Journal of Statistical Software
September 2012, Volume 50, Issue 13. http://www.jstatsoft.org/

ClustOfVar: An R Package for the Clustering of Variables

Marie Chavent
University of Bordeaux

Vanessa Kuentz-Simonet
Irstea

Benoît Liquet
University of Bordeaux

Jérôme Saracco
University of Bordeaux

Abstract

Clustering of variables is as a way to arrange variables into homogeneous clusters, i.e., groups of variables which are strongly related to each other and thus bring the same information. These approaches can then be useful for dimension reduction and variable selection. Several specific methods have been developed for the clustering of numerical variables. However concerning qualitative variables or mixtures of quantitative and qualitative variables, far fewer methods have been proposed. The R package **ClustOfVar** was specifically developed for this purpose. The homogeneity criterion of a cluster is defined as the sum of correlation ratios (for qualitative variables) and squared correlations (for quantitative variables) to a synthetic quantitative variable, summarizing "as good as possible" the variables in the cluster. This synthetic variable is the first principal component obtained with the PCAMIX method. Two clustering algorithms are proposed to optimize the homogeneity criterion: iterative relocation algorithm and ascendant hierarchical clustering. We also propose a bootstrap approach in order to determine suitable numbers of clusters. We illustrate the methodologies and the associated package on small datasets.

Keywords: dimension reduction, hierarchical clustering of variables, k-means clustering of variables, mixture of quantitative and qualitative variables, stability.

1. Introduction

Principal component analysis (PCA) and multiple correspondence analysis (MCA) are appealing statistical tools for multivariate description of respectively numerical and categorical data. Rotated principal components fulfill the need to get more interpretable components.

Clustering of variables is an alternative since it makes possible to arrange variables into homogeneous clusters and thus to obtain meaningful structures. From a general point of view, variable clustering lumps together variables which are strongly related to each other and thus bring the same information. Once the variables are clustered into groups such that attributes in each group reflect the same aspect, the practitioner may be spurred on to select one variable from each group. One may also want to construct a synthetic variable. For instance in the case of quantitative variables, a solution is to realize a PCA in each cluster and to retain the first principal component as the synthetic variable of the cluster.

A simple and frequently used approach for clustering a set of variables is to calculate the dissimilarities between these variables and to apply a classical cluster analysis method to this dissimilarity matrix. We can cite the functions `hclust` of the R package **stats** (R Development Core Team 2012) and `agnes` of the package **cluster** (Maechler, Rousseeuw, Struyf, Hubert, and Hornik 2012) which can be used for single, complete, average linkage hierarchical clustering. The functions `diana` and `pam` of the package **cluster** can also be used for respectively divisive hierarchical clustering and partitioning around medoids (Kaufman and Rousseeuw 1990). But the dissimilarity matrix has to be calculated first. For quantitative variables many dissimilarity measures can be used: correlation coefficients (parametric or nonparametric) can be converted to different dissimilarities depending if the aim is to lump together correlated variables regardless of the sign of the correlation or if a negative correlation coeffcient between two variables shows disagreement between them. For categorical variables, many association measures can be used as χ^2, Rand, Belson, Jaccard, Sokal and Jordan among others. Many strategies can then be applied and it can be difficult for the user to choose one of them. Moreover, no synthetic variables of the clusters are directly provided with this approach.

Besides these classical methods devoted to the clustering of observations, there exists methods specifically devoted to the clustering of variables. The most famous one is the `VARCLUS` procedure of the SAS software (SAS Institute Inc. 2011). Recently specific methods based on PCA were proposed by Vigneau and Qannari (2003) with the name clustering around latent variables (CLV) and by Dhillon, Marcotte, and Roshan (2003) with the name Diametrical Clustering. But all these specific approaches work only with quantitative data and as far as we know, they are not implemented in R.

The aim of the package **ClustOfVar** – available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=ClustOfVar` – is then to propose in R, methods specifically devoted to the clustering of variables with no restriction on the type (quantitative or qualitative) of the variables. The clustering methods developed in the package work with a mixture of quantitative and qualitative variables and also work for a set exclusively containing quantitative (or qualitative) variables. Two methods are proposed for the clustering of variables: a hierarchical clustering algorithm and a k-means type partitioning algorithm. They are implemented in the functions `hclustvar` and `kmeansvar`. These two methods are based on PCAMIX, a principal component method for a mixture of qualitative and quantitative variables (Kiers 1991). This method includes the ordinary PCA and MCA as special cases. Here we use a Singular Value Decomposition (SVD) approach of PCAMIX (Chavent, Kuentz-Simonet, and Saracco 2012). Both clustering algorithms aim at maximizing the same homogeneity criterion: a cluster of variables is defined as homogeneous when the variables in the cluster are strongly linked to a central quantitative synthetic variable. This link is measured by the squared Pearson correlation for the quantitative variables and by the correlation ratio for the qualitative variables. The quantitative central synthetic variable of a

cluster is the first principal component of PCAMIX applied to all the variables in the cluster. Note that the synthetic variables of the clusters can be used for dimension reduction or for recoding purpose. Moreover a method based on a bootstrap approach is also proposed to evaluate the stability of the partitions of variables and can be used to determine a suitable number of clusters. It is implemented in the function `stability`.

In addition note that missing data are allowed: they are replaced by means for quantitative variables and by zeros in the indicator matrix for qualitative variables.

