Package ‘MixGHD’

March 4, 2015

Type Package

Title Model Based Clustering, Classification and Discriminant Analysis
   Using the Mixture of Generalized Hyperbolic Distributions

Version 1.4

Date 2015-2-10

Author Cristina Tortora, Ryan P. Browne, Brian C. Franczak and Paul D. McNicholas.

Maintainer Cristina Tortora <ctortora@math.mcmaster.ca>

Description Carries out model-based clustering, classification and discriminant analysis using five different models. The models are all based on the generalized hyperbolic distribution. The first model MGHD is the classical mixture of generalized hyperbolic distributions. The MGHFA is the mixture of generalized hyperbolic factor analyzers for high dimensional data sets. The MSGHD, mixture of multiple scaled generalized hyperbolic distributions. The cMSGHD is a MSGHD with convex contour plots. The MCGHD, mixture of coalesced generalized hyperbolic distributions is a new more flexible model.

Imports Bessel, stats, mvtnorm, ghyp, numDeriv

Depends MASS, mclust, R (>= 3.1.1)

NeedsCompilation no

License GPL (>= 2)

Repository CRAN

Date/Publication 2015-03-04 21:30:29

R topics documented:

MixGHD-package .................................................. 2
bankruptcy ....................................................... 3
DA ............................................................... 4
MCGHD ........................................................ 6
MGHD .......................................................... 7
MGHFA ......................................................... 9
MSGHD ........................................................ 11
rMSGHD ....................................................... 13
Selection ....................................................... 14
sonar ........................................................... 16
MixGHD-package

Description

Carries out model-based clustering using three different models. The models are all based on the generalized hyperbolic distribution. The first model MGHD is the classical mixture of generalized hyperbolic distributions. The MGHFA is the mixture of generalized hyperbolic factor analyzers for high dimensional data sets. The MCGHD, mixture of coalesced generalized hyperbolic distributions is a new more flexible model.

Details

Package: MixGHD
Type: Package
Version: 1.0
Date: 2014-06-19
License: GPL (>=2)

This package contains the function MGHD for model based clustering and classification using the mixture of generalized hyperbolic distributions; the function MGHFA for model based clustering and classification using the mixture of generalized hyperbolic factor analyzers; the function MCGHD for model based clustering using the mixture of coalesced generalized hyperbolic distributions, and some real data sets.

Author(s)

Cristina Tortora, Ryan P. Browne, Brian C. Franczak and Paul D. McNicholas. Maintainer: Cristina Tortora <ctortora@uoguelph.ca>

References


Examples

```r
## model based clustering using the mixture of generalized hyperbolic distributions
data(crabs)
modelGH=MGHD(data=crabs[,4:8], G=2, max.iter=50)
```
bankruptcy

The data set contain the ratio of retained earnings (RE) to total assets, and the ratio of earnings before interests and taxes (EBIT) to total assets of 66 American firms recorded in the form of ratios. Half of the selected firms had filed for bankruptcy.

Usage

data(bankruptcy)

Format

A data frame with the following variables:

Y the status of the firm: 0 bankruptcy or 1 financially sound.
RE ratio
EBIT ratio

References

DA  

*Discriminant analysis using the mixture of generalized hyperbolic distributions.*

---

**Description**

Carries out model-based discriminant analysis using 5 different models: the mixture of multiple scaled generalized hyperbolic distributions (MGHD), the mixture of generalized hyperbolic factor analyzers (MGHFA), the mixture of multiple scaled generalized hyperbolic distributions (MSGHD), the mixture of convex multiple scaled generalized hyperbolic distributions (cMSGHD) and the mixture of coalesed generalized hyperbolic distributions (MCGHD).

