Package 'ClustVarLV'

October 20, 2018

Title Clustering of Variables Around Latent Variables

Version 1.6.0

Author Evelyne Vigneau [aut, cre],
Mingkun Chen [ctb]

Maintainer Evelyne Vigneau <evelyne.vigneau@oniris-nantes.fr>

Description Functions for the clustering of variables around Latent Variables. Each cluster of variables, which may be defined as a local or directional cluster, is associated with a latent variable. External variables measured on the same observations or/and additional information on the variables can be taken into account. A ``noise" cluster or sparse latent variables can also be defined.

Depends R (>= 3.0.0)

License GPL-2

Encoding UTF-8

LazyData TRUE

Imports Rcpp, doParallel, foreach, parallel, iterators

LinkingTo Rcpp, RcppEigen

Suggests knitr, rmarkdown

VignetteBuilder knitr

RoxygenNote 6.1.0

NeedsCompilation yes

Repository CRAN

Date/Publication 2018-10-20 17:20:03 UTC

R topics documented:

`apples_sh` *apples from southern hemisphere data set*

Description

Sensory characterization and consumers preference for 12 variables of apples

Usage

```
data(apples_sh)
```

Format

A data frame with 12 observations and 2 blocks of variables.

senso 43 sensory attributes

pref hedonic scores given by a panel of 60 consumers

References

Daillant-Spinnler, B, MacFie, H.J.H, Beyts, P.K., Hedderley, D. (1996). Relationships between perceived sensory properties and major preference directions of 12 variaties of apples from the southern hemisphere. Food Quality and Preference, 7(2), 113-126.

Examples

```
data(apples_sh)
names(apples_sh)
apples_sh$senso
apples_sh$pref
```

AUPA_psycho *Psychological eating behavior data set*

Description

The psychological behaviour items in this dataset is a part of French Research Project (AUPALESENS, 2010-2013, http://www2.dijon.inra.fr/aupalesens/) dealing with food behaviour and nutritional status of elderly people. There are 31 psychological items organised into five blocks, each aiming to describe a given behavioural characteristic: emotional eating (E) with six items, external eating (X) with five items, restricted eating (R) with five items, pleasure for food (P) with five items, and self esteem (S) with ten items. Detailed description and analysis of the emotional, external and restricted eating items for this study are available in Bailly, Maitre, Amand, Herve, and Alaphilippe (2012). 559 subjects were considered.

Usage

```
data(AUPA_psycho)
```

Format

A data frame with 559 observations, (row names from 1 to 559) and 31 items. The name of the items refers to the corresponding block (E, X, R, P, S).

References

Bailly N, Maitre I, Amand M, Herve C, Alaphilippe D (2012). The Dutch Eating Behaviour Questionnaire(DEBQ). Assessment of eating behaviour in an aging French population. Appetite, 59(853-858).

Examples

```
X = data(AUPA_psycho)
```

authen_NMR *Authentication data set/ NMR spectra*

Description

Discrimination between authentic and adulterated juices using 1H NMR spectroscopy. 150 samples were prepared by varying the percentage of co-fruit mixed with the fruit juice of interest. The two first characters in the row names represent this percentage. Authentic juice names begin with "00". Samples prepared with the co-fruit alone are identified by "99" (rather than 100). Measurements were done for two spectral ranges. All Spectral values were log-transformed.

Usage

```
data(authen_NMR)
```

Format

150 observations and 2 blocks of variables.

authen_NMR$Xz1 spectral range from 6 to 9 ppm (300 variables)

authen_NMR$Xz2 spectral range from 0.5 to 2.3 ppm (180 variables)

References

Vigneau E, Thomas F (2012). Model calibration and feature selection for orange juice authentication by 1H NMR spectroscopy.Chemometrics and Intelligent Laboratory Systems, 117, 22:30.

