Package ‘SelvarMix’

September 20, 2015

Type Package

Title Regularization for Variable Selection in Model-Based Clustering and Discriminant Analysis

Version 1.1

Date 2015-09-20

Author Mohammed Sedki, Gilles Celeux, Cathy Maugis-Rabusseau

Maintainer Mohammed Sedki <mohammed.sedki@umspud.fr>

Description Performs a regularization approach to variable selection in the model-based clustering and classification frameworks. First, the variables are arranged in order with a lasso-like procedure. Second, the method of Maugis, Celeux, and Martin-Magniette (2009, 2011) is adapted to define the role of variables in the two frameworks.

License GPL (>= 3)

Imports Rcpp (>= 0.11.1), glasso, parallel, Rmixmod, methods

LinkingTo Rcpp, RcppArmadillo

NeedsCompilation yes

Repository CRAN

Date/Publication 2015-09-20 22:51:05

R topics documented:

SelvarMix-package ........................................ 2
scenarioCor ........................................... 4
SelvarClustLasso ....................................... 5
SelvarLearnLasso ...................................... 7
SortvarClust ........................................... 9
SortvarLearn .......................................... 11

Index 13
SelvarMix-package

Regularization for variable selection in model-based clustering and discriminant analysis

Description

SelvarMix is a package where a regularization approach of variable selection is considered in model-based clustering and discriminant analysis frameworks. First, this procedure consists of ranking the variables with a lasso-like procedure. Second, the method of Maugis et al (2009, 2011) is adapted to define the role of variables in the two frameworks. SelvarMix provides a faster variable selection algorithm than the backward stepwise or forward stepwise algorithms of Maugis et al (2009), allowing us to study high-dimensional datasets.

Details

Package: SelvarMix
Type: Package
Version: 1.0
Date: 2014-04-03
License: GPL-3 + file LICENSE
LazyLoad: yes

The general purpose of the package is to perform variable selection in model-based clustering and discriminant analysis. It focus on model-based clustering, where the clusters are assumed to arise from Gaussian distributions. The most achieved model in model-based clustering has been proposed by Maugis et al (2009). This so-called \textit{SRUW} modeling considers three roles of variables: one variable my belong to the relevant clustering set \( S \), the redundant variable set \( U \) or the independent variable set \( W \). Moreover, the redundant variables may be explained by a subset \( R \) of the relevant variables \( S \). In order to avoid the slow of this algorithm when data with numerous variables are studied, the SelvarMix procedure is proposed. It proceeds in two steps: First, the variables are ranked using a lasso-like procedure analogous to the one of Zhou et al (2009); second, the \textit{SRUW} procedure is run on this ranked set of variables.

Author(s)

Author: Mohammed Sedki, Gilles Celeux and Cathy Maugis-Rabusseau

References


Examples

```r
## Not run:
## Simulated data example as shown in Maugis et al. (2009) (correlated scenario 2)
## n = 2000 observations, p = 14 variables
require(Rmixmod)
require(glasso)
data(scenarioCor)
data.cor <- scenarioCor[,1:14]
labels.cor <- scenarioCor[,15]
lambda <- seq(20, 100, by = 10)
rho <- seq(1, 2, length=2)
hybrid.size <- 3
models <- mixmodGaussianModel(family = "spherical", equal.proportions = TRUE)
regModel <- c("LI","LB","LC")
indepModel <- c("LI","LB")

## variable selection in model-based clustering
nbCluster <- c(3,4)
criterion <- "BIC"
simulate.cl <- SelvarClustLasso(data.cor, nbCluster, lambda, rho, hybrid.size, criterion, models, regModel, indepModel)

## variables selection in discriminant analysis
## training sample : n = 1900, p = 14 variables
data.learn <- scenarioCor[1:1900,1:14]
labels.learn <- scenarioCor[1:1900,15]

## testing sample : n = 100, p = 14 variables
data.test <- scenarioCor[1901:2000,1:14]
labels.test <- scenarioCor[1901:2000,15]
lambda <- seq(20, 50, length = 10)
simulate.da <- SelvarLearnLasso(data.learn, labels.learn, lambda, rho, hybrid.size, models, regModel, indepModel, data.test, labels.test)

## End(Not run)
```
Description

