ClustOfVar: an R package for the clustering of variables

Marie Chavent & Vanessa Kuentz
& Benoît Liquet & Jérôme Saracco

IMB, University of Bordeaux, France
INRIA Bordeaux Sud-Ouest, CQFD Team
CEMAGREF, UR ADBX, Bordeaux, France
ISPED, University of Bordeaux, France

The R User Conference 2011
University of Warwick, August 16-18 2011

Outline

Outline

Introduction

- **Clustering of variables** lumps together strongly related variables
- **Usefulness** for case studies, variable selection and dimension reduction
- **A first approach**: apply classical method dedicated to the **clustering of observations**

Introduction

Some specific methods:

- VARCLUS (SAS)
- Likelihood Linkage Analysis (Lerman, 1987)
- Qualitative variable clustering (Abdallah and Saporta, 2001)

Specific methods based on PCA:

- CLV (Vigneau and Qannari, 2003)
- Diametrical clustering (Dhillon et al., 2003)
  → For quantitative variables

Introduction

The goal of the package **ClustOfVar**:

- Propose methods for the clustering of a mixture of quantitative and qualitative variables
- Also suitable for non mixed quantitative or qualitative data

↪ For that purpose we use the PCAMIX method

↪ A hierarchical clustering algorithm and a k-means type partitionning algorithm

↪ A method based on a bootstrap approach to evaluate the stability of the partitions to determine suitable numbers of clusters

Outline

Homogeneity criterion of a partition of variables

- $\mathcal{V}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_{p_1}\}$ of **quantitative** variables
- $\mathcal{V}_2 = \{\mathbf{z}_1, \ldots, \mathbf{z}_{p_2}\}$ of **qualitative** variables
- Let **X** and **Z** be the corresponding quantitative and qualitative data matrices
- Let $P = (C_1, \ldots, C_K)$ be a partition of $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$
- The homogeneity of this partition P:

$$\mathcal{H}(P) = \sum_{k=1}^{K} H(C_k, \mathbf{y}_k)$$

where $\mathbf{y}_k$ is central (quantitative) synthetic variable also called the center of C_k

Homogeneity criterion of a cluster of variables

- The function H measures the adequacy between C_k and $\mathbf{y}_k$:

$$H(C_k,\mathbf{y}_k)=\sum_{\mathbf{x}_j\in C_k} r^2(\mathbf{x}_j,\mathbf{y}_k)+\sum_{\mathbf{z}_j\in C_k}\eta^2(\mathbf{z}_j,\mathbf{y}_k)$$

  where $r^2(\mathbf{x}_j,\mathbf{y}_k)$ is the squared correlation of $\mathbf{x}_j$ with $\mathbf{y}_k$ and $\eta^2(\mathbf{z}_j,\mathbf{y}_k)$ is the correlation ratio between $\mathbf{z}_j$ and $\mathbf{y}_k$

Definition of the synthetic variable of a cluster

- The **center** of C_k is:

$$\mathbf{y}_k = \arg\max_{\mathbf{u}\in\mathbb{R}^n} \left\{ \sum_{\mathbf{x}_j\in C_k} r^2(\mathbf{x}_j,\mathbf{u}) + \sum_{\mathbf{z}_j\in C_k} \eta^2(\mathbf{z}_j,\mathbf{u}) \right\}$$

- $\mathbf{y}_k$ is the first principal component of **PCAMIX** applied to the columns of **X** and **Z** corresponding to the variables in C_k

PCAMIX

- PCAMIX (Kiers, 1991) and AFDM (Pagès, 2004)
- It includes PCA and MCA as special cases
- A Singular Value Decomposition approach is implemented in the package

PCAMIX in a cluster

- Let $\mathbf{X}_k$ and $\mathbf{Z}_k$ be the matrices of the columns of $\mathbf{X}$ and $\mathbf{Z}$ corresponding to the variables in C_k
- Recoding of $\mathbf{X}_k$ and $\mathbf{Z}_k$:
    - $\tilde{\mathbf{X}}_k$ is the standardized version of the quantitative matrix $\mathbf{X}_k$
    - $\tilde{\mathbf{Z}}_k = \mathbf{JGD}^{-1/2}$ is the standardized version of the indicator matrix $\mathbf{G}$ of the qualitative matrix $\mathbf{Z}_k$, where $\mathbf{D}$ is the diagonal matrix of frequencies of the categories and $\mathbf{J} = \mathbf{I} - \mathbf{1}'\mathbf{1}/n$ is the centering operator
    - $\mathbf{M}_k = (\tilde{\mathbf{X}}_k | \tilde{\mathbf{Z}}_k)$

PCAMIX in a cluster

- Singular Value Decomposition of $\mathbf{M}_k$:

$$\mathbf{M}_k = \mathbf{U}_k \Lambda_k \mathbf{V}'_k$$

$\hookrightarrow$ $\sqrt{n}\mathbf{U}_k\Lambda_k$ is the matrix of the PC's scores of PCAMIX
$\hookrightarrow$ $\mathbf{y}_k$ is the first column of this matrix

