Package 'clustrd'

December 6, 2016

Type Package

Title Methods for Joint Dimension Reduction and Clustering

Description A class of methods that combine dimension reduction and clustering of continuous or categorical data. For continuous data, the package contains implementations of factorial K-means (Vichi and Kiers 2001; <DOI:10.1016/S0167-9473(00)00064-5>) and reduced K-means (De Soete and Carroll 1994; <DOI:10.1007/978-3-642-51175-2_24>); both methods that combine principal component analysis with K-means clustering. For categorical data, the package provides MCA K-means (Hwang, Dillon and Takane 2006; <DOI:10.1007/s11336-004-1173-x>), i-FCB (Iodice D'Enza and Palumbo 2013, <DOI:10.1007/s00180-012-0329-x>) and Cluster Correspondence Analysis (van de Velden, Iodice D'Enza and Palumbo 2016; <DOI:10.1007/s11336-016-9514-0>), which combine multiple correspondence analysis with K-means.

Version 1.1.0

Date 2016-12-06

Author Alfonso Iodice D'Enza [aut], Angelos Markos [aut, cre], Michel van de Velden [ctb]

Maintainer Angelos Markos <amarkos@gmail.com>

Depends ggplot2, dummies, grid

Imports corpcor, GGally, fpc, cluster, dplyr, plyr, ggrepel

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2016-12-06 18:27:21

R topics documented:

`clusmca` *Joint dimension reduction and clustering of categorical data.*

Description

This function implements MCA K-means (Hwang, Dillon and Takane, 2006), i-FCB (Iodice D' Enza and Palumbo, 2013) and Cluster Correspondence Analysis (van de Velden, Iodice D' Enza and Palumbo, 2016). The methods combine variants of Correspondence Analysis for dimension reduction with K-means for clustering.

Usage

```
clusmca(data, nclus, ndim, method = "clusCA", alpha = .5,
nstart = 10, smartStart = NULL, gamma = TRUE, seed = 1234)
```

Arguments

`data`	Categorical dataset
`nclus`	Number of clusters
`ndim`	Dimensionality of the solution
`method`	Specifies the method. Options are MCAk for MCA K-means, iFCB for Iterative Factorial Clustering of Binary variables and clusCA for Cluster Correspondence Analysis (default = `"clusCA"`)
`alpha`	Non-negative scalar to adjust for the relative importance of of MCA and K-means in the solution (default = .5). Works only in combination with `method = "MCAk"`
`nstart`	Number of random starts
`smartStart`	If `NULL` then a random cluster membership vector is generated. Alternatively, a cluster membership vector can be provided as a starting solution
`gamma`	Scaling parameter that leads to a similar spread in the object and attribute points (default = `TRUE`)
`seed`	An integer that is used as argument by `set.seed()` for offsetting the random number generator when `smartStart = NULL`. The default value is 1234

Value

obscoord	Object scores
attcoord	Varable scores
centroid	Cluster centroids
cluID	Cluster membership
criterion	Optimal value of the objective criterion
csize	Cluster size
nstart	A copy of nstart in the return object
odata	A copy of data in the return object

References

Hwang, H., Dillon, W. R. and Takane, Y. (2006). An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents, *Psychometrika*, 71, 161-171.

Iodice D'Enza, A. and Palumbo, F. (2013). Iterative factor clustering of binary data. *Computational Statistics*, *28*(2), 789-807.

van de Velden M., Iodice D'Enza, A. and Palumbo, F. (2016). Cluster correspondence analysis.*Psychometrika* (in press) DOI: 10.1007/s11336-016-9514-0

See Also

cluspca, tune_clusmca

Examples

```
data(cmc)
# values of wife's age and number of children were categorized
# into three groups based on quartiles
cmc$W_AGE = ordered(cut(cmc$W_AGE, c(16,26,39,49), include.lowest = TRUE))
levels(cmc$W_AGE) = c("16-26","27-39","40-49")
cmc$NCHILD = ordered(cut(cmc$NCHILD, c(0,1,4,17), right = FALSE))
levels(cmc$NCHILD) = c("0","1-4","5 and above")
outclusMCA = clusmca(cmc[,-c(3,6,7)], 3, 2, method = "clusCA")
```

---

cluspca *Joint dimension reduction and clustering of continuous data.*

Description

This function implements Factorial K-means (Vichi and Kiers, 2001) and Reduced K-means (De Soete and Carroll, 1994), as well as a compromise version of these two methods. The methods combine Principal Component Analysis for dimension reduction with K-means for clustering.