The rest of this paper is organized as follows. Section 2 contains a detailed description of the homogeneity criterion and a description of the PCAMIX procedure for the determination of the central synthetic variable. Section 3 describes the clustering algorithms and the bootstrap procedure. Section 4 provides two data-driven examples in order to illustrate the use of the functions and objects of the package **ClustOfVar**. It also provides computational time examples for simulated data. Finally, Section 5 gives concluding remarks.

2. The homogeneity criterion

Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_{p_1}\}$ be a set of p_1 quantitative variables and $\{\mathbf{y}_1, \ldots, \mathbf{y}_{p_2}\}$ a set of p_2 qualitative variables. Let $\mathbf{X}$ and $\mathbf{Y}$ be the corresponding quantitative and qualitative data matrices of dimensions $n \times p_1$ and $n \times p_2$, where n is the number of observations. For seek of simplicity, we denote $\mathbf{x}_j \in \mathcal{R}^n$ the j-th column of $\mathbf{X}$ and $\mathbf{y}_j \in \mathcal{M}_j^n$ the j-th column of $\mathbf{Y}$ with $\mathcal{M}_j$ the set of categories of $\mathbf{y}_j$. Let $P_K = (C_1, \ldots, C_K)$ be a partition into K clusters of the $p = p_1 + p_2$ variables.

Synthetic variable of a cluster C_k. It is defined as the quantitative variable $\mathbf{c}_k \in \mathcal{R}^n$ the "most linked" to all the variables in C_k:

$$\mathbf{c}_k = \arg\max_{\mathbf{u} \in \mathcal{R}^n} \left\{ \sum_{\mathbf{x}_j \in C_k} r^2_{\mathbf{u},\mathbf{x}_j} + \sum_{\mathbf{y}_j \in C_k} \eta^2_{\mathbf{u}|\mathbf{y}_j} \right\},$$

where r^2 denotes the squared Pearson correlation and η^2 denotes the correlation ratio. More precisely, the correlation ratio $\eta^2_{\mathbf{u}|\mathbf{y}_j} \in [0,1]$ measures the part of the variance of $\mathbf{u}$ explained by the categories of $\mathbf{y}_j$:

$$\eta^2_{\mathbf{u}|\mathbf{y}_j} = \frac{\sum_{s \in \mathcal{M}_j} n_s (\bar{\mathbf{u}}_s - \bar{\mathbf{u}})^2}{\sum_{i=1}^n (u_i - \bar{\mathbf{u}})^2},$$

where n_s is the frequency of category s, $\bar{\mathbf{u}}_s$ is the mean value of $\mathbf{u}$ calculated on the observations belonging to category s and $\bar{\mathbf{u}}$ is the mean of $\mathbf{u}$.

We have the following important results (Escofier (1979), Saporta (1990), Pagès (2004)):

- $\mathbf{c}_k$ is the first principal component of PCAMIX applied to $\mathbf{X}_k$ and $\mathbf{Y}_k$, the matrices made up of the columns of $\mathbf{X}$ and $\mathbf{Y}$ corresponding to the variables in C_k;
- the empirical variance of $\mathbf{c}_k$ is equal to: $\mathsf{VAR}(\mathbf{c}_k) = \sum_{\mathbf{x}_j \in C_k} r^2_{\mathbf{x}_j,\mathbf{c}_k} + \sum_{\mathbf{y}_j \in C_k} \eta^2_{\mathbf{c}_k|\mathbf{y}_j}$.

The determination of $\mathbf{c}_k$ using PCAMIX is carried on according to the following steps:

1. Recoding of $\mathbf{X}_k$ and $\mathbf{Y}_k$:
   (a) $\tilde{\mathbf{X}}_k$ is the standardized version of the quantitative matrix $\mathbf{X}_k$,
   (b) $\tilde{\mathbf{Y}}_k = \mathbf{JGD}^{-1/2}$ is the standardized version of the indicator matrix $\mathbf{G}$ of the qualitative matrix $\mathbf{Y}_k$, where $\mathbf{D}$ is the diagonal matrix of frequencies of the categories. $\mathbf{J} = \mathbf{I} - \mathbf{1}^\top\mathbf{1}/n$ is the centering operator where $\mathbf{I}$ denotes the identity matrix and $\mathbf{1}$ the vector with unit entries.
2. Concatenation of the two recoded matrices: $\mathbf{M}_k = \frac{1}{\sqrt{n}}(\tilde{\mathbf{X}}_k|\tilde{\mathbf{Y}}_k)$.
3. Singular Value Decomposition of $\mathbf{M}_k$: $\mathbf{M}_k = \mathbf{U}_k\Lambda_k\mathbf{V}'_k$.
4. Extraction/calculus of useful outputs:
   - $\sqrt{n}\mathbf{U}_k\Lambda_k$ is the matrix of the principal component scores of PCAMIX;
   - $\mathbf{c}_k = \sqrt{n}\mathbf{u}_k^1\lambda_{C_k}^1$ where $\mathbf{u}_k^1$ is the first eigenvector in $\mathbf{U}_k$ and $\lambda_{C_k}^1$ is the first eigenvalue in Λ_k;
   - $\mathsf{VAR}(\mathbf{c}_k) = \lambda_{C_k}^1$.

Note that we recently developed an R package named **PCAmixdata** with a function `PCAmix` which provides the principal components of PCAMIX and a function `PCArot` which provides the principal component after orthogonal rotation.

