighted. We distinguish clearly the relationships between the categories i.e. the votes which are related

HAC of categories under Tanagra

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

Linking criterion : « average linkage »



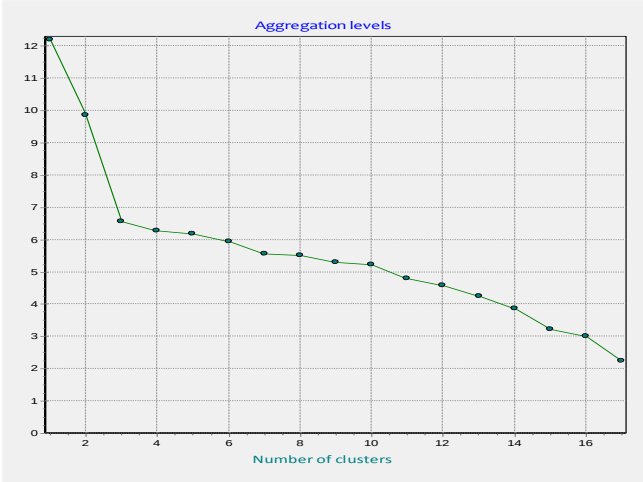
Dendrogram

(height : aggregation distance)

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

Association of the categories to the groups



Evolution of the aggregation distance

(an “elbow” gives an indication about the right number of groups)

HAC of categories, handling the illustrative variables

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

#illustrative variable – dummy coding scheme
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

We understand the influence of the political association on the votes.

δ^2

For a category (from the illustrative variable), mean of the squared distance to the indicator variables of a group

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

Clustering the categories of categorical variables (2)

Using other measures of similarity and dissimilarity

Varclus of the « Hmisc » package for R

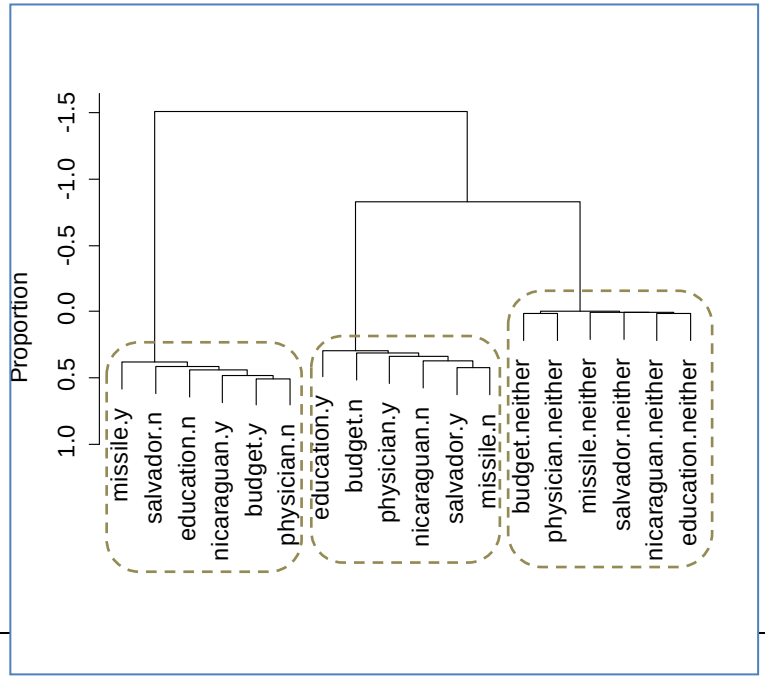
Similarity measure $s_{jj'} = \frac{1}{n}\sum_{i=1}^{n} m_{ij} \times m_{ij'}$

Conjoint frequencies i.e. the proportion of the individuals which belong to the 2 categories (0:no instances belong simultaneously to the 2 studied categories; 1: all the instances have the two categories)

Dissimilarity measure $d_{jj'} = 1 - s_{jj'}$

- Caution, this is not a distance ($d_{jj} \neq 0$), but this does not interfere with the hclust() procedure.
- $d_{jj'} = 1$ necessarily for 2 categories belonging to the same variable. Their merging is only possible at the end of the aggregation process (HAC).

```
# loading the package
library(Hmisc)
# calling the "varclus" function
# see the help file for the parameters
v <-
varclus(as.matrix(disj),type="data.matrix",
similarity="bothpos",method="ward.D")
plot(v)
```



The partition into 3 groups is also obvious here.

