Package ‘SelvarMix’

October 16, 2017

Type Package
Title Regularization for Variable Selection in Model-Based Clustering and Discriminant Analysis
Version 1.2.1
Date 2017-10-16
Author Mohammed Sedki, Gilles Celeux, Cathy Maugis-Rabusseau
Maintainer Mohammed Sedki <mohammed.sedki@u-psud.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) <doi:10.1016/j.csda.2009.04.013>, <doi:10.1016/j.jmva.2011.05.004> is adapted to define the role of variables in the two frameworks.
License GPL (>= 3)
Depends R (>= 3.1.0), glasso, Rmixmod, parallel, base
Imports Rcpp (>= 0.11.1), methods
LinkingTo Rcpp, RcppArmadillo
NeedsCompilation yes
Repository CRAN
Date/Publication 2017-10-16 16:18:03 UTC

R topics documented:

SelvarMix-package ................................................................. 2
scenarioCor ................................................................. 3
SelvarClustLasso .......................................................... 4
SelvarLearnLasso ........................................................... 6
SortvarClust ................................................................. 9
SortvarLearn .............................................................. 10
wine ................................................................. 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 $SR UW$ modeling considers three roles of variables: one variable may 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 greedy algorithms when high-dimensional data 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 $SR UW$ 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:
## wine data set
## n = 178 observations, p = 27 variables
data(wine)
## variable selection in model-based clustering
set.seed(123)
obj <- SelvarClustLasso(x=wine[,1:27], nbcluster=1:5, nbcores=4)
summary(obj)
print(obj)

## variables selection in discriminant analysis
set.seed(123)
a <- seq(1, 178, 10)
b <- setdiff(1:178, a)
obj <- SelvarLearnLasso(x=wine[b,1:27], z=wine[b,28], xt=wine[a,1:27], zt=wine[a,28], nbcores=4)
summary(obj)
print(obj)

## End(Not run)
```

---

### scenarioCor

**Simulated quantitative data according SRUW modeling**

**Description**

The dataset consists of 2000 data points in \( \mathbb{R}^{14} \). 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:

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

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)

SelvarClustLasso  Regularization for variable selection in model-based clustering

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(x, nbcluster, lambda, rho, type, rank, hsize, criterion, models, rmodel, imodel, nbcores)

Arguments

- $x$: matrix or data frame 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 parameters for $\ell_1$ mean penalty.
- rho: numeric listing of the tuning parameters for $\ell_1$ precision matrix penalty.
- type: character defining the type of ranking procedure, must be "lasso" or "likelihood". Default is "lasso".
- rank: integer listing the rank of variables with (the length this vector must be equal to the number of variables in the dataset).
- hsize: optional parameter make less strength the 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)

rmodel 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")

imodel 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 used (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

rmodel The selected covariance form for the regression

imodel The selected covariance form for the independent Gaussian distribution

parameters Rmixmod [Parameter] object containing all mixture parameters

regparameters Matrix containing all regression coefficients, each column is the regression coefficients of one redundant variable on the selected $R$ set

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 wine

Examples

## Not run:
## wine data set
## n = 178 observations, p = 27 variables
data(wine)
set.seed(123)
obj <- SelvarClustLasso(x=wine[,1:27], nbcluster=1:5, nbcores=4)
summary(obj)
print(obj)

## End(Not run)

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

SelvarLearnLasso(x, z, lambda, rho, type, rank, hsize, models, rmodel, imodel, xtest, ztest, nbcores)
Arguments

- **x**: matrix containing quantitative data. Rows correspond to observations and columns correspond to variables.
- **z**: an integer vector or a factor corresponding to labels of data.
- **lambda**: numeric listing of tuning parameters for $\ell_1$ mean penalty.
- **rho**: numeric listing of tuning parameters for $\ell_1$ precision matrix penalty.
- **type**: character defining the type of ranking procedure, must be "lasso" or "likelihood". Default is "lasso".
- **rank**: integer listing the rank of variables with (the length of this vector must be equal to the number of variables in the dataset).
- **hsize**: optional parameter make less strength the 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_Ck, and Gaussian_pk_Lk_Ck are called by default (see mixmodGaussianModel() in Rmixmod package to specify other models).
- **rmodel**: list of character defining the covariance matrix form for the linear regression of $U$ on the $R$ set of variable. 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").
- **imodel**: 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").
- **xtest**: matrix containing quantitative testing data. Rows correspond to observations and columns correspond to variables.
- **ztest**: an integer vector or a factor of size number of testing observations. Each cell corresponds to a cluster affectation.
- **nbcores**: number of CPUs to be used when parallel computing is used (default is 2).

Value

- **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.
- **model**: The selected covariance model.
- **rmodel**: The selected covariance form for the regression.
- **imodel**: The selected covariance form for the independent variables.
- **parameters**: Rmixmod [Parameter] object containing all mixture parameters.
SelvarLearnLasso

regparameters Matrix containing all regression coefficients, each column is the regression coefficients of one redundant variable on the selected R set

proba Optional: matrix containing the conditional probabilities of belonging to each cluster for the testing observations

partition Optional: vector containing the cluster assignments of the testing observations according to the Maximum-a-Posteriori rule. When testing dataset is missed, we use the training dataset as testing one

error Optional: error rate done by the predicted partition (obtained using Maximum-A-Posteriori rule). When testing dataset is missed, we use the training dataset as testing one

Author(s)

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

References


See Also

SelvarClustLasso SortvarLearn SortvarClust wine

Examples

