Package 'CluMix'

September 19, 2017

Type Package

Title Clustering and Visualization of Mixed-Type Data

Version 2.1

Date 2017-09-19

Author M. Hummel, D. Edelmann, A. Kopp-Schneider

Maintainer Manuela Hummel <m.hummel@dkfz.de>

Description Provides utilities for clustering subjects and variables of mixed data types. Similarities between subjects are measured by Gower's general similarity coefficient with an extension of Podani for ordinal variables. Similarities between variables can be assessed i) by combination of appropriate measures of association for different pairs of data types or ii) based on distance correlation. Alternatively, variables can also be clustered by the 'ClustOfVar' approach. The main feature of the package is the generation of a mixed-data heatmap. For visualizing similarities between either subjects or variables, a heatmap of the corresponding distance matrix can be drawn. Associations between variables can be explored by a 'confounder plot', which allows visual detection of possible confounding, collinear, or surrogate factors for some variables of primary interest. Distance matrices and dendrograms for subjects and variables can be derived and used for further visualizations and applications. This work was supported by BMBF grant 01ZX1609B, Germany.

License GPL (>= 2)

Depends

Imports ClustOfVar, Hmisc, DescTools, extracat, marray, FD, gplots, Matrix, Biobase

Suggests devtools, dendextend

NeedsCompilation no

Repository CRAN

Date/Publication 2017-09-19 09:26:31 UTC

R topics documented:

`CluMix-package` *Clustering and Visualization of Mixed-Type Data*

Description

Provides utilities for clustering subjects and variables of mixed data types. Similarities between subjects are measured by Gower's general similarity coefficient with an extension of Podani for ordinal variables. Similarities between variables can be assessed i) by combination of appropriate measures of association for different pairs of data types or ii) based on distance correlation. Alternatively, variables can also be clustered by the 'ClustOfVar' approach. The main feature of the package is the generation of a mixed-data heatmap. For visualizing similarities between either subjects or variables, a heatmap of the corresponding distance matrix can be drawn. Associations between variables can be explored by a 'confounder plot', which allows visual detection of possible confounding, collinear, or surrogate factors for some variables of primary interest. Distance matrices and dendrograms for subjects and variables can be derived and used for further visualizations and applications. This work was supported by BMBF grant 01ZX1609B, Germany.

Details

The DESCRIPTION file: This package was not yet installed at build time.

Index of help topics:

```
CluMix-package          Clustering and Visualization of Mixed-Type Data
association             Function to calculate a measure for association
                        between two variables on arbitrary scales
confounderPlot          Confounder Plot
dendro.subjects         Subjects dendrogram
dendro.variables        Variables dendrogram
dist.subjects           Distance matrix for subjects
dist.variables          Distance matrix for variables
distmap                 Display similarity matrix
mix.heatmap             Heatmap for data with variables of mixed types
mixdata                 Small example dataset with variables of
                        different types
```

```
similarity.subjects     Similarity matrix for subjects
similarity.variables    Similarity matrix for variables
```

The main function `mix.heatmap` of the package generates a mixed-data heatmap. For visualizing similarities between either subjects or variables, a heatmap of the corresponding distance matrix can be drawn (`distmap`). Associations between variables can be explored by the `confounderPlot`, which allows visual detection of possible confounding, collinear, or surrogate factors for some variables of primary interest. Distance matrices and dendrograms for subjects and variables can be derived by functions `dist.subjects`, `dist.variables`, `dendro.subjects`, and `dendro.variables`. Clustering subjects is based on Gower's general similarity coefficient. Variables can be clustered by i) combination of association measures, ii) distance correlation, iii) the ClustOfVar approach.

Author(s)

M. Hummel, D. Edelmann, A. Kopp-Schneider

Maintainer: Manuela Hummel <m.hummel@dkfz.de>

References

Hummel M, Edelmann D, Kopp-Schneider A. Clustering of samples and variables with mixed-type data. Submitted.