Usage

```
cluspca(data, nclus, ndim, alpha = NULL, method = "RKM",
center = TRUE, scale = TRUE, rotation = "none", nstart = 10,
smartStart = NULL, seed = 1234)
```

Arguments

data	Continuous dataset
nclus	Number of clusters
ndim	Dimensionality of the solution
alpha	Adjusts for the relative importance of the two terms of Clustering and Dimension Reduction; `alpha = 1` reduces to PCA, `alpha = 0.5` to reduced K-means, and `alpha = 0` to factorial K-means
method	Specifies the method. Options are RKM for reduced K-means and FKM for factorial K-means (`default = "RKM"`)
center	A logical value indicating whether the variables should be shifted to be zero centered (`default = TRUE`)
scale	A logical value indicating whether the variables should be scaled to have unit variance before the analysis takes place (`default = TRUE`)
rotation	Specifies the method used to rotate the factors. Options are `none` for no rotation, `varimax` for varimax rotation with Kaiser normalization and `promax` for promax rotation (default = "none")
nstart	Number of starts
smartStart	If `NULL` then a random cluster membership vector is generated. Alternatively, a cluster membership vector can be provided as a starting solution
seed	An integer that is used as argument by `set.seed()` for offsetting the random number generator when smartStart = NULL. The default value is 1234

Value

obscoord	Object scores
attcoord	Variable scores
centroid	Cluster centroids
cluID	Cluster membership
criterion	Optimal value of the objective function
csize	Cluster size
scale	A copy of `scale` in the return object
center	A copy of `center` in the return object
nstart	A copy of `nstart` in the return object
odata	A copy of `data` in the return object

References

De Soete, G. and Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. In Diday E. et al. (Eds.), *New Approaches in Classification and Data Analysis*, Heidelberg: Springer, 212-219.

Vichi, M. and Kiers, H.A.L. (2001). Factorial K-means analysis for two-way data. *Computational Statistics and Data Analysis*, 37, 49-64.

See Also

clusmca, tune_cluspca

Examples

```
data(macro)
outRKM = cluspca(macro, 3, 2, method = "RKM", rotation = "varimax")
plot(outRKM, cludesc = TRUE)
```

---

clusval *Distance-based statistics for cluster quality assessment.*

Description

This function computes two distance-based statistics (average silhouette widths and the Calinski-Harabasz index), which can be used for cluster quality assessment and decision about the number of clusters and dimensions.

Usage

```
clusval(x, dst = "full")
```

Arguments

x	An object of class cluspca or clusmca. For cluspca the distance measure used is the Euclidean distance, and for clusmca is Gower's distance
dst	Specifies the data used to compute the distances between objects. Options are full for the original data (after possible scaling) and low for the object scores in the low-dimensional space (default = "full")

Value

ch	Calinski-Harabasz index
asw	Average silhouette width

See Also

tune_cluspca, tune_clusmca

Examples

```
data(USArrests, package = "datasets")
outCDR = cluspca(USArrests, 3, 2, alpha = 0.6, rotation = "varimax")
clusval(outCDR, dst = "full")$asw
```

---

cmc *Contraceptive Choice in Indonesia*

Description

Data of married women in Indonesia who were not pregnant (or did not know they were pregnant) at the time of the survey. The dataset contains demographic and socio-economic characteristics of the women along with their preferred method of contraception (no use, long-term methods, short-term methods).