Examples

```
data(authen_NMR)
xlab=as.numeric(colnames(authen_NMR$Xz2))
plot(xlab, authen_NMR$Xz2[1,], type="l", xlab="ppm",ylab="", ylim=c(14.8,15.8),
                                          xlim=rev(range(xlab)))
for (i in (1:nrow(authen_NMR$Xz2))) lines(xlab,authen_NMR$Xz2[i,])
```

---

boot_clv *Boostrapping for assessing the stability of a CLV result*

Description

Bootstrapping on the samples is performed. Each boostrapped data matrix is submitted to CLV in order to get partitions from 1 to nmax clusters. For each number of clusters, K, the adjusted Rand Index between actual and the bootstrapped partitions are computed and used in order to assess the stability of the solution into K clusters. Parallel computing is performed in order to save time.

Usage

```
boot_clv(object, B = 100, nmax = NULL)
```

Arguments

object	: result of CLV()
B	: the number of bootstrap to be run (100 by default)
nmax	: maximal size of the partitions to be considered (if NULL, the value of nmax used for the object is used)

Value

matARI	a matrix of the Adjusted Rand Index of size (B x nmax).

See Also

CLV

`CLV` *Hierarchical clustering of variables with consolidation*

Description

Hierarchical Cluster Analysis of a set of variables with consolidation. Directional or local groups may be defined. Each group of variables is associated with a latent component. Moreover, the latent component may be constrained using external information collected on the observations or on the variables.

Usage

```
CLV(X, Xu = NULL, Xr = NULL, method = NULL, sX = TRUE,
  sXr = FALSE, sXu = FALSE, nmax = 20, maxiter = 20)
```

Arguments

`X`	: The matrix of variables to be clustered
`Xu`	: The external variables associated with the columns of X
`Xr`	: The external variables associated with the rows of X
`method`	: The criterion to be use in the cluster analysis. 1 or "directional" : the squared covariance is used as a measure of proximity (directional groups). 2 or "local" : the covariance is used as a measure of proximity (local groups)
`sX`	,TRUE/FALSE : standardization or not of the columns X (TRUE by default) (predefined -> cX = TRUE : column-centering of X)
`sXr`	,TRUE/FALSE : standardization or not of the columns Xr (FALSE by default) (predefined -> cXr = TRUE : column-centering of Xr)
`sXu`	,TRUE/FALSE : standardization or not of the columns Xu (FALSE by default) (predefined -> cXu= FALSE : no centering, Xu considered as a weight matrix)
`nmax`	: maximum number of partitions for which the consolidation will be done (by default nmax=20)
`maxiter`	: maximum number of iterations allowed for the consolidation/partitioning algorithm (by default maxiter=20)

Details

If external variables are used, define either Xr or Xu, but not both. Use the LCLV function when Xr and Xu are simultaneously provided.

Value

`tabres` Results of the clustering algorithm. In each line you find the results of one specific step of the hierarchical clustering.

- Columns 1 and 2 : The numbers of the two groups which are merged
- Column 3 : Name of the new cluster
- Column 4 : The value of the aggregation criterion for the Hierarchical Ascendant Clustering (HAC)
- Column 5 : The value of the clustering criterion for the HAC
- Column 6 : The percentage of the explained initial criterion value (method 1 => % var. expl. by the latent comp.)
- Column 7 : The value of the clustering criterion after consolidation
- Column 8 : The percentage of the explained initial criterion value after consolidation
- Column 9 : The number of iterations in the partitioning algorithm. Remark : A zero in columns 7 to 9 indicates that no consolidation was done

`partition K` contains a list for each number of clusters of the partition, K=2 to nmax with

- clusters : in line 1, the groups membership before consolidation; in line 2 the groups membership after consolidation
- comp : The latent components of the clusters (after consolidation)
- loading : if there are external variables Xr or Xu : The loadings of the external variables (after consolidation)

References

Vigneau E., Qannari E.M. (2003). Clustering of variables around latents components. Comm. Stat, 32(4), 1131-1150.