**Usage**

```
DA(train,trainL,test,testL,method="GHD",starting="km",max.iter=10,eps=1e-2,q=2,scale=TRUE)
```

**Arguments**

- **train**: A n1 x p matrix or data frame such that rows correspond to observations and columns correspond to variables of the training data set.
- **trainL**: A n1 dimensional vector of membership for the units of the training set. If trainL[i]=k then observation belongs to group k.
- **test**: A n2 x p matrix or data frame such that rows correspond to observations and columns correspond to variables of the test data set.
- **testL**: A n2 dimensional vector of membership for the units of the test set. If testL[i]=k then observation belongs to group k.
- **method**: (optional) A string indicating the method to be used form discriminant analysis, if not specified GHD is used. Alternative methods are: MGHFA, MSGHD, cMSGHD, MCGHD.
- **starting**: (optional) A string indicating the initialization criterion, if not specified kmeans clustering is used. Alternative methods are: hierarchical "hierarchical" and model based "modelBased" clustering
- **max.iter**: (optional) A numerical parameter giving the maximum number of iterations each EM algorithm is allowed to use.
- **eps**: (optional) A number specifying the epsilon value for the convergence criteria used in the EM algorithms. For each algorithm, the criterion is based on the difference between the log-likelihood at an iteration and an asymptotic estimate of the log-likelihood at that iteration. This asymptotic estimate is based on the Aitken acceleration.
- **q**: (optional) used only if MGHFA method is selected. A numerical parameter giving the number of factors.
- **scale**: (optional) A logical value indicating whether or not the data should be scaled, true by default.
Value

- `model`: A list with the model parameters.
- `testMembership`: A vector of integers indicating the membership of the units in the test set.
- `ARI`: A value indicating the adjusted rand index for the test set.

Author(s)

Cristina Tortora Maintainer: Cristina Tortora <ctortora@mcmaster.ca>

References


See Also

- `mghd` `mghfa` `msghd` `cmsghd` `mcghd`

Examples

```r
# loading crabs data
data(crabs)

# divide the data in training set and test set
lab <- rbind(matrix(1,100,1),matrix(2,100,1))
train <- crabs[,c(1:90,110:200),,drop = FALSE]
trainl <- lab[,c(1:90,110:200),,drop = FALSE]
test <- crabs[91:109,4:8]
testl <- lab[91:109,1]

# model estimation
model <- DA(train,trainl,test,testl,method = "MSGHD",max.iter = 40,starting = "hierarchical")

# result
model$ARI(test)
```
MCGHD

Description

Carries out model-based clustering using the mixture of coalesced generalized hyperbolic distributions.

Usage

\texttt{MCGHD(data=NULL,gpar0=NULL,G=2,max.iter=100,eps=1e-2,label=NULL,method="km",scale=TRUE)}

Arguments

- \texttt{data} A \(n \times p\) matrix or data frame such that rows correspond to observations and columns correspond to variables.
- \texttt{gpar0} (optional) A list containing the initial parameters of the mixture model. See the 'Details' section.
- \texttt{G} A numerical parameter giving the number of clusters.
- \texttt{max.iter} (optional) A numerical parameter giving the maximum number of iterations each EM algorithm is allowed to use.
- \texttt{eps} (optional) A number specifying the epsilon value for the convergence criteria used in the EM algorithms. For each algorithm, the criterion is based on the difference between the log-likelihood at an iteration and an asymptotic estimate of the log-likelihood at that iteration. This asymptotic estimate is based on the Aitken acceleration.
- \texttt{label} (optional) A \(n\) dimensional vector, if label[i]=k then observation belongs to group k, if NULL then the data has no known groups.
- \texttt{method} (optional) A string indicating the initialization criterion, if not specified kmeans clustering is used. Alternative methods are: hierarchical "hierarchical" and model based "modelBased"
- \texttt{scale} (optional) A logical value indicating whether or not the data should be scaled, true by default.

Details

The arguments \texttt{gpar0}, if specified, has to be a list structure containing as much element as the number of components \(G\). Each element must include the following parameters: one \(p\) dimensional vector \(\mu\), \(\alpha\) and \(\phi\), a \(p \times p\) matrix \(\Gamma\), a \(p \times 2\) vector \(cpl\) containing the vectors \(\omega\) and \(\lambda\), and a 2-dimensional vector containing the \(\omega_0\) and \(\lambda_0\).
Value

A list with components

BIC  Bayesian information criterion value.
gpar A list of the model parameters in the rotated space.
loglik The log-likelihood values.
map  A vector of integers indicating the maximum a posteriori classifications for the best model.
par  A list of the model parameters.
z  A matrix giving the raw values upon which map is based.

Author(s)

Cristina Tortora, Ryan P. Browne, Brian C. Franczak and Paul D. McNicholas. Maintainer: Cristina Tortora <ctortora@uoguelph.ca>

References


See Also

MGHD, MSGHD

Examples

```r
# loading bankruptcy data
data(bankruptcy)

# model estimation
model = MCGHD(bankruptcy[,2:3],G=2)

# result
plot(bankruptcy[,2:3], col=model$map)
table(bankruptcy[,1], model$map)
```

---

**MGHD**

*Mixture of generalized hyperbolic distributions (MGHD).*

Description

Carries out model-based clustering and classification using the mixture of generalized hyperbolic distributions.
Usage

MGHD(data=NULL,gpar0=NULL,G=2,max.iter=100,label=NULL,eps=1e-2,method="kmeans",scale=TRUE)

Arguments

data
A n x p matrix or data frame such that rows correspond to observations and columns correspond to variables.

gpar0
(optional) A list containing the initial parameters of the mixture model. See the 'Details' section.

