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

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

1. Introduction
2. The methods in ClustOfVar
3. Illustration on simple examples
4. Concluding remarks
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
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
Homogeneity criterion of a partition of variables

- $\mathcal{V}_1 = \{x_1, \ldots, x_{p_1}\}$ of **quantitative** variables
- $\mathcal{V}_2 = \{z_1, \ldots, z_{p_2}\}$ of **qualitative** variables
- Let $\mathbf{X}$ and $\mathbf{Z}$ 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 $P$:

$$
H(P) = \sum_{k=1}^{K} H(C_k, y_k)
$$

where $y_k$ is central (quantitative) synthetic variable also called the center of $C_k$
The function $H$ measures the adequacy between $C_k$ and $y_k$:

$$H(C_k, y_k) = \sum_{x_j \in C_k} r^2(x_j, y_k) + \sum_{z_j \in C_k} \eta^2(z_j, y_k)$$

where $r^2(x_j, y_k)$ is the squared correlation of $x_j$ with $y_k$ and $\eta^2(z_j, y_k)$ is the correlation ratio between $z_j$ and $y_k$. 

Homogeneity criterion of a cluster of variables
Definition of the synthetic variable of a cluster

- The center of $C_k$ is:

$$y_k = \arg \max_{u \in \mathbb{R}^n} \left\{ \sum_{x_j \in C_k} r^2(x_j, u) + \sum_{z_j \in C_k} \eta^2(z_j, u) \right\}$$

- $y_k$ 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
Let $X_k$ and $Z_k$ be the matrices of the columns of $X$ and $Z$ corresponding to the variables in $C_k$

Recoding of $X_k$ and $Z_k$:

- $\tilde{X}_k$ is the standardized version of the quantitative matrix $X_k$
- $\tilde{Z}_k = JGD^{-1/2}$ is the standardized version of the indicator matrix $G$ of the qualitative matrix $Z_k$, where $D$ is the diagonal matrix of frequencies of the categories and $J = I - 1'1/n$ is the centering operator
- $M_k = (\tilde{X}_k | \tilde{Z}_k)$
PCAMIX in a cluster

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

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

$\sqrt{n} \mathbf{U}_k \Lambda_k$ is the matrix of the PC’s scores of PCAMIX

$\mathbf{y}_k$ is the first column of this matrix

- The homogeneity of $C_k$ is:

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

$$= \lambda^1_k$$

$H(P) = \lambda^1_1 + \ldots + \lambda^1_k$
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

\( \leftrightarrow \) The \texttt{hclustvar} function gives a hierarchy
\( \leftrightarrow \) The \texttt{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
  
  → We defined a similarity measure between two variables of any type (quantitative and/or qualitative)
  → 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 $y_k$ 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)

$\rightarrow$ The \texttt{kmeansvar} R function
The stability of the partitions

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

- $B$ bootstrap samples of the observations are drawn and $B$ "bootstrap" 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
First example: "decathlon" data

```r
> 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

**Aggregation levels**

- Javeline
- High_jump
- Shot_put
- Discus
- Long_jump
- 400m
- 100m
- 110m_hurdle
- Pole_vault
- 1500m

**Cluster Dendrogram**

- Javeline
- High_jump
- Shot_put
- Discus
- Long_jump
- 400m
- 100m
- 110m_hurdle
- Pole_vault
- 1500m
First example: "decathlon" data

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

![Stability of the partitions](image)
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

```r
> part$scores # synthetic variables

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

UseR! 2011

ClustOfVar: an R package for the clustering of variables
Second example: "wine" data

```r
> data(wine) # data of the package FactoMineR
> head(wine[,c(1:4)])

<table>
<thead>
<tr>
<th>Label</th>
<th>Soil</th>
<th>Odor.Intensity</th>
<th>Aroma.quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>2EL</td>
<td>Saumur</td>
<td>Env1</td>
<td>3.07</td>
</tr>
<tr>
<td>1CHA</td>
<td>Saumur</td>
<td>Env1</td>
<td>2.96</td>
</tr>
<tr>
<td>1FON</td>
<td>Bourgueuil</td>
<td>Env1</td>
<td>2.85</td>
</tr>
<tr>
<td>1VAU</td>
<td>Chinon</td>
<td>Env2</td>
<td>2.80</td>
</tr>
<tr>
<td>1DAM</td>
<td>Saumur</td>
<td>Reference</td>
<td>3.60</td>
</tr>
<tr>
<td>2BOU</td>
<td>Bourgueuil</td>
<td>Reference</td>
<td>2.85</td>
</tr>
</tbody>
</table>

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

Cluster Dendrogram

Height

Phenolic Label Spice before shaking Spice Odor Intensity before shaking Odor Intensity Bitterness Soil Astringency before shaking Nuance Aroma persistency intensity Attack intensity Intensity Alcohol before shaking Aroma intensity before shaking Flavor before shaking Flower before shaking Aroma quality before shaking Quality of odour before shaking Fruity before shaking Acidity Balance Smooth Harmony Plante Aroma quality 0.0 1.0 2.0 3.0

UseR! 2011 ClustOfVar: an R package for the clustering of variables
Second example: "wine" data

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

> summary(part)
Cluster 1 :

<table>
<thead>
<tr>
<th></th>
<th>squared loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odor.Intensity.before.shaking</td>
<td>0.76</td>
</tr>
<tr>
<td>Spice.before.shaking</td>
<td>0.62</td>
</tr>
<tr>
<td>Odor.Intensity</td>
<td>0.67</td>
</tr>
<tr>
<td>Spice</td>
<td>0.54</td>
</tr>
<tr>
<td>Bitterness</td>
<td>0.66</td>
</tr>
<tr>
<td>Soil</td>
<td>0.78</td>
</tr>
</tbody>
</table>
```

...
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**)

*UseR! 2011 ClustOfVar: an R package for the clustering of variables*
Some references


A similarity measure between two variables for mixed data

- The R function `mixedvarsim` returns a squared canonical correlation.
- In case of two qualitative variables $z_i$ and $z_j$ 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 $\tilde{Z}_i\tilde{Z}_i'$ $\tilde{Z}_j\tilde{Z}_j'$
  - $r$ then return the first eigenvalue of $\tilde{V}_{ij}\tilde{V}_{ji}$ with $\tilde{V}_{ij} = \tilde{Z}_i\tilde{Z}_j$
  - $s$ then return the first eigenvalue of $\tilde{V}_{ji}\tilde{V}_{ij}$
- The squared correlation $r^2(x_i, x_j)$
- The correlation ratio $\eta^2(x_i, z_j)$