Vigneau E., Chen M., Qannari E.M. (2015). ClustVarLV: An R Package for the clustering of Variables around Latent Variables. The R Journal, 7(2), 134-148

See Also

CLV_kmeans, LCLV

Examples

```
data(apples_sh)
#directional groups
resclvX <- CLV(X = apples_sh$senso, method = "directional", sX = TRUE)
plot(resclvX,type="dendrogram")
plot(resclvX,type="delta")
#local groups with external variables Xr
resclvYX <- CLV(X = apples_sh$pref, Xr = apples_sh$senso, method = "local", sX = FALSE, sXr = TRUE)
```

CLV_kmeans *K-means algorithm for the clustering of variables*

Description

K-means algorithm for the clustering of variables. Directional or local groups may be defined. Each group of variables is associated with a latent component. Moreover external information collected on the observations or on the variables may be introduced.

Usage

```
CLV_kmeans(X, Xu = NULL, Xr = NULL, method, sX = TRUE, sXr = FALSE,
  sXu = FALSE, clust, iter.max = 20, nstart = 100,
  strategy = "none", rho = 0.3)
```

Arguments

X	The matrix of the variables to be clustered
Xu	The external variables associated with the columns of X
Xr	The external variables associated with the rows of X
method	The criterion to use in the cluster analysis. 1 or "directional" : the squared covariance is used as a measure of proximity (directional groups). 2 or "local" : the covariance is used as a measure of proximity (local groups)
sX	TRUE/FALSE : standardization or not of the columns X (TRUE by default) (predefined -> cX = TRUE : column-centering of X)
sXr	TRUE/FALSE : standardization or not of the columns Xr (FALSE by default) (predefined -> cXr = TRUE : column-centering of Xr)
sXu	TRUE/FALSE : standardization or not of the columns Xu (FALSE by default) (predefined -> cXu= FALSE : no centering, Xu considered as a weight matrix)
clust	: a number i.e. the size of the partition, K, or a vector of INTEGERS i.e. the group membership of each variable in the initial partition (integer between 1 and K)
iter.max	maximal number of iteration for the consolidation (20 by default)
nstart	nb of random initialisations in the case where init is a number (100 by default)
strategy	"none" (by default), or "kplusone" (an additional cluster for the noise variables), or "sparselv" (zero loadings for the noise variables)
rho	a threshold of correlation between 0 and 1 (0.3 by default)

Details

The initalization can be made at random, repetitively, or can be defined by the user.

The parameter "strategy" makes it possible to choose a strategy for setting aside variables that do not fit into the pattern of any cluster.

Value

tabres	The value of the clustering criterion at convergence. The percentage of the explained initial criterion value. The number of iterations in the partitioning algorithm.
clusters	the group's membership
comp	The latent components of the clusters
loading	if there are external variables Xr or Xu : The loadings of the external variables

References

Vigneau E., Qannari E.M. (2003). Clustering of variables around latents components. Comm. Stat, 32(4), 1131-1150.

Vigneau E., Chen M., Qannari E.M. (2015). ClustVarLV: An R Package for the clustering of Variables around Latent Variables. The R Journal, 7(2), 134-148

Vigneau E., Chen M. (2016). Dimensionality reduction by clustering of variables while setting aside atypical variables. Electronic Journal of Applied Statistical Analysis, 9(1), 134-153

See Also

CLV, LCLV

Examples

```
data(apples_sh)
#local groups with external variables Xr
resclvkmYX <- CLV_kmeans(X = apples_sh$pref, Xr = apples_sh$senso,method = "local",
          sX = FALSE, sXr = TRUE, clust = 2, nstart = 20)
```

---

data_biplot *biplot for the dataset*

Description

Loading plot of the variables from a Principal Components Analysis. scores of the observations are surimposed

Usage

```
data_biplot(X, sX = TRUE, axeh = 1, axev = 2, cex.lab = 1)
```

Arguments

X	the data matrix
sX	TRUE/FALSE : standardization or not of the columns X (TRUE by default)
axeh	component number for the horizontal axis
axev	component number for the vertical axis
cex.lab	: magnification to be used for labels (1 by default)

get_comp *latent variables associated with each cluster*

Description

To get the latent variables associated with each cluster.