Homogeneity H of a cluster C_k. It is a measure of adequacy between the variables in the cluster and its central synthetic quantitative variable $\mathbf{c}_k$:

$$H(C_k) = \sum_{\mathbf{x}_j \in C_k} r^2_{\mathbf{x}_j,\mathbf{c}_k} + \sum_{\mathbf{y}_j \in C_k} \eta^2_{\mathbf{c}_k|\mathbf{y}_j} = \lambda^1_{C_k}. \tag{1}$$

The first term in (1) (based on the squared Pearson correlation r^2) measures the link between the quantitative variables in C_k and $\mathbf{c}_k$ independently of the sign of the relationship. The second one (based on the correlation ratio η^2) measures the link between the qualitative variables in C_k and $\mathbf{c}_k$. The homogeneity of a cluster is maximum when all the quantitative variables are correlated (or anti-correlated) to $\mathbf{c}_k$ and when all the correlation ratios of the qualitative variables are equal to 1. It means that all the variables in the cluster C_k bring the same information.

Homogeneity $\mathcal{H}$ of a partition P_K. It is defined as the sum of the homogeneities of its clusters:

$$\mathcal{H}(P_K) = \sum_{k=1}^{K} H(C_k) = \lambda^1_{C_1} + \ldots + \lambda^1_{C_K}, \tag{2}$$

where $\lambda^1_{C_1}, \ldots, \lambda^1_{C_K}$ are the first eigenvalues of PCAMIX applied to the K clusters C_k of P_K. This homogeneity is maximum for the partition of the singletons (partition in p clusters) with $\mathcal{H}(P_p) = p$. Indeed if C_k is a singleton (reduced to one variable), we have $H(C_k) = 1$.

3. The clustering algorithms

The aim is to find a partition of a set of quantitative and/or qualitative variables such that

the variables within a cluster are strongly related to each other. In other words the objective is to find a partition P_K which maximizes the homogeneity function $\mathcal{H}$ defined in (2). For this, a hierarchical and a partitioning clustering algorithms are proposed in the package **ClustOfVar**. A bootstrap procedure is also proposed to evaluate the stability of the partitions into $K = 2, 3, \ldots, p-1$ clusters and then to help the user to determine a suitable number of clusters of variables.

The hierarchical clustering algorithm. This algorithm builds a set of p nested partitions of variables in the following way:

1. Step $l = 0$: initialization. Start with the partition in p clusters.
2. Step $l = 1, \ldots, p-2$: aggregate two clusters of the partition in $p-l+1$ clusters to get a new partition in $p-l$ clusters. For this, choose clusters A and B with the smallest dissimilarity d defined as:

$$d(C_1, C_2) = H(C_1) + H(C_2) - H(C_1 \cup C_2) = \lambda^1_{C_1} + \lambda^1_{C_2} - \lambda^1_{C_1 \cup C_2}, \tag{3}$$

   where $H(C_k) = \lambda^1_{C_k}$ is obtained by PCAMIX on the variables in C_k as defined in (1).
3. Step $l = p-1$: stop. The partition in one cluster is obtained.

The dissimilarity d measures the lost of homogeneity observed when the two clusters C_1 and C_2 are merged. Using this aggregation measure the new partition in $p-l$ clusters maximizes $\mathcal{H}$ among all the partitions in $p-l$ clusters obtained by aggregation of two clusters of the partition in $p-l+1$ clusters. This algorithm is implemented in the function `hclustvar` which builds a hierarchy of the p variables. The function `plot.hclustvar` gives the dendrogram of this hierarchy. The height of a cluster $C = A \cup B$ in this dendrogram is defined as $h(C) = d(A, B)$. It is easy to verify that $h(C) \geq 0$ but the property "$A \subset B \Rightarrow h(A) \leq h(B)$" has not been proved yet. Nevertheless, inversions in the dendrogram have never been observed in practice neither on simulated data nor on real data sets. Finally the function `cutreevar` cuts this dendrogram and gives one of the p nested partitions according to the number K of clusters given in input by the user.

The partitioning algorithm. This partitioning algorithm requires the definition of a similarity measure between two variables of any type (quantitative or qualitative). We use for this purpose the squared canonical correlation between two data matrices $\mathbf{E}$ and $\mathbf{F}$ of dimensions $n \times r_1$ and $n \times r_2$. This correlation, denoted by ρ, can be easily calculated with the following procedure:

$$\rho(\mathbf{E}, \mathbf{F}) = \begin{cases} \text{first eigenvalue of the } n \times n \text{ matrix } \mathbf{E}\mathbf{F}^\top\mathbf{F}\mathbf{E}^\top & \text{if} \quad \min(n, r_1, r_2) = n, \\ \text{first eigenvalue of the } r_1 \times r_1 \text{ matrix } \mathbf{E}^\top\mathbf{F}\mathbf{F}^\top\mathbf{E} & \text{if} \quad \min(n, r_1, r_2) = r_1, \\ \text{first eigenvalue of the } r_2 \times r_2 \text{ matrix } \mathbf{F}^\top\mathbf{E}\mathbf{E}^\top\mathbf{F} & \text{if} \quad \min(n, r_1, r_2) = r_2. \end{cases}$$

More precisely:

- For two quantitative variables $\mathbf{x}_i$ and $\mathbf{x}_j$, let $\mathbf{E} = \tilde{\mathbf{x}}_i$ and $\mathbf{F} = \tilde{\mathbf{x}}_j$ where $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$ are the standardized versions of $\mathbf{x}_i$ and $\mathbf{x}_j$. In this case, the squared canonical correlation is the squared Pearson correlation: $\rho(\mathbf{x}_i, \mathbf{x}_j) = r^2_{\mathbf{x}_i, \mathbf{x}_j}$.