G
A numerical parameter giving the number of clusters.

max.iter
(optional) A numerical parameter giving the maximum number of iterations each EM algorithm is allowed to use.

label
( optional) A n dimensional vector, if label[i]=k then observation belongs to group k, If label[i]=0 then observation has no known group, if NULL then the data has no known groups.

eps
(optional) A number specifying the epsilon value for the convergence criteria used in the EM algorithms. For each algorithm, the criterion is based on the difference between the log-likelihood at an iteration and an asymptotic estimate of the log-likelihood at that iteration. This asymptotic estimate is based on the Aitken acceleration.

method
( optional) A string indicating the initialization criterion, if not specified kmeans clustering is used. Alternative methods are: hierarchical "hierarchical" and model based "modelBased" clustering

scale
( optional) A logical value indicating whether or not the data should be scaled, true by default.

Details

The arguments gpar0, if specified, is a list structure containing at least one p dimensional vector mu, and alpha, a p x p matrix sigma, and a 2 dimensional vector containing omega and lambda.

Value

A list with components

BIC
Bayesian information criterion value.

gpar
A list of the model parameters.

loglik
The log-likelihood values.

map
A vector of integers indicating the maximum a posteriori classifications for the best model.

z
A matrix giving the raw values upon which map is based.

Author(s)

Ryan P. Browne, Cristina Tortora Maintainer: Cristina Tortora <ctortora@uoguelph.ca>
References


Examples

```r
# loading crabs data
data(crabs)

# model estimation
model=mghd(data=crabs[,4:8], G=2)

# result
plot(model$loglik)
table(model$map, crabs[,2])

## Classification
# loading bankruptcy data
data(bankruptcy)
# 70% belong to the training set
label=bankruptcy[,1]
# for a classification purpose the label cannot be 0
label[1:33]=2
a=round(runif(20)*65+1)
label[a]=0

## model estimation
model=mghd(data=bankruptcy[,2:3], G=2, label=label)

# result
table(model$map,bankruptcy[,1])
```

### MGHFA

**Mixture of generalized hyperbolic factor analyzers (MGHFA).**

**Description**

Carries out model-based clustering and classification using the mixture of generalized hyperbolic factor analyzers.

**Usage**

`MGHFA(data=NULL, gpar0=NULL, G=2, max.iter=100, label =NULL ,q=2,eps=1e-2 , method="kmeans", scale=TRUE )`
Arguments

data  A matrix or data frame such that rows correspond to observations and columns correspond to variables.
gpar0  (optional) A list containing the initial parameters of the mixture model. See the 'Details' section.
G  A numerical parameter giving the number of clusters.
max.iter  (optional) A numerical parameter giving the maximum number of iterations each EM algorithm is allowed to use.
label  (optional) A n dimensional vector, if label[i]=k then observation belongs to group k, If label[i]=0 then observation has no known group, if NULL then the data has no known groups.
q  A numerical parameter giving the number of factors.
eps  (optional) A number specifying the epsilon value for the convergence criteria used in the EM algorithms. For each algorithm, the criterion is based on the difference between the log-likelihood at an iteration and an asymptotic estimate of the log-likelihood at that iteration. This asymptotic estimate is based on the Aitken acceleration.
method  (optional) A string indicating the initialization criterion, if not specified kmeans clustering is used. Alternative methods are: hierarchical "hierarchical" and model based "modelBased" clustering
scale  (optional) A logical value indicating whether or not the data should be scaled, true by default.

Details

The arguments gpar0, if specified, is a list structure containing at least one p dimensional vector mu, alpha and phi, a pxp matrix gamma, a 2 dimensional vector cpl containing omega and lambda.

Value

A list with components

- **BIC**  Bayesian information criterion value.
- **gpar**  A list of the model parameters.
- **loglik**  The log-likelihood values.
- **map**  A vector of integers indicating the maximum a posteriori classifications for the best model.
- **z**  A matrix giving the raw values upon which map is based.