Usage

```
data("cmc")
```

Format

A data frame containing 1,437 observations on the following 10 variables.

`W_AGE` wife's age in years.

`W_EDU` ordered factor indicating the wife's education, with levels `"low"`, `"2"`, `"3"` and `"high"`.

`H_EDU` ordered factor indicating the wife's education, with levels `"low"`, `"2"`, `"3"` and `"high"`.

`NCHILD` number of children.

`W_REL` factor indicating the wife's religion, with levels `"non-Islam"` and `"Islam"`.

`W_WORK` factor indicating if the wife is working.

`H_OCC` ordered factor indicating the husbands occupation, with levels `"low"`, `"2"`, `"3"` and `"high"`.

`SOL` ordered factor indicating the standard of living index with levels `"low"`, `"2"`, `"3"` and `"high"`.

`MEDEXP` factor indicating media exposure, with levels `"good"` and `"not good"`.

`CM` factor indicating the contraceptive method used, with levels `"no-use"`, `"long-term"` and `"short-term"`.

Source

This dataset is part of the 1987 National Indonesia Contraceptive Prevalence Survey and was created by Tjen-Sien Lim. It has been taken from the UCI Machine Learning Repository at http://archive.ics.uci.edu/ml/.

References

Lim, T.-S., Loh, W.-Y. & Shih, Y.-S. (1999). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. *Machine Learning*, *40*(3), 203-228.

Examples

```
data(cmc)
```

`hsq` *Humor Styles*

Description

The dataset was collected with an interactive online version of the Humor Styles Questionnaire (HSQ) which assesses four independent ways in which people express and appreciate humor (Martin et al. 2003): affiliative, defined as the benign uses of humor to enhance one's relationships with others; self-enhancing, indicating uses of humor to enhance the self; aggressive, the use of humor to enhance the self at the expense of others; self-defeating the use of humor to enhance relationships at the expense of oneself. The main part of the questionnaire consisted of 32 statements rated from 1 to 5 according to the respondents' level of agreement. Three more questions were included (age, gender and self-reported accuracy of answer). The number of respondents is 993, after removing the cases with missing values in the 32 statements.

Usage

```
data("hsq")
```

Format

A data frame with 993 observations on 35 variables. The first 32 variables are Likert-type statements with 5 response categories, ranging from 1 (strong agreement) to 5 (strong disagreement).

`AF1` I usually don't laugh or joke around much with other people

`AF2` If I am feeling depressed, I can usually cheer myself up with humor

`AF3` If someone makes a mistake, I will often tease them about it

`AF4` I let people laugh at me or make fun at my expense more than I should

`AF5` I don't have to work very hard at making other people laugh - I seem to be a naturally humorous person

`AF6` Even when I'm by myself, I'm often amused by the absurdities of life

`AF7` People are never offended or hurt by my sense of humor

`AF8` I will often get carried away in putting myself down if it makes my family or friends laugh

`SE1` I rarely make other people laugh by telling funny stories about myself

`SE2` If I am feeling upset or unhappy I usually try to think of something funny about the situation to make myself feel better

`SE3` When telling jokes or saying funny things, I am usually not very concerned about how other people are taking it

`SE4` I often try to make people like or accept me more by saying something funny about my own weaknesses, blunders, or faults

SE5 I laugh and joke a lot with my closest friends

SE6 My humorous outlook on life keeps me from getting overly upset or depressed about things

SE7 I do not like it when people use humor as a way of criticizing or putting someone down

SE8 I don't often say funny things to put myself down

AG1 I usually don't like to tell jokes or amuse people

AG2 If I'm by myself and I'm feeling unhappy, I make an effort to think of something funny to cheer myself up

AG3 Sometimes I think of something that is so funny that I can't stop myself from saying it, even if it is not appropriate for the situation

AG4 I often go overboard in putting myself down when I am making jokes or trying to be funny

AG5 I enjoy making people laugh

AG6 If I am feeling sad or upset, I usually lose my sense of humor

AG7 I never participate in laughing at others even if all my friends are doing it

