Package 'VarSelLCM'

June 10, 2015

Type Package

Version 1.2

Date 2015-06-08

Author Matthieu Marbac and Mohammed Sedki

Title Variable Selection for Model-Based Clustering using the Integrated Complete-Data Likelihood of a Latent Class Model

Description
Uses a finite mixture model for performing the cluster analysis with variable selection of continuous data by assuming independence between classes. The package deals dataset with missing values by assuming that values are missing at random. The one-dimensional marginals of the components follow Gaussian distributions for facilitating both model interpretation and model selection. The variable selection is led by the Maximum Integrated Complete-Data Likelihood criterion. The maximum likelihood inference is done by an EM algorithm for the selected model. This package also performs the imputation of missing values.

Maintainer Mohammed Sedki <mohammed.sedki@u-psud.fr>

License GPL (>= 2)

Imports methods, Rcpp (>= 0.11.1), parallel

LinkingTo Rcpp, RcppArmadillo

ByteCompile true

LazyLoad yes

Collate 'CheckInputs.R' 'DataCstr.R' 'VSLCMGrlClasses.R' 'ICLexact.R' 'DesignOutput.R' 'Summary.R' 'Print.R' 'VarSelLCM.R' 'RcppExports.R' 'Imputation.R' 'withoutmixture.R'

NeedsCompilation yes

Repository CRAN

Date/Publication 2015-06-10 19:00:07

R topics documented:

`VarSelLCM-package` *Variable Selection in model-based clustering managed by the Latent Class Model for analyzing continuous data with missing values.*

Description

The R package VarSelLCM uses a finite mixture model for performing the cluster analysis with variable selection of continuous data by assuming independence between classes. The package deals dataset with missing values by assuming that values are missing at random. The one-dimensional marginals of the components follow Gaussian distributions for facilitating both model interpretation and model selection. The variable selection is led by an alternated optimization procedure for maximizing the MICL criterion. The maximum likelihood inference is done by an EM algorithm for the selected model. This package also performs the imputation of missing values.

Details

Package:	VarSelLCM
Type:	Package
Version:	1.2
Date:	2015-06-08
License:	GPL (>= 2)

The main functions to use are `VarSelCluster` and `VarSelImputation`.

Function `VarSelCluster` carries out the model selection by maximizing the MICL criterion, then it performs the maximum likelihood estimation of the selected model via an EM algorithm.

Function `VarSelImputation` performs the imputation of missing values by taking the expectation of the missing values conditionally on the model, its parameters and on the observed variables.

Tool methods `summary` and `print` are available for facilitating the interpretation.

Author(s)

Matthieu Marbac and Mohammed Sedki Maintainer: Mohammed Sedki <mohammed.sedki@u-psud.fr>

References

M. Marbac and M. Sedki (2015). Variable Selection for Model-Based Clustering using the Integrated Completed-Data Likelihood. *Preprint*.

Examples

```
## Not run:
require(VarSelLCM)
data(banknote)

results <- VarSelCluster(banknote[,-1], 2, nbcores=2, initModel=40)

summary(results)

print(results)

## End(Not run)
```

---

banknote *Swiss banknotes data*

Description

The data set contains six measurements made on 100 genuine and 100 counterfeit old-Swiss 1000-franc bank notes.

Usage

```
data(banknote)
```

Format

A data frame with the following variables:

Status the status of the banknote: `genuine` or `counterfeit`

Length Length of bill (mm)

Left Width of left edge (mm)

Right Width of right edge (mm)

Bottom Bottom margin width (mm)

Top Top margin width (mm)

Diagonal Length of diagonal (mm)

Source

Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics: A practical approach.* London: Chapman & Hall, Tables 1.1 and 1.2, pp. 5-8.

`print` *Print an object of class VSLCMresults*

Description

Print an object of class VSLCMresultsContinuous

Arguments

`x`	An object of class `VSLCMresultsContinuous`

Value

NULL.

See Also

`print`

Examples

```
## Not run:
require(VarSelLCM)
data(banknote)
results <- VarSelCluster(banknote[,-1], 2, initModel=40)
print(results)

## End(Not run)
```

`summary` *Produce result summary of a VSLCMresultsContinuous class*

Description

Produce result summary of a VSLCMresultsContinuous class

Arguments

`object`	An object of class `VSLCMresultsContinuous`

Value

NULL. Summaries to standard out.

See Also

summary

Examples

```
## Not run:
require(VarSelLCM)
data(banknote)
results <- VarSelCluster(banknote[,-1], 2, initModel=40)
summary(results)

## End(Not run)
```

VarSelCluster *VarSelCluster*

Description

Function VarSelCluster performs the cluster analysis of a continuous data set with missing values. If option vbleSelec=TRUE, then a variable selection is performed by an iterative algorithm to provide the model maximizing the MICL criterion. If option paramEstim=TRUE, then the maximum likelihood inference is performed by an EM algorithm for the selected model. Missing data are managed by assuming that data are missing at random.

