Package 'ClusPred'

June 18, 2021

Type Package

Title Simultaneous Semi-Parametric Estimation of Clustering and Regression

Version 1.0.0

Maintainer Matthieu Marbac <matthieu.marbac-lourdelle@ensai.fr>

Description Parameter estimation of regression models with fixed group effects, when the group variable is missing while group-related variables are available. Parametric and semi-parametric approaches described in Marbac et al. (2020) <arXiv:2012.14159> are implemented.

Imports Rcpp, parallel, ALDqr, ald, quantreg, VGAM

LinkingTo Rcpp, RcppArmadillo

ByteCompile true

URL https://arxiv.org/abs/2012.14159

LazyLoad yes

Author Matthieu Marbac [aut, cre, cph],
Mohammed Sedki [aut],
Christophe Biernacki [aut],
Vincent Vandewalle [aut]

Collate 'cluspred.R' 'RcppExports.R' 'Singleblock_algo_NP.R' 'Singleblock_algo_Param.R' 'Singleblock_prediction.R' 'tool.R' 'TwoSteps_algo.R' 'Twosteps_computelogPDFwithZ.R' 'TwoSteps_Mstep.R'

License GPL (>= 2)

Encoding UTF-8

Depends R (>= 3.5)

RoxygenNote 7.1.0

NeedsCompilation yes

Repository CRAN

Date/Publication 2021-06-18 09:30:06 UTC

R topics documented:

`ClusPred-package` *ClusPred.*

Description

Parameter estimation of regression models with fixed group effects, when the group variable is missing while group-related variables are available.

Details

Package:	ClusPred
Type:	Package
Version:	1.0.0
Date:	2021-06-01
License:	GPL-3
LazyLoad:	yes

References

Simultaneous semi-parametric estimation of clustering and regression, Matthieu Marbac and Mohammed Sedki and Christophe Biernacki and Vincent Vandewalle (2020) <arXiv:2012.14159>.

`cluspred` *Function used for clustering and fitting the regression model*

Description

Estimation of the group-variable Z based on covariates X and estimation of the parameters of the regression of Y on (U, Z)

Usage

```
cluspred(
  y,
  x,
  u = NULL,
  K = 2,
  model.reg = "mean",
  tau = 0.5,
  simultaneous = TRUE,
  np = TRUE,
  nbinit = 20,
  nbCPU = 1,
  tol = 0.01,
  band = (length(y)^(-1/5)),
  seed = 134
)
```

Arguments

y	numeric vector of the traget variable (must be numerical)
x	matrix used for clustering (can contain numerical and factors)
u	matrix of the covariates used for regression (can contain numerical and factors)
K	number of clusters
model.reg	indicates the type of the loss ("mean", "quantile", "expectile", "logcosh", "huber"). Only the losses "mean" and "quantile" are implemented if simultaneous=FALSE or np=FALSE
tau	specifies the level for the loss (quantile, expectile or huber)
simultaneous	oolean indicating whether the clustering and the regression are performed simultaneously (TRUE) or not (FALSE)
np	boolean indicating whether nonparameteric model is used (TRUE) or not (FALSE)
nbinit	number of random initializations
nbCPU	number of CPU only used for linux
tol	to specify the stopping rule
band	bandwidth selection
seed	value of the seed (used for drawing the starting points)

Value

cluspred returns a list containing the model parameters (param), the posterior probabilities of cluster memberships (tik), the partition (zhat) and the (smoothed) loglikelihood)

References

Simultaneous semi-parametric estimation of clustering and regression, Matthieu Marbac and Mohammed Sedki and Christophe Biernacki and Vincent Vandewalle (2020) <arXiv:2012.14159>.

Examples

```
require(ClusPred)
# data loading
data(simdata)

# mean regression with two latent groups in parametric framework and two covariates
res <- cluspred(simdata$y, simdata$x, simdata$u, K=2,
 np=FALSE, nbCPU = 1, nbinit = 10)
# coefficient of the regression
res$param$beta
# proportions of the latent groups
res$param$pi
# posterior probability of the group memberships
head(res$tik)
# partition
res$zhat
# loglikelihood
res$loglike
# prediction (for possible new observations)
pred <- predictboth(simdata$x, simdata$u, res, np = FALSE)
# predicted cluster membreships
pred$zhat
# predicted value of the target variable
pred$yhat


# median regression with two latent groups in nonparametric framework and two covariates
res <- cluspred(simdata$y, simdata$x, simdata$u, K=2,
model.reg = "quantile", tau = 0.5, nbinit = 10)
# coefficient of the regression
res$param$beta
# proportions of the latent groups
res$param$pi
# posterior probability of the group memberships
head(res$tik)
# partition
res$zhat
# smoothed loglikelihood
res$logSmoothlike
# prediction (for possible new observations)
pred <- predictboth(simdata$x, simdata$u, res, np = TRUE)
# predicted cluster membreships
pred$zhat
# predicted value of the target variable
pred$yhat
```

`predictboth` *Prediction (clustering and target variable)*

Description

Prediction for new observations

Usage

```
predictboth(x, u = NULL, result, np = FALSE)
```

Arguments

x	covariates used for clustering
u	covariates of the regression (can be null)
result	results provided by function cluspred
np	boolean indicating whether nonparametric estimation is used (TRUE) or not (FALSE)

Value

predictboth returns a list containing the predicted cluster membership (zhat) and the predicted value of the target variable (yhat).

Examples

```
require(ClusPred)
# data loading
data(simdata)

# mean regression with two latent groups in parametric framework and two covariates
res <- cluspred(simdata$y, simdata$x, simdata$u, K=2,
np=FALSE, nbCPU = 1, nbinit = 10)
# coefficient of the regression
res$param$beta
# proportions of the latent groups
res$param$pi
# posterior probability of the group memberships
head(res$tik)
# partition
res$zhat
# loglikelihood
res$loglike
# prediction (for possible new observations)
pred <- predictboth(simdata$x, simdata$u, res, np = FALSE)
# predicted cluster membreships
pred$zhat
# predicted value of the target variable
pred$yhat


# median regression with two latent groups in nonparametric framework and two covariates
res <- cluspred(simdata$y, simdata$x, simdata$u, K=2,
 model.reg = "quantile", tau = 0.5, nbinit = 10)
```

```
# coefficient of the regression
res$param$beta
# proportions of the latent groups
res$param$pi
# posterior probability of the group memberships
head(res$tik)
# partition
res$zhat
# smoothed loglikelihood
res$logSmoothlike
# prediction (for possible new observations)
pred <- predictboth(simdata$x, simdata$u, res, np = TRUE)
# predicted cluster membreships
pred$zhat
# predicted value of the target variable
pred$yhat
```

---

simdata *Simulated data*

Description

simulated data used for the pacakge examples.

Examples

```
data(simdata)
```

Index