Package 'CPMA'

January 13, 2022

Type Package

Title Change plane (point) model average for subgroup identification

Version 1.0

Date 2022-01-22

Author Pan Liu, Jialiang Li, Yaguang Li

Maintainer Pan Liu <e0647249@u.nus.edu>

Depends R (>= 4.0.0), stats, grpreg, plus, BB, ncvreg, nloptr

Description Tools for subgroup identification and response prediction through change plane (point) model average method.

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

R topics documented:

`CPdata` *simulated data*

Description

This data file contains 500 simulated subjects of 1 response variable (Y) and 5 predictors (X1-X5). They are generated with underlying subgroup structures.

Usage

```
data(CPdata)
```

Format

a data set containig 500 observations of 6 variables.

Source

simulated data

`CPlane` *Change Plane method for subgroup identification.*

Description

Idenfiy subgroups marked by multiple parallel change planes.

Usage

```
CPlane(
  data.tr,
  data.te = NULL,
  yind,
  xind,
  zind,
  ini.theta = rep(1, length(zind)),
  tol = 0.001,
  K = 10
)
```

Arguments

`data.tr`	a matrix or data frame of the training dataset.
`data.te`	a matrix or data frame of the testing dataset.
`yind`	the column number or column name of the response variable.
`xind`	a numeric or character vector, containing the column numbers or column names of the predictors (covariates).
`zind`	a numeric or character vector, containing the column numbers or column names of the threshold variables.
`ini.theta`	a numeric vector, containing the initial values of change plane parameters (the linear combination coefficients of the threshold variables). For better performance of this function, it is highly recommended that the user provide an input based on preknowledge or a reasonable guess of the change plane structure. Default is ini.theta=rep(1, length(zind)).
`tol`	error tolerance in parameter estimation and optimization. Default is tol=1e-3.
`K`	maximum time of iterations in parameter estimation and optimization. Dault is K=10.

Details

This function identifies subgroups and makes predictions by fitting a multithreshold change plane regression model to the given training dataset. Subgroups are thus charaterised by a linear combination of the specified threshold variables. The change plane parameters, thresholds and regression coefficients are estimated by a two stage approach proposed by Li et al (2018). Note that this function should only be used when there are multiple threshold variables so that a linear combination of them makes sense, i.e. length(zind) > 1. Otherwise, one should turn to the CPoint() function instead.

Value

A list consisting of the following components:

`train.res`	estimation results for the training dataset, including estimated response value (Yhat), subgroup id (subgroup), mean squared error (mse), estimated change plane parameter (theta), thresholds on the linear combination of threshold variables (threshold) and the segment regression coefficients (coefficient).
`test.res`	estimation results for the testing dataset, including estimated response value (Yhat), subgroup id (subgroup) and mean squared error (mse).

References

Li, J., Y. Li, B. Jin, and M. R. Kosorok (2021). Multithreshold change plane model: estimation theory and applications in subgroup identification. Statistics in Medicine 40(15), 3440-3459.

Examples

```
out <- CPlane(data.tr=CPdata[1:350,], data.te=CPdata[-(1:350),], yind = 1, xind = 2:6, zind = 2:6,
ini.theta = c(sqrt(0.5), -sqrt(0.5), 0, 0, 0))
```

`CPlaneMA` *Change Plane Model Average method for subgroup identification.*

Description

While CPlane() assumes that subgroups are charaterised by parallel change planes, this function admits change planes that are no necessarily parallel and yields multiple vectors of change plane parameters through model averaging.

Usage

```
CPlaneMA(
  data.tr,
  data.te = NULL,
  yind,
  xind,
  zind,
  ini.theta = matrix(1, 1, length(zind)),
  tol = 0.001,
  K = 10,
  subm.vol = rep(1, length(xind) + 1)
)
```

Arguments

data.tr	a matrix or data frame of the training dataset.
data.te	a matrix or data frame of the testing dataset.
yind	the column number or column name of the response variable.
xind	a numeric or character vector, containing the column numbers or column names of the predictors (covariates).
zind	a numeric or character vector, containing the column numbers or column names of the threshold variables.
ini.theta	a numeric matrix, with each row containing the initial values of change plane parameters for one submodel. Note that the initial values (the rows) should be arranged to match the order of submodels. If there is no enough initial value entered, i.e. the row number of ini.theta is smaller than the number of submodels, the provided initial values will be repeated for use. For better performance of this function, it is highly recommended that the user provide an input based on preknowledge or a reasonable guess of the change plane structure. Default is ini.theta=matrix(1, 1, length(zind)).
tol	error tolerance in parameter estimation and optimization. Default is tol=1e-3.
K	maximum time of iterations in parameter estimation and optimization. Dault is K=10.
subm.vol	a numeric vector, containing the number of predictors (starting from the constant term) whose varying covariate effect should be considered in the corresponding submodel. Default is subm.vol=rep(1, length(xind)+1).

Details

The methodology of this function consists of two levels. In the first level, a number of individual change plane regression submodels are fitted to model the varying covariate effect of some of the given predictors (the constant term may also be included) in the training dataset. In the second level, their model averaging weights are estimated so that a weighted ensemble of these submodels can be used to further approximate the true model. Since the change plane parameters yielded by different submodels are typically not the same, the averaged full model admits change planes that are no necessarily parallel. The structure of submodels can be specified by the user through the input parameter subm.vol.

