Package ‘BOSO’

July 1, 2021

Type Package

Title Bilevel Optimization Selector Operator

Version 1.0.3

Date 2021-06-15


In order to run the vignette, it is recommended to install the 'bestsubset' package, using the following command: devtools::install_github(repo=``ryantibs/best-subset'', subdir=``bestsubset'').

If you do not have gurobi, run devtools::install_github(repo=``lvalcarcel/best-subset'', subdir=``bestsubset'').

SystemRequirements IBM ILOG CPLEX (>= 12.1)

Depends R (>= 4.0)

Imports Matrix, MASS, methods

Suggests cplexAPI, testthat, glmnet, knitr, rmarkdown, ggplot2, ggpubr, dplyr, kableExtra, devtools, BiocStyle, bestsubset

License GPL-3

LazyData true

RoxygenNote 7.1.1

Encoding UTF-8

VignetteBuilder knitr

NeedsCompilation no

Author Luis V. Valcarcel [aut, cre, ctb] (<https://orcid.org/0000-0003-3769-5419>), Edurne San Jose-Eneriz [aut] (<https://orcid.org/0000-0001-5786-5273>)
BOSO

Description

Fit a ridge linear regression by a feature selection model conducted by BOSO MILP. The package 'cplexAPI' is necessary to run it.

