Package 'mixture'

February 13, 2018

Type Package

Title Finite Gaussian Mixture Models for Clustering and Classification

Version 1.5

Date 2018-02-13

Author Ryan P. Browne, Aisha ElSherbiny and Paul D. McNicholas

Maintainer Ryan Browne <rbrowne@math.mcmaster.ca>

Description An implementation of all 14 Gaussian parsimonious clustering models (GPCMs) for model-based clustering and model-based classification.

License GPL (>= 2)

LazyLoad yes

NeedsCompilation yes

Repository CRAN

SystemRequirements GNU make

Date/Publication 2018-02-13 22:07:56 UTC

R topics documented:

e.step *E-Step*

Description

Carries out the E-step for EM algorithm

Usage

```
e.step(data=NULL, gpar=NULL, labels=NULL, v=1)
```

Arguments

data	A matrix or data frame such that rows correspond to observations and columns correspond to variables. Note that this function currently only works with multivariate data p > 1.
gpar	A list of the model parameters.
labels	A vector of groups labels. If NULL none are known.
v	The value for deterministic annealing. If v=1 the standard estimate is used.

Details

Carries out the E-step for EM algorithm

Value

A nxG matrix of weights.

Examples

```
data("x2")
u0 = runif(nrow(x2))
m0 = m.step(data=x2, covtype="VVV", w=cbind(u0,1-u0), D=NULL, mtol=1e-8, mmax=10)
w0 = e.step(data=x2, gpar=m0, labels=NULL, v=1)
```

gpcm *Gaussian Parsimonious Clustering Models*

Description

Carries out model-based clustering or classification using some or all of the 14 parsimonious Gaussian clustering models (GPCM).

Usage

```
gpcm(data=NULL,  G=1:3, mnames=NULL, start=0, label=NULL, veo=FALSE,
nmax=1000, atol=1e-8, mtol=1e-8, mmax=10, pprogress=FALSE, pwarning=FALSE)
```

Arguments

data	A matrix or data frame such that rows correspond to observations and columns correspond to variables. Note that this function currently only works with multivariate data p > 1.
G	A sequence of integers giving the number of components to be used.
mnames	The models (i.e., covariance structures) to be used. If NULL then all 14 are fitted.
start	If 0 then the kmeans function is used for initialization. If a positive value is inputted then best out of ceiling(k) random initializations are used. If is.vector then deterministic annealing is used with the given sequence of values in [0,1]; cf. Zhou and Lange (2010). If is.matrix then matrix is used as an initialization matrix as along as it has non-negative elements. Note: only models with the same number of columns of this matrix will be fit. If is.function then this function is used for building an initialization matrix. See Examples.
label	If NULL then the data has no known groups. If is.integer then some of the observations have known groups. If label[i]=k then observation belongs to group k. If label[i]=0 then observation has no known group. See Examples.
veo	If TRUE then if the number variables in the model exceeds the number of observations the model is still fitted.
nmax	The maximum number of iterations each EM algorithm is allowed to use.
atol	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 and details are given in the References.
mtol	A number specifying the epsilon value for the convergence criteria used in the M-step in the GEM algorithms.
mmax	The maximum number of iterations each M-step is allowed in the GEM algorithms.
pprogress	If TRUE print the progress of the function.
pwarning	If TRUE print the warnings.

Details

The data x are either clustered or classified using Gaussian mixture models with some or all of the 14 parsimonious covariance structures described in Celeux & Govaert (1995). The algorithms given by Celeux & Govaert (1995) is used for 12 of the 14 models; the "EVE" and "VVE" models use the algorithms given in Browne & McNicholas (2012, 2013). Starting values are very important to the successful operation of these algorithms and so care must be taken in the interpretation of results.

Value

An object of class gpcm is a list with components:

map	A vector of integers indicating the maximum *a posteriori* classifications for the best model.
gpar	A list of the model parameters.
bicModel	A list containing; the number of groups for the best model, the covariance structure, and Bayesian Information Criterion (BIC) value.
loglik	The log-likelihood values from fitting the best model.
z	A matrix giving the raw values upon which map is based.
BIC	An array containing the log-likelihood (loglik), number of model parameters (npar) and BIC indexed by the covariance structure and number of components.
start	The value inputted into start.
startobject	The type of object inputted into start.