Gower J (1971). A general coefficient of similarity and some of its properties. Biometrics, 27:857-871.

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

Szekely GJ, Rizzo ML, Bakirov NK (2007). Measuring and testing dependence by correlation of distances. The Annals of Statistics, 35.6:2769-2794.

Lyons R (2013). Distance covariance in metric spaces. The Annals of Probability, 41.5:3284-3305.

See Also

`mix.heatmap`

Examples

```
data(mixdata)

mix.heatmap(mixdata, rowmar=7)
```

`association` *Function to calculate a measure for association between two variables on arbitrary scales*

Description

Similarities between variables used for clustering are calculated by this function. For each combination of different data types, an appropriate measure for association is chosen, see details.

Usage

```
association(x, y)
```

Arguments

`x`	vector of class `numeric`, `factor`, `ordered`, or `logical`
`y`	vector of class `numeric`, `factor`, `ordered`, or `logical`

Details

The following association measures for respective types of variables are chosen:

- quantitative vs quantitative/ordinal: absolute Spearman correlation coefficient
- ordinal vs ordinal: absolute Goodman and Kruskal's gamma coefficient (Goodman and Kruskal, 1954)
- quantitative/ordinal vs binary: absolute Goodman and Kruskal's gamma coefficient
- quantitative vs categorical (>2 categories): The categories of the categorical variable are reordered with respect to average ranks of the quantitative variable within those categories. Then absolute Spearman correlation coefficient is calculated as if it was an ordered factor. To avoid over-optimism, the reordering is only applied if a Kruskal-Wallis pre-test of association yields a significant result ($p<0.05$).
- ordinal vs categorical (>2 categories): as for 'quantitative vs categorical', but instead of Spearman correlation the absolute Goodman and Kruskal's gamma coefficient is calculated
- categorical vs categorical: Also in this case the reordering strategy is applied by "diagonalizing" the cross-table between the two factors (see `optile` from the `extracat` package). Association is then measured by absolute Goodman and Kruskal's gamma coefficient. To avoid over-optimism, the reordering is only applied if a chi-square pre-test of association yields a significant result ($p<0.05$).

Value

Estimated value of association between `x` and `y`

Author(s)

Manuela Hummel

References

Goodman LA and Kruskal WH (1954). Measures of association for cross classifications. Journal of the American Statistical Association, 49:732-764.

See Also

similarity.variables, dist.variables,

Examples

```
x <- rnorm(100)
y <- as.factor(sample(1:3, 100, replace=TRUE))
association(x, y)
```

---

confounderPlot *Confounder Plot*

Description

Plots similarities of all variables to an outcome variable against similarities of all variables to a predictor of interest

Usage

```
confounderPlot(data, S, x, y, labels, method = c("associationMeasures", "distcor"),
returnS = FALSE, plotLegend = TRUE, col, pch, font, cex.text, xlim, ylim, ...)
```

Arguments

data	data frame with variables of interest
S	similarity matrix; if missing it will be calculated from data by similarity.variables
x	name of the predictor variable (as used in data and S) of main interest, for which confounders / collinearities shall be detected
y	name of the outcome variable (as used in data and S)
labels	variable names used for plotting; have to be in corresponding order with columns of data; if missing, names of data are used
method	method to calculate similarities: combination of association measures ('associationMeasures') or distance correlation ('distcor')
returnS	shall similarity matrix be returned?
plotLegend	shall (default) legend be shown, indicating categorical and continuous variables
col	symbol and label color; by default categorical variables are shown in purple, continuous variables in black
pch	plotting symbol, default 16
font	font of plotted labels; by default names of variables x and y are shown in bold

`cex.text`	size of plotted labels
`xlim, ylim`	axis limits
`...`	graphical parameters passed to `plot`

Details

The similarities of all variables in a dataset with two variables of special interest (i.e. predictor and outcome of a regression model) are simultaneously visualized in a scatter plot, where the x-axis shows similarities to the predictor and the y-axis similarities to the outcome. The height of the predictor variable's point indicates its association with the outcome and hence its predicting ability. Variables in the upper right part are potential confounders for which prediction model should be adjusted, or collinear variables that should be removed. Variables in the lower right part are strongly related to the predictor, but not associated with the outcome. Variables very close to the outcome variable's point are potential surrogate outcomes.