Usage

```r
BOSO(
  x,
  y,
  xval,  # Optional argument
  yval=FALSE,  # Default is FALSE
  IC = "eBIC",
  IC.blocks = NULL,
  nlambda = 100,
  nlambda.blocks = 10,
  lambda.min.ratio = ifelse(nrow(x) < ncol(x), 0.01, 1e-04),
  lambda = NULL,
  intercept = TRUE,
  standardize = TRUE,
)```

R topics documented:

- BOSO .......................................................... 2
- BOSO.multiple.coldstart .................................. 5
- BOSO.multiple.warmstart .................................. 7
- BOSO.single ................................................ 9
- coef.BOSO .................................................. 11
- predict.BOSO ................................................ 12
- sim.xy ....................................................... 12
- SimResultsVignette ....................................... 13

Index 14
Arguments

\textbf{x} \hspace{1cm} \text{Input matrix, of dimension ‘n’ x ‘p’. This is the data from the training partition. It’s recommended to be class "matrix".}

\textbf{y} \hspace{1cm} \text{Response variable for the training dataset. A matrix of one column or a vector, with ‘n’ elements.}

\textbf{xval} \hspace{1cm} \text{Input matrix, of dimension ‘n’ x ‘p’. This is the data from the validation partition. Its recommended to be class "matrix".}

\textbf{yval} \hspace{1cm} \text{Response variable for the validation dataset. A matrix of one column or a vector, with ‘n’ elements.}

\textbf{IC} \hspace{1cm} \text{information criterion to be used. Default is ‘eBIC’.}

\textbf{IC.blocks} \hspace{1cm} \text{information criterion to be used in the block strategy. Default is the same as IC, but eBIC uses BIC for the block strategy.}

\textbf{nlambda} \hspace{1cm} \text{The number of lambda values. Default is 100.}

\textbf{nlambda.blocks} \hspace{1cm} \text{The number of lambda values in the block strategy part. Default is 10.}

\textbf{lambda.min.ratio} \hspace{1cm} \text{Smallest value for lambda, as a fraction of lambda.max, the (data derived) entry value.}

\textbf{lambda} \hspace{1cm} \text{A user supplied lambda sequence. Typical usage is to have the program compute its own lambda sequence based on nlambda and lambda.min.ratio. Supplying a value of lambda overrides this. WARNING: use with care.}

\textbf{intercept} \hspace{1cm} \text{Boolean variable to indicate if intercept should be added or not. Default is false.}

\textbf{standardize} \hspace{1cm} \text{Boolean variable to indicate if data should be scaled according to mean(x) mean(y) and sd(x) or not. Default is false.}

\textbf{dfmax} \hspace{1cm} \text{Maximum number of variables to be included in the problem. The intercept is not included in this number. Default is min(p,n).}

\textbf{maxVarsBlock} \hspace{1cm} \text{maximum number of variables in the block strategy.}

\textbf{costErrorVal} \hspace{1cm} \text{Cost of error of the validation set in the objective function. Default is 1. WARNING: use with care, changing this value changes the formulation presented in the main article.}
costErrorTrain  Cost of error of the training set in the objective function. Default is 0. WARNING: use with care, changing this value changes the formulation presented in the main article.

costVars  Cost of new variables in the objective function. Default is 0. WARNING: use with care, changing this value changes the formulation presented in the main article.

Threads  CPLEX parameter, number of cores that CPLEX is allowed to use. Default is 0 (automatic).

timeLimit  CPLEX parameter, time limit per problem provided to CPLEX. Default is 1e75 (infinite time).

verbose  print progress, different levels: 1) print simple progress. 2) print result of blocks. 3) print each k in blocks Default is FALSE.

seed  set seed for random number generator for the block strategy. Default is system default.

warmstart  warmstart for CPLEX or use a different problem for each k. Default is False.

TH_IC  is the ratio over one that the information criterion must increase to be STOP. Default is 1e-3.

indexSelected  array of pre-selected variables. WARNING: debug feature.

Details

Compute the BOSO for use one block. This function calls cplexAPI to solve the optimization problem

Value

A ‘BOSO’ object which contains the following information:

betas  estimated betas
x  training x set used in BOSO (input parameter)
y  training x set used in BOSO (input parameter)
xval  validation x set used in BOSO (input parameter)
yval  validation x set used in BOSO (input parameter)
nlambda  nlambda used by ‘BOSO’ (input parameter)
intercept  if ‘BOSO’ has used intercept (input parameter)
standardize  if ‘BOSO’ has used standardization (input parameter)
xm  Mean value of each variable. 0 if data has not been standarized
sx  Standard deviation value of each variable. 0 if data has not been standarized
my  Mean value of output variable. 0 if data has not been standarized
dfmax  Maximum number of variables set to be used by ‘BOSO’ (input parameter)
result.final  list with the results of the final problem for each K
errorTrain  error in training set in the final problem
errorVal error in Validation set in the final problem of used by
lambda.selected lambda selected in the final problem of
p number of initial variables
n number of events in the training set
nval number of events in the validation set
blockStrategy index of variables which were stored in each iteration by ‘BOSO‘ in the block strategy

Author(s)
Luis V. Valcarcel

Examples

#This first example is a basic
#example of how to execute BOSO

data("sim.xy", package = "BOSO")
obj <- BOSO(x = sim.xy[['low']]$x,
            y = sim.xy[['low']]$y,
            xval = sim.xy[['low']]$xval,
            yval = sim.xy[['low']]$yval,
            IC = 'eBIC',
            nlambda=50,
            intercept= 0, standardize = 0,
            Threads=1, verbose = 3, seed = 2021)
coef(obj)  # extract coefficients at a single value of lambda
predict(obj, newx = sim.xy[['low']]$x[1:20, ])  # make predictions

Description

Function to run a single block BOSO problem, generating for each K a different CPLEX object.
Usage

BOSO.multiple.coldstart(
  x, y, xval, yval,
  nlambda = 100,
  IC = "eBIC",
  n.IC = NULL,
  p.IC = NULL,
  lambda.min.ratio = ifelse(nrow(x) < ncol(x), 0.01, 1e-04),
  lambda = NULL,
  intercept = TRUE,
  standardize = FALSE,
  dfmin = 0,
  dfmax = NULL,
  costErrorVal = 1,
  costErrorTrain = 0,
  costVars = 0,
  Threads = 0,
  timeLimit = 1e+75,
  verbose = F,
  TH.IC = 0.001
)

Arguments

x          Input matrix, of dimension 'n' x 'p'. This is the data from the training partition. Its recommended to be class "matrix".

y          Response variable for the training dataset. A matrix of one column or a vector, with 'n' elements

xval       Input matrix, of dimension 'n' x 'p'. This is the data from the validation partition. Its recommended to be class "matrix".

yval       Response variable for the validation dataset. A matrix of one column or a vector, with 'n' elements.

nlambda    The number of lambda values. Default is 100.

IC         information criterion to be used. Default is 'eBIC'.

n.IC       number of events for the information criterion.

p.IC       number of initial variables for the information criterion.