Clustering the categories of categorical variables (3)

Tandem clustering

Tandem clustering

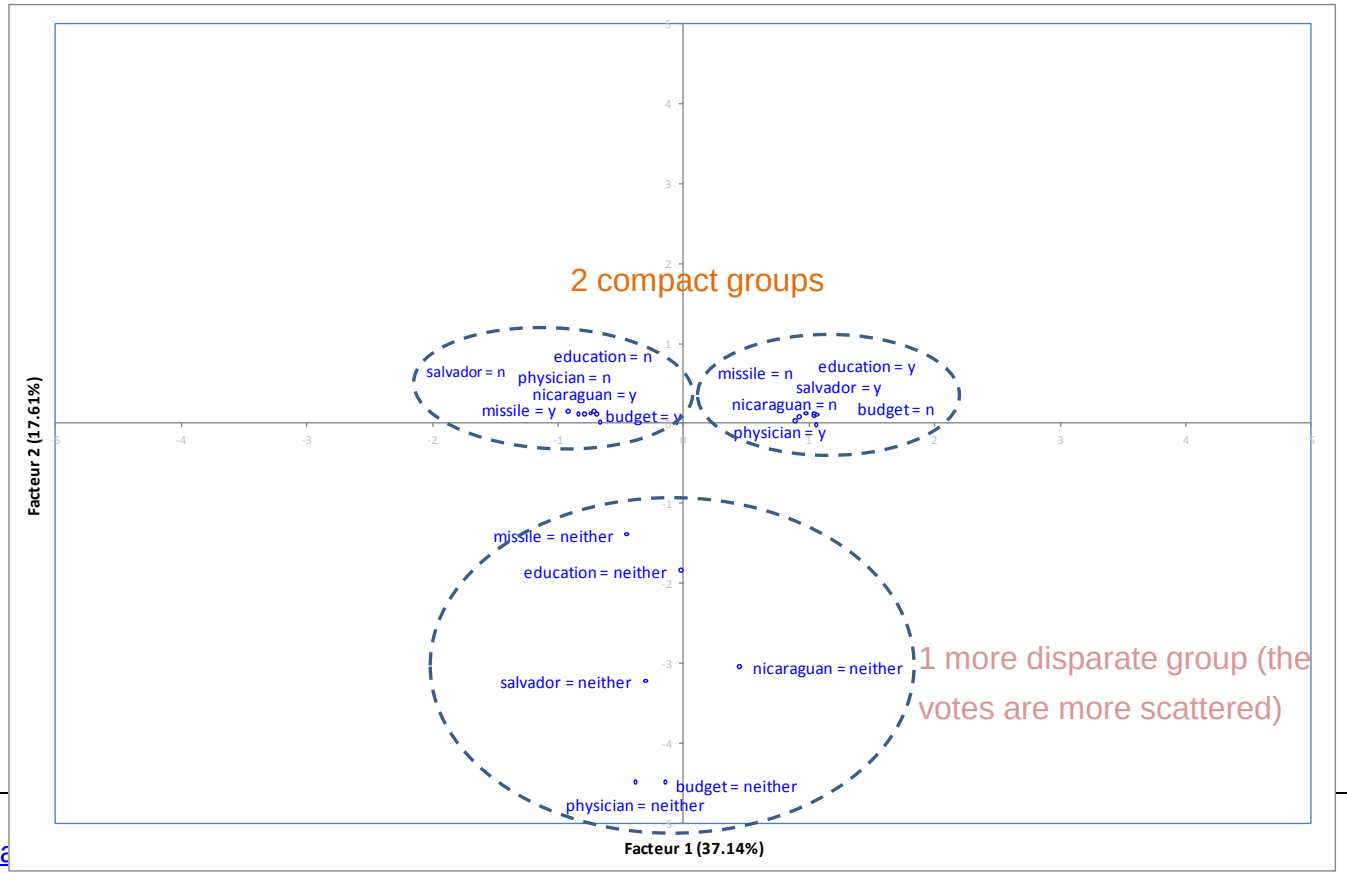
Factor scores from the MCA (multiple correspondence analysis)

Two steps:

1. Calculating the coordinates of the categories in a new representation space.
2. Performing the clustering with the Euclidean distance

Individuals = categories. Performing the HAC (or another clustering approach) in the new representation space. We can use only a few number of factors. This can be viewed as a regularization strategy.

