Package 'GPCSIV'

February 19, 2015

Type Package

Title GPCSIV, Generalized Principal Component of Symbolic Interval variables

Version 0.1.0

Date 2013-06-06

Author Brahim Brahim and Sun Makosso-Kallyth <sun.makosso-kallyth@crchuq.ulaval.ca>

Maintainer Brahim Brahim <brahim.brahim@bigdatavisualizations.com>

Description This package implements an extension of principal component analysis (PCA) tailored to handle multiple data tables. It can handle Big Data in the sense that the variation in massive data can be described by intervals [a, b] and multiple tables.

License GPL (>= 2)

Depends scatterplot3d, sqldf

NeedsCompilation no

Repository CRAN

Date/Publication 2013-09-15 09:53:08

R topics documented:

GPCSIV-package *Generalized Principal Component of Symbolic Interval Variables.*

Description

This package implements an extension of principal component analysis (PCA) tailored to handle multiple data tables. These multiple data tables contain the same number of Interval variables and the same observations. This package can handle Big Data in the sense that the variation in massive data can be described by intervals [a, b] and multiple tables. If only one data table is specified, in this case this package performs a PCA of interval data.

Details

Package:	GPCSIV
Type:	Package
Version:	1.0
Date:	2013-06-06
License:	GPL (>= 2)

Each dataset can be in csv, excel, access, txt,...,file. The only constraint is that for each variable, the maximum column must follow the minimum column. The Resdata class implemented returns two list of data frames (list of minimums and maximums). These lists of data frames are the inputs of the gpca function.

Author(s)

Brahim Brahim and Sun Makosso-Kallyth Maintainer : Brahim Brahim <brahim.brahim@bigdatavisualizations.com>

References

Billard, L. and E. Diday (2006). Symbolic Data Analysis: conceptual statistics and data Mining. Berlin: Wiley series in computational statistics.

Diday, E. and M. Noirhomme-Fraiture (2008). Symbolic Data Analysis and the SODAS Software. Chichester: Wiley Interscience.

Makosso-Kallyth, S (2013). Analysis of m sets of symbolic interval variables. Revue des Nouvelles Technologies de l'Information, vol. RNTI-E25. pp. 97-108.

gpca *Main function gpca, Generalized Principal Component of Symbolic Interval variables*

Description

Performs an analysis in principal axes of multiple tables of symbolic interval variables. The function uses a class "Resdata" object.

Usage

```
gpca(xmin = list, xmax = list, reduire = 0, nomVar = NULL,
axes = c(1, 2), axes2=c(1,2,3), nomInd = NULL, legend = NULL, xlim = NULL,
 ylim = NULL, nametable = NULL, plot3d.table=NULL)
```

Arguments

`xmin` List of all data frames containing all min of initial data. These data have the same number of rows and columns.

`xmax` List of all data frames containing all max of initial data. These data have the same number of rows and columns.

`reduire` is a logical argument of the Centrage function. To centering without scaling by standard deviation, use reduire=0. Otherwise use reduire=1.

`nomVar` Set the column names of all data frames

`axes` a length 2 vector specifying the components to plot

`axes2` a length 2 vector specifying the components to plot

`nomInd` Set the column row names of all data frames

`legend` This function could be used to add legends to plots.

`xlim` range for the plotted "x" values, defaulting to the range of the finite values of "x"

`ylim` range for the plotted "y" values, defaulting to the range of the finite values of "y"

`nametable` Set the column names of the tables

`plot3d.table` for visualization in 2D and 3D of tables

Value

Returns a list including:

`PC` array containing the projections of the min and max of the average of input interval datasets.

`Correl` Correlations based on interval variables - dimensions

`Pval2` a matrix containing all the eigenvalues, the percentage of variance and the cumulative percentage of variance

`PCinterval` array list containing the coordinates of the individuals on the principal axes

Author(s)

Brahim Brahim and Sun Makosso-Kallyth

References

S.Makosso-Kallyth, Analysis of m sets of symbolic interval variables. Revue des Nouvelles Technologies de l"Informatique, vol. RNTI-E25. pp. 97-108, 2013.