AG8 When I am with friends or family, I often seem to be the one that other people make fun of or joke about

SD1 I don't often joke around with my friends

SD2 It is my experience that thinking about some amusing aspect of a situation is often a very effective way of coping with problems

SD3 If I don't like someone, I often use humor or teasing to put them down

SD4 If I am having problems or feeling unhappy, I often cover it up by joking around, so that even my closest friends don't know how I really feel

SD5 I usually can't think of witty things to say when I'm with other people

SD6 I don't need to be with other people to feel amused - I can usually find things to laugh about even when I'm by myself

SD7 Even if something is really funny to me, I will not laugh or joke about it if someone will be offended

SD8 Letting others laugh at me is my way of keeping my friends and family in good spirits

Source

Martin, R. A., Puhlik-Doris, P., Larsen, G., Gray, J., & Weir, K. (2003). Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of Research in Personality*, *37*(1), 48-75.

Examples

```
data(hsq)
```

macro *Economic Indicators of 20 OECD countries for 1999*

Description

Data on the macroeconomic performance of national economies of 20 countries, members of the OECD (September 1999). The performance of the economies reflects the interaction of six main economic indicators (percentage change from the previous year): gross domestic product (GDP), leading indicator (LI), unemployment rate (UR), interest rate (IR), trade balance (TB), net national savings (NNS).

Usage

```
data(macro)
```

Format

A data frame with 20 observations on the following 6 variables.

`GDP` numeric

`LI` numeric

`UR` numeric

`IR` numeric

`TB` numeric

`NNS` numeric

Source

Vichi, M. & Kiers, H. A. (2001). Factorial k-means analysis for two-way data. *Computational Statistics & Data Analysis*, *37*(1), 49-64.

plot.clusmca *Plotting function for* `clusmca()` *output.*

Description

Plotting function that creates a `ggplot2` based map of the object scores and a scatter plot of both the attribute scores and the centroids.

Usage

```
## S3 method for class 'clusmca'
plot(x, dims = c(1,2), disp = TRUE, cludesc = FALSE, what = c(TRUE,TRUE),
attlabs = NULL, binary = FALSE, ...)
```

Arguments

x	Object returned by clusmca()
dims	Numerical vector of length 2 indicating the dimensions to plot on horizontal and vertical axes respectively; default is first dimension horizontal and second dimension vertical
disp	A logical value indicating whether the plots are shown in the R window or saved as PDF files in the working directory (default = TRUE)
what	Vector of two logical values specifying the contents of the plots. First entry indicates whether a scatterplot of the objects is displayed in principal coordinates. Second entry indicates whether a scatterplot of the attribute categories is displayed in principal coordinates. The default is c(TRUE, TRUE) and the resultant plot is a biplot of both objects and attribute categories with gamma-based scaling (see van de Velden et al. (2016))
cludesc	A logical value indicating whether a series of barplots is produced showing the largest (in absolute value) standardized residuals per attribute for each cluster (default = FALSE)
attlabs	Vector of attribute labels; if not provided, default labeling is applied
binary	Vector of attribute labels; if not provided, default labeling is applied
...	Further arguments to be transferred to clusmca()

References

Hwang, H., Dillon, W. R. and Takane, Y. (2006). An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents, *Psychometrika*, 71, 161-171.

Iodice D' Enza, A. and Palumbo, F. (2013). Iterative factor clustering of binary data. *Computational Statistics*, *28*(2), 789-807.

Van de Velden M., Iodice D' Enza, A. and Palumbo, F. (2016). Cluster correspondence analysis.*Psychometrika* (in press) DOI: 10.1007/s11336-016-9514-0

See Also

plot.cluspca

Examples

```
data("hsq")
outclusMCA = clusmca(hsq[,1:8], 3, 2, method = "iFCB")
plot(outclusMCA, cludesc = TRUE)
```

`plot.cluspca` *Plotting function for* `cluspca()` *output.*

Description

Plotting function that creates a `ggplot2` scatterplot of the objects, a correlation circle of the variables or a biplot of both objects and variables.