Usage

```
get_comp(resclv, K = NULL)
```

Arguments

resclv	: result of CLV(), CLV_kmeans() or LCLV()
K	: the number of groups chosen (already defined if CLV_kmeans is used)

Value

comp	the group latent variables (centered, but not standardized) For results of LCLV, two types of latent variables are available : compt : The latent variables of the clusters defined according to the Xr variables, compc : The latent variables of the clusters defined according to the Xu variables

Examples

```
data(apples_sh)
resclvX <- CLV(X = apples_sh$senso, method = "directional", sX = TRUE)
comp4G<-get_comp(resclvX, K = 4)
```

get_load *Loadings of the external variables in each cluster*

Description

To get the loadings of the external variables regarding the latent variable in each cluster Applies only when external variables (Xr, Xu or both) are involved.

Usage

```
get_load(resclv, K = NULL)
```

Arguments

resclv	: result of CLV(), CLV_kmeans() or LCLV()
K	: the number of groups chosen (already defined if CLV_kmeans is used

Value

loading	the loadings of the external variables For results of LCLV, two types of ladings are defined : loading_v : loadings of the external Xr variables, loading_u : loadings of the external Xu variables.

`get_partition` *Clusters of variables.*

Description

To get the clusters of variables. This function returns the group's membership for the p variables. The output can be a vector p x 1 of integers between 1 and K, or a binary matrix of size p x n.

Usage

```
get_partition(resclv, K = NULL, type = "vector")
```

Arguments

resclv	: result of CLV(), CLV_kmeans() or LCLV()
K	: the number of groups chosen (already defined if CLV_kmeans is used)
type	: presented in the form of a "vector" (by default) or a "matrix"

Value

partition	the group's membership for the variables)

Examples

```
data(apples_sh)
resclvX <- CLV(X = apples_sh$senso, method = "directional", sX = TRUE)
parti4G<-get_partition(resclvX, K = 4)
```

get_sparseload *sparse loadings in each cluster when using the "sparselv" strategy*

Description

Applies only on CLV_kmeans output with strategy="sparselv".

Usage

```
get_sparseload(resclv, type = "list")
```

Arguments

resclv	: result of CLV_kmeans()
type	: presented in the form of a "list" (one element by cluster, by default) or a "vector"

Value

sparse_loadings

the loadings of the variables for each latent variables when the "sparselv strategy is used.

imput_clv *Imputation of a data matrix based on CLV results*

Description

For each variable, its missing data will be imputed according to the values of the latent variable of the group in which the variable belong to.

Usage

```
imput_clv(x, X0, K = NULL)
```

Arguments

x	: an object of class clv
X0	: the initial data matrix with missing values (NA)
K	: the number of Latent Variables to be considered, each of them being associated with a group of variables.

Details

It is adviced to use a larger number of latent variables, on the basis of which the imputation will be done, than the suspected 'true' number of groups of variables

Value

X0imput : the imputed data matrix, in the original scale

Ximput : the imputed matrix, centered and scaled according to the pretratment parameters chosen in CLV

LCLV *L-CLV for L-shaped data*

Description

Define clusters of X-variables aroud latent components. In each cluster, two latent components are extracted, the first one is a linear combination of the external information collected for the rows of X and the second one is a linear combination of the external information associated with the columns of X.

