Package ‘ClustOfVar’

May 20, 2015

Type Package

Title Clustering of variables

Version 1.1

Date 2015-04-21

Author
Marie Chavent and Vanessa Kuentz and Amaury Labenne and Benoit Liquet and Jerome Saracco

Maintainer <marie.chavent@u-bordeaux.fr>

Description Cluster analysis of a set of variables. Variables can be quantitative, qualitative or a mixture of both.

License GPL (>=2.0)

Depends R(>= 2.10.0),
PCAmixdata,

R topics documented:

cutreevar *Cut a tree into groups of variables*

Description

Cuts a hierarchical tree of variables resulting from `hclustvar` into several clusters by specifying the desired number of clusters.

Usage

```
cutreevar(obj, k = NULL, matsim = FALSE)
```

Arguments

`obj`	an object of class 'hclustvar'.
`k`	an integer scalar with the desired number of clusters.
`matsim`	boolean, if 'TRUE', the matrices of similarities between variables in same cluster are calculated.

Value

`var`	a list of matrices of squared loadings i.e. for each cluster of variables, the squared loadings on first principal component of PCAmix. For quantitative variables (resp. qualitative), squared loadings are the squared correlations (resp. the correlation ratios) with the first PC (the cluster center).
`sim`	a list of matrices of similarities i.e. for each cluster, similarities between their variables. The similarity between two variables is defined as a square cosine: the square of the Pearson correlation when the two variables are quantitative; the correlation ratio when one variable is quantitative and the other one is qualitative; the square of the canonical correlation between two sets of dummy variables, when the two variables are qualitative. `sim` is 'NULL' if 'matsim' is 'FALSE'.
`cluster`	a vector of integers indicating the cluster to which each variable is allocated.
`wss`	the within-cluster sum of squares for each cluster: the sum of the correlation ratio (for qualitative variables) and the squared correlation (for quantitative variables) between the variables and the center of the cluster.
`E`	the pourcentage of homogeneity which is accounted by the partition in k clusters.
`size`	the number of variables in each cluster.
`scores`	a `n` by `k` numerical matrix which contains the `k` cluster centers. The center of a cluster is a synthetic variable: the first principal component calculated by PCAmix. The `k` columns of `scores` contain the scores of the `n` observations units on the first PCs of the `k` clusters.
`coef`	a list of the coefficients of the linear combinations defining the synthetic variable of each cluster.

Author(s)

`<Marie.Chavent@u-bordeaux.fr>`, Vanessa Kuentz, Amaury Labenne Benoit Liquet, Jerome Saracco

See Also

`hclustvar`, `summary.clustvar`,`predict.clustvar`,`stability`

Examples

```
data(decathlon)
tree <- hclustvar(decathlon[,1:10])
plot(tree)
#choice of the number of clusters
stability(tree,B=40)
part <- cutreevar(tree,4)
print(part)
summary(part)
```

hclustvar *Hierarchical clustering of variables*

Description

Ascendant hierarchical clustering of a set of variables. Variables can be quantitative, qualitative or a mixture of both. The aggregation criterion is the decrease in homogeneity for the clusters being merged. The homogeneity of a cluster is the sum of the correlation ratio (for qualitative variables) and the squared correlation (for quantitative variables) between the variables and the center of the cluster which is the first principal component of PCAmix. PCAmix is defined for a mixture of qualitative and quantitative variables and includes ordinary principal component analysis (PCA) and multiple correspondence analysis (MCA) as special cases. Missing values are replaced by means for quantitative variables and by zeros in the indicator matrix for qualitative variables.

Usage

```
hclustvar(X.quanti = NULL, X.quali = NULL, init = NULL)
```

Arguments

`X.quanti`	a numeric matrix of data, or an object that can be coerced to such a matrix (such as a numeric vector or a data frame with all numeric columns).
`X.quali`	a categorical matrix of data, or an object that can be coerced to such a matrix (such as a character vector, a factor or a data frame with all factor columns).
`init`	an initial partition (a vector of integers indicating the cluster to which each variable is allocated).

Details

If the quantitative and qualitative data are in a same dataframe, the function `splitmix` can be used to extract automatically the qualitative and the quantitative data in two separated dataframes.