- For one qualitative variable $\mathbf{y}_i$ and one quantitative variable $\mathbf{x}_j$, let $\mathbf{E} = \tilde{\mathbf{Y}}_i$ and $\mathbf{F} = \tilde{\mathbf{x}}_j$ where $\tilde{\mathbf{Y}}_i$ is the standardized version of the indicator matrix $\mathbf{G}_i$ of the qualitative variable $\mathbf{y}_i$. In this case, the squared canonical correlation is the correlation ratio: $\rho(\mathbf{y}_i, \mathbf{x}_j) = \eta^2_{\mathbf{x}_j|\mathbf{y}_i}$.

- For two qualitative variables $\mathbf{y}_i$ and $\mathbf{y}_j$ having r and s categories, let $\mathbf{E} = \tilde{\mathbf{Y}}_i$ and $\mathbf{F} = \tilde{\mathbf{Y}}_j$. In this case, the squared canonical correlation $s(\mathbf{y}_i, \mathbf{y}_j)$ does not correspond to a well known association measure. Its interpretation is geometrical: the closer to one is $\rho(\mathbf{y}_i, \mathbf{y}_j)$, the closer are the two linear subspaces og $\mathcal{R}^n$ spanned by the matrices $\mathbf{E}$ and $\mathbf{F}$. Then the two qualitative variables $\mathbf{y}_i$ and $\mathbf{y}_j$ bring similar information.

This similarity measure is implemented in the function `mixedVarSim`.

The clustering algorithm implemented in the function `kmeansvar` builds then a partition in K clusters in the following way:

1. Initialization step: two possibilities are available.
   (a) A non random initialization: an initial partition in K clusters is given in input (for instance the partition obtained by cutting the dendrogram of the hierarchy).
   (b) A random initialization:
      i. K variables are randomly selected among the p variables as initial central synthetic variables (named centers hereafter).
      ii. An initial partition into K clusters is built by allocating each variable to the cluster with the closest initial center: the similarity between a variable and an initial center is calculated using the function `mixedVarSim`.
2. Repeat
   (a) A representation step: the quantitative central synthetic variable $\mathbf{c}_k$ of each cluster C_k is calculated with PCAMIX as defined in Section 2.
   (b) An allocation step: a partition is constructed by assigning each variable to the closest cluster. The similarity between a variable and the central synthetic quantitative variable of the corresponding cluster is calculated with the function `mixedVarSim`: it is either a squared correlation (if the variable is quantitative) or a correlation ratio (if the variable is qualitative).
3. Stop if there is no more changes in the partition or if a maximum number of iterations (fixed by the user) is reached.

This iterative procedure `kmeansvar` provides a partition P_K into K clusters which maximizes $\mathcal{H}$ but this optimum is local and may depend on the initial partition. A solution to overcome this problem and to avoid the influence of the choice of an arbitrary initial partition is to consider multiple random initializations. In this case, steps 1(b), 2 and 3 are repeated, and we propose to retain as final partition the one which provides the highest value of $\mathcal{H}$.

Stability of partitions of variables. We propose a procedure which evaluates the stability of the p nested partitions of the dendrogram obtained with `hclustvar`. It works as follows:

1. B boostrap samples of the n observations are drawn and the corresponding B dendrograms are obtained with the function `hclustvar`.

2. The partitions of these B dendrograms are compared with the partitions of the initial hierarchy using the corrected Rand index. The Rand and the adjusted Rand indices are implemented in the function `Rand` (see Hubert and Arabie 1985 for details on these indices).

3. The stability of a partition is evaluated by the mean of the B adjusted Rand indices.

The plot of this stability criterion according to the number of clusters can help the user in the choice of a sensible and suitable number of clusters. Note that an error message may appear with this function in some case of rare categories of qualitative variable. Indeed, if a rare category disappears in a bootstrap sample of observations, a column of identical values is then formed and the standardization of this variable is not possible in PCAMIX step.

4. Illustration on simple examples

First, we illustrate our R package **ClustOfVar** on two real datasets. The first one only concerns quantitative variables, the second one is a mixture of quantitative and qualitative variables. Then we give computational times examples for the functions `kmeansvar` and `hclustvar` applied on simulated data.

4.1. First example: Quantitative data

We use the dataset `decathlon` which contains $n = 41$ athletes described according to their performances in $p = 10$ different sports of decathlon.

```
R> library("ClustOfVar")
R> data("decathlon")
R> 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
```

In order to have an idea of the links between these 10 quantitative variables, one usually plots the correlation circle of the two first PCA dimensions. Figure 1 gives a first sight of groups of correlated or anti-correlated variables. However it does not provide a strict partition of variables. To go further we construct a hierarchy with the function `hclustvar`.

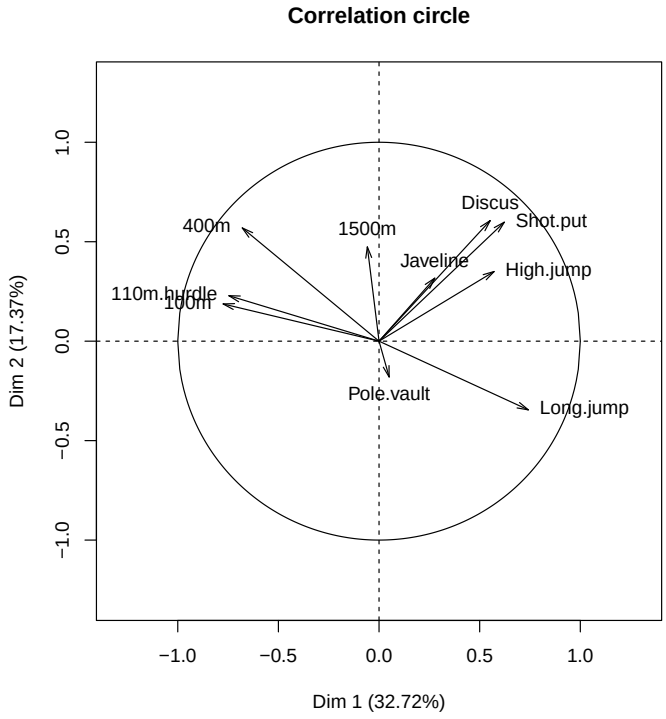

Figure 1: Correlation circle of the two first PCA dimensions.