Value

Scatterplot of variable similarities. Chosen predictor and outcome variables are highlighted in bold. Categorical/quantitative variables are shown in purple/black by default.

Author(s)

Manuela Hummel

See Also

`similarity.variables`

Examples

```
data(mixdata)

confounderPlot(mixdata, x="X2.quant", y="X1.cat")
```

`dendro.subjects` *Subjects dendrogram*

Description

Get dendrogram for subjects (observations) based on variables of mixed data types

Usage

```
dendro.subjects(data, weights, linkage="ward.D2")
```

Arguments

data	data frame
weights	optional vector of weights for variables in data
linkage	agglomeration method used for hierarchical clustering; corresponds to parameter method of hclust

Details

Distances between subjects are based on Gower's general similarity coefficient with an extension of Podani for ordinal variables, see gowdis. In the case that all variables are quantitative, Euclidean distances are used. Then a dendrogram is derived by standard hierarchical clustering (hclust with agglomeration method = "ward.D2" by default).

Value

An object of class dendrogram

Author(s)

Manuela Hummel

References

Gower J (1971). A general coefficient of similarity and some of its properties. Biometrics, 27:857-871.

Podani J (1999). Extending Gower's general coefficient of similarity to ordinal characters. Taxon, 48(2):331-340.

See Also

dendro.variables, dist.subjects, mix.heatmap

Examples

```
data(mixdata)

dend <- dendro.subjects(mixdata)
plot(dend)

## example with weights
w <- rep(1:2, each=5)
dend <- dendro.subjects(mixdata, weights=w)
plot(dend)
```

`dendro.variables` *Variables dendrogram*

Description

Get dendrogram for variables of mixed types

Usage

```
dendro.variables(data, method = c("associationMeasures", "distcor", "ClustOfVar"),
  linkage="ward.D2", associationFun = association, check.psd = TRUE)
```

Arguments

`data`	data frame with variables of interest
`method`	If `"associationMeasures"`, similarities between variables are assessed by combination of appropriate measures of association for different pairs of data types. If `"distcor"`, distances between variables are calculated based on distance correlation. In both cases, then a dendrogram is derived by standard hierarchical clustering (`hclust`). If `"ClustOfVar"`, variables are clustered by `hclustvar` from the `ClustOfVar` package.
`linkage`	agglomeration method used for hierarchical clustering when `dist.variables.method = "associationM` corresponds to parameter `method` of `hclust`
`associationFun`	By default, appropriate association measures are chosen for each pair of variables, see `association` for details. But the user can also define a function that for any two variables calculates a similarity measure. Ignored if `dist.variables.method = "ClustOfVa`
`check.psd`	If `TRUE`, it is checked if the variable's similarity matrix S is positive semi-definite (p.s.d.), and if not it is transformed to a p.s.d. one by `nearPD`, see `dist.variables` for details. Ignored if `dist.variables.method = "ClustOfVar"`

Details

Clustering of variables can either be done i) similarity-based using measures of association, ii) similarity-based using distance correlation, or iii) by the ClustOfVar approach, which uses principal components analysis for mixed data.

Value

An object of class `dendrogram`

Author(s)

Manuela Hummel

References

Hummel M, Edelmann D, Kopp-Schneider A. Clustering of samples and variables with mixed-type data. Submitted.