- The homogeneity of C_k is:

$$\begin{aligned} H(C_k, \mathbf{y}_k) &= \sum_{\mathbf{x}_j \in C_k} r^2(\mathbf{x}_j, \mathbf{y}_k) + \sum_{\mathbf{z}_j \in C_k} \eta^2(\mathbf{z}_j, \mathbf{y}_k) \\ &= \lambda_k^1 \end{aligned}$$

$\hookrightarrow$ $\mathcal{H}(\mathcal{P}) = \lambda_1^1 + \ldots + \lambda_K^1$

The hierarchical clustering method

The algorithm:

- Starts with the partition in p clusters
- Successively aggregate the two clusters with the smallest dissimilarity d:
  $d(A,B) = H(A) + H(B) - H(A \cup B) = \lambda_A^1 + \lambda_B^1 - \lambda_{A \cup B}^1$
  $d(A,B) = h(A \cup B)$ is the height of the cluster $A \cup B$ in the dendrogram of the hierarchy
- Stop when the partition in one cluster is obtained

$\hookrightarrow$ The **hclustvar** function gives a hierarchy
$\hookrightarrow$ The **cutreevar** function cuts the hierarchy

The partitionning method of *K*-means type

The algorithm:

- Initialization step:
  - An initial partition given in input
  - Multiple random initializations
    - Random selection of K variables as initial centers
    - Construct the initial partition by allocating each variable to the cluster with the closest initial center

$\hookrightarrow$ We defined a similarity measure between two variables of any type (quantitative and/or qualitative)
$\hookrightarrow$ The function **mixedvarsim** returns a squared canonical correlation (squared correlation or correlation ratio as special cases)

The partitionning method of K-means type

- Repeat
    - Representation step: the central synthetic variable $\mathbf{y}_k$ of each cluster C_k is calculated with PCAMIX
    - Allocation step: a partition is constructed by assigning each variable to the closest cluster
- Stop if no more changes in the partition (or a maximum number of iterations reached)

↪ The **kmeansvar** R function

The stability of the partitions

The procedure evaluates the stability of the partitions of the hierarchy:

- B boostrap samples of the observations are drawn and B "boostrap" hierarchies are obtained
- The partitions of the B bootstrap hierarchies are compared with the partitions of the initial hierarchy with the corrected Rand index
- The stability of a partition is the mean value of the corrected Rand indices

↪ **Stability** R function

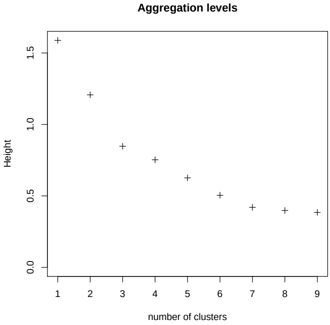
Outline

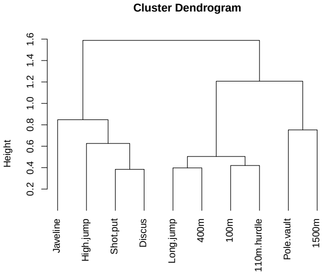
1. Introduction
2. The methods in ClustOfVar
3. Illustration on simple examples
4. Concluding remarks

First example: "decathlon" data

```
> data(decathlon) #data of the package FactoMineR
> head(decathlon[,1:4])
           100m   Long.jump   Shot.put   High.jump
 SEBRLE    11.04  7.58        14.83      2.07
 CLAY      10.76  7.40        14.26      1.86
 KARPOV    11.02  7.30        14.77      2.04
 BERNARD   11.02  7.23        14.25      1.92
 YURKOV    11.34  7.09        15.19      2.10
 WARNERS   11.11  7.60        14.31      1.98
> tree <- hclustvar(X.quanti=decathlon[,1:10])
> plot(tree)
```

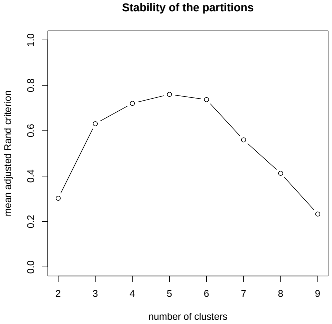
First example: "decathlon" data





First example: "decathlon" data

```
> stab<-stability(tree,B=40)
> plot(stab,main="Stability of the partitions")
```



First example: "decathlon" data

```
> part<-cutreevar(tree,5) #cut of the tree

> print(part)
Call:
cutreevar(obj = tree, k = 5)

name description
"$var" "list of variables in each cluster"
"$sim" "similarity matrix in each cluster"
"$cluster" "cluster memberships"
"$wss" "within-cluster sum of squares"
"$E" "gain in cohesion (in %)"
"$size" "size of each cluster"
"$scores" "score of each cluster"
```

First example: "decathlon" data

```
> summary(part)
Call:
cutreevar(obj = tree, k = 5)

Cluster 1 :
                squared loading
 100m                      0.68
 Long.jump                 0.69
 400m                      0.67
 110m.hurdle               0.64

...

Gain in cohesion (in %):  65.33
```