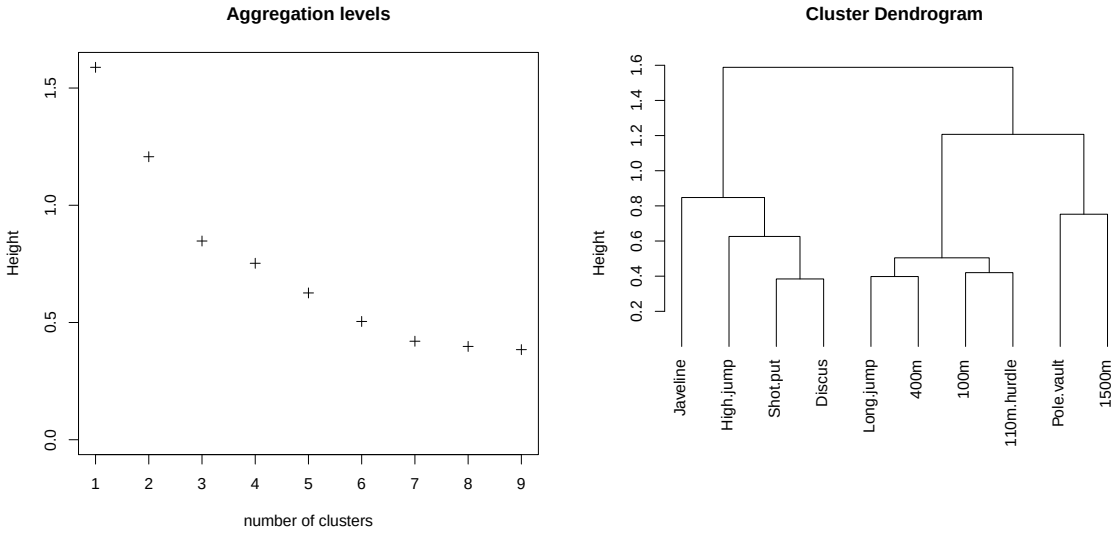

Figure 2: Graphical output of the function `plot.hclustvar`.

```
R> tree <- hclustvar(decathlon[, 1:10])
R> plot(tree)
```

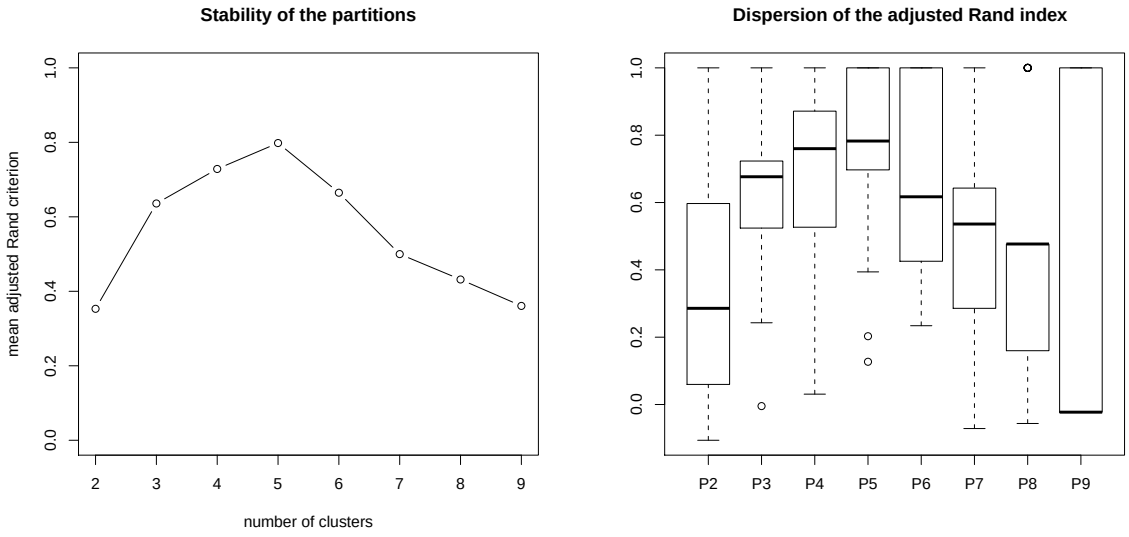

Figure 3: Graphical output of the functions `stability` and `plot.clustab`.

In Figure 2, the plot of the aggregation levels suggests to choose 3 clusters of variables. The dendrogram, on the right hand side of this figure, shows the link between the variables in terms of r^2. For instance, the two variables `Discus` and `Shot.put` are linked as well as the two variables `Long.jump` and `400m`, but the user must keep in mind that the dendrogram does not indicate the sign of these relationships: a careful study of these variables shows that `Discus` and `Shot.put` are correlated whereas `Long.jump` and `400m` are anti-correlated.

