Package ‘pmcgd’

February 20, 2015

Type Package

Title pmcgd

Version 1.1

Date 2013-01-27

Author Antonio Punzo and Paul D. McNicholas

Maintainer Antonio Punzo <Antonio.Punzo@unict.it>

Description Parsimonious Mixtures of Contaminated Gaussian Distributions

License GPL-2

LazyLoad yes

Depends R (>= 2.15.0)

Imports mixture, mnormt

NeedsCompilation no

Repository CRAN

Date/Publication 2013-12-21 21:21:38

R topics documented:

`pmcgd-package` *pmcgd - Parsimonious Mixtures of Contaminated Gaussian Distributions*

Description

This package allows for fitting, according to the expectation-conditional maximization algorithm, of the family of 14 parsimonious mixtures of contaminated Gaussian distributions discussed in Punzo & McNicholas (2013). Some likelihood-based model selection criteria can be adopted to select the best model in the family and the best number of mixture components.

Details

Package:	pmcgd
Type:	Package
Version:	1.0
Date:	2013-12-10
License:	GNU-2

Author(s)

Punzo A., McNicholas, P. D.

Maintainer: Punzo Antonio <antonio.punzo@unict.it>

References

Punzo, A., and McNicholas, P. D. (2013). Outlier Detection via Parsimonious Mixtures of Contaminated Gaussian Distributions. *arXiv.org* e-print **1305.4669**, available at: `http://arxiv.org/abs/1305.4669`.

See Also

`MS`,`class`

`class` *Matrix of Indicator Variables given Classification*

Description

Converts a classification vector into a matrix of indicator variables.

Usage

```
class(groups, k)
```

Arguments

`groups` A numeric vector of integers. Typically the distinct entries of this vector would represent a classification of observations in a data set.

`k` An integer indicating the number of groups.

Value

An n (number of observations) by k (number of groups) matrix of (0,1) indicator variables. The `[i,j]`th entry is 1 if `groups[i]` is `j` and 0 otherwise.

Author(s)

Punzo A. and McNicholas P. D.

References

Punzo, A., and McNicholas, P. D. (2013). Outlier Detection via Parsimonious Mixtures of Contaminated Gaussian Distributions. *arXiv.org* e-print **1305.4669**, available at: `http://arxiv.org/abs/1305.4669`.

See Also

`pmcgd-package`,`MS`

Examples

```
n <- 20
k <- 3
prob <- c(0.5,0.3,0.2)
groups <- sample(1:k, size=n, replace = TRUE, prob = prob)
matclass <- class(groups, k)
matclass
```

`MS` *Fitting for the Parsimonious Mixtures of Contaminated Gaussian Distributions*

Description

Carries out model-based clustering or model-based classification using some or all of the 14 parsimonious mixtures of contaminated Gaussian Distributions by using the ECM algorithm. Likelihood-based model-selection criteria are used to select the best model and the number of mixture components.

Usage

```
MS(X, k, model = NULL, initialization = "mclust",
   alphacon = TRUE, alphamin = NULL, alphafix = FALSE, alpha = NULL,
   etacon = TRUE, etafix = FALSE, eta = NULL, etamax = 200,
 start.z = NULL, start.v = NULL, start = 0,
   ind.label = NULL, label = NULL, iter.max = 1000, threshold = 1.0e-03)
```

Arguments

`X` 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 ($p > 1$).

`k` a vector containing the numbers of groups to be tried.

`model` vector indicating the models (i.e., the covariance structures: "EII", "VII", "EEI", "VEI", "EVI", "VVI", "EEE", "VEE", "EVE", "EEV", "VVE", "VEV", "EVV", "VVV") to be used. If `NULL`, then all 14 models are fitted.

`initialization` initialization strategy for the ECM-algorithm. It can be:

- `"mclust"`: posterior probabilities from mixtures of Gaussian distributions are used for initialization;
- `"random.soft"`: initial posterior probabilities are random generated;
- `"random.hard"`: initial classification matrix is random generated;
- `"manual"`: the user must specify, via the arguments `start.z` and `start.v`, posterior probabilities or classification matrix for the mixture components and the 3D array of membership to the “good” and “bad” groups in each mixture component, respectively.

Default value is `"mclust"`.

`alphacon` if `TRUE`, the vector with proportions of good observations is constrained to be greater than the vector specified by the `alphamin` argument.

`alphamin` when `alphacon=TRUE`, vector with minimum proportions of good observations in each group.

`alphafix` when `alphafix=TRUE`, the vector of proportions of good observations is fixed to the vector specified in the `alpha` argument.

`alpha` vector of proportions of good observations in each group to be considered when `alphafix=TRUE`.

