Package 'GraphPCA'

April 13, 2018

Type Package

Title Graphical Tools of Histogram PCA

Version 1.1

Depends R (>= 2.15.1)

Imports ggplot2, FactoMineR, scatterplot3d,ggplot2movies

Date 2018-04-04

Author Brahim Brahim <brahim.brahim@bigdatavisualizations.com> and Sun Makosso-Kallyth <makosso.sun@gmail.com>

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

Description Histogram principal components analysis is the generalization of the PCA. Histogram data are adapted to design complex and big data which histograms used as variables (big data adapter). Functions implemented provides numerical and graphical tools of an extension of PCA. Sun Makosso Kallyth (2016) <doi:10.1002/sam.11270>. Sun Makosso Kallyth and Edwin Diday (2012) <doi:10.1007/s11634-012-0108-0>.

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2018-04-13 20:20:04 UTC

R topics documented:

GraphPCA-package *Graphical Tools of Histogram PCA.*

Description

Histogram principal components analysis is the generalization of the PCA, Histogram data are adapted to design complex and Big data which histograms used as variables (Big Data adapter). Functions implemented provides numerical and graphical tools of an extension of PCA.

Details

Package:	GraphPCA
Type:	Package
Version:	1.1
Date:	2018-04-04
License:	GPL (>= 2)

PrepHistogram, HistPCA, Visu.

Author(s)

Brahim Brahim <brahim.brahim@bigdatavisualizations.com> and Sun Makosso-Kallyth <makosso.sun@gmail.com>

References

Sun Makosso Kallyth (2016) principal axes analysis of symbolic histogram variables statistical snalysis and data mining. Edwin Diday, Sun Makosso Kallyth (2012) adaptation of interval PCA to symbolic histogram variables data analysis and classification.

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

Diday, E., Rodriguez O. and Winberg S. (2000). Generalization of the Principal Components Analysis to Histogram Data, 4th European Conference on Principles and Practice of Knowledge Discovery in Data Bases, September 12-16, 2000, Lyon, France.

Donoho, D., Ramos, E. (1982). Primdata: Data Sets for Use With PRIM-H. Version for second (15-18, Aug, 1983) Exposition of Statistical Graphics Technology, by American Statistical Association.

Le-Rademacher J., Billard L. (2013). Principal component histograms from interval-valued observations, Computational Statistics, v.28 n.5, p.2117-2138.

Makosso-Kallyth S. and Diday E. (2012). Adaptation of interval PCA to symbolic histogram variables, Advances in Data Analysis and Classification July, Volume 6, Issue 2, pp 147-159.

HistPCA *HistPCA*

Description

Performs a PCA of multiple tables of histogram variables.

Usage

```
HistPCA(Variable = list, score = NULL, t = 1.1, axes = c(1, 2),
       Row.names = NULL, xlim = NULL, ylim = NULL, xlegend = NULL, ylegend = NULL,
         Col.names = NULL, transformation = 1, method = "hypercube", proc = 0,
         plot3d.table = NULL, axes2 = c(1, 2, 3), ggplot = 1)
```

Arguments

Variable	List of all data frames containing initial histogram variable. Every histogram is a data frames and every columns of data frame contains histogram bins.
score	List of bins score of every histogram variable. By default these scores are the ranks of histogram bins.
t	t is a real number used for transforming histogram to interval via Tchebytchev's inequality. By default, t=1.1.
axes	a length 2 vector specifying the components to plot
Row.names	Retrieve or set the row names of a matrix-like object.
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".
xlegend	This function could be used to add legends to plots.
ylegend	This function could be used to add legends to plots.
Col.names	Retrieve or set the row names of a matrix-like object.
transformation	type of tranformation for data. If transformation=2, angular is used.
method	method used (method='hypercube',method='longueur')
proc	option valid when method='longueur'. If proc=1, the procuste analysis is used.
plot3d.table	specification for the scatterplot3d. if plot3d.table=1, the scatterplot3d will appear.
axes2	a length 2 vector specifying the components to plot
ggplot	