The user can use the `stability` function in order to have an idea of the stability of the partitions of the dendrogram represented in Figure 2.

```
R> stab <- stability(tree, B = 40)
R> plot(stab, main = "Stability of the partitions")
R> boxplot(stab$matCR, main = "Dispersion of the adjusted Rand index")
```

On the left of Figure 3, the plot of the mean (over the $B = 40$ bootstrap samples) of the adjusted Rand indices is obtained with the function `plot.clustab`. It clearly suggests to choose 5 clusters. The boxplots on the right of Figure 3 show the dispersion of these indices over the $B = 40$ bootstrap replications for partition, and they also suggest to retain 5 clusters.

In the following we choose $K = 3$ clusters because PCA applied to each of the 3 clusters gives each time only one eigenvalue greater than 1. The function `cutree` cuts the dendrogram of the hierarchy and gives a partition into $K = 3$ clusters of the $p = 10$ variables:

```
R> P3 <- cutreevar(tree, 3, matsim = TRUE)
R> cluster <- P3$cluster
R> X <- decathlon[, 1:10]
R> princomp(X[, which(cluster==1)], cor = TRUE)$sdev^2
R> princomp(X[, which(cluster==2)], cor = TRUE)$sdev^2
R> princomp(X[, which(cluster==3)], cor = TRUE)$sdev^2
```

The partition `P3` is contained in an object of class `clustvar`. Note that partitions obtained with the `kmeansvar` function are also objects of class `clustvar`. The function `print` method gives a description of the values of this object.

```
R> print(P3)

Call:
cutreevar(obj = tree, k = 3)

 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"
```

The value `$wss` is $\mathcal{H}(P_K)$ where the homogeneity function $\mathcal{H}$ is defined in (2). The gain in cohesion `$E` is the percentage of homogeneity which is accounted by the partition P_K. It is defined by:

$$E(P_K) = \frac{\mathcal{H}(P_K) - \mathcal{H}(P_1)}{\mathcal{H}(P_p) - \mathcal{H}(P_1)}. \tag{4}$$

where $\mathcal{H}(P_p) = p$. The value `$sim` provides the similarity matrices of the variables in each cluster (calculated with the function `mixedVarSim`). Note that it is time consuming to perform these similarity matrices when the number of variables is large. Thus they are not calculated by default: `matsim=TRUE` must be specified in the parameters of the function `cutreevar` if the user wants this output. We provide below the similarity matrix for the first cluster of this partition into 3 clusters.

```
> round(P3$sim$cluster1, digit = 2)

            100m Long.jump 400m 110m.hurdle
100m        1.00      0.36 0.27        0.34
Long.jump   0.36      1.00 0.36        0.26
400m        0.27      0.36 1.00        0.30
110m.hurdle 0.34      0.26 0.30        1.00
```

The value `$cluster` is a vector of integers indicating the cluster to which each variable is allocated.

```
R> P3$cluster

       100m   Long.jump    Shot.put   High.jump        400m 110m.hurdle
          1           1           2           2           1           1
     Discus  Pole.vault    Javeline       1500m
          2           3           2           3
```

The value `$var` gives a description of each cluster of the partition. More precisely it provides for each cluster the squared loadings with the central synthetic variable of the cluster (which is the first principal component of PCAMIX). For quantitative variables (resp. qualitative), the squared loadings are squared correlations (resp. correlation ratio) with this central synthetic variable. For instance the squared correlation between the variable `100m` and the central synthetic variable of `cluster1` is 0.68.

```
R> P3$var

$cluster1
            squared loading
100m              0.6822349
Long.jump         0.6873076
400m              0.6652279
110m.hurdle       0.6427661

$cluster2
          squared loading
Shot.put        0.7861012
High.jump       0.4991778
Discus          0.6023186
Javeline        0.2546550

$cluster3
           squared loading
Pole.vault       0.6237239
1500m            0.6237239
```

The value `$scores` is the $n \times K$ matrix of the scores of the n observations on the first principal components of PCAMIX applied to the K clusters: PCAMIX is applied 3 times here, one time in each cluster. Each column is then the synthetic variable of the cluster. The central synthetic variable of `cluster1` for instance is the first column of the 41×3 matrix above. This column gives the scores of the 41 athletes on the first component of PCAMIX applied to the variables of `cluster1` (`100m`, `Long.jump`, `400m`, `110m.hurdle`).

```
R> head(P3$scores)

          cluster1   cluster2   cluster3
SEBRLE   0.2640687 -1.0353928 -1.4405915
CLAY     1.3816943 -0.3454687 -1.7840860
KARPOV   1.1098485 -0.7209119 -1.7043603
BERNARD -0.1949061  0.7082857 -1.5017373
YURKOV  -2.0319539 -1.8850107  0.2702640
WARNERS  1.1385110  1.0929346 -0.3490226
```

Note that this 41×3 matrix of the scores of the 41 athletes in each cluster of variables is of course different from the 41×3 matrix of the scores of the athletes on the first 3 principal

components of PCAMIX (here PCA) applied to the initial dataset. The 3 synthetic variables for instance can be correlated whereas the first 3 principal components of PCAMIX are not correlated by construction. Moreover this matrix of the synthetic variables in `$scores` can be used as the matrix of the principal components of PCAMIX for dimension reduction purpose.