Value

`height` a set of p-1 non-decreasing real values: the values of the aggregation criterion.

`clusmat` a p by p matrix with group memberships where each column k corresponds to the elements of the partition in k clusters.

`merge` a p-1 by 2 matrix. Row i of `merge` describes the merging of clusters at step i of the clustering. If an element j in the row is negative, then observation -j was merged at this stage. If j is positive then the merge was with the cluster formed at the (earlier) stage j of the algorithm. Thus negative entries in `merge` indicate agglomerations of singletons, and positive entries indicate agglomerations of non-singletons.

Author(s)

Marie Chavent <`marie.chavent@u-bordeaux.fr`>, Vanessa Kuentz, Amaury Labenne Benoit Liquet, Jerome Saracco

References

Chavent, M., Liquet, B., Kuentz, V., Saracco, J. (2012), ClustOfVar: An R Package for the Clustering of Variables. Journal of Statistical Software, Vol. 50, pp. 1-16.

Chavent, M., Kuentz, V., Saracco, J. (2011), Orthogonal Rotation in PCAMIX. Advances in Classification and Data Analysis, Vol. 6, pp. 131-146.

See Also

`cutreevar`, `plot.hclustvar`, `stability`,`kmeansvar`,`splitmix`

Examples

```
#quantitative variables
data(decathlon)
tree <- hclustvar(X.quanti=decathlon[,1:10], X.quali=NULL, init=NULL)
plot(tree)

#qualitative variables with missing values
data(vnf)
tree_NA <- hclustvar(X.quali=vnf, X.quanti=NULL)
plot(tree_NA)
dev.new()
vnf2<-na.omit(vnf)
tree <- hclustvar(X.quali=vnf2, X.quanti=NULL)
plot(tree)

#mixture of quantitative and qualitative variables
data(wine)
X.quanti <- splitmix(wine)$X.quanti[,1:27]
X.quali <- splitmix(wine)$X.quali
tree <- hclustvar(X.quanti,X.quali)
plot(tree)

#combined clustering
#library(mixOmics)
#data(breast.tumors)
#X.quanti <- breast.tumors$gene.exp
```

```
#init<- kmeansvar(X.quanti,init=100)$cluster
#tree <- hclustvar(X.quanti,init=init)
#plot(tree)

#data(yeast)
#X.quanti <- yeast$data
#tree <- hclustvar(X.quanti)
#plot(tree)
#init<- cutreevar(tree,k=10)$cluster
#tree2 <- hclustvar(X.quanti,init=init)
#plot(tree2)

#cutreevar(tree,3)$cluster
#cutreevar(tree2,3)$cluster
```

---

kmeansvar *k-means clustering of variables*

Description

Iterative relocation algorithm of k-means type which performs a partitionning of a set of variables. Variables can be quantitative, qualitative or a mixture of both. The center of a cluster of variables is a synthetic variable but is not a 'mean' as for classical k-means. This synthetic variable is the first principal component calculated by PCAmix. PCAmix is defined for a mixture of qualitative and quantitative variables and includes ordinary principal component analysis (PCA) and multiple correspondence analysis (MCA) as special cases. The homogeneity of a cluster of variables is defined as the sum of the correlation ratio (for qualitative variables) and the squared correlation (for quantitative variables) between the variables and the center of the cluster, which is in all cases a numerical variable. Missing values are replaced by means for quantitative variables and by zeros in the indicator matrix for qualitative variables.

Usage

```
kmeansvar(X.quanti = NULL, X.quali = NULL, init, iter.max = 150,
  nstart = 1, matsim = FALSE)
```

Arguments

`X.quanti`	a numeric matrix of data, or an object that can be coerced to such a matrix (such as a numeric vector or a data frame with all numeric columns).
`X.quali`	a categorical matrix of data, or an object that can be coerced to such a matrix (such as a character vector, a factor or a data frame with all factor columns).
`init`	either the number of clusters or an initial partition (a vector of integers indicating the cluster to which each variable is allocated). If `init` is a number, a random set of (distinct) columns in `X.quali` and `X.quanti` is chosen as the initial cluster centers.
`iter.max`	the maximum number of iterations allowed.
`nstart`	if `init` is a number, `nstart` corresponds with the number of random sets used in the process.
`matsim`	boolean, if 'TRUE', the matrices of similarities between variables in same cluster are calculated.