MCA: the two first factors seem enough here

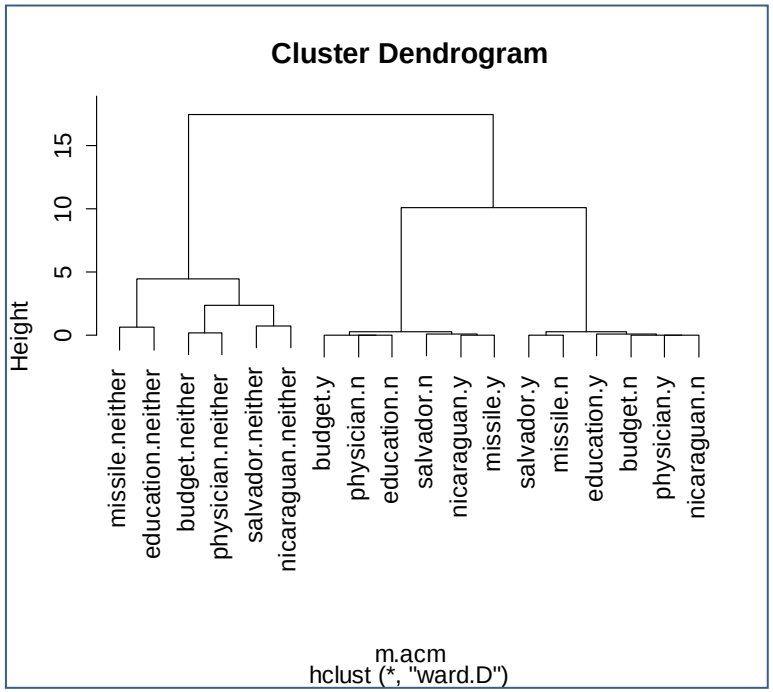


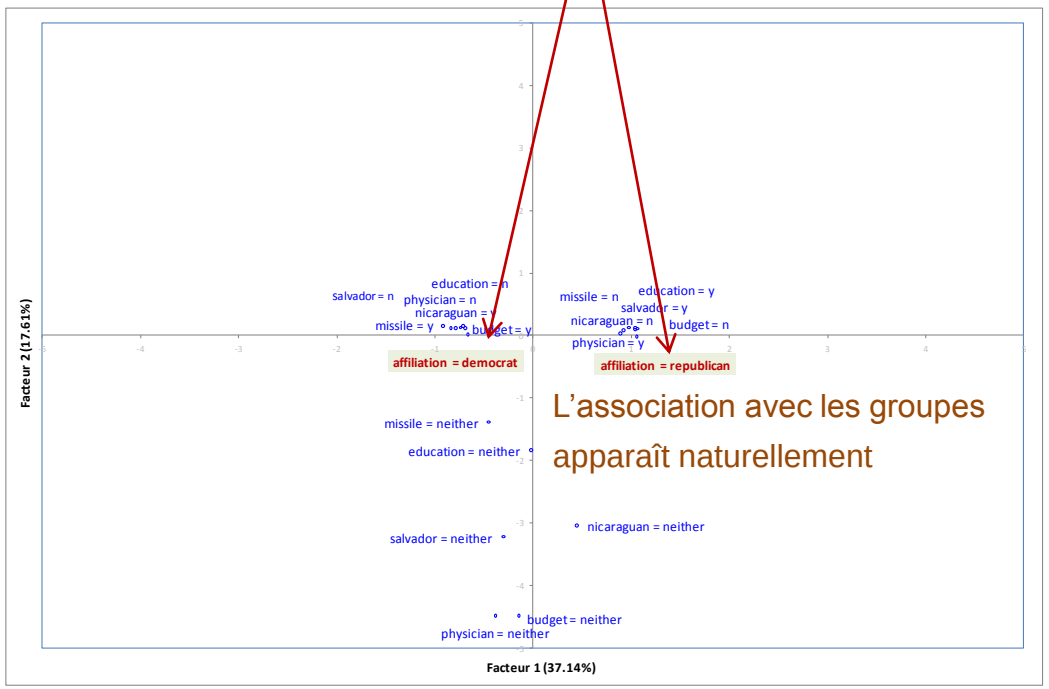
HAC from the factor scores – Euclidean distance

```
#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)
```

The individuals (categories) have not the same frequency (weight). If they are very different, we should take into account that in the clustering process (see. "members" parameter of hclust).

Coordinates of the categories of the illustrative variable in the factorial representation space.





L'association avec les groupes apparaît naturellement

Conclusion

Conclusion

The clustering of qualitative variables seeks to gather together variables into clusters: variables in the same group are strongly related each other; variables in different groups are weakly related.

The method is interesting if we try to detect redundancies, e.g. to help the variable selection process in a supervised learning task.

But it does not give indications about the nature of the associations between the variables.

In this context, it is more relevant to perform a clustering of categories of categorical variables.

The approach is mainly based on the definition of a similarity measure between categories.

But other approaches are possible e.g. a tandem clustering: in a first step, we calculate the scores of the categories in a new representation space; in a second step, we perform the clustering process using these new variables.

References

H. Abdallah, G. Saporta, « Classification d'un ensemble de variables qualitatives » (Clustering of a set of categorical variables), 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.