Usage

```
VarSelCluster(x, g, initModel=50, vbleSelec=TRUE,
paramEstim=TRUE, nbcores=1, nbSmall=250, iterSmall=20,
nbKeep=50, iterKeep=10**3, tolKeep=10**(-3))
```

Arguments

x	A data frame or a matrix where rows correspond to observations and columns correspond to variables. Variables must be "numeric".
g	An integer specifying the number of mixture components.
initModel	The number of random initializations of the algorithm. The default is 50.
vbleSelec	Logical indicating if a variable selection is performed. The default is TRUE.
paramEstim	Logical indicating if the maximum likelihood (ML) inference is performed for the best model (according to the MICL criterion). The default is TRUE.
nbcores	Number of cores. The default is 1.
nbSmall	Numeric indicating the number of SmallEM algorithms performed for the ML inference. The default is 250.
iterSmall	Numeric indicating the number of iterations for each SmallEM algorithm. The default is 20.

nbKeep	Numeric indicating the number of chains used for the final EM algorithm. The default is 50.
iterKeep	Numeric indicating the maximal number of iterations for each EM algorithm. The default is 1000.
tolKeep	Numeric indicating the maximal gap between two successive iterations of EM algorithm which stops the algorithm. The default is 0.001.

Value

An object of class VSLCMresultsContinuous containing the model maximizing the MICL criterion. Thus, it indicates which variables are relevant to the clustering. It also provides the maximum likelihood estimates associated to this model.

References

M. Marbac and M. Sedki (2015). Variable Selection for Model-Based Clustering using the Integrated Completed-Data Likelihood. *Preprint*.

Examples

```
## Not run:
require(VarSelLCM)
data(banknote)

# Cluster analysis without variable selection
resultswithout <- VarSelCluster(banknote[,-1], 2, vbleSelec=FALSE)
summary(resultswithout)

# Cluster analysis with variable selection
results <- VarSelCluster(banknote[,-1], 2, nbcores=2, initModel=40)
summary(results)

## End(Not run)
```

VarSelImputation *VarSelImputation*

Description

This function performs the imputation on individuals having missing values. It uses the parameters resulting from function VarSelCluster. Imputation is made by taking the expectation of the missing variables conditionally on the observed variables and on the parameters.

Usage

```
VarSelImputation(obj, ind)
```

Arguments

obj	An object resulting from function VarSelCluster (i.e. an object of VSLCMresultsContinuous).
ind	An vector of indicating which individuals are imputated (by default, all the individuals having missing values are imputated).

Value

A data.frame contaning the imputation made on the selected individuals.

Examples

```
## Not run:
require(VarSelLCM)
data(banknote)
banknote[1:3,2:3] <- NA

results <- VarSelCluster(banknote[,-1], 2, initModel=40)


ximput <- VarSelImputation(results, c(1:3))

# Individuals without imputation
print(banknote[c(1:3),-1])

# Individuals with imputation
print(ximput)

## End(Not run)
```

VSLCMcriteria-class *Class* "VSLCMcriteria"

Description

This class contains the information criteria

Slots

loglikelihood: Value of the log-likelihood

BIC: Values of the BIC criterion

ICL: Values of the ICL criterion

MICL: Values of the MICL criterion

degeneracyrate: Rate of EM having degenerated

Examples

```
showClass("VSLCMcriteria")
```

VSLCMdataContinuous-class *Class* "VSLCMdataContinuous"

Description

This class contains the data information

Slots

n: Number of individuals.

d: Number of continuous variables.

data: Whole data set.

notNA: Logical indicating if the realization is missing.

priors: Values of the hyper-parameters. One row corresponds to one variable, and hyperparameters are set as follows: alpha=beta=1, lambda=mean(x[,j]), delta=1/(100g).

Examples

```
showClass("VSLCMdataContinuous")
```

VSLCMmodel-class *Class* "VSLCMmodel"

Description

This class contains the model information

Slots

g: Number of components

omega: Boolean vector indicating if each variable is relevant (1) or not (0) to the clustering

Examples

```
showClass("VSLCMmodel")
```

VSLCMparamContinuous-class *Class* "VSLCMparamContinuous"

Description

This class contains the model information

Slots

pi: proportions.

mu: means.

sd: standart deviations.

Examples

```
showClass("VSLCMparamContinuous")
```

VSLCMpartitions-class *Class* "VSLCMpartitions"

Description

This class contains the estimated partitions

Slots

zMAP: A vector indicating the component membership of each individual by using the MAP rule computed for the best model with its maximum likelihood estimates

zOPT: Partition maximizing the integrated complete-data likelihood of the selected model

tik: Fuzzy partition computed for the best model with its maximum likelihood estimates

Examples

```
showClass("VSLCMpartitions")
```

VSLCMresultsContinuous-class
Class "VSLCMresultsContinuous"

Description

This class contains the model information

Slots

data: An object of class VSLCMdataContinuous

criteria: An object of class VSLCMcriteria

partitions: An object of class VSLCMpartitions

model: An object of class VSLCMmodel

strategy: An object of class VSLCMstrategy

param: An object of class VSLCMparamContinuous

cvrate: Occurrence where the couple (m,z) maximizing the MICL criterion has been found.

Examples

```
showClass("VSLCMresultsContinuous")
```

VSLCMstrategy-class *Class* "VSLCMstrategy"

Description

This class contains the estimated partitions

Slots

initModel: number of initialisations for the model selection algorithm

vbleSelec: logical indicating if the variable selection is performed

paramEstim: logical indicating if the parameter estimation is performed

parallel: logical indicating if a parallelisation is done

nbSmall: number of small EM

iterSmall: number of iteration for the small EM

nbKeep: number of chains kept for the EM

iterKeep: maximum number of iterations for the EM

tolKeep: value of the difference between successive iterations of EM stopping the EM

Examples

```
showClass("VSLCMstrategy")
```

Index