Package 'MSclust'

April 22, 2024

Type Package

Title Multiple-Scaled Clustering

Version 1.0.4

Date 2024-04-21

Maintainer Cristina Tortora <grikris1@gmail.com>

Description Model based clustering using the multivariate multiple Scaled t (MST) and multivariate multiple scaled contaminated normal (MSCN) distributions. The MST is an extension of the multivariate Student-t distribution to include flexible tail behaviors, Forbes, F. & Wraith, D. (2014) <doi:10.1007/s11222-013-9414-4>. The MSCN represents a heavy-tailed generalization of the multivariate normal (MN) distribution to model elliptical contoured scatters in the presence of mild outliers (also referred to as ``bad'' points) and automatically detect bad points, Punzo, A. & Tortora, C. (2021) <doi:10.1177/1471082X19890935>.

Depends R (>= 3.5)

Imports gtools, Matrix, mclust, mnormt, mvtnorm, psych, cluster, ggplot2, GGally

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2024-04-22 03:22:43 UTC

Author Cristina Tortora [aut, cre, cph] (<https://orcid.org/0000-0001-8351-3730>), Antonio Punzo [aut] (<https://orcid.org/0000-0001-7742-1821>), Louis Tran [aut]

R topics documented:

`mscn` *Mixtures of Multiple Scaled Contaminated Normal Distributions.*

Description

Fits a mixture of multiple scaled contaminated normal distributions to the given data.

Usage

```
mscn(X,k,ini="km",sz=NULL,al=c(0.5,0.99),eta.min=1.01,m="BFGS",stop=c(10^-5,200),VB=FALSE)
```

Arguments

`X`	A matrix or data frame such that rows correspond to observations and columns correspond to variables.
`k`	The number of clusters.
`ini`	Using kmeans by default or `"pam"` for partition around medoids, `"mclust"` for Gaussian mixture models, `"random.soft"` or `"random.hard"` for random or manual; if `"manual"`, a partition (`sz`) must be provided.
`sz`	If initialization is `"manual"`, this matrix contains the starting values for `z`.
`al`	2-dimensional vector containing minimum and maximum proportion of good points in each group for the contaminated normal distribution.
`eta.min`	Minimum value for inflation parameter for the covariance matrix for the bad points.
`m`	Method for the optimization of the eigenvector matrix, see `optim` for other options.
`stop`	2-dimensional vector with the Aitken criterion stopping rule and maximum number of iterations.
`VB`	If `TRUE`, tracing information on the progress of the optimization is produced; see `optim` for details and plotting of the log-likelihood versus iterations.

Value

`X`	Data used for clustering.
`n`	The number of observations in the data.
`d`	The number of features in the data.
`k`	Value corresponding to the number of components.
`cluster`	Vector of group membership as determined by the model.

detect	Detect if the point is bad or not per each principal component given the cluster membership.
npar	The number of parameters.
mu	Either a vector of length d, representing the mean value, or (except for rmscn) a matrix whose rows represent different mean vectors; if it is a matrix, its dimensions must match those of x.
Lambda	Orthogonal matrix whose columns are the normalized eigenvectors of Sigma.
Gamma	Diagonal matrix of the eigenvalues of Sigma.
Sigma	A symmetric positive-definite matrix representing the scale matrix of the distribution.
alpha	Proportion of good observations.
eta	Degree of contamination.
z	The component membership of each observations.
v	The indicator if an observation is good or bad with respect to each dimension; 1 is good, and 0 means bad.
weight	The matrix of the expected value of the characteristic weights; corespond to the value of v+(1-v)/eta.
iter.stop	The number of iterations until convergence for the model.
loglik	The log-likelihood corresponding to the model.
AIC	The Akaike's Information Criterion of the model.
BIC	The Bayesian Information Criterion of the model.
ICL	The Integrated Completed Likelihood of the model.
KIC	The Kullback Information Criterion of the model.
KICc	The Bias correction of the Kullback Information Criterion of the model.
AWE	The Approximate Weight of Evidence of the model.
AIC3	Another version of Akaike's Information Criterion of the model.
CAIC	The Consistent Akaike's Information Criterion of the model.
AICc	The AIC version which is used when sample size n is small relative to d.
CLC	The Classification Likelihood Criterion of the model.

Author(s)

Cristina Tortora and Antonio Punzo

References

Punzo, A. & Tortora, C. (2021). *Multiple scaled contaminated normal distribution and its application in clustering. Statistical Modelling*, **21**(4): 332–358.

Examples

```
## Not run:
## Not run:
data(sim)
result <- mscn(X = sim, k = 2)
plot(result)
summary(result)
## End(Not run)
## End(Not run)
```

`mst` *Mixture of Multiple Scaled Student-t Distributions*

Description

Fits the mixture of multiple scaled Student-t distributions to the given data.

