Package ‘FPDclustering’

January 28, 2020

Type Package

Title PD-Clustering and Factor PD-Clustering

Version 1.4.1

Date 2020-01-28

Author Cristina Tortora [aut, cre, cph], Noe Vidales [aut], Francesco Palumbo [aut], and Paul D. McNicholas [fnd]

Maintainer Cristina Tortora <grikris1@gmail.com>

Description Probabilistic distance clustering (PD-clustering) is an iterative, distribution free, probabilistic clustering method. PD-clustering assigns units to a cluster according to their probability of membership, under the constraint that the product of the probability and the distance of each point to any cluster centre is a constant. PD-clustering is a flexible method that can be used with non-spherical clusters, outliers, or noisy data. PDQ is an extension of the algorithm for clusters of different size. GPDC and TPDC uses a dissimilarity measure based on densities. Factor PD-clustering (FPDC) is a recently proposed factor clustering method that involves a linear transformation of variables and a cluster optimizing the PD-clustering criterion. It works on high dimensional datasets.

Depends ThreeWay ,mvtnorm,R (>= 3.5)

Imports ExPosition,cluster,rootSolve

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2020-01-28 20:50:05 UTC

R topics documented:

`ais` *Australian institute of sport data*

Description

Data obtained to study sex, sport and body-size dependency of hematology in highly trained athletes.

Usage

```
data(asymmetric3)
```

Format

A data frame with 202 observations and 13 variable.

rcc red blood cell count, in

wcc while blood cell count, in per liter

hc hematocrit, percent

hg hemaglobin concentration, in g per decaliter

ferr plasma ferritins, ng

bmi Body mass index, kg

ssf sum of skin folds

pcBfat percent Body fat

lbm lean body mass, kg

ht height, cm

wt weight, kg

sex a factor with levels f m

sport a factor with levels B_Ball Field Gym Netball Row Swim T_400m T_Sprnt Tennis W_Polo

Source

R package DAAG

References

Telford, R.D. and Cunningham, R.B. 1991. Sex, sport and body-size dependency of hematology in highly trained athletes. Medicine and Science in Sports and Exercise 23: 788-794.

Examples

```
data(ais)
pairs(ais[,1:11],col=ais$sex)
```

asymmetric20 *Asymmetric data set shape=20*

Description

Each cluster has been generated according to a multivariate asymmetric Gaussian distribution, with shape=20, covariance matrix equal to the identity matrix and randomly generated centres.

Usage

```
data(asymmetric20)
```

Format

A data frame with 800 observations on the following 101 variables. The first variable is the membership.

Source

Generated with R using the package sn (The skew-normal and skew-t distributions), function rsn

Examples

```
data(asymmetric20)
plot(asymmetric20[,2:3])
```

asymmetric3 *Asymmetric data set shape=3*

Description

Each cluster has been generated according to a multivariate asymmetric Gaussian distribution, with shape=3, covariance matrix equal to the identity matrix and randomly generated centres.

Usage

```
data(asymmetric3)
```

Format

A data frame with 800 observations on 101 variables. The first variable is the membership labels.

Source

Generated with R using the package sn (The skew-normal and skew-t distributions), function rsn

Examples

```
data(asymmetric3)
plot(asymmetric3[,2:3])
```

FPDC *Factor probabilistic distance clustering*

Description

An implementation of FPDC, a probabilistic factor clustering algorithm that involves a linear transformation of variables and a cluster optimizing the PD-clustering criterion

Usage

```
FPDC(data = NULL, k = 2, nf = 2, nu = 2)
```

Arguments

data	A matrix or data frame such that rows correspond to observations and columns correspond to variables.
k	A numerical parameter giving the number of clusters
nf	A numerical parameter giving the number of factors for variables
nu	A numerical parameter giving the number of factors for units

Value

A list with components

label	A vector of integers indicating the cluster membership for each unit
centers	A matrix of cluster centers
probability	A matrix of probability of each point belonging to each cluster
JDF	The value of the Joint distance function
iter	The number of iterations
explained	The explained variability

Author(s)

Cristina Tortora and Paul D. McNicholas

References

Tortora, C., M. Gettler Summa, M. Marino, and F. Palumbo. *Factor probabilistic distance clustering (fpdc): a new clustering method for high dimensional data sets*. Advanced in Data Analysis and Classification, 10(4), 441-464, 2016. doi:10.1007/s11634-015-0219-5.