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

See Also

association, similarity.variables, dist.variables, dendro.subjects, mix.heatmap, hclustvar

Examples

```
data(mixdata)

dend1 <- dendro.variables(mixdata, method="associationMeasures")
plot(dend1)

dend2 <- dendro.variables(mixdata, method="distcor")
plot(dend2)

dend3 <- dendro.variables(mixdata, method="ClustOfVar")
plot(dend3)
```

---

dist.subjects *Distance matrix for subjects*

Description

Get distance matrix for subjects (observations) based on variables of mixed data types

Usage

```
dist.subjects(data, weights, alwaysGower = FALSE)
```

Arguments

data	data frame
weights	optional vector of weights for variables in data
alwaysGower	controls the way distances are calculated in case of exclusively continuous data; if FALSE (default), Euclidean distances, if TRUE Gower's distances

Details

Distances between subjects are based on Gower's general similarity coefficient with an extension of Podani for ordinal variables, see gowdis. In the case that all variables are quantitative, either Euclidean distances or still Gower's distances can be used.

Value

An object of class dist

Author(s)

Manuela Hummel

References

Gower J (1971). A general coefficient of similarity and some of its properties. Biometrics, 27:857-871.

Podani J (1999). Extending Gower's general coefficient of similarity to ordinal characters. Taxon, 48(2):331-340.

See Also

dendro.subjects, similarity.subjects, dist.variables, mix.heatmap

Examples

```
data(mixdata)

D <- dist.subjects(mixdata)

## example with weights
w <- rep(1:2, each=5)
D <- dist.subjects(mixdata, weights=w)
```

---

dist.variables *Distance matrix for variables*

Description

Get distance matrix for variables of mixed types

Usage

```
dist.variables(data, method = c("associationMeasures", "distcor"),
associationFun = association, check.psd = TRUE)
```

Arguments

data	data frame with variables of interest
method	method to calculate distances: combination of association measures ('associationMeasures') or distance correlation ('distcor')

`associationFun` only applies if `method = 'associationMeasures'`: by default, appropriate association measures are chosen for each pair of variables, see `association` for details. But the user can also define a function that for any two variables calculates a similarity measure.

`check.psd` only applies if `method = 'associationMeasures'`: if `TRUE`, it is checked if the variable's similarity matrix S is positive semi-definite (p.s.d.), and if not it is transformed to a p.s.d. one by `nearPD`.

Details

A distance matrix for variables can be derived by combining different measures of association or by a distance correlation approach. For the association measure approach, for each pair of variables, similarity coefficients s_ij are calculated, see `association` for details. Distances are then calculated as d_ij = sqrt(1 - s_ij). If the similarity matrix is (made) positive semi-definite, those distances have metric properties (Gower, 1971), which means for instance that the triangular inequality holds. The distance correlation approach uses generalized distance correlation based on Gower's similarity coefficient between sample elements. The distance is then defined by 1 minus the square root of the distance correlation matrix.

Value

An object of class `dist`

Author(s)

Manuela Hummel, Dominic Edelmann

References

Hummel M, Edelmann D, Kopp-Schneider A. Clustering of samples and variables with mixed-type data. Submitted.

Gower J (1971). A general coefficient of similarity and some of its properties. Biometrics, 27:857-871.

Szekely GJ, Rizzo ML, Bakirov NK (2007). Measuring and testing dependence by correlation of distances. The Annals of Statistics, 35.6:2769-2794.

Lyons R (2013). Distance covariance in metric spaces. The Annals of Probability, 41.5:3284-3305.