The dataset consists of 2000 data points in $R^{14}$. On the subset of relevant clustering variables $S = \{1, 2\}$, data are distributed from a mixture of four equiprobable spherical Gaussian distributions with means $(0,0),(4,0),(0,2)$ and $(4,2)$. The subset of redundant variables is $U = \{3-11\}$ that are explained by the subset of predictor variables $R = \{1, 2\}$. The last three variables are independent $W = \{11, 12, 13\}$.

Format

A data matrix with 2000 observations on 14 variables and the last column contains the labels.

scenarioCor[,1:14] a numeric matrix containing the observations
scenarioCor[,15] an integer vector containing the labels

Details

The subset $U$ of redundant variables is simulated as follows:

$$x^U = (0, 0, 0, 0.4, 0.8, ..., 2) + x^Sb + \varepsilon, \text{ with } \varepsilon \sim N(0, \Omega)$$

The subset $W$ of independent variables is simulated as follows:

$$x^W \sim N((3.2, 3.6, 4), I_3)$$

For more details on the regression coefficients $b$ and the covariance matrix $\Omega$ see Maugis et al.(2009).

References


Examples

data(scenarioCor)
Description

This function implements the variable selection in model-based clustering using a lasso ranking on the variables as described in Sedki et al (2014). The variable ranking step uses the penalized EM algorithm of Zhou et al (2009).

Usage

SelvarClustLasso(data, nbCluster, lambda, rho, hybrid.size, criterion, models, regModel, indepModel, nbCores)

Arguments

data matrix containing quantitative data. Rows correspond to observations and columns correspond to variables

nbCluster numeric listing of the number of clusters (must be positive integers)

lambda numeric listing of the tuning parameter for $\ell_1$ mean penalty

rho numeric listing of the tuning parameter for $\ell_1$ precision matrix penalty

hybrid.size optional parameter make less strength the hybrid forward and backward algorithms to select $S$ and $W$ sets

criterion list of character defining the criterion to select the best model. The best model is the one with the highest criterion value. Possible values: "BIC", "ICL", c("BIC", "ICL"). Default is "BIC"

models a Rmixmod [Model] object defining the list of models to run. The models Gaussian_pk_L_C, Gaussian_pk_Lk_C, Gaussian_pk_L_Ck, and Gaussian_pk_Lk_Ck are called by default (see mixmodGaussianModel() in Rmixmod package to specify other models)

regModel list of character defining the covariance matrix form for the linear regression of $U$ on the $R$ set of variables. Possible values: "LI" for spherical form, "LB" for diagonal form and "LC" for general form. Possible values: "LI", "LB", "LC", c("LI", "LB"), c("LI", "LC"), c("LB", "LC") and c("LI", "LB", "LC"). Default is c("LI", "LB", "LC")

indepModel list of character defining the covariance matrix form for independent variables $W$. Possible values: "LI" for spherical form and "LB" for diagonal form. Possible values: "LI", "LB", c("LI", "LB"). Default is c("LI", "LB")

nbCores number of CPUs to be used when parallel computing is utilized (default is 2)
Value

for each criterion BIC or ICL

S The selected set of relevant clustering variables
R The selected subset of regressors
U The selected set of redundant variables
W The selected set of independent variables
criterionValue The criterion value for the selected model
nbCluster The selected number of clusters
model The selected Gaussian mixture form
regModel The selected covariance form for the regression
indepModel The selected covariance form for the independent gaussian distribution
proba Matrix containing the conditional probabilities of belonging to each cluster for all observations
partition Vector of length n containing the cluster assignments of the n observations according to the Maximum-a-Posteriori rule