4.2. Second example: A mixture of quantitative and qualitative data

We use the dataset `wine` which contains $n = 21$ french wines described by $p = 31$ variables. The first two variables `Label` and `Soil` are qualitative with respectively 3 and 4 categories. The other 29 variables are quantitative.

```
R> data("wine")
R> head(wine[, 1:4])
```

```
          Label      Soil Odor.Intensity Aroma.quality
2EL      Saumur      Env1          3.074         3.000
1CHA     Saumur      Env1          2.964         2.821
1FON Bourgueuil      Env1          2.857         2.929
1VAU     Chinon      Env2          2.808         2.593
1DAM     Saumur Reference          3.607         3.429
2BOU Bourgueuil Reference          2.857         3.111
```

In order to have an idea of the links between the 29 first quantitative variables and the two qualitative variables, we construct a hierarchy using the function `hclustvar`.

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

In Figure 4, we plot the dendrogram. It shows for instance that the qualitative variable `Label` is linked (in term of correlation ratio) with the quantitative variable `Phenolic`. The user chooses according to this dendrogram to cut this dendrogram into $K = 6$ clusters:

```
R> part_hier <- cutreevar(tree, 6)
R> part_hier$var$cluster1
```

```
                    squared loading
Odor.Intensity            0.7617528
Spice.before.shaking      0.6160243
Odor.Intensity.1          0.6663325
Spice                     0.5357837
Bitterness                0.6620632
Soil                      0.7768805
```

A close reading of the output for `cluster1` shows that the correlation ratio between the qualitative variable `Soil` and the synthetic variable of the cluster is about 0.78. The squared correlation between `Odor.Intensity` and the synthetic variable of the cluster is 0.76.

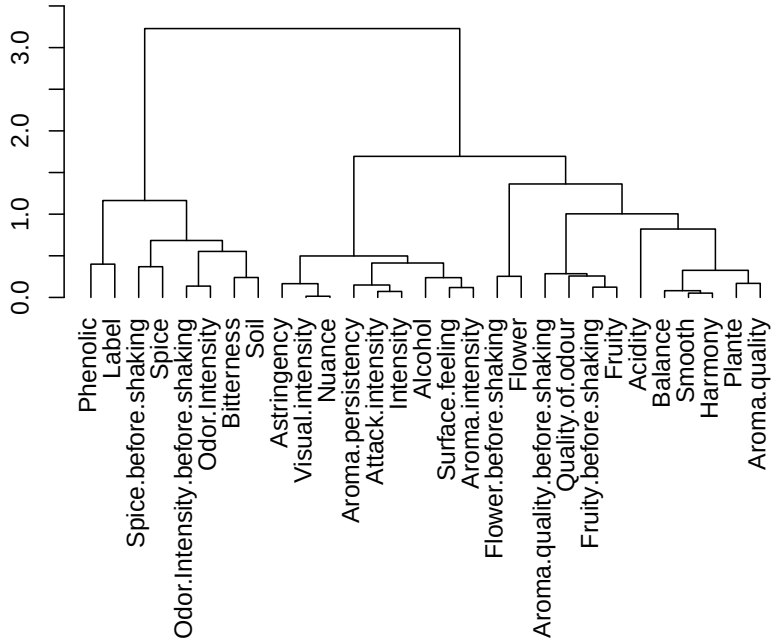

Figure 4: Dendrogram of the hierarchy of the 29 variables of the wine dataset.

The central synthetic variables of the 6 clusters are in `part_hier$scores`. This 21×6 quantitative matrix can replace the original 21×31 data matrix mixing qualitative and quantitative variables. Usually this matrix of the synthetic variables can then be used for recoding a mixed data matrix (or a qualitative data matrix) into a quantitative data matrix, as is usually done with the matrix of the principal components of PCAMIX.

The function `kmeansvar` can also provide a partition into $K = 6$ clusters of the 31 variables.

```
R> part_km <- kmeansvar(X.quanti, X.quali, init = 6, nstart = 10)
```

The gain in cohesion of the partition obtained with the k-means type partitioning algorithm and 10 random initializations is smaller than that of the partition obtained with the hierarchical clustering algorithm (51.02 versus 56.84):

```
R> part_km$E
[1] 51.02414
R> part_hier$E
[1] 56.84082
```

4.3. Computational time examples for simulated data

The time complexity of the proposed variable clustering algorithms is obtained from the expression of complexity for objects clustering by swapping the number of objects n by the number of variables p. Concerning the hierarchical ascendant algorithm, we use the Nearest Neighbours Chain algorithm implemented in the `hclust` R function. The complexity of this algorithm is then quadratic in p that is $o(p^2)$. On the other side, the complexity of the

	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 10000$	$n = 20000$
`kmeansvar` ($K = 10$)	0.004	0.012	0.020	0.104	0.215	0.441
`hclustvar`	0.072	0.106	0.141	0.497	0.867	1.644

Table 1: CPU time in minutes (with 3.06 GHz processor) for $p = 100$ variables and varying number of objects.