Usage

```
## S3 method for class 'cluspca'
plot(x, dims = c(1, 2), disp = TRUE, cludesc = FALSE, what = c(TRUE,TRUE), ...)
```

Arguments

`x`	Object returned by `cluspca()`
`dims`	Numerical vector of length 2 indicating the dimensions to plot on horizontal and vertical axes respectively; default is first dimension horizontal and second dimension vertical
`disp`	A logical value indicating whether the plots are shown in the `R` window or saved as PDF files in the working directory (`default = TRUE`)
`what`	Vector of two logical values specifying the contents of the plots. First entry indicates whether a scatterplot of the objects is displayed and the second entry whether a correlation circle of the variables is displayed. The default is `c(TRUE, TRUE)` and the resultant plot is a biplot of both objects and variables
`cludesc`	A logical value indicating if a parallel plot showing cluster means is produced (default = `FALSE`)
`...`	Further arguments to be transferred to `cluspca()`

References

De Soete, G. and Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. In Diday E. et al. (Eds.), *New Approaches in Classification and Data Analysis*, Heidelberg: Springer, 212-219.

Vichi, M. and Kiers, H.A.L. (2001). Factorial K-means analysis for two-way data. *Computational Statistics and Data Analysis*, 37, 49-64.

See Also

`plot.clusmca`

Examples

```
data("iris", package = "datasets")
outclusPCA = cluspca(iris[,-5], 3, 2, alpha = 0.3, rotation = "varimax")
table(outclusPCA$cluID,iris[,5])
plot(outclusPCA, cludesc = TRUE)
```

tune_clusmca *Methods for categorical data with cluster quality assessment.*

Description

This function facilitates the selection of the appropriate number of clusters and dimensions for joint dimension reduction and clustering of categorical data.

Usage

```
tune_clusmca(data, nclusrange = 2:7, ndimrange = 2:4, method = "clusCA",
criterion = "asw", dst = "full", alpha = .5, nstart = 10,
smartStart = NULL, seed = 1234)
```

Arguments

data	Categorical dataset
nclusrange	An integer vector with the range of numbers of clusters which are to be compared by the cluster validity criteria
ndimrange	An integer vector with the range of dimensions which are to be compared by a cluster quality criterion
criterion	One of asw, ch or crit. Determines whether average silhouette width, Calinski-Harabasz index or objective value of the selected method is used (default = "asw")
dst	Specifies the data used to compute the distances between objects. Options are full for the original data (after possible scaling) and low for the object scores in the low-dimensional space (default = "full")
method	Specifies the method. Options are MCAk for MCA K-means, iFCB for Iterative Factorial Clustering of Binary variables and clusCA for Cluster Correspondence Analysis (default = clusCA).
alpha	Non-negative scalar to adjust for the relative importance of of MCA and K-means in the solution (default = .5). Works only in combination with method = "MCAk".
nstart	Number of random starts.
smartStart	If NULL then a random cluster membership vector is generated. Alternatively, a cluster membership vector can be provided as a starting solution
seed	An integer that is used as argument by set.seed() for offsetting the random number generator when smartStart = NULL. The default value is 1234.

Value

clusmcaobj	The output of the optimal run of the clusmca() function
nclusbest	The optimal number of cluster
ndimbest	The optimal number of dimensions

critbest	The optimal criterion value for nclusbest clusters and ndimbest dimensions
critgrid	Matrix of size nclusrange x ndimrange with criterion values for the specified ranges of numbers of clusters and numbers of dimensions (values are calculated for the number of clusters greater than the number of dimensions; otherwise values are left blank)

See Also

clusmca, tune_cluspca

Examples

```
data(underwear)
bestclusCA = tune_clusmca(underwear[,2:3], 3:4, 2:3, criterion = "asw", nstart = 20)
plot(bestclusCA$clusmcaobj)
```

---

tune_cluspca *Methods for continuous data with cluster quality assessment.*

Description

This function facilitates the selection of the appropriate number of clusters and dimensions for joint dimension reduction and clustering of continuous data.