Value

A list consisting of the following components:

train.res	estimation results for the training dataset, including estimated response value (Yhat), subgroup id (subgroup), mean squared error (mse), estimated change plane parameter (theta), thresholds on the linear combination of threshold variables (threshold) and the segment regression coefficients (coefficient).
test.res	estimation results for the testing dataset, including estimated response value (Yhat), subgroup id (subgroup) and mean squared error (mse).
submodel.res	estimation results for each of the submodels.

References

Li, J., Y. Li, B. Jin, and M. R. Kosorok (2021). Multithreshold change plane model: estimation theory and applications in subgroup identification. Statistics in Medicine 40(15), 3440-3459.

Examples

```
Theta <- matrix(0, nrow = 6, ncol = 5)
Theta[1,] <- rep(1, 5)
Theta[2,] <- c(sqrt(0.5), -sqrt(0.5), 0, 0, 0)
Theta[3,] <- c(0.75, 0, -0.5, 0, sqrt(1-(0.75)^2-(-0.5)^2))
Theta[4,] <- c(sqrt(0.5), -sqrt(0.5), 0, 0, 0)
Theta[5,] <- c(0.75, 0, -0.5, 0, sqrt(1-(0.75)^2-(-0.5)^2))
Theta[6,] <- rep(1, 5)
out <- CPlaneMA(data.tr=CPdata[1:350,], data.te=CPdata[-(1:350),], yind=1, xind=2:6, zind=2:6,
ini.theta=Theta, subm.vol=rep(1, 6))
```

CPoint *Change Point method for subgroup identification.*

Description

Subgroup identification by change point detection on a (single) specified threshold variable Z.

Usage

```
CPoint(data.tr, data.te = NULL, yind, xind, zind, c = seq(0.5, 1.5, 0.1))
```

Arguments

data.tr	a matrix or data frame of the training dataset.
data.te	a matrix or data frame of the testing dataset.
yind	the column number or column name of the response variable.
xind	a numeric or character vector, containing the column numbers or column names of the predictors (covariates).
zind	the column number or column name of the threshold variable.
c	a numeric vector to determine the initial segment length in the splitting stage of TSMCD, i.e. the tentative choices of initial segment length m = c * sqrt(n), where n is the sample size of the training dataset. Default is c = seq(0.5, 1.5, 0.1)

Details

This function identifies subgroups and makes predictions by fitting a change point (threshold regression) model to the given training dataset. The change points (thresholds) and regression coefficients are estimated by a two stage multiple change-points detection (TSMCD) method proposed by Li and Jin (2018).

Value

A list consisting of the following components:

train.res	estimation results for the training dataset, including estimated response value (Yhat), subgroup id (subgroup), mean squared error (mse), detected change points (threshold) and the segment regression coefficients (coefficient).
test.res	estimation results for the testing dataset, including estimated response value (Yhat), subgroup id (subgroup) and mean squared error (mse).

References

Li, J. and B. Jin (2018). Multi-threshold accelerated failure time model. The Annals of Statistics 46(6A), 2657-2682.

Examples

```
out <- CPoint(data.tr=CPdata[1:350,], data.te=CPdata[-(1:350),], yind=1, xind=2:6, zind=2)
```

`CPointMA` *Change Point Model Average method for subgroup identification.*

Description

While CPoint() identifies subgroups based on a single threshold variable, this function detects change points each threshold variable separately and then assembles the subgrouping results through model averaging.

Usage

```
CPointMA(
  data.tr,
  data.te = NULL,
  yind,
  xind,
  zind,
  c = seq(0.5, 1.5, 0.1),
  penalty = c("SCAD", "MCP", "LASSO")
)
```

Arguments

`data.tr`	a matrix or data frame of the training dataset.
`data.te`	a matrix or data frame of the testing dataset.
`yind`	the column number or column name of the response variable.
`xind`	a numeric or character vector, containing the column numbers or column names of the predictors (covariates).
`zind`	a numeric or character vector, containing the column numbers or column names of the threshold variables.
`c`	a numeric vector to determine the initial segment length in the splitting stage of TSMCD for submodel estimation, i.e. the tentative choices of initial segment length m = c * sqrt(n), where n is the sample size of the training dataset. Default is c = seq(0.5, 1.5, 0.1)
`penalty`	the penalty to be used in the model average step, including 'SCAD', 'MCP' and 'LASSO'. Default is 'SCAD'.

Details

The methodology of this function consists of two levels. An individual change point (threshold regression) submodel is fitted based on each of the given threshold variables in the training dataset. Then their model averaging weights are estimated so that a weighted ensemble of these submodels can be used to further approximate the true model. Note that this function should only be used when there are multiple threshold variables, i.e. length(zind) > 1. Otherwise, one should turn to the CPoint() function instead.

Value

A list consisting of the following components:

`train.res`	estimation results for the training dataset, including estimated response value (Yhat), subgroup id (subgroup), mean squared error (mse), detected change points (threshold) and the segment regression coefficients (coefficient).
`test.res`	estimation results for the testing dataset, including estimated response value (Yhat), subgroup id (subgroup) and mean squared error (mse).
`submodel.res`	estimation results for each of the submodels.

References

Li, J. and B. Jin (2018). Multi-threshold accelerated failure time model. The Annals of Statistics 46(6A), 2657-2682.

Examples

```
out <- CPointMA(data.tr=CPdata[1:350,], data.te=CPdata[-(1:350),], yind=1, xind=2:6, zind=2:6)
```

Index