Details

See Examples

Value

`Correlation`	Correlations between means of histogram and their principal components
`Tableaumean`	Table containing the average of histogram mean
`VecteurPropre`	eigen vector of PCA of histogram mean
`PourCentageComposante`	a matrix containing all the eigenvalues, the percentage of variance and the cumulative percentage of variance
`PCinterval`	Data frame containing the coordinates of the individuals on the principal axes

Author(s)

Brahim Brahim <brahim.brahim@bigdatavisualizations.com> and Sun Makosso-Kallyth <makosso.sun@gmail.com>

References

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

Diday, E., Rodriguez O. and Winberg S. (2000). Generalization of the Principal Components Analysis to Histogram Data, 4th European Conference on Principles and Practice of Knowledge Discovery in Data Bases, September 12-16, 2000, Lyon, France.

Donoho, D., Ramos, E. (1982). Primdata: Data Sets for Use With PRIM-H. Version for second (15-18, Aug, 1983) Exposition of Statistical Graphics Technology, by American Statistical Association.

Le-Rademacher J., Billard L. (2013). Principal component histograms from interval-valued observations, Computational Statistics, v.28 n.5, p.2117-2138.

Makosso-Kallyth S. and Diday E. (2012). Adaptation of interval PCA to symbolic histogram variables, Advances in Data Analysis and Classification July, Volume 6, Issue 2, pp 147-159.

Examples

```
data(movies)
ab = movies
ab = na.omit(ab)
Action = subset(ab,Action==1)
Action$genre = as.factor("Action")
Drama = subset(ab,Drama==1)
Drama$genre = as.factor("Drama")

Animation = subset(ab,Animation==1)
Animation$genre = as.factor("Animation")

Comedy = subset(ab,Comedy==1)
Comedy$genre = as.factor("Comedy")

Documentary = subset(ab,Documentary ==1)
Documentary $genre = as.factor("Documentary")


Romance = subset(ab,Romance==1)
```

```
Romance$genre = as.factor("Romance")

Short = subset(ab,Short==1)
Short$genre = as.factor("Short")

 ab = rbind(Action,Drama,Animation,Comedy,Documentary,Romance,Short)
 Hist1=PrepHistogram(X=sapply(ab[,3],unlist),Z=ab[,25],k=5)$Vhistogram
Hist2=PrepHistogram(X=sapply(ab[,4],unlist),Z=ab[,25],k=5)$Vhistogram
 Hist3=PrepHistogram(X=sapply(ab[,5],unlist),Z=ab[,25],k=5)$Vhistogram
Hist4=PrepHistogram(X=sapply(ab[,6],unlist),Z=ab[,25],k=5)$Vhistogram
 Hist5=PrepHistogram(X=sapply(ab[,7],unlist),Z=ab[,25],k=5)$Vhistogram

 ss1=Ridi(Hist1)$Ridit
 ss2=Ridi(Hist2)$Ridit
 ss3=Ridi(Hist3)$Ridit
 ss4=Ridi(Hist4)$Ridit
 ss5=Ridi(Hist5)$Ridit


HistPCA(list(Hist1,Hist2,Hist3,Hist4,Hist5),score=list(ss1,ss2,ss3,ss4,ss5))

res_pca=HistPCA(list(Hist1,Hist2,Hist3,Hist4,Hist5),score=list(ss1,ss2,ss3,ss4,ss5))

 Visu(res_pca$PCinterval)
```

`movies` *movies*

Description

A movies data frame with

Usage

```
data("movies")
```

Format

A data frame with 58788 observations on the following 24 variables.

`title` a character vector

`year` a numeric vector

`length` a numeric vector

`budget` a numeric vector

`rating` a numeric vector

`votes` a numeric vector

`r1` a numeric vector

`r2` a numeric vector

`r3` a numeric vector

`r4` a numeric vector

`r5` a numeric vector

`r6` a numeric vector

`r7` a numeric vector

`r8` a numeric vector

`r9` a numeric vector

`r10` a numeric vector

`mpaa` a character vector

`Action` a numeric vector

`Animation` a numeric vector

`Comedy` a numeric vector

`Drama` a numeric vector

`Documentary` a numeric vector

`Romance` a numeric vector

`Short` a numeric vector

Details

Initial movies data frame on which Histogram variables are built/

Source

https://cran.r-project.org/web/packages/ggplot2movies/index.html

References

Makosso-Kallyth, Sun; Diday, Edwin. Adaptation of interval PCA to symbolic histogram variables. Advances in Data Analysis and Classification. Volume 6. n 2. 2012. pages 147-159. Springer.