See Also

`association`, `similarity.variables`, `dendro.variables`, `dist.subjects`, `mix.heatmap`

Examples

```
data(mixdata)

D1 <- dist.variables(mixdata, method="association")
D2 <- dist.variables(mixdata, method="distcor")
```

distmap *Display similarity matrix*

Description

Calculates and visualizes a similarity matrix for subjects or variables in an image plot

Usage

```
distmap(data, what = c("subjects", "variables"), variables.method =
c("associationMeasures", "distcor"), varweights, linkage = "ward.D2",
reorderdend, col, ...)
```

Arguments

data	data.frame with original data or similarity matrix
what	Shall similarity matrix of subjects or variables be visualized?; ignored if data is a similarity matrix
variables.method	method to calculate similarities if what = "variables": combination of association measures ("associationMeasures") or distance correlation ("distcor")
varweights	optional vector of variable weights, used for calculating Gower's distances between subjects; ignored if what = "associationMeasures"
linkage	agglomeration method used for hierarchical clustering; corresponds to parameter method of hclust
reorderdend	optional numeric values for reordering the dendrogram (maintaining the constraints on the dendrogram), see wts option of reorder.dendrogram
col	Color palette; defaults to blue-scale palette, where darker blue indicates higher similarity
...	graphical parameters passed to heatmap.2

Details

If data is a data.frame, the similarity matrix is calculated for subjects (if what = "subjects") or variables (if what = "variables"). Similarities for subjects are calculated by similarity.subjects. \ Similarities for variables are derived by similarity.variables. Alternatively, data can also be a previously calculated similarity matrix.

Value

Image plots and dendrograms

Author(s)

Manuela Hummel

See Also

similarity.variables, dist.variables, similarity.subjects, dist.subjects, mix.heatmap

Examples

```
data(mixdata)

## subjects
distmap(mixdata, what="subjects")

# example with variable weights
w <- rep(1:2, each=5)
distmap(mixdata, what="subjects", varweights=w)

## variables
distmap(mixdata, what="variables", method="association")
distmap(mixdata, what="variables", method="distcor")
```

---

mix.heatmap *Heatmap for data with variables of mixed types*

Description

Produces a heatmap visualizing all samples and variables of a dataset. Both samples and variables are clustered using methods suitable for mixed-type data. Different types of variables are indicated by different color schemes.

Usage

```
mix.heatmap(data, D.subjects, D.variables, dend.subjects, dend.variables, varweights,
dist.variables.method = c("associationMeasures", "distcor", "ClustOfVar"),
linkage="ward.D2", associationFun = association, rowlab, rowmar = 3, lab.cex = 1.5,
ColSideColors, RowSideColors,
col.cont = marray::maPalette(low = "lightblue", high = "darkblue", k = 50),
cont.fixed.range = FALSE, cont.range,
col.ord = list(low = "lightgreen", high = "darkgreen"),
col.cat = c("indianred1","darkred","orangered","orange","palevioletred1",
          "violetred4","red3","indianred4"),
legend.colbar, legend.rowbar, legend.mat = FALSE, legend.cex = 1, legend.srt = 0)
```

Arguments

data	data frame where columns are variables (of different data types) and rows are observations (subjects, samples)
D.subjects	A previously calculated distance matrix (class dissimilarity) for subjects can be given. If missing, it is calculated by dist.subjects. If set to NULL, no clustering is done and original order in data will be preserved.

D.variables A previously calculated distance matrix (of class dissimilarity) for variables can be given. If missing, it is calculated by dist.variables. If set to NULL, no clustering is done and original order in data will be preserved.

dend.subjects A dendrogram for subjects can be given; then no distances between subjects will be calculated and D.subjects will be ignored.

dend.variables A dendrogram for variables can be given; then no distances between variables will be calculated and D.variables will be ignored.

varweights optional vector of variable weights, used for calculating Gower's distances between subjects

dist.variables.method
If "associationMeasures", similarities between variables are assessed by combination of appropriate measures of association for different pairs of data types. If "distcor", distances between variables are calculated based on distance correlation. In both cases, then a dendrogram is derived by standard hierarchical clustering (hclust). If "ClustOfVar", variables are clustered by hclustvar from the ClustOfVar package.