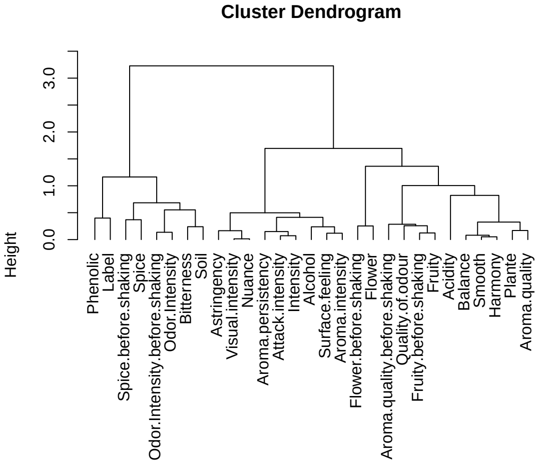
First example: "decathlon" data

```
> part$scores # synthetic variables

         cluster1  cluster2  cluster3  cluster4  cluster5
 SEBRLE      0.26     -0.72      0.94      1.02      1.10
 CLAY        1.38     -0.25      0.57      0.38      1.95
 KARPOV      1.11     -1.41      0.57     -1.68      1.84
 BERNARD    -0.19      1.12      2.03      0.93      0.09
 YURKOV     -2.03     -1.62     -0.15      1.07     -0.23
 WARNERS     1.14      0.67      0.57     -1.37     -0.08
 ...
```

Second example: "wine" data

```
> data(wine) #data of the package FactoMineR
> head(wine[,c(1:4)])
       Label       Soil        Odor.Intensity   Aroma.quality
 2EL   Saumur      Env1        3.07             3.00
 1CHA  Saumur      Env1        2.96             2.82
 1FON  Bourgueuil  Env1        2.85             2.92
 1VAU  Chinon      Env2        2.80             2.59
 1DAM  Saumur      Reference   3.60             3.42
 2BOU  Bourgueuil  Reference   2.85             3.11

> X.quanti <- wine[,c(3:29)]
> X.quali <- wine[,c(1,2)]
> tree <- hclustvar( X.quanti, X.quali)
> plot(tree)
```

Second example: "wine" data



Second example: "wine" data

```
> part<-cutreevar(tree,6) #cut of the tree

> summary(part)
Cluster 1 :
                                squared loading
 Odor.Intensity.before.shaking             0.76
 Spice.before.shaking                      0.62
 Odor.Intensity                            0.67
 Spice                                     0.54
 Bitterness                                0.66
 Soil                                      0.78
...
```

Outline

1. Introduction
2. The methods in ClustOfVar
3. Illustration on simple examples
4. Concluding remarks

Concluding remarks

- A package for the clustering of a **mixture of quantitative and qualitative variables**
- Bootstrap approach to help for the choice of the number of clusters (stability of the partition)
- Clustering of variables: alternative to MCA (resp. PCA) for dimension reduction
- PCAMIX with rotation will soon be available in an R package (named **PCAmixdata**)

Some references

Chavent, M., Kuentz, V., Liquet B., Saracco, J., (2010), The ClustOfVar R package, The CRAN R Project.

Dhillon, I.S, Marcotte, E.M., Roshan, U., (2003), Diametrical clustering for identifying anti-correlated gene clusters, *Bioinformatics*, **19**(13), 1612-1619.

Kiers, H.A.L., (1991), Simple structure in Component Analysis Techniques for mixtures of qualitative and quantitative variables, *Psychometrika*, **56**, 197-212.

Pagès, J., (2004), Analyse Factorielle de Données Mixtes [Factor Analysis for Mixed Data], *Revue de Statistique Appliquée*, **52**(4), 93-11.

Vigneau, E., Qannari, E.M., (2003), Clustering of variables around latent components, *Communications in statistics Simulation and Computation*, **32**(4), 1131-1150.

A similarity measure between two variables for mixed data

- The R function **mixedvarsim** returns a squared canonical correlation
- In case of two qualitative variables $\mathbf{z}_i$ and $\mathbf{z}_j$ having r and s categories the squared canonical correlation is calculated as follows: if $min(n, r, s)$ is equal to
  - n then return the first eigenvalue of $\tilde{\mathbf{Z}}_i\tilde{\mathbf{Z}}_i'\tilde{\mathbf{Z}}_j\tilde{\mathbf{Z}}_j'$
  - r then return the first eigenvalue of $\tilde{\mathbf{V}}_{ij}\tilde{\mathbf{V}}_{ji}$ with $\tilde{\mathbf{V}}_{ij} = \tilde{\mathbf{Z}}_i'\tilde{\mathbf{Z}}_j$
  - s then return the first eigenvalue of $\tilde{\mathbf{V}}_{ji}\tilde{\mathbf{V}}_{ij}$
- The squared correlation $r^2(\mathbf{x}_i, \mathbf{x}_j)$
- The correlation ratio $\eta^2(\mathbf{x}_i, \mathbf{z}_j)$