Usage

```
mst(X,k,ini="km",sz=NULL,df.min=1,dfU="num",frm="dir",m="BFGS",stop=c(10^-5,200),VB=FALSE)
```

Arguments

`X`	A matrix or data frame such that rows correspond to observations and columns correspond to variables.
`k`	The number of clusters.
`ini`	Using kmeans by default or `"pam"` for partition around medoids, `"mclust"` for Gaussian mixture models, `"random.soft"` or `"random.hard"` for random or manual; if `"manual"`, a partition (`sz`) must be provided.
`sz`	If initialization is manual, this matrix contains the starting value for `z`.
`df.min`	Minimum proportion of good points in each group for the contaminated normal distribution.
`dfU`	Criterion to update the degrees of freedom.
`frm`	Direct by default or indirect, technique used to compute the density function.
`m`	Method for the optimization of the eigenvector matrix, see optim for other options.
`stop`	2-dimensional vector with the Aitken criterion stopping rule and Maximum number of iterations.
`VB`	If true, tracing information on the progress of the optimization is produced; see optim() for details and plotting of the log-likelihood versus iterations.

Value

X	Data used for clustering.
n	The number of observations in the data.
d	The number of features in the data.
k	Value corresponding to the number of components.
cluster	Vector of group membership as determined by the model.
detect	Detect if the point is bad or not per each principal component given the cluster membership.
npar	The number of parameters.
mu	Either a vector of length d, representing the mean value, or a matrix whose rows represent different mean vectors; if it is a matrix, its dimensions must match those of x.
Lambda	Orthogonal matrix whose columns are the normalized eigenvectors of Sigma.
Gamma	Diagonal matrix of the eigenvalues of Sigma.
Sigma	A symmetric positive-definite matrix representing the scale matrix of the distribution.
df	vector containing the degrees of freedom for each component.
z	The component membership of each observations.
v	The indicator if an observation is good or bad with respect to each dimension; 1 is good, and 0 means bad.
weight	The matrix of the expected value of the characteristic weights; corespond to the value of v+(1-v)/eta.
iter.stop	The number of iterations until convergence for the model.
loglik	The log-likelihood corresponding to the model.
AIC	The Akaike's Information Criterion of the model.
BIC	The Bayesian Information Criterion of the model.
ICL	The Integrated Completed Likelihood of the model.
KIC	The Kullback Information Criterion of the model.
KICc	The Bias correction of the Kullback Information Criterion of the model.
AWE	The Approximate Weight of Evidence of the model.
AIC3	Another version of Akaike's Information Criterion of the model.
CAIC	The Consistent Akaike's Information Criterion of the model.
AICc	The AIC version which is used when sample size n is small relative to d.
CLC	The Classification Likelihood Criterion of the model.

Author(s)

Cristina Tortora and Antonio Punzo

References

Forbes, F. & Wraith, D. (2014). *A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. Statistics and Computing*, **24**(6), 971–984.

Examples

```
## Not run:
## Not run:
data(sim)
result <- mst(X = sim, k = 2)
plot(result)
## End(Not run)

## End(Not run)
```

`plot.MSclust` *MSclust Plotting*

Description

MSclust Plotting

Usage

```
## S3 method for class 'MSclust'
plot(x, ...)
```

Arguments

`x`	A `MSclust` object.
`...`	Arguments to be passed to methods, such as graphical parameters.

Value

No return value, called to visualize the fitted model's results

Examples

```
## Not run:
## Not run:
data(sim)
result <- mscn(X = sim, k = 2)
plot(result)
## End(Not run)
## End(Not run)
```

rmscn *Multiple Scaled Contaminated Normal Distribution*

Description

Probability density function and pseudo random number generation for the multiple scaled contaminated normal distribution.

Usage

```
  dmscn(x,  mu = NULL, L = NULL, G = NULL, Sigma = NULL, alpha = NULL, eta = NULL)
rmscn(n,d=2,mu=rep(0,d),L=NULL,G=NULL,Sigma=diag(d),alpha=rep(0.99,d),eta=rep(1.01,d))
```

Arguments

x	A matrix or data frame such that rows correspond to observations and columns correspond to variables.
n	The number of random vectors to be generated.
d	A number specifing the dimenstion.
mu	Either a vector of length d, representing the mean value, or (except for rmscn) a matrix whose rows represent different mean vectors; if it is a matrix, its dimensions must match those of x.
L	Lambda diagonal d-dimensional matrix of the eigenvalues of Sigma.
G	Gamma orthogonal d-dimensional matrix whose columns are the normalized eigenvectors of Sigma.
Sigma	A symmetric positive-definite d-dimensional matrix representing the scale matrix of the distribution; a vector of length 1 is also allowed (in this case, d = 1 is set). Identity matrix by default.
alpha	d-dimensional vector containing the proportion of good observations; it must be a number between 0 and 1.
eta	d-dimensional vector containing the degree of contamination; it should be a number greater than 1.