Usage

```
LCLV(X, Xr, Xu, ccX = FALSE, sX = TRUE, sXr = FALSE, sXu = FALSE,
  nmax = 20)
```

Arguments

X	The matrix of variables to be clustered
Xr	The external variables associated with the rows of X
Xu	The external variables associated with the columns of X
ccX	TRUE/FALSE : double centering of X (FALSE, by default) If FALSE this implies that cX = TRUE : column-centering of X
sX	TRUE/FALSE : standardization or not of the columns X (TRUE by default)
sXr	TRUE/FALSE : standardization or not of the columns Xr (FALSE by default) (predefined -> cXr = TRUE : column-centering of Xr)
sXu	TRUE/FALSE : standardization or not of the columns Xu (FALSE by default) (predefined -> cXu= FALSE : no centering, Xu considered as a weight matrix)
nmax	maximum number of partitions for which the consolidation will be done (by default nmax=20)

Value

tabres — Results of the clustering algorithm. In each line you find the results of one specific step of the hierarchical clustering.

- Columns 1 and 2 : The numbers of the two groups which are merged
- Column 3 : Name of the new cluster
- Column 4 : The value of the aggregation criterion for the Hierarchical Ascendant Clustering (HAC)
- Column 5 : The value of the clustering criterion for the HAC

- Column 6 : The percentage of the explained initial criterion value
- Column 7 : The value of the clustering criterion after consolidation
- Column 8 : The percentage of the explained initial criterion value after consolidation
- Column 9 : number of iterations in the partitioning algorithm.
  Remark: A zero in columns 7 to 9 indicates that no consolidation was done

`partition K` a list for each number of clusters of the partition, K=2 to nmax with

- clusters : in line 1, the groups membership before consolidation; in line 2 the groups membership after consolidation
- compt : The latent components of the clusters (after consolidation) defined according to the Xr variables
- compc : The latent components of the clusters (after consolidation) defined according to the Xu variables
- loading_v : loadings of the external Xr variables (after consolidation)
- loading_u : loadings of the external Xu variables (after consolidation)

References

Vigneau E., Qannari E.M. (2003). Clustering of variables around latents components. Comm. Stat, 32(4), 1131-1150.

Vigneau, E., Charles, M.,& Chen, M. (2014). External preference segmentation with additional information on consumers: A case study on apples. Food Quality and Preference, 32, 83-92.

Vigneau E., Chen M., Qannari E.M. (2015). ClustVarLV: An R Package for the clustering of Variables around Latent Variables. The R Journal, 7(2), 134-148

`lm_CLV` *linear model based on CLV : prediction of a response variable, y, based on clusters of explanatory variables, X.*

Description

boosted-liked procedure for identifying groups of explanatory variables (and the associated latent variable) specifically correlated with the response variable, y. Directional groups are considered. Discarding spurious variables is allowed using the strategy and rho parameters.

Usage

```
lm_CLV(X, y, method = "directional", sX = sX, nbiter = 100,
  strategy = "none", rho = 0.3, shrinkp = 1, validation = FALSE,
  id.test = NULL)
```

Arguments

X	: The matrix of the explanatory variables, to be clustered
y	: The response variable (numeric)
method	: The criterion to be use in the cluster analysis. 1 or "directional" : the squared covariance is used as a measure of proximity (directional groups). 2 or "local" : the covariance is used as a measure of proximity (local groups)
sX	: TRUE/FALSE, i.e. standardization or not of the columns X (TRUE by default) (predefined -> cX = TRUE : column-centering of X)
nbiter	: maximum number of steps (by default nbiter=100)
strategy	: "none" (by default), or "kplusone" (an additional cluster for the noise variables), or "sparselv" (zero loadings for the noise variables)
rho	: a threshold of correlation between 0 and 1 (used for kplusone or sparselv strategies, 0.3 by default)
shrinkp	: shrinkage paramater used in the boosting (1 by default)
validation	TRUE/FALSE i.e. using a test set or not. By default no validation
id.test	: if validation==TRUE, the number of the observations used as test set

Value

Group	a list of the groups of variables X iteratively extracted.
Comp	a list of the latent variables associated with the groups of X variables extracted.
Load	a list for the loadings of the X variables according to the associated latent variable.
Beta	a list for the coefficients associated to the scaled predictors (X must to be scaled accordind to sX) regarding y .