Tortora C., Gettler Summa M., and Palumbo F.. Factor pd-clustering. In Lausen et al., editor, *Algorithms from and for Nature and Life, Studies in Classification*, Data Analysis, and Knowledge Organization DOI 10.1007/978-3-319-00035-011, 115-123, 2013.

Tortora C., *Non-hierarchical clustering methods on factorial subspaces*, 2012.

See Also

`PDclust`

Examples

```
## Not run:
# Asymmetric data set clustering example (with shape=3).
data('asymmetric3')
x<-asymmetric3[,-1]
fpdas3=FPDC(x,4,3,3)
table(asymmetric3[,1],fpdas3$label)
Silh(fpdas3$probability)

## End(Not run)

## Not run:
# Asymmetric data set clustering example (with shape=20).
data('asymmetric20')
x<-asymmetric20[,-1]
fpdas20=FPDC(x,4,3,3)
table(asymmetric20[,1],fpdas20$label)
Silh(fpdas20$probability)

## End(Not run)

## Not run:
# Clustering example with outliers.
data('outliers')
x<-outliers[,-1]
fpdout=FPDC(x,4,5,4)
table(outliers[,1],fpdout$label)
Silh(fpdout$probability)

## End(Not run)
```

GPDC *Gaussian PD-Clustering*

Description

An implementation of Gaussian PD-Clustering GPDC, an extention of PD-clustering adjusted for cluster size that uses a dissimilarity measure based on the Gaussian density.

Usage

```
GPDC(data=NULL,k=2,method="kmedoids", nr=5,iter=100)
```

Arguments

`data`	A matrix or data frame such that rows correspond to observations and columns correspond to variables.
`k`	A numerical parameter giving the number of clusters
`method`	A parameter that selects center starts. Options available are random, kmedoid, and PDclust
`nr`	Number of random starts
`iter`	Maximum number of iterations

Value

A list with components

`label`	A vector of integers indicating the cluster membership for each unit
`centers`	A matrix of cluster means
`sigma`	A list of K elements, with the variance-covariance matrix per cluster
`probability`	A matrix of probability of each point belonging to each cluster
`JDF`	The value of the Joint distance function
`iter`	The number of iterations

Author(s)

Cristina Tortora and Francesco Palumbo

References

Tortora C., McNicholas P.D., and Palumbo F. *A probabilistic distance clustering algorithm using Gaussian and Student-t multivariate density distributions.* SN Computer Science (to appear) 2020.

C. Rainey, C. Tortora and F.Palumbo. *A parametric version of probabilistic distance clustering*. In: Greselin F., Deldossi L., Bagnato L., Vichi M. (eds) Statistical Learning of Complex Data. CLADAG 2017. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Cham, 33-43 2019. doi.org/10.1007/978-3-030-21140-0_4

See Also

`PDclust`,`PDQ`

Examples

```
data(ais)
dataSEL=ais[,c(10,3,5,8)]
res=GPDC(dataSEL,k=2,method = "kmedoids")
table(res$label,ais$sex)
pairs(dataSEL,col=res$label,pch=res$label)
```

`outliers` *Data set with outliers*

Description

Each cluster has been generated according to a multivariate Gaussian distribution, with centers c randomly generated. For each cluster, 20% of uniform distributed outliers have been generated at a distance included in max(x-c) and max(x-c)+5 form the center.

Usage

```
data(outliers)
```

Format

A data frame with 960 observations on the following 101 variables. The first variable corresponds to the membership

Source

generated with R

Examples

```
data(outliers)
 plot(outliers[,2:3])
```

PDclust *Probabilistic Distance Clustering*

Description

Probabilistic distance clustering (PD-clustering) is an iterative, distribution free, probabilistic clustering method. PD clustering assigns units to a cluster according to their probability of membership, under the constraint that the product of the probability and the distance of each point to any cluster centre is a constant.