Author(s)

Cristina Tortora, Ryan P. Browne, and Paul D. McNicholas. Maintainer: Cristina Tortora <ctortora@uoguelph.ca>
References

Examples
```r
## Classification
#70% belong to the training set
data(sonar)
label=sonar[,61]
set.seed(135)
a=round(runif(62)*207+1)
label[a]=0

## model estimation
model=MGHFA(data=sonar[,1:60], G=2, max.iter=50, q=2,label=label )

# result
table(model$map,sonar[,61])
```

---

**MSGHD**

*Mixture of multiple scaled generalized hyperbolic distributions (MSGHD).*

Description
Carries out model-based clustering using the mixture of multiple scaled generalized hyperbolic distributions.

Usage

```r
MSGHD(data=NULL,gpar0=NULL,G=2,max.iter=100,label=NULL,eps=1e-2,method="km",scale=TRUE)
```

Arguments
- `data`: A n x p matrix or data frame such that rows correspond to observations and columns correspond to variables.
- `gpar0`: (optional) A list containing the initial parameters of the mixture model. See the 'Details' section.
- `G`: A numerical parameter giving the number of clusters.
- `max.iter`: (optional) A numerical parameter giving the maximum number of iterations each EM algorithm is allowed to use.
- `label`: (optional) A n dimensional vector, if label[i]=k then observation belongs to group k, if NULL then the data has no known groups.
eps (optional) A number specifying the epsilon value for the convergence criteria used in the EM algorithms. For each algorithm, the criterion is based on the difference between the log-likelihood at an iteration and an asymptotic estimate of the log-likelihood at that iteration. This asymptotic estimate is based on the Aitken acceleration.

method (optional) A string indicating the initialization criterion, if not specified kmeans clustering is used. Alternative methods are: hierarchical "hierarchical" and model based "modelBased" clustering.

scale (optional) A logical value indicating whether or not the data should be scaled, true by default.

Details
The arguments gpar0, if specified, is a list structure containing at least one p dimensional vector mu, alpha and phi, a pxp matrix gamma, and a px2 matrix cpl containing the vector omega and the vector lambda.

Value
A list with components

- `BIC` Bayesian information criterion value.
- `gpar` A list of the model parameters
- `loglik` The log-likelihood values.
- `map` A vector of integers indicating the maximum a posteriori classifications for the best model.
- `z` A matrix giving the raw values upon which map is based.

Author(s)
Cristina Tortora, Ryan P. Browne, Brian C. Franczak and Paul D. McNicholas. Maintainer: Cristina Tortora <ctortora@mcmaster.ca>

References

See Also
MGHD

Examples
```r
# loading bankruptcy data
data(bankruptcy)

# model estimation
model <- MSGHD(bankruptcy[,2:3], G=2)
```
Convex mixture of multiple scaled generalized hyperbolic distributions (MSGHD).

Description

Carries out model-based clustering using the convex mixture of multiple scaled generalized hyperbolic distributions. The cMSGHD only allows convex level sets.

Usage

cMSGHD(data=NULL,gpar0=NULL,G=2,max.iter=100,label=NULL,eps=1e-2,method="km",scale=TRUE)

Arguments

data A n x p matrix or data frame such that rows correspond to observations and columns correspond to variables.
gpar0 (optional) A list containing the initial parameters of the mixture model. See the 'Details' section.
G A numerical parameter giving the number of clusters.
max.iter (optional) A numerical parameter giving the maximum number of iterations each EM algorithm is allowed to use.
label (optional) A n dimensional vector, if label[i]=k then observation belongs to group k, if NULL then the data has no known groups.
eps (optional) A number specifying the epsilon value for the convergence criteria used in the EM algorithms. For each algorithm, the criterion is based on the difference between the log-likelihood at an iteration and an asymptotic estimate of the log-likelihood at that iteration. This asymptotic estimate is based on the Aitken acceleration.
method (optional) A string indicating the initialization criterion, if not specified kmeans clustering is used. Alternative methods are: hierarchical "hierarchical" and model based "modelBased"
scale (optional) A logical value indicating whether or not the data should be scaled, true by default.

Details

The arguments gpar0, if specified, is a list structure containing at least one p dimensional vector mu, alpha and phi, a p x p matrix gamma, and a p x 2 matrix cpl containing the vector omega and the vector lambda.
**Value**

A list with components

- **BIC** Bayesian information criterion value.
- **gpar** A list of the model parameters
- **loglik** The log-likelihood values.
- **map** A vector of integers indicating the maximum a posteriori classifications for the best model.
- **z** A matrix giving the raw values upon which map is based.