Examples

```
data(movies)
ab = movies
ab = na.omit(ab)
Action = subset(ab,Action==1)
Action$genre = as.factor("Action")
Drama = subset(ab,Drama==1)
Drama$genre = as.factor("Drama")

Animation = subset(ab,Animation==1)
Animation$genre = as.factor("Animation")

Comedy = subset(ab,Comedy==1)
```

```
Comedy$genre = as.factor("Comedy")

Documentary = subset(ab,Documentary ==1)
Documentary $genre = as.factor("Documentary")


Romance = subset(ab,Romance==1)
Romance$genre = as.factor("Romance")

Short = subset(ab,Short==1)
Short$genre = as.factor("Short")

 ab = rbind(Action,Drama,Animation,Comedy,Documentary,Romance,Short)
 Hist1=PrepHistogram(X=sapply(ab[,3],unlist),Z=ab[,25],k=5)$Vhistogram
Hist2=PrepHistogram(X=sapply(ab[,4],unlist),Z=ab[,25],k=5)$Vhistogram
 Hist3=PrepHistogram(X=sapply(ab[,5],unlist),Z=ab[,25],k=5)$Vhistogram
Hist4=PrepHistogram(X=sapply(ab[,6],unlist),Z=ab[,25],k=5)$Vhistogram
 Hist5=PrepHistogram(X=sapply(ab[,7],unlist),Z=ab[,25],k=5)$Vhistogram

 ss1=Ridi(Hist1)$Ridit
 ss2=Ridi(Hist2)$Ridit
 ss3=Ridi(Hist3)$Ridit
 ss4=Ridi(Hist4)$Ridit
 ss5=Ridi(Hist5)$Ridit


HistPCA(list(Hist1,Hist2,Hist3,Hist4,Hist5),score=list(ss1,ss2,ss3,ss4,ss5))

res_pca=HistPCA(list(Hist1,Hist2,Hist3,Hist4,Hist5),score=list(ss1,ss2,ss3,ss4,ss5))

 Visu(res_pca$PCinterval)
```

PrepHistogram *This function transform standard variables into histogram-valued variable (PrepHistogram)*

Description

This function transforms quantitative variable into histogram-valued variable.

Usage

```
PrepHistogram(X, Z = NULL, k = 3,group=NULL)
```

Arguments

X	X quantitative variable that need to be transformed into symbolic histogram
Z	Z categorical variable that need to be used for the purpose of clustering
k	k number of bins of histogram-valued variables
group	Data group

Value

Returns a list including :

`mido`	return the class centres of class and the min and the max of histograms
`Vhistogram`	dataframe containing the relative frequency of histogram variables

Author(s)

Brahim Brahim Brahim Brahim <brahim.brahim@bigdatavisualizations.com> and Sun Makosso-Kallyth <makosso.sun@gmail.com>

References

Makosso-Kallyth, Sun; Diday, Edwin. Adaptation of interval PCA to symbolic histogram variables. Advances in Data Analysis and Classification. Volume 6. n 2. 2012. pages 147-159. Springer.

See Also

http://www.ivisualizations.com

Examples

```
###  example1 from iris data

## preparation of histogram-valued variables (k= 3 bins)

Sepal.LengthHistogram=PrepHistogram(X=iris[,1],Z=iris[,5])$Vhistogram
Sepal.WidthHistogram=PrepHistogram(X=iris[,2],Z=iris[,5])$Vhistogram
Petal.LengthHistogram=PrepHistogram(X=iris[,3],Z=iris[,5])$Vhistogram
Petal.WidthHistogram=PrepHistogram(X=iris[,4],Z=iris[,5])$Vhistogram

############################     Hitsogram PCA   ################################
HistPCA(Variable=list(Sepal.LengthHistogram,Sepal.WidthHistogram,
        Petal.LengthHistogram,Petal.WidthHistogram),
        Row.names=names(table(iris[,5])),
Col.names=colnames(iris)[1:4])


###  example2 from iris data

## preparation of histogram-valued variables (k= 4 bins)

Sepal.LengthHistogram=PrepHistogram(X=iris[,1],Z=iris[,5],k=2)$Vhistogram
Sepal.WidthHistogram=PrepHistogram(X=iris[,2],Z=iris[,5],k=2)$Vhistogram
Petal.LengthHistogram=PrepHistogram(X=iris[,3],Z=iris[,5],k=2)$Vhistogram
Petal.WidthHistogram=PrepHistogram(X=iris[,4],Z=iris[,5],k=2)$Vhistogram


############################     Hitsogram PCA   ################################
```

```
HistPCA(Variable=list(Sepal.LengthHistogram,Sepal.WidthHistogram,
Petal.LengthHistogram,Petal.WidthHistogram),
Row.names=names(table(iris[,5])),Col.names=colnames(iris)[1:4])
```