Author(s)

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

References


See Also

SelvarLearnLasso SortvarClust SortvarLearn scenarioCor

Examples

```r
## Not run:
## Simulated data example as shown in Maugis et al. (2009)
## n = 2000 observations, p = 14 variables
require(Rmixmod)
require(glasso)
data(scenarioCor)
data.cor <- scenarioCor[,1:14]
```
**SelvarLearnLasso**

Regularization for variable selection in discriminant analysis

**Description**

This function implements the variable selection in discriminant analysis using a lasso ranking on the variables as described in Sedki et al (2014). The variable ranking step uses the penalized EM algorithm of Zhou et al (2009) (adapted in Sedki et al (2014) for the discriminant analysis settings). A testing sample can be used to compute the averaged classification error rate.

**Usage**

```r
SelvarLearnLasso(data, knownlabels, lambda, rho, hybrid.size, models, regModel, indepModel, dataTest, labelsTest, nbCores)
```

**Arguments**

- `data` matrix containing quantitative data. Rows correspond to observations and columns correspond to variables
- `knownlabels` an integer vector or a factor of size number of observations. Each cell corresponds to a cluster affectation
- `lambda` numeric listing of tuning parameter for $\ell_1$ mean penalty
- `rho` numeric listing of tuning parameter for $\ell_1$ precision matrix penalty
- `hybrid.size` optional parameter make less strength the hybrid forward and backward algorithms to select $S$ and $W$ sets
- `models` a Rmixmod [Model] object defining the list of models to run. The models Gaussian_pk_L_C, Gaussian_pk_Lk_C, Gaussian_pk_L-C, and Gaussian_pk_Lk_Ck are called by default (see mixmodGaussianModel() in Rmixmod package to specify other models)
**Value**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>The selected set of relevant clustering variables</td>
</tr>
<tr>
<td>R</td>
<td>The selected set of regressors</td>
</tr>
<tr>
<td>U</td>
<td>The selected set of redundant variables</td>
</tr>
<tr>
<td>W</td>
<td>The selected set of independent variables</td>
</tr>
<tr>
<td>criterionValue</td>
<td>The criterion value for the selected model</td>
</tr>
<tr>
<td>nbCluster</td>
<td>The selected number of clusters</td>
</tr>
<tr>
<td>model</td>
<td>The selected covariance model</td>
</tr>
<tr>
<td>regModel</td>
<td>The selected covariance form for the regression</td>
</tr>
<tr>
<td>indepModel</td>
<td>The selected covariance form for the independent variables</td>
</tr>
<tr>
<td>proba</td>
<td>Optional: matrix containing the conditional probabilities of belonging to each cluster for the testing observations</td>
</tr>
<tr>
<td>partition</td>
<td>Optional: vector containing the cluster assignments of the testing observations according to the Maximum-a-Posteriori rule</td>
</tr>
<tr>
<td>error</td>
<td>Optional: error rate done by the predicted partition (obtained using Maximum-A-Posteriori rule)</td>
</tr>
</tbody>
</table>

**Author(s)**

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

**References**


Sedki, M., Celeux, G., Maugis-Rabusseau, C., 2014. "SelvarMix: A R package for variable selection in model-based clustering and discriminant analysis with a regularization approach". Inria Research Report available at [http://hal.inria.fr/hal-01053784](http://hal.inria.fr/hal-01053784)
See Also

SelvarClustLasso SelvarLearn SortvarClust scenarioCor

Examples

```R
## Not run:
## Simulated data example as shown in Sedki et al (2014)
require(Rmixmod)
require(glasso)
data(scenarioCor)

lambda <- seq(20, 50, length = 10)
rho <- seq(1, 2, length=2)
hybrid.size <- 3
models <- mixmodGaussianModel(family = “spherical”, equal.proportions = TRUE)
regModel <- c("LI","LB","LC")
indepModel <- c("LI","LB")

## variables selection in discriminant analysis
## training sample : n = 1900 observations, p = 14 variables
data.learn <- scenarioCor[1:1900,1:14]
labels.learn <- scenarioCor[1:1900,15]

data.test <- scenarioCor[1901:2000,1:14]
labels.test <- scenarioCor[1901:2000,15]
simulate.da <- SelvarLearnLasso(data.learn, labels.learn, lambda, rho, hybrid.size,
models, regModel, indepModel, data.test, labels.test)

## End(Not run)
```

