Package 'Kmedians'

December 18, 2023

Type Package

Title K-Medians

Version 2.2.0

Description Online, Semi-online, and Offline K-medians algorithms are given. For both methods, the algorithms can be initialized randomly or with the help of a robust hierarchical clustering. The number of clusters can be selected with the help of a penalized criterion. We provide functions to provide robust clustering. Function gen_K() enables to generate a sample of data following a contaminated Gaussian mixture. Functions Kmedians() and Kmeans() consists in a K-median and a K-means algorithms while Kplot() enables to produce graph for both methods.
Cardot, H., Cenac, P. and Zitt, P-A. (2013). ``Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm". Bernoulli, 19, 18-43. <doi:10.3150/11-BEJ390>.
Cardot, H. and Godichon-Baggioni, A. (2017). ``Fast Estimation of the Median Covariation Matrix with Application to Online Robust Principal Components Analysis". Test, 26(3), 461-480 <doi:10.1007/s11749-016-0519-x>.
Godichon-Baggioni, A. and Surendran, S. ``A penalized criterion for selecting the number of clusters for K-medians" <arXiv:2209.03597>
Vardi, Y. and Zhang, C.-H. (2000). ``The multivariate L1-median and associated data depth". Proc. Natl. Acad. Sci. USA, 97(4):1423-1426. <doi:10.1073/pnas.97.4.1423>.

License GPL (>= 2)

Encoding UTF-8

Imports foreach, doParallel,parallel, genieclust, Gmedian,mvtnorm, capushe, ggplot2, reshape2

RoxygenNote 7.1.2

NeedsCompilation no

Author Antoine Godichon-Baggioni [aut, cre, cph], Sobihan Surendran [aut]

Maintainer Antoine Godichon-Baggioni <antoine.godichon_baggioni@upmc.fr>

Repository CRAN

Date/Publication 2023-12-18 13:40:05 UTC

R topics documented:

`Kmedians-package` *K-Medians*

Description

We provide functions to provide robust clustering. Function `gen_K` enables to generate a sample of data following a contaminated Gaussian mixture. Functions `Kmedians` and `Kmeans` consists in a K-median and a K-means algorithms while `Kplot` enables to produce graph for both methods.

Author(s)

Antoine Godichon-Baggioni [aut, cre, cph], Sobihan Surendran [aut]

Maintainer: Antoine Godichon-Baggioni <antoine.godichon_baggioni@upmc.fr>

References

Cardot, H., Cenac, P. and Zitt, P-A. (2013). Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19, 18-43.

Cardot, H. and Godichon-Baggioni, A. (2017). Fast Estimation of the Median Covariation Matrix with Application to Online Robust Principal Components Analysis. *Test*, 26(3), 461-480

Godichon-Baggioni, A. and Surendran, S. A penalized criterion for selecting the number of clusters for K-medians. *arxiv.org/abs/2209.03597*

Vardi, Y. and Zhang, C.-H. (2000). The multivariate L1-median and associated data depth. *Proc. Natl. Acad. Sci. USA*, 97(4):1423-1426.

gen_K *gen_K*

Description

Generate a sample of a Gaussian Mixture Model whose centers are generate randomly on a sphere of radius `radius`.

Usage

```
gen_K(n=500,d=5,K=3,pcont=0,df=1,
      cont="Student",min=-5,max=5,radius=5)
```

Arguments

`n`	A positive integer giving the number of data per cluster. Default is `500`.
`d`	A positive integer giving the dimension. Default is `5`.
`K`	A positive integer giving the number of clusters. Default is `3`.
`pcont`	A scalar between `0` and `1` giving the proportion of contaminated data.
`df`	A positive integer giving the degrees of freedom of the law of the contaminated data if `cont='Student'`. Default is `1`.
`cont`	The law of the contaminated data. Can be `'Student'` (default) and `'Unif'`.
`min`	A scalar giving the lower bound of the uniform law if `cont='Unif'`. Default is `-5`.
`max`	A scalar giving the upper bound of the uniform law if `cont='Unif'`. Default is `5`.
`radius`	The radius of the sphere on each the centers of the class are generated. Default is `5`.

Value

A list with:

`X`	A numerical matrix giving the generated data.
`cluster`	An character vector specifying the true classification.

See Also

See also `Kmedians` and `Kmeans`.

Examples

```
n <- 500
K <- 3
pcont <- 0.2
ech <- gen_K(n=n,K=K,pcont=pcont)
X=ech$X
```

Kmeans *Kmeans*

Description

A K-means algorithm.