	$p = 100$	$p = 500$	$p = 1000$	$p = 5000$	$p = 10000$	$p = 20000$
`kmeansvar` ($K = 10$)	0.004	0.024	0.069	0.632	1.387	4.034
`hclustvar`	0.072	3.078	18.282	–	–	–

Table 2: CPU time in minutes (with 3.06 GHz processor) for $n = 100$ objects and varying number of variables.

iterative relocation algorithm is classical and linear in p. The main difference lies in the fact that we do not calculate some gravity centers but we realize Singular Value Decompositions, which complexity for a $(n \times p)$ matrix of rank r is equal to $o(npr)$. But to give a concrete idea of computational times for `hclustvar` and `kmeansvar` R functions, we simulate quantitative data matrices drawn from a uniform distribution with varying number of observations and variables. We calculate the CPU time in minutes (with 3.06 GHz processor) for $p = 100$ variables and varying number of objects (from $n = 100$ to $n = 20000$) and then for $n = 100$ objects and varying number of variables (from $p = 100$ to $p = 20000$ for `kmeansvar` and from $p = 100$ to $p = 1000$ for `hclustvar`)

Table 1 shows that for $p = 100$ variables, both clustering functions remain fast even if the number n of objects increases. For growing number of variables with fixed number of objects ($n = 100$), Table 2 shows that `kmeansvar` takes from few seconds with $p = 100$ variables to few minutes with $p = 20000$ variables. Not surprisingly, the `hclustvar` function is slower. For instance it takes 18 minutes for $p = 1000$ variables. We stopped our numerical experiments at this number of variables because it should have taken about at least 7 hours ($5^2 \times 18$mn) for $p = 5000$ variables. In this case, `kmeansvar` and `hclustvar` can be combined in the following way. First we apply `kmeansvar` with $K = 1000$ clusters for instance and then we build with `hclustvar` the hierarchy from the previously obtained 1000 synthetic variables. The idea was previously and widely used in the context of objects clustering by several authors (see for instance Wong 1982).

For datasets with simultaneously large number of objects and variables, both functions encounter problems of computation time and storage capacity.

5. Concluding remarks

The R package **ClustOfVar** proposes hierarchical and k-means type algorithms for the clustering of variables of any type (quantitative and/or qualitative).

This package proposes useful tools to visualize the links between the variables and the redundancy in a data set. It is also an alternative to PCA or MCA methods for dimension reduction and for recoding qualitative or mixed data matrices into quantitative data matrices. Let us recall that the main difference between PCA (for instance) and the approach of clustering of variables presented in this paper, is that the synthetic variables of the clusters

can be correlated whereas the principal components are not correlated by construction.

To deal with datasets having huge number of variables, a future work is to propose a new version of the package with the functions `hclustvar`, `kmeansvar` and `stability` developed for parallel computing.

We mentioned that the package **ClustOfVar** can deal with missing data. However let us note that the imputation method used in the code is simple and may not perform well when the proportion of missing data is too large. In that case, one of the numerous R packages devoted to missing data imputation should be used prior to **ClustOfVar**.

References

Chavent M, Kuentz-Simonet V, Saracco J (2012). "Orthogonal Rotation in PCAMIX." *Advances in Data Analysis and Classification*, **6**(2), 131–146.

Dhillon IS, Marcotte EM, Roshan U (2003). "Diametrical Clustering for Identifying Anti-Correlated Gene Clusters." *Bioinformatics*, **19**(13), 1612–1619.

Escofier B (1979). "Traitement Simultané de Variables Qualitatives et Quantitatives en Analyse Factorielle [Simultaneous Treatment of Qualitative and Quantitative Variables in Factor Analysis]." *Les Cahiers de l'Analyse des Données*, **4**(2), 137–146.

Hubert L, Arabie P (1985). "Comparing Partitions." *Journal of Classification*, **2**(1), 193–208.

Kaufman L, Rousseeuw PJ (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Hoboken.

Kiers HAL (1991). "Simple Structure in Component Analysis Techniques for Mixtures of Qualitative and Quantitative Variables." *Psychometrika*, **56**, 197–212.

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2012). ***cluster**: Cluster Analysis Basics and Extensions*. R package version 1.14.2, URL http://CRAN.R-project.org/package=cluster.

Pagès J (2004). "Analyse Factorielle de Données Mixtes [Factor Analysis for Mixed Data]." *Revue de Statistique Appliquée*, **52**(4), 93–111.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Saporta G (1990). "Simultaneous Treatment of Quantitative and Qualitative Data." In *Atti dela XXXV Riunone Scientifica; Società Italiana di Statistica*, pp. 63–72.

SAS Institute Inc (2011). *SAS/STAT Software, Version 9.2, The* `VARCLUS` *Procedure*. SAS Institute Inc., Cary, NC. URL http://support.sas.com/documentation/onlinedoc/stat/930/varclus.pdf.

Vigneau E, Qannari EM (2003). "Clustering of Variables around Latent Components." *Communications in Statistics Simulation and Computation*, **32**(4), 1131–1150.

Wong MA (1982). "A Hybrid Clustering Method for Identifying High-Density Clusters." *Journal of the American Statistical Association*, **77**(380), 841–847.

Affiliation:

Marie Chavent
University of Bordeaux, IMB, UMR 5251
33400 Talence, France
and
CNRS, IMB, UMR 5251
33400 Talence, France
and
INRIA
33400 Talence, France
E-mail: Marie.Chavent@u-bordeaux2.fr

Vanessa Kuentz-Simonet
Irstea, UR ADBX
33612 Cestas Cedex, France
E-mail: vanessa.kuentz-simonet@irstea.fr

Benoît Liquet
INSERM, ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique
33000 Bordeaux, France
and
University of Bordeaux, ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique
33000 Bordeaux, France
E-mail: Benoit.Liquet@isped.u-bordeaux2.fr

Jérôme Saracco
IPB, IMB, UMR 5251,
33400 Talence, France
and
CNRS, IMB, UMR 5251
33400 Talence, France
and
INRIA
33400 Talence, France
E-mail: jerome.saracco@math.u-bordeaux1.fr

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 50, Issue 13 *Submitted:* 2011-11-04
September 2012 *Accepted:* 2012-08-22