Usage

```
PDclust(data = NULL, k = 2)
```

Arguments

`data`	A matrix or data frame such that rows correspond to observations and columns correspond to variables.
`k`	A numerical parameter giving the number of clusters

Value

A list with components

`label`	A vector of integers indicating the cluster membership for each unit
`centers`	A matrix of cluster centers
`probability`	A matrix of probability of each point belonging to each cluster
`JDF`	The value of the Joint distance function
`iter`	The number of iterations

Author(s)

Cristina Tortora and Paul D. McNicholas

References

Ben-Israel C. and Iyigun C. Probabilistic D-Clustering. *Journal of Classification*, **25**(1), 5–26, 2008.

Examples

```
#Normally generated clusters
c1 = c(+2,+2,2,2)
c2 = c(-2,-2,-2,-2)
c3 = c(-3,3,-3,3)
n=200
```

```
x1 = cbind(rnorm(n, c1[1]), rnorm(n, c1[2]), rnorm(n, c1[3]), rnorm(n, c1[4]) )
x2 = cbind(rnorm(n, c2[1]), rnorm(n, c2[2]),rnorm(n, c2[3]), rnorm(n, c2[4]) )
x3 = cbind(rnorm(n, c3[1]), rnorm(n, c3[2]),rnorm(n, c3[3]), rnorm(n, c3[4]) )
x = rbind(x1,x2,x3)
pdn=PDclust(x,3)
plot(x[,1:2],col=pdn$label)
plot(x[,3:4],col=pdn$label)
```

`PDQ` *Probabilistic Clustering Adjusted for Cluster Size*

Description

An implementation of PDQ, a probabilistic distance clustering algorithm that involves optimizing the PD-clustering criterion with the option of Euclidean and Chi as dissimilarity measurments.

Usage

```
PDQ(data=NULL,K=2,method="random", distance="euc", cent=NULL)
```

Arguments

`data` A matrix or data frame such that rows correspond to observations and columns correspond to variables.

`K` A numerical parameter giving the number of clusters

`method` A parameter that selects center starts. Options available are random, kmedoid, and center(user inputs center starts)

`distance` A parameter that selects the distance measure used. Options available are Eucledean euc and chi square chi

`cent` User inputed centers if method selected is "random"

Value

A list with components

`label` A vector of integers indicating the cluster membership for each unit

`centers` A matrix of cluster centers

`probability` A matrix of probability of each point belonging to each cluster

`JDF` The value of the Joint distance function

`iter` The number of iterations

`jdfvector` collection of all jdf calculations at each iteration

Author(s)

Cristina Tortora and Noe Vidales

References

Iyigun, Cem, and Adi Ben-Israel. *Probabilistic distance clustering adjusted for cluster size.* Probability in the Engineering and Informational Sciences 22.4 (2008): 603-621. doi.org/10.1017/S0269964808000351.

See Also

`PDclust`

Examples

```
# Gaussian Generated Data  no  overlap
x<-rmvnorm(100, mean=c(1,5,10), sigma=diag(1,3))
y<-rmvnorm(100, mean=c(4,8,13), sigma=diag(1,3))
data<-rbind(x,y)
pdq1=PDQ(data,2,method="random",distance="euc")
table(rep(c(2,1),each=100),pdq1$label)
Silh(pdq1$probability)



# Gaussian Generated Data with  overlap
x2<-rmvnorm(100, mean=c(1,5,10), sigma=diag(1,3))
y2<-rmvnorm(100, mean=c(2,6,11), sigma=diag(1,3))
data2<-rbind(x2,y2)
pdq2=PDQ(data2,2,method="random",distance="euc")
table(rep(c(1,2),each=100),pdq2$label)
Silh(pdq2$probability)
```

`Silh` *Probabilistic silhouette plot*

Description

Graphical tool to see how well each point belongs to the cluster.

Usage

```
Silh(p)
```

Arguments

`p` A matrix of probabilities such that rows correspond to observations and columns correspond to clusters.

Details

The probabilistic silhouettes are an adaptation of the ones proposed by Menardi(2011) according to the following formula:

$$dbs_i = (log(p_{im_k}/p_{im_1}))/max_i|log(p_{im_k}/p_{im_1})|$$

,

where m_k is such that x_i belongs to cluster k and m_1 is such that p_{im_1} is maximum for m different fromm_k.