```r
## Not run:
## wine data set
## n = 178 observations, p = 27 variables
data(wine)
set.seed(123)
a <- seq(1, 178, 10)
b <- setdiff(1:178, a)
obj <- SelvarLearnLasso(x=wine[b,1:27], z=wine[b,28], xt=wine[a,1:27], zt=wine[a,28], nbcores=4)
summary(obj)
print(obj)

## End(Not run)
```
Description

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

Usage

`SortvarClust(x, nbcluster, type, lambda, rho, nbcores)`

Arguments

- `x` matrix containing quantitative data. Rows correspond to observations and columns correspond to variables
- `nbcluster` numeric listing of the number of clusters (must be integers)
- `type` character defining the type of ranking procedure, must be "lasso" or "likelihood". Default is "lasso"
- `lambda` numeric listing of the tuning parameters for $\ell_1$ mean penalty
- `rho` numeric listing of the tuning parameters for $\ell_1$ precision matrix penalty
- `nbcores` number of CPUs to be used when parallel computing is utilized (default is 2)

Value

matrix where rows correspond to variable ranking. Each row corresponds to a competing value of `nbcluster`.

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

SortvarLearn
Examples

```r
## Not run:
## wine data set
## n = 178 observations, p = 27 variables
require(Rmixmod)
require(glasso)
data(wine)
set.seed(123)
obj <- SortvarClust(x=wine[,1:27], nbcluster=1:5, nbcores=4)

## End(Not run)
```

---

**SortvarLearn**

Variable ranking with LASSO in discriminant analysis

**Description**


**Usage**

`SortvarLearn(x, z, type, lambda, rho, nbcores)`

**Arguments**

- `x` matrix containing quantitative data. Rows correspond to observations and columns correspond to variables.
- `z` an integer vector or a factor corresponding to labels of data.
- `type` character defining the type of ranking procedure, must be "lasso" or "likelihood". Default is "lasso".
- `lambda` numeric listing of tuning parameters for $\ell_1$ mean penalty.
- `rho` numeric listing of tuning parameters 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:
## wine data set
## n = 178 observations, p = 27 variables
require(Rmixmod)
require(glasso)
data(wine)
set.seed(123)
obj <- SortvarLearn(x=wine[,1:27], z=wine[,28], nbcores=4)

## End(Not run)
```

wine 

Wine data set

Description

This data set is made of 178 observations (Italian wines) described by 27 variables (physicochemical measures). These wines come from three different regions of Italy.

Usage

data("wine")

Format

We have labels and data as follows:

The last column of the data frame (wine[,28]): it indicates the class label 1, 2 or 3.
The data involving columns 1 to 27:
Alcohol
Sugar-free_extract
Fixed_acidity
Tartaric_acid
Malic_acid
Uronic_acids
pH
Ash
Alcalinity_of_ash
Potassium
Calcium
Magnesium
Phosphate
Chloride
Total_phenols
Flavanoids
Nonflavanoid_phenols
Proanthocyanins
Color_Intensity
Hue
OD280/OD315_of_diluted_wines
OD280/OD315_of_flavanoids
Glycerol
2-3-butanediol
Total_nitrogen
Proline
Methanol

Examples

data(wine)

head(wine)
Index

*Topic Penalized discriminant analysis
  SortvarLearn, 10
*Topic Penalized model-based clustering
  SortvarClust, 9
*Topic Variable ranking
  SortvarClust, 9
  SortvarLearn, 10
*Topic datasets
  scenarioCor, 3
  wine, 11
*Topic discriminant analysis, variable selection, lasso ranking and graphical lasso
  SelvarLearnLasso, 6
*Topic model-based clustering, discriminant analysis, variable selection, lasso ranking and graphical lasso
  SelvarClustLasso, 4
*Topic package
  SelvarMix-package, 2

Model, 5, 7

Parameter, 5, 7

scenarioCor, 3
SelvarClustLasso, 4, 8
SelvarLearnLasso, 6, 6
SelvarMix-package, 2
SortvarClust, 6, 8, 9, 11
SortvarLearn, 6, 8, 9, 10

wine, 6, 8, 11