Details

If the quantitative and qualitative data are in a same dataframe, the function `splitmix` can be used to extract automatically the qualitative and the quantitative data in two separated dataframes.

Value

`var` a list of matrices of squared loadings i.e. for each cluster of variables, the squared loadings on first principal component of PCAmix. For quantitative variables (resp. qualitative), squared loadings are the squared correlations (resp. the correlation ratios) with the first PC (the cluster center).

`sim` a list of matrices of similarities i.e. for each cluster, similarities between their variables. The similarity between two variables is defined as a square cosine: the square of the Pearson correlation when the two variables are quantitative; the correlation ratio when one variable is quantitative and the other one is qualitative; the square of the canonical correlation between two sets of dummy variables, when the two variables are qualitative. `sim` is 'NULL' if 'matsim' is 'FALSE'.

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

`wss` the within-cluster sum of squares for each cluster: the sum of the correlation ratio (for qualitative variables) and the squared correlation (for quantitative variables) between the variables and the center of the cluster.

`E` the pourcentage of homogeneity which is accounted by the partition in k clusters.

`size` the number of variables in each cluster.

`scores` a n by k numerical matrix which contains the k cluster centers. The center of a cluster is a synthetic variable: the first principal component calculated by PCAmix. The k columns of `scores` contain the scores of the n observations units on the first PCs of the k clusters.

`coef` a list of the coefficients of the linear combinations defining the synthetic variable of each cluster.

Author(s)

Marie Chavent <`Marie.Chavent@u-bordeaux.fr`>, Vanessa Kuentz, Amaury Labenne, Benoit Liquet, Jerome Saracco

References

Chavent, M., Liquet, B., Kuentz, V., Saracco, J. (2012), ClustOfVar: An R Package for the Clustering of Variables. Journal of Statistical Software, Vol. 50, pp. 1-16.

Chavent, M., Kuentz, V., Saracco, J. (2011), Orthogonal Rotation in PCAMIX. Advances in Classification and Data Analysis, Vol. 6, pp. 131-146.

See Also

`splitmix`, `summary.clustvar`,`print.clustvar`,`stability`,`cutreevar`,`predict.clustvar`

Examples

```
data(decathlon)
#choice of the number of clusters
tree <- hclustvar(X.quanti=decathlon[,1:10])
stab <- stability(tree,B=60)
#a random set of variables is chosen as the initial cluster centers, nstart=10 times
part1 <- kmeansvar(X.quanti=decathlon[,1:10],init=5,nstart=10)
summary(part1)
#the partition from the hierarchical clustering is chosen as initial partition
part_init<-cutreevar(tree,5)$cluster
part2<-kmeansvar(X.quanti=decathlon[,1:10],init=part_init,matsim=TRUE)
summary(part2)
part2$sim
```

`mixedVarSim` *Similarity between two variables*

Description

Returns the similarity between two quantitative variables, two qualitative variables or a quantitative variable and a qualitative variable. The similarity between two variables is defined as a square cosine: the square of the Pearson correlation when the two variables are quantitative; the correlation ratio when one variable is quantitative and the other one is qualitative; the square of the canonical correlation between two sets of dummy variables, when the two variables are qualitative.

Usage

```
mixedVarSim(X1, X2)
```

Arguments

`X1`	a vector or a factor
`X2`	a vector or a factor

Author(s)

Marie Chavent <`Marie.Chavent@u-bordeaux.fr`>, Vanessa Kuentz, Benoit Liquet, Jerome Saracco

---

`plot.clustab` *Plot of an index of stability of partitions of variables*

Description

Plot of the index of stability of the partitions against the number of clusters.

Usage

```
## S3 method for class clustab
plot(x, nmin = NULL, nmax = NULL, ...)
```

Arguments

`x`	an object of class `clustab`.
`nmin`	the minimum number of clusters in the plot.
`nmax`	the maximum number of clusters in the plot.
`...`	further arguments passed to or from other methods.

