

Package ‘TVsMiss’

April 5, 2018

Type Package

Title Variable Selection for Missing Data

Version 0.1.1

Date 2018-04-05

Author Jiwei Zhao, Yang Yang, and Ning Yang

Maintainer Yang Yang <yyang39@buffalo.edu>

Description Use a regularization likelihood method to achieve variable selection purpose. Likelihood can be worked with penalty lasso, smoothly clipped absolute deviations (SCAD), and mini-max concave penalty (MCP).

Tuning parameter selection techniques include cross validation (CV), Bayesian information criterion (BIC) (low and high), stability of variable selection (sVS), stability of BIC (sBIC), and stability of estimation (sEST).

More details see Jiwei Zhao, Yang Yang, and Yang Ning (2018) <arXiv:1703.06379> ``Penalized pairwise pseudo likelihood for variable selection with nonignorable missing data." *Statistica Sinica*.

License GPL (>= 2)

Imports glmnet, Rcpp

NeedsCompilation yes

Encoding UTF-8

LazyData true

RoxygenNote 6.0.1

LinkingTo Rcpp

URL <https://github.com/yang0117/TVsMiss>

BugReports <https://github.com/yang0117/TVsMiss/issues>

Repository CRAN

Date/Publication 2018-04-05 05:16:08 UTC

R topics documented:

pairdata	2
plot.TVsMiss	2
tvsmis	3

Index	5
--------------	----------

pairdata	<i>Generate paired dataset from original dataset</i>
----------	--

Description

Generate the paired data used for logistics regression

Usage

```
pairdata(data1)
```

Arguments

data1 the original unpaired data

Examples

```
n <- 50
p <- 8
beta <- c(3,0,1.5,0,2,rep(0,p-5))
xm <- matrix(rnorm(n*p),ncol = p, nrow = n)
y <- xm %*% beta + rnorm(n)
colnames(xm) <- paste0("Var_",1:p)
pair01 <- pairdata(cbind(y,xm))
nrow(pair01) == choose(n,2)
```

plot.TVsMiss	<i>plot solution path from the fitted "TVsMiss" object</i>
--------------	--

Description

solution path is generated, the x-axis can be either in log or normal scale, the variable names of each predictors can be chosen to show or not

Usage

```
## S3 method for class 'TVsMiss'
plot(x, label = FALSE, log = TRUE, ...)
```

Arguments

x	fitted "TVsMiss" object
label	If TRUE, the name of each predictor variable will be showed
log	If TRUE, x-axis is log scale; if FALSE, x-axis is in normal scale
...	graphical parameters to plot

Examples

```

n <- 50
p <- 8
beta <- c(3,0,1.5,0,2,rep(0,p-5))
xm <- matrix(rnorm(n*p),ncol = p, nrow = n)
y <- xm %*% beta + rnorm(n)
colnames(xm) <- paste0("Var_",1:p)

fit01 <- tvsmiiss(x=xm,y=y)
fit01$selection_beta
fit01$beta_matrix
plot(fit01)
plot(fit01,x.log=TRUE,label = FALSE)
plot(fit01,x.log=TRUE,label = TRUE)

fit04 <- tvsmiiss(x=xm,y=y,penalty = "SCAD",method = "BIC")
fit04$selection_beta
fit04$beta_matrix
plot(fit04)
plot(fit04,x.log = TRUE)
plot(fit04,x.log = TRUE,label = TRUE)

```

tvsmiiss

fit and select variable(s) for data with missing value

Description

Fit a model based on a pseudo likelihood and select variable(s) through one of multiple techniques. The regularization path is computed for lasso, SCAD, or MCP. Three steps are used to finish this the variable selection purpose: 1. remove missing and pair each observations; 2. use penalty to get lambda path and corresponding beta matrix; 3. use specific method to finish variable selection.

Usage

```

tvsmiiss(x, y, penalty = c("lasso", "MCP", "SCAD"), method = c("CV", "BIC",
  "BIC1", "BIC2", "sBIC", "sBIC1", "sBIC2", "sVS", "sEST"), lambda = NULL,
  fold = 5, cv.ind = NULL, repeat_b = 20, alpha_n = 0.1, refit = F,
  gamma = switch(penalty, SCAD = 3.7, MCP = 3, lasso = NA), use.penalty = T)

```

Arguments

x	the covariate matrix, should be in matrix format and at least two columns, each row is an observation
y	the response variable
penalty	the penalty used for regularization, can be lasso, SCAD, or MCP. The default is lasso.
method	the variable selection method, can be cross-validation (CV), Bayesian information criterion (BIC), BIC1 and BIC2 are adapted for the consistency in the high dimension, sBIC is the information stability, sBIC1 and sBIC2 are information stability for high dimension data, sVS is the variable selection stability, sEST is the estimation stability
lambda	lambda path used in the regularization path. If not specified by user, the path will be generated automatically
fold	the number of folds used to divided data, will be used in CV, sBIC, sBIC1, sBIC2, sVS, and sEST method
cv.ind	a vector to indicate what fold each observations belong, useful to make reproducible research
repeat_b	B parameter in sVS method, the repeating time to calculate selection stability criteria
alpha_n	the parameter used to take care of variables with weak effect in sVS method
refit	If TRUE, refit technique will be used to get estimation, i.e., use selection variable to refit the model to get estimation
gamma	the tuning parameter of the SCAD/MCP. Default is 3.7 for SCAD and 3 for MCP
use.penalty	If TRUE, use penalty and variable selection techniques; if FALSE, just fit a logistic regression model with paired data

Examples

```

n <- 50
p <- 8
beta <- c(3,0,1.5,0,2,rep(0,p-5))
xm <- matrix(rnorm(n*p),ncol = p, nrow = n)
y <- xm %*% beta + rnorm(n)
colnames(xm) <- paste0("Var_",1:p)

fit01 <- tvsmis(x=xm,y=y)
fit01$selection_beta

fit02 <- tvsmis(x=xm,y=y,method = "BIC")
fit02$selection_beta
fit02$beta_matrix

fit06 <- tvsmis(x=xm,y=y,penalty = "SCAD",method = "sVS",fold = 5)
fit06$selection_beta
fit06$beta_matrix

```

Index

pairedata, [2](#)
plot.TVSMiss, [2](#)
tvsmisss, [3](#)