Examples

```
data(Judge1)
data(Judge2)
data(Judge3)

preparation1<-Resdata(list(Judge1,Judge2,Judge3))
List1min<-preparation1$tablemin
List1max<-preparation1$tablemax

# example 1 with the use of some parameters by default
example1<-gpca(xmin=List1min,xmax=List1max,nomInd=paste('Region',1:6),
              nomVar=c('Banana','Coffee','Thea','Cocoa'))

# example 1 with visualization of table containing the coordinates
# of the individuals onto the principal axes
example1<-gpca(xmin=List1min,xmax=List1max,nomInd=paste('Region',1:6),nomVar=c('Banana',
              'Coffee','Thea','Cocoa'),axes=c(1,2),axes2=c(1,2,3),plot3d.table=c(1:3),
   nametable=paste('Expert',1:3,sep='-'))

# example 1 with visualization of the table 2 and 3 containing
#the coordinates of the individuals onto the principal axes
example1<-gpca(xmin=List1min,xmax=List1max,nomInd=paste('Region',1:6),
              nomVar=c('Banana','Coffee','Thea','Cocoa'),axes=c(1,2),
  axes2=c(1,2,3),plot3d.table=c(2:3))

#### print numeric output of example1
# input tables onto the axes of the compromise
example1$PCinterval

# Principal components of the compromise
example1$PCCompromise

# Correlation between initial interval variables and principal
#component of the compromise
example1$Correl

# print eigenvalue, % of variance, cumulative % percentage
# of PCA of the compromise
example1$Pval


data(video1)
data(video2)
data(video3)
preparation2<-Resdata(list(video1,video2,video3))
List2min<-preparation2$tablemin
List2max<-preparation2$tablemax
```

```
# example2 : analysis of video dataset
example2<-gpca(xmin=List2min,xmax=List2max,nomVar=c('nvisit','nwatch',
'nlike','ncoment','nshare'),
nametable=paste('Video', 1:3))

# example2 : analysis of video dataset with the 3D graphics
example2<-gpca(xmin=List2min,xmax=List2max,nomVar=c('nvisit',
'nwatch','nlike','ncoment','nshare'),nametable=paste('Video', 1:3),
nomInd=paste('Obs',1:10),plot3d.table=c(1,2,3))


data(oils)
preparation3<-Resdata(list(oils))
List3min<-preparation3$tablemin
List3max<-preparation3$tablemax

# example3 Interval Principal component analysis based on min and max
example3<-gpca(xmin=List3min,xmax=List3max,nomInd=rownames(oils),
nomVar=c('Gravity','Freezing','Iodine','Saponification'))

#### print numeric output of example3

# interval Principal components
example3$PCinterval

# Correlation between initial interval variables and principal
#components
example3$Correl

# print eigenvalue, % of variance, cumulative % percentage
# of PCA of the compromise
example3$Pval

# example3 Interval Principal component analysis based on min and max
#with standardisation of variables
example3bis<-gpca(xmin=List3min,xmax=List3max,nomInd=rownames(oils),
nomVar=c('Gravity','Freezing','Iodine','Saponification'),reduire=1)

# interval Principal components
example3bis$PCinterval

# Correlation between initial interval variables and principal
#components
example3bis$Correl

# print eigenvalue, % of variance, cumulative % percentage
# of PCA of the compromise
example3bis$Pval
```

Judge1 *GPCSIV, dataset Judge 1*

Description

This is a simulated dataset containing information on agricultural products from six regions graded by 3 experts. This dataset (Judge1) contains 4 symbolic interval variables. In these data, the five successive pairs of columns represented the five symbolic interval variables . Each pair consists of a min and a max.

Usage

```
data(Judge1)
```

Format

A data frame with 6 observations on the following 8 variables.