`Ridi` *Mantel Hansen's Scores computation using cumulative distribution function*

Description

This function computes Mantel Hansen's scores

Usage

```
Ridi(X)
```

Arguments

`X`	histogram-valued variable built via PrepHistogram function.

Details

This function computes Mantel Hansen's scores

Value

Returns a list including :

`p`	bins number of histograms
`entree`	Means table of Mantel's Hansen scores
`Ridit`	Mantel's Hansen scores

Note

Perform scores

Author(s)

Brahim Brahim Brahim Brahim <brahim.brahim@bigdatavisualizations.com> and Sun Makosso-Kallyth <makosso.sun@gmail.com>

References

Makosso-Kallyth, Sun; Diday, Edwin. Adaptation of interval PCA to symbolic histogram variables. Advances in Data Analysis and Classification. Volume 6. n 2. 2012. pages 147-159. Springer.

Examples

```
Hist1=PrepHistogram(X=iris[,1],Z=iris[,5])$Vhistogram
Hist2=PrepHistogram(X=iris[,2],Z=iris[,5])$Vhistogram
Hist3=PrepHistogram(X=iris[,3],Z=iris[,5])$Vhistogram
Hist4=PrepHistogram(X=iris[,4],Z=iris[,5])$Vhistogram

s1=Ridi(Hist1)$Ridit
s2=Ridi(Hist2)$Ridit
s3=Ridi(Hist3)$Ridit
s4=Ridi(Hist4)$Ridit
```

Ridi2 *Standardized Mantel Hansen's Scores computation using cumulative distribution function*

Description

This function computes Standardized Mantel Hansen's scores

Usage

```
Ridi2(X)
```

Arguments

X	histogram-valued variable built via PrepHistogram function.

Details

This function computes Standardized Mantel Hansen's scores

Value

Returns a list including :

p	bins number of histograms
entree	Means table of Mantel's Hansen scores
Ridit	Mantel's Hansen scores

Note

Computes Standardized Mantel Hansen's scores

Author(s)

Brahim Brahim Brahim Brahim <brahim.brahim@bigdatavisualizations.com> and Sun Makosso-Kallyth <makosso.sun@gmail.com>

References

Makosso-Kallyth, Sun; Diday, Edwin. Adaptation of interval PCA to symbolic histogram variables. Advances in Data Analysis and Classification. Volume 6. n 2. 2012. pages 147-159. Springer.

Examples

```
Hist1=PrepHistogram(X=iris[,1],Z=iris[,5])$Vhistogram
Hist2=PrepHistogram(X=iris[,2],Z=iris[,5])$Vhistogram
Hist3=PrepHistogram(X=iris[,3],Z=iris[,5])$Vhistogram
Hist4=PrepHistogram(X=iris[,4],Z=iris[,5])$Vhistogram

ss1=Ridi2(Hist1)$Ridit
ss2=Ridi2(Hist2)$Ridit
ss3=Ridi2(Hist3)$Ridit
ss4=Ridi2(Hist4)$Ridit
```

Ridi3 *Normalized Mantel Hansen's Scores computation using cumulative distribution function*

Description

This function computes Normalized Mantel Hansen's scores

Usage

```
Ridi3(X)
```

Arguments

X	histogram-valued variable built via PrepHistogram function.