See Also

`stability`

Examples

```
data(decathlon)
tree <- hclustvar(X.quanti=decathlon[,1:10])
stab<-stability(tree,B=20)
plot(stab,nmax=7)
```

`plot.clustvar` *Plot scores of variables on synthetic variables.*

Description

Plot scores of variables of cluster k on the synthetic variable associated to the same cluster. For each cluster two plots are displayed: one for numerical variables (if available in the cluster) and one for levels of categorical variables (if available in the cluster).

Usage

```
## S3 method for class clustvar
plot(x, ...)
```

Arguments

`x`	an object of class `clustvar` obtained with `cutreevar` or `kmeansvar`.
`...`	Further arguments to be passed to or from other methods. They are ignored in this function.

Value

`coord.quanti`	coordinates of quantitative variables belonging to cluster k on the synthetic variable associate to the same cluster k
`coord.levels`	coordinates of levels of categorical variables belonging to cluster k on the synthetic variable associate to the same cluster k

Examples

```
data(wine)
X.quanti <- wine[,c(3:29)]
X.quali <- wine[,c(1,2)]
tree <- hclustvar(X.quanti,X.quali)
tree.cut<-cutree(tree,6)

#plot of scores on synthetic variables
res.plot<-plot(tree.cut)
res.plot$coord.quanti
res.plot$coord.levels
```

---

plot.hclustvar *Dendrogram of the hierarchy of variables*

Description

Dendrogram of the hierarchy of variables resulting from `hclustvar` and aggregation levels plot.

Usage

```
## S3 method for class hclustvar
plot(x, type = "tree", which = c(1:2),
  ask = prod(par("mfcol")) < length(which) && dev.interactive(), sub = "",
  ...)
```

Arguments

x	an object of class `hclustvar`.
type	if type="tree" plot of the dendrogram and if type="index" aggregation levels plot.
which	if one of the two plots is required, specify a subset of the numbers 1:2
ask	logical; if TRUE, the user is _ask_ed before each plot.
sub	a sub title for the plot.
...	further arguments passed to or from other methods.

See Also

`hclustvar`

Examples

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

# 2 plots on 1 page
par(mfrow = c(1, 2))
plot(tree)
```

```
 # plot just the dendrogram
plot(tree,which=1)
```

predict.clustvar *Scores of new objects on the synthetic variables of a given partition*

Description

A partition of variables obtained with kmeansvar or with cutreevar is given in input. Each cluster of this partition is associated with a synthetic variable which is a linear combination of the variables of the cluster. The coefficients of these k linear combinations (one for each cluster) are used here to calculate new scores of a objects described in a new dataset (with the same variables). The output is the matrix of the scores of these new objects on the k synthetic variables.

Usage

```
## S3 method for class clustvar
predict(object, X.quanti = NULL, X.quali = NULL, ...)
```

Arguments

`object`	an object of class clustvar
`X.quanti`	numeric matrix of data for the new objects
`X.quali`	a categorical matrix of data for the new objects
`...`	Further arguments to be passed to or from other methods. They are ignored in this function.

Value

Returns the matrix of the scores of the new objects on the k syntetic variables of the k-clusters partition given in input.

Author(s)

Marie Chavent <Marie.Chavent@u-bordeaux.fr>, Vanessa Kuentz, Amaury Labenne, Benoit Liquet, Jerome Saracco

Examples

```
data(wine)
n <- nrow(wine)
sub <- 10:20
X.quanti <- wine[sub,c(3:29)] #learning sample
X.quali <- wine[sub,c(1,2)]
part <-kmeansvar(X.quanti,X.quali,init=5)
X.quanti.t <- wine[-sub,c(3:29)]
X.quali.t <- wine[-sub,c(1,2)]
new <- predict(part,X.quanti.t,X.quali.t)
```

print.clustab *Print a 'clustab' object*

Description

This is a method for the function print for objects of the class clustab.

Usage

```
## S3 method for class clustab
print(x, ...)
```

Arguments

x	An object of class clustab generated by the function stability.
...	Further arguments to be passed to or from other methods. They are ignored in this function.

See Also

stability

print.clustvar *Print a 'clustvar' object*

Description

This is a method for the function print for objects of the class clustvar.

Usage