SortvarClust Variable ranking with LASSO in model-based clustering

Description

This function implements variable ranking procedure in model-based clustering using the penalized EM algorithm of Zhou et al (2009).

Usage

SortvarClust(data, nbCluster, lambda, rho, nbCores)

Arguments

data matrix containing quantitative data. Rows correspond to observations and columns correspond to variables

nbCluster numeric listing of the number of clusters (must be integers)
lambda: numeric listing of the tuning parameter for $\ell_1$ mean penalty
rho: numeric listing of the tuning parameter for $\ell_1$ precision matrix penalty
nbCores: number of CPUs to be used when parallel computing is utilized (default is 2)

Value

matrix with rows corresponding to variable ranking. Each row corresponds to a value nbCluster.

Author(s)

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

References


See Also

SortvarLearn

Examples

```r
## Not run:
## Simulated data example as shown in Sedki et al (2014)
## n = 2000 observations, p = 14 variables
require(glm)
data(scenarioCor)
data.cor <- scenarioCor[,1:14]

lambda <- seq(20, 100, by = 10)
rho <- seq(1, 2, length=2)
nbCluster <- c(3, 4)

## variable ranking in model-based clustering
var.ranking.cl <- SortvarClust(data.cor, nbCluster, lambda, rho)

## End(Not run)
```
SortvarLearn Variable ranking with LASSO in discriminant analysis

Description

Usage
SortvarLearn(data, knownlabels, lambda, rho, nbCores)

Arguments
- data: matrix containing quantitative data. Rows correspond to observations and columns correspond to variables
- knownlabels: an integer vector or a factor of size number of observations. Each cell corresponds to a cluster affectation. So the maximum value is the number of clusters.
- lambda: numeric listing of tuning parameter for $\ell_1$ mean penalty
- rho: numeric listing of tuning parameter for $\ell_1$ precision matrix penalty
- nbCores: number of CPUs to be used when parallel computing is utilized (default is 2)

Value
vector of integers corresponding to variable ranking.

Author(s)
Mohammed Sedki <mohammed.sedki@u-psud.fr>

References

See Also
SortvarClust
Examples

```r
## Not run:
## Simulated data example as shown in Sedki et al (2014)
## n = 2000 observations, p = 14 variables
require(glasso)
data(scenariocor)
data.cor <- scenariocor[,1:14]
labels.cor <- scenariocor[,15]

lambda <- seq(20, 50, length = 10)
rho <- seq(1, 2, length=2)

## variable ranking in discriminant analysis
var.ranking.da <- SortvarLearn(data.cor, labels.cor, lambda, rho)

## End(Not run)
```
Index

-* Topic Penalized discriminant analysis
  SortvarLearn, 11
-* Topic Penalized model-based clustering
  SortvarClust, 9
-* Topic Variable ranking
  SortvarClust, 9
  SortvarLearn, 11
-* Topic datasets
  scenarioCor, 4
-* Topic discriminant analysis, variable selection, lasso ranking and graphical lasso
  SelvarLearnLasso, 7
-* Topic model-based clustering, discriminant analysis, variable selection, lasso ranking and graphical lasso
  SelvarClustLasso, 5
-* Topic package
  SelvarMix-package, 2

Model, 5, 7

scenarioCor, 4, 6, 9
SelvarClustLasso, 5, 9
SelvarLearnLasso, 6, 7
SelvarMix-package, 2
SortvarClust, 6, 9, 11
SortvarLearn, 6, 9, 10, 11