Note

Dedicated print, plot and summary functions are available for objects of class gpcm.

Author(s)

Ryan P. Browne, Aisha ElSherbiny and Paul D. McNicholas.

Maintainer: Ryan Browne <rbrowne@uoguelph.ca>

References

Browne, R.P. and McNicholas, P.D. (2014). Estimating common principal components in high dimensions. Advances in Data Analysis and Classification **8**(2), 217-226.

Celeux, G., Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition* **28**(5), 781-793.

Examples

```
data("x2")

# use k-means starts
ax0 = gpcm(x2, G=1:5, mnames=c("VVV", "EVE"),start=0, pprogress=TRUE, atol=1e-2)
summary(ax0)
ax0

# use 6 random values for starting values
ax6 = gpcm(x2, G=1:5, mnames=c("VVV", "EVE"),start= 2, atol=1e-2)
summary(ax6)
ax6

# use deterministic annealing for starting values
#axNULL = gpcm(x2, G=1:5, mnames=c("VVV", "EVE"), start=NULL, atol=1e-2)
#summary(axNULL)
```

```
#axNULL

# use your own deterministic annealing values for starting values
#vseq0 = rep(seq(.05, 1, length.out=5),each=2)
#axv = gpcm(x2, G=1:5, mnames=c("VVV", "EVE"), start=vseq0, atol=1e-2)
#summary(axv)
#axv

# Initialization using your own function
igparhc <-  function(data=NULL, g=NULL,covtype=NULL) {
lw = cutree(hclust(dist(data), "complete"),k=g)
w = matrix(0, nrow=nrow(data), ncol=g)
for (j in 1:ncol(w)) w[,j] = as.numeric( lw ==j )
return(w)
}
axhclust = gpcm(x2, G=1:5, mnames=c("VVV", "EVE"),start= igparhc, atol=1e-2)
summary(axhclust)
axhclust

# Estimate all 14 covariance structures from k-means starts
ax = gpcm(x2, G=1:5, start=0, atol=1e-2)
summary(ax)
ax

# model based classification
x2.label = numeric(nrow(x2))
x2.label[c(10,50, 110, 150, 210, 250)] = c(1,1,2,2,3,3)
plot(x2, col=x2.label)
axl = gpcm(x2, G=3:5, mnames=c("VVV", "EVE"), label=x2.label, atol=1e-2)
```

`m.step` *M-Step*

Description

Carries out the M-step for EM algorithm

Usage

```
m.step(data=NULL, covtype=NULL, w=NULL, D=NULL, mtol=NULL, mmax=NULL)
```

Arguments

`data`	A matrix or data frame such that rows correspond to observations and columns correspond to variables. Note that this function currently only works with multivariate data p > 1.
`covtype`	A three letter sequence indicating the covariance structure.
`w`	A nxG matrix of weights.

D	An initial value for D. If NULL then the identity matrix is used.
mtol	The convergence criteria for the m.step if an iterative procedure is necessary.
mmax	The maximum number of iterations for an iterative procedure.

Details

Carries out the M-step for EM algorithm

Value

A list of the model parameters with the mu, sigma, invsigma and logdet for each group.

Examples

```
data("x2")
u0 = runif(nrow(x2))
m.step(data=x2, covtype="VVV", w=cbind(u0,1-u0), D=NULL, mtol=1e-8, mmax=10)
```

---

mixture *Mixture Models for Clustering and Classification*

Description

An implementation of all 14 Gaussian parsimonious clustering models (GPCMs) for model-based clustering and model-based classification.

Details

Package:	mixture
Type:	Package
Version:	1.3
Date:	2015-03-10
License:	GPL (>=2)

This package contains the functions gpcm, e.step, and m.step as well as one simulated data set.

Author(s)

Ryan P. Browne, Aisha ElSherbiny and Paul D. McNicholas.

Maintainer: Ryan Browne <rbrowne@math.mcmaster.ca>

See Also

Details, examples, and references are given under gpcm.

x2 *Simulated Data*

Description

Simulated data, with two variables with three groups, used to illustrate gpcm.

Usage

```
data(x2)
```

Format

A data frame with 300 observations and 2 columns.

Source

These data were simulated using R.

Index