```
## S3 method for class clustvar
print(x, ...)
```

Arguments

x	An object of class clustvar generated by the functions cutreevar and kmeansvar.
...	Further arguments to be passed to or from other methods. They are ignored in this function.

See Also

cutreevar , kmeansvar

`print.hclustvar` *Print a 'hclustvar' object*

Description

This is a method for the function print for objects of the class `hclustvar`.

Usage

```
## S3 method for class hclustvar
print(x, ...)
```

Arguments

`x`	An object of class `hclustvar` generated by the function `hclustvar`.
`...`	Further arguments to be passed to or from other methods. They are ignored in this function.

See Also

`hclustvar`

`rand` *Rand index between two partitions*

Description

Returns the Rand index, the corrected Rand index or the asymmetrical Rand index. The asymmetrical Rand index (corrected or not) measures the inclusion of a partition 'P' into and partition 'Q' with the number of clusters in 'P' greater than the number of clusters in 'Q'.

Usage

```
rand(P, Q, symmetric = TRUE, adj = TRUE)
```

Arguments

`P`	a factor, e.g., the first partition.
`Q`	a factor, e.g., the second partition.
`symmetric`	a boolean. If 'FALSE' the asymmetrical Rand index is calculated.
`adj`	a boolean. If 'TRUE' the corrected index is calculated.

Author(s)

Marie Chavent <marie.chavent@u-bordeaux.fr>, Vanessa Kuentz, Amaury Labenne, Benoit Liquet, Jerome Saracco

See Also

`stability`

stability *Stability of partitions from a hierarchy of variables*

Description

Evaluates the stability of partitions obtained from a hierarchy of p variables. This hierarchy is performed with hclustvar and the stability of the partitions of 2 to p-1 clusters is evaluated with a bootstrap approach. The boostrap approch is the following: hclustvar is applied to B boostrap samples of the n rows. The partitions of 2 to p-1 clusters obtained from the B bootstrap hierarchies are compared with the partitions from the initial hierarchy . The mean of the corrected Rand indices is plotted according to the number of clusters. This graphical representation helps in the determination of a suitable numbers of clusters.

Usage

```
stability(tree, B = 100, graph = TRUE)
```

Arguments

tree	an object of class hclustvar.
B	the number of bootstrap samples.
graph	boolean, if 'TRUE' a graph is displayed.

Value

matCR	matrix of corrected Rand indices.
meanCR	vector of mean corrected Rand indices.

Author(s)

Marie Chavent <marie.chavent@u-bordeaux.fr>, Vanessa Kuentz, Amaury Labenne, Benoit Liquet, Jerome Saracco

See Also

plot.clustab, hclustvar

Examples

```
data(decathlon)
tree <- hclustvar(X.quanti=decathlon[,1:10])
stab<-stability(tree,B=20)
plot(stab,nmax=7)
dev.new()
boxplot(stab$matCR[,1:7])
```

summary.clustab *Summary of a 'clustab' object*

Description

This is a method for the function summary for objects of the class `clustab`.

Usage

```
## S3 method for class clustab
summary(object, ...)
```

Arguments

`object`	An object of class `clustab` generated by the function `stability`.
`...`	Further arguments passed to or from other methods.

See Also

`stability`

summary.clustvar *Summary of a 'hclustvar' object*

Description

This is a method for the function summary for objects of the class `clustvar`.

Usage

```
## S3 method for class clustvar
summary(object, ...)
```

Arguments

`object`	an object of class `clustvar`.
`...`	further arguments passed to or from other methods.

Value

Returns a list of matrices of squared loadings i.e. for each cluster of variables, the squared loadings on first principal component of PCAmix. For quantitative variables (resp. qualitative), squared loadings are the squared correlations (resp. the correlation ratios) with the first PC (the cluster center). If the partition of variables has been obtained with `kmeansvar` the number of iteration until convergence is also indicated.

See Also

`kmeansvar`, `cutreevar`

Examples

```
data(decathlon)
part<-kmeansvar(X.quanti=decathlon[,1:10],init=5)
summary(part)
```

Index