See Also

CLV, CLV_kmeans

---

plot.clv *Graphical representation of the CLV clustering stages*

Description

This function plots either the CLV dendrogram or the variations of the consolidated CLV criterion.

Usage

```
## S3 method for class 'clv'
plot(x, type = "dendrogram", cex = 0.8, ...)
```

Arguments

x	: an object of class clv
type	: What to plot. "dendrogram" : the dendrogram of the hierchical clustering algorithm, "delta" : a barplot showing the variation of the clustering criterium after consolidation.
cex	: Character expansion for labels.
...	further arguments passed to or from other methods

See Also

CLV

---

plot.lclv *Graphical representation of the LCLV clustering stages*

Description

This function plots either the CLV dendrogram or the variations of the consolidated CLV criterion.

Usage

```
## S3 method for class 'lclv'
plot(x, type = "dendrogram", cex = 0.8, ...)
```

Arguments

x	: an object of class lclv
type	: What to plot. "dendrogram" : the dendrogram of the hierchical clustering algorithm, "delta" : a barplot showing the variation of the clustering criterium after consolidation.
cex	: Character expansion for labels.
...	further arguments passed to or from other methods

See Also

LCLV

plot_var *Representation of the variables and their group membership*

Description

Loading plot of the variables from a Principal Components Analysis. The group membership of the variables is superimposed.

Usage

```
plot_var(resclv, K = NULL, axeh = 1, axev = 2, label = FALSE,
  cex.lab = 1, v_colors = NULL, v_symbol = FALSE, beside = FALSE)
```

Arguments

resclv	results of CLV(), CLV_kmeans() or LCLV()
K	the number of groups in the partition (already defined if CLV_kmeans is used)
axeh	component number for the horizontal axis
axev	component number for the vertical axis
label	= TRUE :the column names in X are used as labels / = FALSE: no labels (by default)
cex.lab	: magnification to be used for labels (1 by default)
v_colors	default NULL. If missing colors are given, by default
v_symbol	=TRUE : symbols are given isntead of colors for the identification of the groups/ =FALSE: no symbol (by default).
beside	=TRUE : a plot per cluster of variables, side-by-side/ =FALSE :an unique plot with all the variables with the identification of their group membership (by default).

Examples

```
data(apples_sh)
resclvX <- CLV(X = apples_sh$senso, method = 1, sX = TRUE)
plot_var(resclvX, K = 4, axeh = 1, axev = 2)
```

print.clv *Print the CLV results*

Description

Print the CLV results

Usage

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

Arguments

x	an object of class clv
...	further arguments passed to or from other methods

See Also

CLV

print.lclv *Print the LCLV results*

Description

Print the LCLV results

Usage

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

Arguments

x	an object of class lclv
...	further arguments passed to or from other methods

See Also

LCLV

stand_quali *Standardization of the qualitative variables*

Description

pretreatment of qualitative variables

Usage

```
stand_quali(X.quali, metric = "chisq")
```

Arguments

X.quali	: a factor or a data frame with several factors
metric	: the metric to be used, i.e. each category is weighted by the inverse of the square-root of its relative frequency

Value

Xdisj.sd : a standardized matrix with as many columns as categories associated with the qualitative variables.

summary.clv *summary and description of the clusters of variables*

Description

This function provides the list of the variables within each group and complementary informations. Users will be asked to specify the number of clusters,

Usage

```
## S3 method for class 'clv'
summary(object, K = NULL, ...)
```

Arguments

object	: result of CLV() or CLV_kmeans()
K	: the number of clusters (unless if CLV_kmeans was used)
...	further arguments passed to or from other methods

Details

The ouputs include :

- the size of the groups,
- the list of the variables within each group. FFor each cluster, the correlation of the each variable with its group latent component and the correlation with the next neighbouring group latent component are given.
- the proportion of the variance within each group explained by its latent variable,
- the proportion of the whole dataset account by the group latent variables
- the matrix of correlation between the latent variables.

Index