linkage agglomeration method used for hierarchical clustering; corresponds to parameter method of hclust

associationFun By default, appropriate association measures are chosen for each pair of variables, see association for details. But the user can also define a function that for any two variables calculates a similarity measure. Ignored if dist.variables.method = "ClustOfVa
or "distcor"

rowlab row (variable) labels; if missing, column names of data are used

rowmar margin for row (variable) labels

lab.cex size of row (variable) labels

ColSideColors vector of length nrow(data) specifying colors for a color bar added on top of the heatmap

RowSideColors vector of length ncol(data) specifying colors for a color bar added to the left of the heatmap

col.cont color palette for continuous variables; defaults to red-blue color palette

cont.fixed.range
If FALSE, color range of each continuous variable is defined by respective individual variable's range. If TRUE, all continuous variables are assumed to have similar range and hence shall have the same color range; "extreme colors" then correspond to extreme values over all continuous variables and are applied to all of them equally. In any case, in order to prevent outlier values to dominate the color scale, "extreme colors" are restricted to 2.5% and 97.5% quantiles. Defaults to FALSE

cont.range if cont.fixed.range=TRUE, extreme value limits for coloring continuous variables can be specified; if missing, extreme values are taken from the data; ignored if cont.fixed.range=FALSE

col.ord List with names of colors for the lowest and highest categories of ordinal variables. A color palette will be created correspondingly based on the number of categories. Defaults to a green color palette

col.cat	vector of colors for categorical variables
legend.colbar	class labels for subject groups defined by ColSideColors
legend.rowbar	class labels for variable groups defined by RowSideColors
legend.mat	shall legend matrix for heatmap be shown?
legend.cex	size of legend text
legend.srt	legend matrix label string rotation in degrees; i.e. legend.srt = 90 produces vertical labels

Details

If no dendrograms or distance matrices are given, subjects and/or samples are clustered with methods for mixed-type data. Similarities between subjects are measured by Gower's general similarity coefficient with an extension of Podani for ordinal variables, see gowdis. Similarities between variables can be assessed by combination of appropriate measures of association for different pairs of data types, see association, or based on distance correlation. Then standard hierarchical clustering with by default Ward's minimum variance method is applied. Alternatively, variables can also be clustered by the ClustOfVar approach.

Variables are shown as rows of the heatmap, samples as columns.

Value

A mixed-data heatmap with dendrograms and annotation

Note

The heatmap is currently limited to 200 variables = columns of data = heatmap rows.

Author(s)

Manuela Hummel, m.hummel@dkfz.de

References

Hummel M, Edelmann D, Kopp-Schneider A. Clustering of samples and variables with mixed-type data. Submitted.

Gower J (1971). A general coefficient of similarity and some of its properties. Biometrics, 27:857-871.

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

Szekely GJ, Rizzo ML, Bakirov NK (2007). Measuring and testing dependence by correlation of distances. The Annals of Statistics, 35.6:2769-2794.

Lyons R (2013). Distance covariance in metric spaces. The Annals of Probability, 41.5:3284-3305.

See Also

dist.variables, dist.subjects, dendro.variables, dendro.subjects, distmap

Examples

```
data(mixdata)

mix.heatmap(mixdata, rowmar=7, legend.mat=TRUE)

## with distance correlation
mix.heatmap(mixdata, dist.variables.method="distcor", rowmar=7, legend.mat=TRUE)

## with (random) color bars
colbar <- rep(5:7, nrow(mixdata))
rowbar <- rep(c("darkorange","grey"), ncol(mixdata))
mix.heatmap(mixdata, ColSideColors=colbar, RowSideColors=rowbar,
legend.colbar=c("1","2","3"), legend.rowbar=c("a","b"), rowmar=7)

## example with variable weights
w <- rep(1:2, each=5)
mix.heatmap(mixdata, varweights=w, rowmar=7)
```

`mixdata` *Small example dataset with variables of different types*

Description

Simulated dataset with quantitative, ordinal and categorical variables. Some variables are correlated and some are associated to sample groups.