Value

Probabilistic silhouette plot

Author(s)

Cristina Tortora

References

Menardi G. Density-based Silhouette diagnostics for clustering methods.*Statistics and Computing*, **21**, 295-308, 2011.

Examples

```
## Not run:
# Asymmetric data set silhouette example (with shape=3).
data('asymmetric3')
x<-asymmetric3[,-1]
fpdas3=FPDC(x,4,3,3)
Silh(fpdas3$probability)

## End(Not run)

## Not run:
# Asymmetric data set shiluette example (with shape=20).
data('asymmetric20')
x<-asymmetric20[,-1]
fpdas20=FPDC(x,4,3,3)
Silh(fpdas20$probability)

## End(Not run)

## Not run:
# Shiluette example with outliers.
data('outliers')
x<-outliers[,-1]
fpdout=FPDC(x,4,4,3)
Silh(fpdout$probability)

## End(Not run)
```

TPDC Student-t PD-Clustering

Description

An implementation of Student-t PD-Clustering TPDC, an extention of PD-clustering adjusted for cluster size that uses a dissimilarity measure based on the multivariate Student-t density.

Usage

```
TPDC(data=NULL,k=2,method="kmedoids", nr=5,iter=100)
```

Arguments

data	A matrix or data frame such that rows correspond to observations and columns correspond to variables.
k	A numerical parameter giving the number of clusters
method	A parameter that selects center starts. Options available are random, kmedoid, and PDclust
nr	Number of random starts
iter	Maximum number of iterations

Value

A list with components

label	A vector of integers indicating the cluster membership for each unit
centers	A matrix of cluster means
sigma	A list of K elements, with the variance-covariance matrix per cluster
df	A vector of K degrees of freedom
probability	A matrix of probability of each point belonging to each cluster
JDF	The value of the Joint distance function
iter	The number of iterations

Author(s)

Cristina Tortora and Francesco Palumbo

References

Tortora C., McNicholas P.D., and Palumbo F. *A probabilistic distance clustering algorithm using Gaussian and Student-t multivariate density distributions.* SN Computer Science (to appear) 2020.

C. Rainey, C. Tortora and F.Palumbo. *A parametric version of probabilistic distance clustering*. In: Greselin F., Deldossi L., Bagnato L., Vichi M. (eds) Statistical Learning of Complex Data. CLADAG 2017. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Cham, 33-43 2019. doi.org/10.1007/978-3-030-21140-0_4

See Also

`PDclust`,`PDQ`

Examples

```
data(ais)
dataSEL=ais[,c(10,3,5,8)]
res=TPDC(dataSEL,k=2,method = "kmedoids")
table(res$label,ais$sex)
pairs(dataSEL,col=res$label,pch=res$label)
```

`TuckerFactors` *Choice of the number of Tucker 3 factors*

Description

An empirical way of choosing the number of factors. The algorithm returns a graph and a table representing the explained variability varying the number of factors.

Usage

```
TuckerFactors(data = NULL, nc = 2)
```

Arguments

`data`	A matrix or data frame such that rows correspond to observations and columns correspond to variables.
`nc`	A numerical parameter giving the number of clusters

Value

A table containing the explained variability varying the number of factors for units (column) and for variables (row) and a plot

Author(s)

Cristina Tortora

References

Kiers H, Kinderen A. A fast method for choosing the numbers of components in Tucker3 analysis.*British Journal of Mathematical and Statistical Psychology*, **56**(1), 119-125, 2003.

Kroonenberg P. *Applied Multiway Data Analysis*. Ebooks Corporation, Hoboken, New Jersey, 2008.

Tortora C., Gettler Summa M., and Palumbo F.. Factor pd-clustering. In Lausen et al., editor, *Algorithms from and for Nature and Life, Studies in Classification*, Data Analysis, and Knowledge Organization DOI 10.1007/978-3-319-00035-011, 115-123, 2013.

See Also

T3

Examples

```
## Not run:
# Asymmetric data set example (with shape=3).
data('asymmetric3')
xp=TuckerFactors(asymmetric3[,-1], nc = 4)


## End(Not run)

## Not run:
# Asymmetric data set example (with shape=20).
data('asymmetric20')
xp=TuckerFactors(asymmetric20[,-1], nc = 4)

## End(Not run)
```

Index