Details

This function computes Normalized Mantel Hansen's scores

Value

Returns a list including :

p	bins number of histograms
entree	Means table of Mantel's Hansen scores
Ridit	Mantel's Hansen scores

Note

Computes Normalized Mantel Hansen's scores

Author(s)

Brahim Brahim Brahim Brahim <brahim.brahim@bigdatavisualizations.com> and Sun Makosso-Kallyth <makosso.sun@gmail.com>

References

Makosso-Kallyth, Sun; Diday, Edwin. Adaptation of interval PCA to symbolic histogram variables. Advances in Data Analysis and Classification. Volume 6. n 2. 2012. pages 147-159. Springer.

Examples

```
Hist1=PrepHistogram(X=iris[,1],Z=iris[,5])$Vhistogram
Hist2=PrepHistogram(X=iris[,2],Z=iris[,5])$Vhistogram
Hist3=PrepHistogram(X=iris[,3],Z=iris[,5])$Vhistogram
Hist4=PrepHistogram(X=iris[,4],Z=iris[,5])$Vhistogram

sss1=Ridi3(Hist1)$Ridit
sss2=Ridi3(Hist2)$Ridit
sss3=Ridi3(Hist3)$Ridit
sss4=Ridi3(Hist4)$Ridit
```

Visu *Visu*

Description

This function plots a scatterplot of histogram variables, using the R package ggplot2.

Usage

```
Visu(PC, Row.names = NULL, labs = NULL, axes = c(1, 2),size = 4)
```

Arguments

PC	PCA data frame scores that contains (PCxmin, PCxmax, PCymin, PCymax).
Row.names	Row names of concepts.
labs	set the names of the axes.
axes	a length 2 vector specifying the components to plot.
size	Default value

Author(s)

Brahim Brahim Brahim Brahim <brahim.brahim@bigdatavisualizations.com> and Sun Makosso-Kallyth <makosso.sun@gmail.com>

References

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

Le-Rademacher J., Billard L. (2013). Principal component histograms from interval-valued observations, Computational Statistics, v.28 n.5, p.2117-2138.

Makosso-Kallyth S. and Diday E. (2012). Adaptation of interval PCA to symbolic histogram variables, Advances in Data Analysis and Classification July, Volume 6, Issue 2, pp 147-159.

Examples

```
data(movies)
ab = movies
ab = na.omit(ab)
Action = subset(ab,Action==1)
Action$genre = as.factor("Action")
Drama = subset(ab,Drama==1)
Drama$genre = as.factor("Drama")

Animation = subset(ab,Animation==1)
Animation$genre = as.factor("Animation")

Comedy = subset(ab,Comedy==1)
Comedy$genre = as.factor("Comedy")

Documentary = subset(ab,Documentary ==1)
Documentary $genre = as.factor("Documentary")


Romance = subset(ab,Romance==1)
Romance$genre = as.factor("Romance")

Short = subset(ab,Short==1)
Short$genre = as.factor("Short")

 ab = rbind(Action,Drama,Animation,Comedy,Documentary,Romance,Short)
```

```
 Hist1=PrepHistogram(X=sapply(ab[,3],unlist),Z=ab[,25],k=5)$Vhistogram
Hist2=PrepHistogram(X=sapply(ab[,4],unlist),Z=ab[,25],k=5)$Vhistogram
 Hist3=PrepHistogram(X=sapply(ab[,5],unlist),Z=ab[,25],k=5)$Vhistogram
Hist4=PrepHistogram(X=sapply(ab[,6],unlist),Z=ab[,25],k=5)$Vhistogram
 Hist5=PrepHistogram(X=sapply(ab[,7],unlist),Z=ab[,25],k=5)$Vhistogram

 ss1=Ridi(Hist1)$Ridit
 ss2=Ridi(Hist2)$Ridit
 ss3=Ridi(Hist3)$Ridit
 ss4=Ridi(Hist4)$Ridit
 ss5=Ridi(Hist5)$Ridit


HistPCA(list(Hist1,Hist2,Hist3,Hist4,Hist5),score=list(ss1,ss2,ss3,ss4,ss5))

res_pca=HistPCA(list(Hist1,Hist2,Hist3,Hist4,Hist5),score=list(ss1,ss2,ss3,ss4,ss5))

 Visu(res_pca$PCinterval)
```

Index