`etacon` if `TRUE`, the contaminated parameters are all constrained to be greater than one.

`etafix` if `TRUE`, the vector of contaminated parameters is fixed to the vector specified by the `eta` argument.

`eta` vector of contaminated parameters to be considered when `etafix`.

`etamax` maximum value for the contamination parameters to be considered in the estimation phase when `etafix=FALSE`.

`start.z` matrix of soft or hard classification; it is used only if `initialization="manual"`.

`start.v` 3D array of soft or hard classification to the good and bad groups in each mixture components. It is used as initialization when `initialization="manual"`.

`start` when `initialization="manual"`, initialization used for the `gpcm()` function of the **mixture** package (see `mixture:gpcm` for details).

`ind.label` vector of positions (rows) of the labeled observations.

`label` vector, of the same dimension as `ind.label`, with the group of membership of the observations indicated in the `ind.label` argument.

`iter.max` maximum number of iterations in the ECM-algorithm. Default value is 1000.

`threshold` threshold for Aitken's acceleration procedure. Default value is 1.0e-03.

Details

The multivariate data contained in `X` are either clustered or classified using parsimonious mixtures of contaminated Gaussian densities with some or all of the 14 parsimonious covariance structures described in Punzo & McNicholas (2013). The algorithms given by Browne & McNicholas (2013) are considered (see also Celeux & Govaert, 1995, for all the models apart from "EVE" and "VVE"). 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 `pmcgd` is a list with components:

`call` an object of class `call`

`best` a data frame with the best number of mixture components (first column) and the best model (second column) with respect to the three model selection criteria adopted (AIC, BIC, and ICL)

`bestAIC,bestBIC,bestICL`
for the best AIC, BIC, and ICL models, these are three lists (of the same type) with components:

- `modelname`: the name of the best model.
- `npar`: number of free parameters.
- `X`: matrix of data.
- `k`: number of mixture components.
- `p`: number of variables.
- `prior`: weights for the mixture components.
- `priorgood`: weights for the good observations in each of the `k` groups.
- `mu`: component means.
- `Sigma`: component covariance matrices for the good observations.
- `lambda`: component volumes for the good observations.
- `Delta`: component shape matrices for the good observations.
- `Gamma`: component orientation matrices for the good observations.
- `eta`: component contamination parameters.
- `iter.stop`: final iteration of the ECM algorithm.
- `z`: matrix with posterior probabilities for the outer groups.
- `v`: matrix with posterior probabilities for the inner groups.

- `group`: vector of integers indicating the maximum a posteriori classifications for the best model.
- `loglik`: log-likelihood value of the best model.
- `AIC`: AIC value
- `BIC`: BIC value
- `ICL`: ICL value
- `call`: an object of class `call` for the best model.

Author(s)

Punzo A. and McNicholas P. D.

References

Punzo, A., and McNicholas, P. D. (2013). Outlier Detection via Parsimonious Mixtures of Contaminated Gaussian Distributions. *arXiv.org* e-print **1305.4669**, available at: `http://arxiv.org/abs/1305.4669`.

Browne, R. P. and McNicholas, P. D. (2013). **mixture**: Mixture Models for Clustering and Classification. R package version 1.0.

Celeux, G. and Govaert, G. (1995). Gaussian Parsimonious Clustering Models. *Pattern Recognition*. **28**(5), 781-793.

See Also

`pmcgd-package`, `class`

Examples

```
# Artificial data from an EEI model with k=2 components

library(mnormt)
p   <- 2
k   <- 2
eta <- c(8,8) # contamination parameters
set.seed(12345)
X1good <- rmnorm(n = 300, mean = rep(3,p), varcov = diag(c(5,0.5)))
X2good <- rmnorm(n = 300, mean = rep(-3,p), varcov = diag(c(5,0.5)))
X1bad  <- rmnorm(n = 30, mean = rep(3,p), varcov = eta[1]*diag(c(5,0.5)))
X2bad  <- rmnorm(n = 30, mean = rep(-3,p), varcov = eta[2]*diag(c(5,0.5)))
X      <- rbind(X1good,X1bad,X2good,X2bad)
plot(X, pch = 16, cex = 0.8)

# model-based clustering with the whole family of 14
# parsimonious models and number of groups ranging from 1 to 3

overallfit <- MS(X, k = 1:2, model = c("EEI","VVV"), initialization = "mclust")

# to see the best BIC results
```

```
bestBIC <- overallfit$bestBIC

# plot of the best BIC model

plot(X, xlab = expression(X[1]), ylab = expression(X[2]), col = "white")
text(X, labels = bestBIC$detection$innergroup, col = bestBIC$group, cex = 0.7, asp = 1)
box(col = "black")
```

Index