Usage

```
Kmeans(X,nclust=1:15,ninit=1,niter=20,par=TRUE)
```

Arguments

X	A numerical matrix giving the data.
nclust	A vector of positive integers giving the possible numbers of clusters. Default is 1:15.
ninit	A non negative integer giving the number of random initializations. Default is 1.
niter	A positive integer giving the number of iterations for the EM algorirthms. Default is 20.
par	A logical argument telling if the parallelization of the algorithm is allowed. Default is TRUE.

Value

A list with:

bestresults	A list giving all the results for the clustering selected by 'capushe'.
allresults	A list containing all the results.
SE	A vector giving the Sum of Errors for each considered number of clusters.
cap	The results given by the function 'capushe' if nclust is of length larger than 10.
Ksel	An integer giving the number of clusters selected by capushe if nclust is of length larger than 10.
data	A numerical matrix giving the data.
nclust	A vector of positive integers giving the considered numbers of clusters.

For the lists bestresult and allresults:

cluster	A vector of positive integers giving the clustering.
centers	A numerical matrix giving the centers of the clusteres.
SE	An integer giving the Sum of Errors.

See Also

See also Kmedians, Kplot and gen_K.

Examples

```
## Not run:
n <- 500
K <- 3
pcont <- 0.2
ech <- gen_K(n=n,K=K,pcont=pcont)
X <-ech$X
res <- Kmeans(X,par=FALSE)
Kplot(res)

## End(Not run)
```

`Kmedians` *Kmedians*

Description

K-medians algorithms.

Usage

```
Kmedians(X,nclust=1:15,ninit=0,niter=20,
         method='Offline', init=TRUE,par=TRUE)
```

Arguments

`X`	A numerical matrix giving the data.
`nclust`	A vector of positive integers giving the possible numbers of clusters. Default is `1:15`.
`ninit`	A non negative integer giving the number of random initializations. Default is `0`.
`niter`	A positive integer giving the number of iterations for the EM algorirthms. Default is `20`.
`method`	The selected method for the K-medians algorithm. Can be `'Offline'` (default), `'Semi-Online'` or `'Online'`.
`init`	A logical argument telling if the function `'genie'` is used for initializing the algorithm. Default is `TRUE`.
`par`	A logical argument telling if the parallelization of the algorithm is allowed. Default is `TRUE`.

Value

A list with:

`bestresults`	A list giving all the results for the clustering selected by `'capushe'`.
`allresults`	A list containing all the results.

SE	A vector giving the Sum of Errors for each considered number of clusters.
cap	The results given by the function 'capushe' if nclust is of length larger than 10.
Ksel	An integer giving the number of clusters selected by 'capushe' if nclust is of length larger than 10.
data	A numerical matrix giving the data.
nclust	A vector of positive integers giving the considered numbers of clusters.

For the lists bestresult and allresults:

cluster	A vector of positive integers giving the clustering.
centers	A numerical matrix giving the centers of the clusteres.
SE	An integer giving the Sum of Errors.

References

Godichon-Baggioni, A. and Surendran, S. A penalized criterion for selecting the number of clusters for K-medians. *arxiv.org/abs/2209.03597*

See Also

See also Kmeans, Kplot and gen_K.

Examples

```
## Not run:
n <- 500
K <- 3
pcont <- 0.2
ech <- gen_K(n=n,K=K,pcont=pcont)
X <-ech$X
res <- Kmedians(X,par=FALSE)
Kplot(res)

## End(Not run)
```

Kplot *Kplot*

Description

A plot function for K-medians and K-means

Usage

```
Kplot(a,propplot=0.95,graph=c('Two_Dim','Capushe','Profiles','SE','Criterion'),
      bestresult=TRUE,Ksel=FALSE,bycluster=TRUE)
```

Arguments

a	Output from Kmedians or Kmeans.
propplot	A scalar between 0 and 1 giving the propotion of data considered for the different graphs.
graph	A string specifying the type of graph requested. Default is c('Two_Dim','Capushe','Profiles','SE',
bestresult	A logical indicating if the graphs must be done for the result chosen by the selected criterion. Default is TRUE.
Ksel	A logical or positive integer giving the chosen number of clusters for each the graphs should be drawn.
bycluster	A logical indicating if the data selected for 'Two_Dim' and 'Profiles' graphs should be selected by cluster or not. Default is TRUE.

Value

No return value.

See Also

See also Kmedians and Kmeans.

Examples

```
## Not run:
n <- 500
K <- 3
pcont <- 0.2
ech <- gen_K(n=n,K=K,pcont=pcont)
X <-ech$X
res <- Kmedians(X,par=FALSE)
Kplot(res)

## End(Not run)
```

Index