Usage

```
data(mixdata)
```

Format

`mixdata` is a data frame with 40 samples (rows) and 10 variables (columns). The variable names indicate their type, i.e. '.quant' (quantitative), '.ord' (ordinal), '.cat' (categorical).

Examples

```
data(mixdata)
str(mixdata)
```

`similarity.subjects` *Similarity matrix for subjects*

Description

Get similarity matrix for subjects (observations) based on variables of mixed data types

Usage

```
similarity.subjects(data, weights)
```

Arguments

`data`	data frame
`weights`	optional vector of weights for variables in `data`

Details

Distances d.ij between subjects are calculated based on Gower's general similarity coefficient with an extension of Podani for ordinal variables, see `gowdis`. In the case that all variables are quantitative, Euclidean distances are used. Similarities s.ij are calculated as s.ij = 1 - d.ij.

Value

Matrix of similarity values for each pair of subjects

Author(s)

Manuela Hummel

References

Gower J (1971). A general coefficient of similarity and some of its properties. Biometrics, 27:857-871.

Podani J (1999). Extending Gower's general coefficient of similarity to ordinal characters. Taxon, 48(2):331-340.

See Also

`dendro.subjects`, `dist.subjects`, `mix.heatmap`

Examples

```
data(mixdata)

S <- similarity.subjects(mixdata)

## example with weights
w <- rep(1:2, each=5)
S <- similarity.subjects(mixdata, weights=w)
```

`similarity.variables` *Similarity matrix for variables*

Description

Get similarity matrix for variables of mixed types

Usage

```
similarity.variables(data, method = c("associationMeasures", "distcor"),
associationFun = association, check.psd = TRUE, make.psd = TRUE)
```

Arguments

`data`	data frame with variables of interest
`method`	method to calculate distances: combination of association measures (`'associationMeasures'`) or distance correlation (`'distcor'`)
`associationFun`	only applies if `method = 'associationMeasures'`: appropriate association measures are chosen for each pair of variables, see `association` for details. But the user can also define a function that for any two variables calculates a similarity measure.
`check.psd`	only applies if `method = 'associationMeasures'`: if `TRUE`, it is checked if the variable's similarity matrix S is positive semi-definite (p.s.d.), and if not it is transformed to a p.s.d. one by `nearPD`.
`make.psd`	only applies if `method = 'associationMeasures'`: if `TRUE`, and if the similarity matrix is not positive semi-definite, it is transformed to a p.s.d. one by `nearPD`. Ignored if `check.psd = FALSE`

Details

A similarity matrix for variables can be derived by combining different measures of association or by a distance correlation approach. For the association measure approach, for each pair of variables, similarity coefficients s_ij are calculated, see `association` for details. If the similarity matrix is (made) positive semi-definite, distances d_ij = sqrt(1 - s_ij) have metric properties (Gower, 1971), which means for instance that the triangular inequality holds. The distance correlation approach uses generalized distance correlation based on Gower's similarity coefficient between sample elements.

Value

Matrix of similarity values for each pair of variables

Author(s)

Manuela Hummel, Dominic Edelmann

References

Hummel M, Edelmann D, Kopp-Schneider A. Clustering of samples and variables with mixed-type data. Submitted.

Gower J (1971). A general coefficient of similarity and some of its properties. Biometrics, 27:857-871.

Szekely GJ, Rizzo ML, Bakirov NK (2007). Measuring and testing dependence by correlation of distances. The Annals of Statistics, 35.6:2769-2794.

Lyons R (2013). Distance covariance in metric spaces. The Annals of Probability, 41.5:3284-3305.

See Also

association, dist.variables, dendro.variables, dist.subjects, mix.heatmap

Examples

```
data(mixdata)

S1 <- similarity.variables(mixdata)
S2 <- similarity.variables(mixdata, method="distcor")
```

Index