Package 'cemco'

April 29, 2020

Version 0.1

Date 2020-04-05

Title Fit 'CemCO' Algorithm

Description Functions to fit the 'CemCO' algorithm, a model-based (Gaussian) clustering algorithm that removes/minimizes the effects of undesirable covariates during the clustering process both in cluster centroid and in cluster covariance structures (Relvas C. & Fujita A., (2020) <arXiv:2004.02333>).

Depends R (>= 3.1.0)

License GPL (>= 2)

LazyData true.

Imports clusteval, doParallel, nnet, rootSolve, foreach, MASS, mclust, mvtnorm

RoxygenNote 6.1.1.9000

NeedsCompilation no

Author Carlos Relvas [aut], Andre Fujita [aut, cre]

Maintainer Andre Fujita <andrefujita@usp.br>

Repository CRAN

Date/Publication 2020-04-29 14:30:03 UTC

R topics documented:

CemCO *Fit CemCO algorithm using multiple threads of the machine*

Description

Model-based clustering based on parameterized finite Gaussian mixture models with covariates effects on the distribution means. Models are estimated by an EM algorithm running in multiple threads of the machine

Usage

```
CemCO(data, y, G, max_iter=100, n_start=20, cores=4)
```

Arguments

data	A numeric vector, matrix, or data frame of observations. Non-numerical values should be converted to integer or float (e.g. dummies). If matrix or data frame, rows and columns correspond to observations (n) and variables (P).
y	numeric matrix of data to use as covariates. Non-numerical values should be converted to integer or float (e.g. dummies).
G	An integer specifying the numbers of mixture components (clusters)
max_iter	maximum number of iterations of the EM optimization (default value equals to 100)
n_start	how many random sets should be chosen? (default value equals to 20)
cores	number of cores for EM optimization (default value equals to 4)

Details

This function optimizes the log likelihood of the CemCO algorithm using a implementation of the EM algorithm. If categorial features need to be used, please create dummies or use another encode method.

Value

The function output is a list

fitted parameters	The estimated parameters of the CemCO algorithm, including clusters centroids, covariance matrix, covariate effects of each cluster and a priori probability of each cluster.
log likelihood	The optimal log likelihood estimated by the model

Author(s)

Relvas, C. & Fujita, A.

References

Stage I non-small cell lung cancer stratification by using a model-based clustering algorithm with covariates, Relvas et al.

Examples

```
set.seed(42)
X = cbind(rnorm(60), rnorm(60))
Y = cbind(rnorm(60), rnorm(60))
K = 2

fit <- CemCO(X, Y, K, max_iter=10, n_start=1, cores=1)
params <- fit[[1]]  ## fitted parameters
ll <- fit[[2]]  ## log likelihood
```

`CemCOVar` *Fit CemCO algorithm with covariates effects on cluster centroids and covariance matrices.*

Description

Model-based clustering based on parameterized finite Gaussian mixture models with covariates effects on the distribution means and the distribution covariance matrices. Models are estimated by an EM algorithm.

Usage

```
CemCOVar(data, y, G, y_cov, max_iter=100, n_start=20, cores=4)
```

Arguments

`data`	A numeric vector, matrix, or data frame of observations. Non-numerical values should be converted to integer or float (e.g. dummies). If matrix or data frame, rows and columns correspond to observations (n) and variables (P).
`y`	numeric matrix of data to use as covariates. Non-numerical values should be converted to integer or float (e.g. dummies).
`G`	An integer specifying the numbers of mixture components (clusters)
`y_cov`	numeric matrix of data to use as covariates for the covariance effect. Non-numerical values should be converted to integer or float (e.g. dummies)
`max_iter`	maximum number of iterations of the EM optimization (default value equals to 100)
`n_start`	how many random sets should be chosen? (default value equals to 20)
`cores`	number of cores for EM optimization (default value equals to 4)

Details

This function optimizes the log likelihood of the CemCO algorithm with covariates effects on cluster centroids and covariance matrices using a implementation of the EM algorithm. The covariates associated with the distributions means and with the distributions covariance matrices do not need to be the same.

Value

The function output is a list

`fitted parameters`
: The estimated parameters of the CemCO algorithm, including clusters centroids, covariance matrix, covariate effects of each cluster and a priori probability of each cluster.

`log likelihood`
: The optimal log likelihood estimated by the model

Author(s)

Relvas, C. & Fujita, A.

References

Stage I non-small cell lung cancer stratification by using a model-based clustering algorithm with covariates, Relvas et al.

Examples

```
set.seed(42)
X = cbind(rnorm(20), rnorm(20))
Y = cbind(rnorm(20), rnorm(20))
K = 2


fit <- CemCOVar(X, Y, K, Y[,1], max_iter=5, n_start=1, cores=1)
params <- fit[[1]]  ## fitted parameters
ll <- fit[[2]]  ## log likelihood
```

EStep

Calculate the E step of the CemCO algorithm with covariates effects on distributions means.

Description

Implements the expectation step of EM algorithm for parameterized Gaussian mixture models with covariates effects on the distribution means. It is also used to calculate the a posteriori probability of each observation belong to each cluster.

Usage

```
EStep(data, Y, phi, G)
```

Arguments

data	A numeric vector, matrix, or data frame of observations. Non-numerical values should be converted to integer or float (e.g. dummies). If matrix or data frame, rows and columns correspond to observations (n) and variables (P).
Y	numeric matrix of data to use as covariates. Non-numerical values should be converted to integer or float (e.g. dummies).
phi	list of fitted parameters in the same format as the output of the CemCO function
G	An integer specifying the numbers of mixture components (clusters)

Details

Calculate the a posteriori probability of each observation belong to each cluster given the data and the current parameters estimation.