`Banana.min` a numeric vector, minimum grade banana proposed by the expert 1

`Banana.max` a numeric vector, maximum grade banana proposed by the expert 1

`Coffee.min` a numeric vector, minimum grade coffee proposed by the expert 1

`Coffee.max` a numeric vector, maximum grade coffee proposed by the expert 1

`Thea.min` a numeric vector, minimum grade thea proposed by the expert 1

`Thea.max` a numeric vector, maximum grade thea proposed by the expert 1

`Cocoa.min` a numeric vector, minimum grade cocoa proposed by the expert 1

`Cocoa.max` a numeric vector, maximum grade banana proposed by the expert 1

Examples

```
data(Judge1)
```

Judge2 *GPCSIV, dataset Judge 2*

Description

This is a simulated dataset which contains informations about assessed by 3 experts to evaluate various product from six regions. This dataset (Judge2) contains 4 symbolic interval variables. In these data, the five successive pairs of columns represented the five symbolic variables of type interval. Each pair consists of a min and a max.

Usage

```
data(Judge2)
```

Format

A data frame with 6 observations on the following 8 variables.

`Banana.min` a numeric vector, minimum grade banana proposed by the expert 2

`Banana.max` a numeric vector, maximum grade banana proposed by the expert 2

`Coffee.min` a numeric vector, minimum grade coffee proposed by the expert 2

`Coffee.max` a numeric vector, maximum grade coffee proposed by the expert 2

`Thea.min` a numeric vector, minimum grade thea proposed by the expert 2

`Thea.max` a numeric vector, maximum grade thea proposed by the expert 2

`Cocoa.min` a numeric vector, minimum grade cocoa proposed by the expert 2

`Cocoa.max` a numeric vector, maximum grade banana proposed by the expert 2

Examples

```
data(Judge2)
```

`Judge3` *GPCSIV, dataset Judge 3*

Description

This is a simulated dataset which contains informations about assessed by 3 experts to evaluate various product from six regions. This dataset (Judge1) contains 4 symbolic interval variables. In these data, the five successive pairs of columns represented the five symbolic variables of type interval. Each pair consists of a min and a max.

Usage

```
data(Judge1)
```

Format

A data frame with 6 observations on the following 8 variables.

`Banana.min` a numeric vector, minimum grade banana proposed by the expert 3

`Banana.max` a numeric vector, maximum grade banana proposed by the expert 3

`Coffee.min` a numeric vector, minimum grade coffee proposed by the expert 3

`Coffee.max` a numeric vector, maximum grade coffee proposed by the expert 3

`Thea.min` a numeric vector, minimum grade thea proposed by the expert 3

`Thea.max` a numeric vector, maximum grade thea proposed by the expert 3

`Cocoa.min` a numeric vector, minimum grade cocoa proposed by the expert 3

`Cocoa.max` a numeric vector, maximum grade banana proposed by the expert 3

Examples

```
data(Judge3)
```

oils *Oils data proposed by Ichino*

Description

Each row in this table represents a class of oil described by four variables interval quantitives : 'specific gravity', 'freezing points', 'iodine value', 'saponification' The four successive pairs of columns contain the five symbolic interval variables. Each pair consists of a min and a max.

Usage

```
data(Judge1)
```

Format

A data frame with 8 observations on the following 8 variables.

`GRA.MIN` a numeric vector, minimum of specific gravity

`GRA.MAX` a numeric vector, maximum of specific gravity

`FRE.MIN` a numeric vector, minimum of freezing points

`FRE.MAX` a numeric vector, maximum of freezing points

`IOD.MIN` a numeric vector, minimum of iodine value

`IOD.MAX` a numeric vector, maximum of iodine value

`SAP.MIN` a numeric vector, minimum of saponification

`SAP.MAX` a numeric vector, maximum saponification

References

Cazes P., Chouakria A., Diday E. et Schektman Y. (1997). Extension de l'analyse en composantes principales a des donnees de type intervalle, Rev. Statistique Appliquee, Vol. XLV Num. 3 pag. 5-24, France.

Ichino M. (1994). Generalized Minkowski metrics for mixed featuretype data analysis. IEEE , transactions on systems, man and cybermetrics, vol.24, n 4.