**Author(s)**

Cristina Tortora, Ryan P. Browne, and Paul D. McNicholas. Maintainer: Cristina Tortora <corttora@mcmaster.ca>

**References**


**See Also**

MGHD MSGHD

**Examples**

```r
# loading banknote data
data(banknote)

# model estimation
model=cMSGHD(banknote[,2:7],G=2,max.iter=30)

# result
table(banknote[,1],model$map)
```

---

**Selection**

*Selection of parameter using the mixture of generalized hyperbolic distributions.*

**Description**

Replicate a routine n times and gives as result the best solution, the available routine are: the mixture of multiple scaled generalized hyperbolic distributions (MGHD), the mixture of generalized hyperbolic factor analyzers (MGHFA), the mixture of multiple scaled generalized hyperbolic distributions (MSGHD), the mixture of convex multiple scaled generalized hyperbolic distributions (cMSGHD) and the mixture of coalesed generalized hyperbolic distributions (MCGHD).
Selection

Usage

Selection(data, label, G=2, niter=50, method="GHD", starting="kmeans", max.iter=10, eps=1e-2, labelcl=NULL, q=2, scale=TRUE, criterion="ARI")

Arguments

data A n x p matrix or data frame such that rows correspond to observations and columns correspond to variables.

label A n dimensional vector of known membership, if label[i]=k unit i belongs to cluster k. It is used to select the solution.

G A numerical parameter giving the number of clusters.

niter A numerical parameter giving the number of replications.

method (optional) A string indicating the method to be used, if not specified GHD is used. Alternative methods are: MGHFA, MSGHD, cMSGHD, MCGHD.

starting (optional) A string indicating the initialization criterion, if not specified kmeans clustering is used. Alternative methods are: hierarchical "hierarchical" and model based "modelBased"

max.iter (optional) A numerical parameter giving the maximum number of iterations each EM algorithm is allowed to use.

eps (optional) A number specifying the epsilon value for the convergence criteria used in the EM algorithms. For each algorithm, the criterion is based on the difference between the log-likelihood at an iteration and an asymptotic estimate of the log-likelihood at that iteration. This asymptotic estimate is based on the Aitken acceleration.

labelcl (optional) A n dimensional vector of membership, if labelcl[i]=k then observation belongs to group k, if NULL then the data has no known groups.

q (optional) used only if MGHFA method is selected. A numerical parameter giving the number of factors.

scale (optional) A logical value indicating whether or not the data should be scaled, true by default.

criterion (optional) A string indicating the selection criterion, if not specified the adjusted rand index (ARI) is used. The alternative criterion is the BIC.

Details

Select the best model according to the adjusted rand index or the BIC

Value

A list with components

model A list with the model parameters.

ARI A value indicating the adjusted rand index.
Author(s)

Cristina Tortora

Maintainer: Cristina Tortora <ctortora@mcmaster.ca>

References


See Also

MGHD MGHFA MSGHD cMSGHD MCIGHD

Examples

```r
# loading bankruptcy data
# data(bankruptcy)

# model estimation
# mod=Selection(bankruptcy[,2:3],bankruptcy[,1],method="MCIGHD")

# result
# mod$ARI
```

Description

The data report the patterns obtained by bouncing sonar signals at various angles and under various conditions. There are 208 patterns in all, 111 obtained by bouncing sonar signals off a metal cylinder and 97 obtained by bouncing signals off rocks. Each pattern is a set of 60 numbers (variables) taking values between 0 and 1.

Usage

data(sonar)

Format

A data frame with 208 observations and 61 columns. The first 60 columns contain the variables. The 61st column gives the material: 1 rock, 2 metal.

Source

UCI machine learning repository
References

Index

*Topic **Classification**
  MGH, 7
  MGHFA, 9
  MixGHD-package, 2

*Topic **Clustering**
  DA, 4
  MCGHD, 6
  MGH, 7
  MGHFA, 9
  MixGHD-package, 2
  MSGHD, 11
  rMSGHD, 13
  Selection, 14

*Topic **Expectation Maximization algorithm**
  MixGHD-package, 2

*Topic **Generalized hyperbolic distribution**
  DA, 4
  MCGHD, 6
  MGH, 7
  MGHFA, 9
  MixGHD-package, 2
  MSGHD, 11
  rMSGHD, 13
  Selection, 14

*Topic **datasets**
  bankruptcy, 3
  sonar, 16

  bankruptcy, 3

cMSGHD, 5, 16
  cMSGHD (rMSGHD), 13

  DA, 4

  MCGHD, 5, 6, 16
  MGH, 5, 7, 7, 12, 14, 16
  MGHFA, 5, 9, 16