lambda.min.ratio       Smallest value for lambda, as a fraction of lambda.max, the (data derived) entry value.

lambda       A user supplied lambda sequence. Typical usage is to have the program compute its own lambda sequence based on nlambda and lambda.min.ratio. Supplying a value of lambda overrides this. WARNING: use with care
BOSO.multiple.warmstart

intercept  Boolean variable to indicate if intercept should be added or not. Default is false.
standardize  Boolean variable to indicate if data should be scaled according to mean(x) mean(y) and sd(x) or not. Default is false.

dfmin  Minimum number of variables to be included in the problem. The intercept is not included in this number. Default is 0.
dfmax  Maximum number of variables to be included in the problem. The intercept is not included in this number. Default is min(p,n).
costErrorVal  Cost of error of the validation set in the objective function. Default is 1. WARNING: use with care, changing this value changes the formulation presented in the main article.
costErrorTrain  Cost of error of the training set in the objective function. Default is 0. WARNING: use with care, changing this value changes the formulation presented in the main article.
costVars  Cost of new variables in the objective function. Default is 0. WARNING: use with care, changing this value changes the formulation presented in the main article.

Threads  CPLEX parameter, number of cores that IBM ILOG CPLEX is allowed to use. Default is 0 (automatic).
timeLimit  CPLEX parameter, time limit per problem provided to CPLEX. Default is 1e75 (infinite time).
verbose  print progress. Default is FALSE.

TH_IC  is the ratio over one that the information criterion must increase to be STOP. Default is 1e-3.

Details

Compute the BOSO for use one block. This function calls ILOG IBM CPLEX with 'cplexAPI' to solve the optimization problem.

Value

A ‘BOSO’ object.

Author(s)

Luis V. Valcarcel

---

BOSO.multiple.warmstart

BOSO.single and associates functions

Description

Function to run a single block BOSO problem, generating one CPLEX object and re-running it for the different K.
Usage

BOSO.multiple.warmstart(
  x,
  y,
  xval,
  yval,
  nlambda = 100,
  IC = "eBIC",
  n.IC = NULL,
  p.IC = NULL,
  lambda.min.ratio = ifelse(nrow(x) < ncol(x), 0.01, 1e-04),
  lambda = NULL,
  intercept = TRUE,
  standardize = FALSE,
  dfmin = 0,
  dfmax = NULL,
  costErrorVal = 1,
  costErrorTrain = 0,
  costVars = 0,
  Threads = 0,
  timeLimit = 1e+75,
  verbose = F,
  TH_IC = 0.001
)

Arguments

x  
   Input matrix, of dimension 'n' x 'p'. This is the data from the training partition. Its recommended to be class "matrix".

y  
   Response variable for the training dataset. A matrix of one column or a vector, with 'n' elements

xval 
   Input matrix, of dimension 'n' x 'p'. This is the data from the validation partition. Its recommended to be class "matrix".

yval 
   Response variable for the validation dataset. A matrix of one column or a vector, with 'n' elements

nlambda  
   The number of lambda values. Default is 100.

IC  
   information criterion to be used. Default is 'eBIC'.

n.IC  
   number of events for the information criterion.

p.IC  
   number of initial variables for the information criterion.

lambda.min.ratio  
   Smallest value for lambda, as a fraction of lambda.max, the (data derived) entry value

lambda  
   A user supplied lambda sequence. Typical usage is to have the program compute its own lambda sequence based on nlambda and lambda.min.ratio. Supplying a value of lambda overrides this. WARNING: use with care
intercept: Boolean variable to indicate if intercept should be added or not. Default is false.
standardize: Boolean variable to indicate if data should be scaled according to mean(x) mean(y) and sd(x) or not. Default is false.
dfmin: Minimum number of variables to be included in the problem. The intercept is not included in this number. Default is 0.
dfmax: Maximum number of variables to be included in the problem. The intercept is not included in this number. Default is min(p,n).
costErrorVal: Cost of error of the validation set in the objective function. Default is 1. WARNING: use with care, changing this value changes the formulation presented in the main article.
costErrorTrain: Cost of error of the training set in the objective function. Default is 0. WARNING: use with care, changing this value changes the formulation presented in the main article.
costVars: Cost of new variables in the objective function. Default is 0. WARNING: use with care, changing this value changes the formulation presented in the main article.
Threads: CPLEX parameter, number of cores that cplex is allowed to use. Default is 0 (automatic).
timeLimit: CPLEX parameter, time limit per problem provided to CPLEX. Default is 1e75 (infinite time).
verbose: print progress. Default is FALSE
TH_IC: is the ratio over one that the information criterion must increase to be STOP. Default is 1e-3.

Details
Compute the BOSO for use one block. This function calls ILOG IBM CPLEX with 'cplexAPI' to solve the optimization problem.

Value
A ‘BOSO’ object.

Author(s)
Luis V. Valcarcel
Usage

BOSO.single(
  x,
  y,
  xval,
  yval,
  nlambda = 100,
  lambda.min.ratio = ifelse(nrow(x) < ncol(x), 0.01, 1e-04),
  lambda = NULL,
  intercept = TRUE,
  standardize = TRUE,
  dfmin = 0,
  dfmax = NULL,
  costErrorVal = 1,
  costErrorTrain = 0,
  costVars = 0,
  Threads = 0,
  timeLimit = 1e+75
)

Arguments

x  Input matrix, of dimension 'n' x 'p'. This is the data from the training partition. Its recommended to be class "matrix".
y  Response variable for the training dataset. A matrix of one column or a vector, with 'n' elements
xval  Input matrix, of dimension 'n' x 'p'. This is the data from the validation partition. Its recommended to be class "matrix".
yval  Response variable for the validation dataset. A matrix of one column or a vector, with 'n' elements
nlambda  The number of lambda values. Default is 100.
lambda.min.ratio  Smallest value for lambda, as a fraction of lambda.max, the (data derived) entry value
lambda  A user supplied lambda sequence. Typical usage is to have the program compute its own lambda sequence based on nlambda and lambda.min.ratio. Supplying a value of lambda overrides this. WARNING: use with care
intercept  Boolean variable to indicate if intercept should be added or not. Default is false.
standardize  Boolean variable to indicate if data should be scaled according to mean(x) mean(y) and sd(x) or not. Default is false.
dfmin  Minimum number of variables to be included in the problem. The intercept is not included in this number. Default is 0.
dfmax  Maximum number of variables to be included in the problem. The intercept is not included in this number. Default is min(p,n).
coef.BOSO

**DETAILS**

Compute the BOSO for ust one block. This function calls ILOG IBM CPLEX with cplexAPI to solve the optimization problem.

**AUTHOR(S)**

Luis V. Valcarcel

---

**FUNCTIONALITY**

`coef.BOSO` *Extract coefficients from a BOSO object*

**DESCRIPTION**

This is an equivalent function to the one offered by `coef.glmnet` for extraction of coefficients.

**USAGE**