Examples

```
data(oils)
```

Resdata *class of objects 'Resdata'*

Description

This class return two list of data frames (list of minimum data and list of maximum

Usage

```
Resdata(enter = list)
```

Arguments

`enter`	list of data input

Value

`tablemin`	List of all arrays containing all minimum of initial data
`tablemax`	List of all arrays containing all maximum of initial data

Author(s)

Brahim Brahim and Sun Makosso-Kallyth

Examples

```
data(video1)
data(video2)
data(video3)
preparation2<-Resdata(list(video1,video2,video3))
List2min<-preparation2$tablemin
List2max<-preparation2$tablemax
```

video1 *Video data, GPCSIV, Generalized Principal Component of Symbolic Interval variables*

Description

This is a simulated dataset which contains information about the behaviour of internauts on a video published on the web. This dataset (video1) contains 5 symbolic interval variables. In these data, the five successive pairs of columns represented the five symbolic interval variables. Each pair consists of a min and a max.

Usage

```
data(video1)
```

Format

A data frame with 10 observations on the following 10 variables.

`nvisitmin` a numeric vector, minimum daily average number of visits.

`nvisitmax` a numeric vector, maximum daily average number of visits.

`nwatchmin` a numeric vector, minimum daily average number of people who clicked on play.

`nwatchmax` a numeric vector, maximum daily average number of people who clicked on play.

`nlikemin` a numeric vector, minimum daily average number of people who clicked on 'I like'.

`nlikemax` a numeric vector, maximum daily average number of people who clicked on 'I like'.

`ncomntmin` a numeric vector, minimum daily average number of people who commented.

`ncomntmax` a numeric vector, maximum daily average number of people who commented.

`nsharemin` a numeric vector, minimum daily average number of people who shared the video.

`nsharemax` a numeric vector, maximum daily average number of people who shared the video.

Examples

```
data(video1)
```

`video2` *Video data, GPCSIV, Generalized Principal Component of Symbolic Interval variables*

Description

This is a simulated dataset which contains informations about behaviour of internauts concerning one video published on the web. This dataset (video2) contains 5 symbolic interval variables. In these data, the five successive pairs of columns represented the five symbolic variables of type interval. Each pair consists of a min and a max.

Usage

```
data(video2)
```

Format

A data frame with 10 observations on the following 10 variables.

`nvisitmin` a numeric vector, minimum daily average number of visits.

`nvisitmax` a numeric vector, maximum daily average number of visits.

`nwatchmin` a numeric vector, minimum daily average number of people who clicked on play.

`nwatchmax` a numeric vector, maximum daily average number of people who clicked on play.

`nlikemin` a numeric vector, minimum daily average number of people who clicked on 'I like'.

`nlikemax` a numeric vector, maximum daily average number of people who clicked on 'I like'.

`ncomntmin` a numeric vector, minimum daily average number of people who commented.

`ncomntmax` a numeric vector, maximum daily average number of people who commented.

`nsharemin` a numeric vector, minimum daily average number of people who shared the video.

`nsharemax` a numeric vector, maximum daily average number of people who shared the video.

Examples

```
data(video2)
```

`video3` *Video data, GPCSIV, Generalized Principal Component of Symbolic Interval variables*

Description

This is a simulated dataset which contains informations about behaviour of internauts concerning one video published on the web. This dataset (video3) contains 5 symbolic interval variables. In these data, the five successive pairs of columns represented the five symbolic variables of type interval. Each pair consists of a min and a max.

Usage

```
data(video3)
```

Format

A data frame with 10 observations on the following 10 variables.

`nvisitmin` a numeric vector, minimum daily average number of visits.

`nvisitmax` a numeric vector, maximum daily average number of visits.

`nwatchmin` a numeric vector, minimum daily average number of people who clicked on play.

`nwatchmax` a numeric vector, maximum daily average number of people who clicked on play.

`nlikemin` a numeric vector, minimum daily average number of people who clicked on 'I like'.

`nlikemax` a numeric vector, maximum daily average number of people who clicked on 'I like'.

`ncomntmin` a numeric vector, minimum daily average number of people who commented.

`ncomntmax` a numeric vector, maximum daily average number of people who commented.

`nsharemin` a numeric vector, minimum daily average number of people who shared the video.

`nsharemax` a numeric vector, maximum daily average number of people who shared the video.

Examples

```
data(video3)
```

Index