Value

Returns a n x G numeric matrix where n represents the number of observations (number of rows of data) and G (the number of clusters). The value i, j represents the probability of the i-th observation belong to j-th cluster.

Author(s)

Relvas, C. & Fujita, A.

References

Stage I non-small cell lung cancer stratification by using a model-based clustering algorithm with covariates, Relvas et al.

Examples

```
set.seed(42)
X = cbind(rnorm(60), rnorm(60))
Y = cbind(rnorm(60), rnorm(60))
K = 2

fit <- CemCO(X, Y, K, max_iter=10, n_start=1, cores=1)
prob <- EStep(X, Y, fit[[1]], K)
```

EStepVar *Calculate the E step of the CemCO algorithm with covariates effects on distributions means and distributions covariance matrices.*

Description

Implements the expectation step of EM algorithm for parameterized Gaussian mixture models with covariates effects on the distribution means and the distribution covariance matrices. It is also used to calculate the posteriori probability of each observation belong to each cluster.

Usage

```
EStepVar(data, Y, phi, G, y_cov)
```

Arguments

data	A numeric vector, matrix, or data frame of observations. Non-numerical values should be converted to integer or float (e.g. dummies). If matrix or data frame, rows and columns correspond to observations (n) and variables (P).
Y	numeric matrix of data to use as covariates. Non-numerical values should be converted to integer or float (e.g. dummies).
phi	list of fitted parameters in the same format as the output of the CemCO function
G	An integer specifying the numbers of mixture components (clusters)
y_cov	numeric matrix of data to use as covariates for the covariance effect. Non-numerical values should be converted to integer or float (e.g. dummies).

Details

Calculate the a posteriori probability of each observation belong to each cluster given the data and the current parameters estimation.

Value

Returns a n x G numeric matrix where n represents the number of observations (number of rows of data) and G (the number of clusters). The value i, j represents the probability of the i-th observation belong to j-th cluster.

Author(s)

Relvas, C. & Fujita, A.

References

Stage I non-small cell lung cancer stratification by using a model-based clustering algorithm with covariates, Relvas et al.

Examples

```
set.seed(42)
X = cbind(rnorm(10), rnorm(10))
Y = cbind(rnorm(10), rnorm(10))
K = 2

fit <- CemCOVar(X, Y, K, Y[,1], max_iter=2, n_start=1, cores=1)
prob <- EStepVar(X, Y, fit[[1]], K, Y[,1])
```

`LogLike` *Log likelihood of the CemCO algorithm with covariates effects on distributions means.*

Description

Returns the log-likelihood of the CemCO algorithm with covariates effects on distributions means.

Usage

```
LogLike(data, Y, phi, G)
```

Arguments

`data`	A numeric vector, matrix, or data frame of observations. Non-numerical values should be converted to integer or float (e.g. dummies). If matrix or data frame, rows and columns correspond to observations (n) and variables (P).
`Y`	numeric matrix of data to use as covariates. Non-numerical values should be converted to integer or float (e.g. dummies).
`phi`	list of fitted parameters in the same format as the output of the CemCO function
`G`	An integer specifying the numbers of mixture components (clusters)

Details

Calculate the log likelihood of the mixture guassian with covariates effects on distributions means. This function is used in the optimization process of the EM algorithm used to estimate the CemCO parameters.

Value

Return the value of the log likelihood.

Author(s)

Relvas, C. & Fujita, A.

References

Stage I non-small cell lung cancer stratification by using a model-based clustering algorithm with covariates, Relvas et al.

Examples

```
set.seed(42)
X = cbind(rnorm(60), rnorm(60))
Y = cbind(rnorm(60), rnorm(60))
K = 2

fit <- CemCO(X, Y, K, max_iter=10, n_start=1, cores=1)
ll <- LogLike(X, Y, fit[[1]], K)
```

LogLikeVar *Log likelihood of the CemCO algorithm with covariates effects on distributions means and distributions covariance matrices.*

Description

Returns the log-likelihood of the CemCO algorithm with covariates effects on distributions means and distributions covariance matrices.

Usage

```
LogLikeVar(data, Y, phi, G, y_cov)
```

Arguments

data	A numeric vector, matrix, or data frame of observations. Non-numerical values should be converted to integer or float (e.g. dummies). If matrix or data frame, rows and columns correspond to observations (n) and variables (P).
Y	numeric matrix of data to use as covariates. Non-numerical values should be converted to integer or float (e.g. dummies).
phi	list of fitted parameters in the same format as the output of the CemCO function.
G	An integer specifying the numbers of mixture components (clusters).
y_cov	numeric matrix of data to use as covariates for the covariance effect. Non-numerical values should be converted to integer or float (e.g. dummies).

Details

Calculate the log likelihood of the mixture guassian distribution given by the CemCO algorithm with covariates effects on distributions means and distributions covariance matrices. This function is used in the optimization process of the EM algorithm used to estimate the CemCO parameters.

Value

Return the value of the log likelihood.

Author(s)

Relvas, C. & Fujita, A.

References

Stage I non-small cell lung cancer stratification by using a model-based clustering algorithm with covariates, Relvas et al.

Examples

```
set.seed(42)
X = cbind(rnorm(10), rnorm(10))
Y = cbind(rnorm(10), rnorm(10))
K = 2

fit <- CemCOVar(X, Y, K, Y[,1], max_iter=2  , n_start=1, cores=1)
ll <- LogLikeVar(X, Y, fit[[1]], K, Y[,1])
```

Index