Usage

```
tune_cluspca(data, nclusrange = 2:7, ndimrange = 2:4, criterion = "asw",
dst = "full", alpha = NULL, method = "RKM", center = TRUE, scale = TRUE,
rotation = "none", nstart = 10, smartStart = NULL, seed = 1234)
```

Arguments

data	Continuous dataset
nclusrange	An integer vector with the range of numbers of clusters which are to be compared by the cluster validity criteria
ndimrange	An integer vector with the range of dimensions which are to be compared by the cluster validity criteria
criterion	One of asw, ch or crit. Determines whether average silhouette width, Calinski-Harabasz index or objective value of the selected method is used (default = "asw")
dst	Specifies the data used to compute the distances between objects. Options are full for the original data (after possible scaling) and low for the object scores in the low-dimensional space (default = "full")
alpha	Adjusts for the relative importance of the two terms of Clustering and Dimension Reduction; alpha = 1 reduces to PCA, alpha = 0.5 to reduced K-means, and alpha = 0 to factorial K-means

method	Specifies the method. Options are RKM for reduced K-means and FKM for factorial K-means (default = "RKM").
center	A logical value indicating whether the variables should be shifted to be zero centered (default = TRUE)
scale	A logical value indicating whether the variables should be scaled to have unit variance before the analysis takes place (default = TRUE)
rotation	Specifies the method used to rotate the factors. Options are none for no rotation, varimax for varimax rotaion with Kaiser normalization and promax for promax rotation (default = "none")
nstart	Number of starts
smartStart	If NULL then a random cluster membership vector is generated. Alternatively, a cluster membership vector can be provided as a starting solution
seed	An integer that is used as argument by set.seed() for offsetting the random number generator when smartStart = NULL. The default value is 1234

Value

cluspcaobj	The output of the optimal run of the cluspca() function
nclusbest	The optimal number of cluster
ndimbest	The optimal number of dimensions
critbest	The optimal criterion value for nclusbest clusters and ndimbest dimensions
critgrid	Matrix of size nclusrange x ndimrange with criterion values for the specified ranges of numbers of clusters and numbers of dimensions (values are calculated for the number of clusters greater than the number of dimensions; otherwise values are left blank)

See Also

cluspca, tune_clusmca

Examples

```
data(macro)
bestRKM = tune_cluspca(macro, 3:4, 2:3, method = "RKM", criterion = "asw", dst = "low")
plot(bestRKM$cluspcaobj)
```

---

underwear *South Korean Underwear*

Description

The dataset comes from a large survey conducted by a South Korean underwear manufacturer in 1997. 664 South Korean consumers were asked to provide responses for three multiple-choice items: attributes when considering a brand of underwear to purchase (15 attributes), preferred brand of underwear (8 brands) and consumer age (3 levels).

Usage

```
data(underwear)
```

Format

A data frame with 664 observations on the following variables.

`brand` categorical: `1. BYC`, `2. TRY`, `3. VICMAN`, `4. James Dean`, `5. Michiko-London`, `6. Benetton`, `7. Bodyguard`, `8. Calvin Klein`

`atts` categorical: `1. Comfortable`, `2. Smooth`, `3. Superior fabrics`, `4. Reasonable price`, `5. Fashionable design`, `6. Favorable advertisements`, `7. Trendy color`, `8. Good design`, `9. Various colors`, `10. Elastic`, `11. Store is near`, `12. Excellent fit`, `13. Design quality`, `14. Youth appeal`, `15. Various sizes`

`age` categorical: `1. 10-29`, `2. 30-49`, `3. 50 and over`

Source

Hwang, H., Dillon, W. R. & Takane, Y. (2006). An extension of multiple correspondence analysis for identifying heterogenous subgroups of respondents. *Psychometrika*, 71, 161-171.

Examples

```
data(underwear)
```

Index