```r
## S3 method for class 'BOSO'
coef(object, beta0 = F, ...)
```

**ARGUMENTS**

- **object**: Fitted `BOSO` or `BOSO.single` object
- **beta0**: Force beta0 to appear (output of `p+1` features)
- **...**: extra arguments for future updates

**VALUE**

A `matrix` object with the corresponding beta values estimated.
predict.BOSO  
*Predict function for BOSO object.*

**Description**

This is an equivalent function to the one offered by `coef.glmnet` for extraction of coefficients.

**Usage**

```r
## S3 method for class 'BOSO'
predict(object, newx, ...)
```

**Arguments**

- `object`: Fitted 'BOSO' or 'BOSO.single' object
- `newx`: Matrix with new data for prediction with BOSO
- `...`: extra arguments for future updates

**Value**

A ‘matrix’ object with the corresponding beta values estimated.

---

**sim.xy**  
*High-5 and Low setting data*

**Description**

Simmelulated data for the high-5-sized scenario and low-sized. It contains a list with the who cases, each of them with the following fields:

- `x`: X matrix for training set
- `y`: Y vector for training set
- `xval`: X matrix for validation set
- `yval`: Y vector for validation set
- `beta`: true beta array

**Usage**

```r
data("sim.xy")
```

**Source**

https://github.com/ryantibs/best-subset

**References**

SimResultsVignette

**Description**

Results from all the algorithms using the simulated data Simulated data for the high-5-sized scenario.

**Usage**

data("SimResultsVignette")

**References**

Index

* datasets
  - sim.xy, 12
  - SimResultsVignette, 13

BOSO, 2
BOSO.multiple.coldstart, 5
BOSO.multiple.warmstart, 7
BOSO.single, 9

coef.BOSO, 11
coeef glmnet, 11, 12

predict.BOSO, 12

sim.results (SimResultsVignette), 13
sim.xy, 12
SimResultsVignette, 13