Value

dmscn	returns a vector of density values.
rmscn	returns a matrix of n rows of observations.

Author(s)

Cristina Tortora and Antonio Punzo

References

Punzo, A. & Tortora, C. (2021). *Multiple scaled contaminated normal distribution and its application in clustering. Statistical Modelling*, **21**(4): 332–358.

Examples

```
x <- matrix(c(0,0),1,2)
alpha <- c(0.8,0.6)
eta <- c(2,4)
density <- dmscn(x = x, alpha = alpha, eta = eta)
density

n <- 100
random <- rmscn(n = n, alpha = alpha, eta = eta)
plot(random)
```

`rmst` *Multiple Scaled Student-t Distribution*

Description

Probability density function and pseudo-random number generation for the multiple scaled Student-t distribution.

Usage

```
 dmst(x, mu = NULL, L = NULL, G = NULL, Sigma = NULL, theta = NULL, formula = "direct")
rmst(n,d=2,mu=rep(0,d),L=NULL,G=NULL,Sigma=diag(d),theta=rep(100,d),n.dens="dmnorm")
```

Arguments

`x`	A matrix or data frame such that rows correspond to observations and columns correspond to variables.
`n`	The number of observations to be generated.
`d`	A number specifing the dimension.
`mu`	Either a vector of length `d`, representing the mean value, or (except for `rmst`) a matrix whose rows represent different mean vectors; if it is a matrix, its dimensions must match those of `x`.
`L`	Lambda diagonal `d`-dimensional matrix of the eigenvalues of `Sigma`.
`G`	Gamma orthogonal `d`-dimensional matrix whose columns are the normalized eigenvectors of `Sigma`.
`Sigma`	A symmetric positive-definite `d`-dimensional matrix representing the scale matrix of the distribution; a vector of length 1 is also allowed (in this case, `d = 1` is set). Identity matrix by default.
`theta`	Vector of dimesion `d` containing the degrees of freedom.
`n.dens`	`"dmnorm"` or `"dmvnorm"`, dependening on the way the density of the normal distribution is computed.
`formula`	`"direct"` or `"indirect"`; if `"direct"`, then Equation (5) in Peel & McLachlan (2000), Statistics & Computing is used.

Value

`dmscn`	returns a vector of density values.
`rmscn`	returns a matrix of `n` rows of observations.

Author(s)

Cristina Tortora and Antonio Punzo

References

Punzo, A., & Tortora, C. (2021). *Multiple scaled contaminated normal distribution and its application in clustering. Statistical Modelling*, **21**(4): 332–358.

Forbes, F. & Wraith, D. (2014). *A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. Statistics and Computing*, **24**(6), 971–984.

Examples

```
x <- matrix(c(0,0),1,2)
theta <- c(5,20)
density <- dmst(x = x, theta = theta)
density

n <- 100
random <- rmst(n = n, theta = theta)
plot(random)
```

`sim` *A Mixture of Two Normal Distributions with outliers*

Description

A simulated mixture of two normal distributions with mean (0,0) and (2,4), respectively, the fist covariance matrix has diagonlas equal to 2 and covariance 1.5, the second 0.5 and 0.1 respectively. Moreover, 18 observations have been transformed in outliers. For details, see Punzo, A. & Tortora, C. (2021). *Multiple scaled contaminated normal distribution and its application in clustering. Statistical Modelling*, **21**(4): 332–358.

Usage

```
data(sim)
```

Format

A matrix with 600 observations (rows) and 2 variables (columns). The first 180 rows belong to cluster 1, and the last 420 rows belong to cluster 2.

V1 variable 1.

V2 variable 2.

Source

Punzo, A. & Tortora, C. (2021). *Multiple scaled contaminated normal distribution and its application in clustering. Statistical Modelling*, **21**(4): 332–358.

`summary.MSclust` *Summary for MSclust*

Description

Summarizes main information regarding a `MSclust` object.

Usage

```
## S3 method for class 'MSclust'
summary(object, ...)
```

Arguments

`object`	A `MSclust` object.
`...`	Arguments to be passed to methods, such as graphical parameters.

Details

Information includes clustering table, total outliers, outliers per cluster, mixing proportions, component means and variances.

Value

No return value, called to summarize the fitted model's results

Examples

```
## Not run:
## Not run:
data(sim)
result <- mscn(X = sim, k = 2, initialization = "kmeans", method = "BFGS")
summary(result)
## End(Not run)
## End(Not run)
```

Index