Package ‘POCRE’

March 16, 2022

Title Penalized Orthogonal-Components Regression
Version 0.6.0
Date 2022-03-15
Author Dabao Zhang, Zhongli Jiang, Zeyu Zhang, Yu-ting Chen
Maintainer Dabao Zhang <zhangdb@purdue.edu>
Description Penalized orthogonal-components regression (POCRE) is a supervised dimension reduction method for high-dimensional data. It sequentially constructs orthogonal components (with selected features) which are maximally correlated to the response residuals. POCRE can also construct common components for multiple responses and thus build up latent-variable models.
Imports stats,utils,ggplot2 (&gt;= 2.2.0),pracma,EbayesThresh
License GPL-2
NeedsCompilation no
Repository CRAN
Date/Publication 2022-03-16 18:10:02 UTC

R topics documented:
cvpocre ................................................................. 2
gps ................................................................. 3
plot.pocre ........................................................... 5
plot.pocrepacth .................................................... 6
pocre ................................................................. 7
pocrepacth ........................................................ 9
pocrescreen ......................................................... 10
selectmodel ......................................................... 12
sim5ydata .......................................................... 13
simbin ............................................................... 14
simdata ............................................................. 15
simpoi ............................................................... 16
sipocre ............................................................. 17

Index 19
cvpocre  

Use k-Fold Cross-Validation to Choose the Tuning Parameter for POCRE

Description

Choose the optimal tuning parameter via k-fold cross-validation for POCRE.

Usage

cvpocre(y, x, n.folds=10, delta=0.1, maxvar=dim(x)[1]/2,  
  ptype=c('ebtz','ebt','l1','scad','mcp'), maxit=100,  
  maxcmp=10, gamma=3.7, lambda.init=1, tol=1e-6,  
  crit=c('press','Pearson','Spearman','Kendall'))

Arguments

y  n*q matrix, values of q response variables (allow for multiple response variables).

x  n*p matrix, values of p predicting variables (excluding the intercept).

n.folds  number of folds to split the data (10-fold CV by default).

delta  step size of different values of the tuning parameter.

maxvar  maximum number of selected variables.

ptype  a character to indicate the type of penalty: 'ebtz' (empirical Bayes thresholding after Fisher’s z-transformation, by default), 'ebt' (empirical Bayes thresholding by Johnstone & Silverman (2004)), 'l1' (L_1 penalty), 'scad' (SCAD by Fan & Li (2001)), 'mcp' (MCP by Zhang (2010)).

maxit  maximum number of iterations to be allowed.

maxcmp  maximum number of components to be constructed.

gamma  a parameter used by SCAD and MCP (=3.7 by default).

lambda.init  initial value of the tuning parameter (=1 by default).

tol  tolerance of precision in iterations.

crit  a character to indicate the validation criterion: 'press' (prediction residual error sum of squares, by default), 'Pearson' (Pearson correlation coefficient), 'Spearman' (Spearman’s rank correlation coefficient), 'Kendall' (Kendall’s rank correlation coefficient).

Details

Use k-folds cross-validation to find the optimal value for the tuning parameter. The validation criterion can be chosen from PRESS, or different types of correlation coefficients, such as Pearson’s, Spearman’s, or Kendall’s.
Value

The optimal value of the tuning parameter.

Author(s)

Dabao Zhang, Zhongli Jiang, Zeyu Zhang, Department of Statistics, Purdue University

References


See Also

pocrescreen, pocrepath, pocre.

Examples

```r
## Not run:
data(simdata)
n <- dim(simdata)[1]
xx <- simdata[,-1]
yy <- simdata[,1]
# tp <- cvpocre(yy,xx,delta=0.01)
tp <- cvpocre(yy,xx)
print(paste(" pocre: Optimal Tuning Parameter = ", tp))
cvpres <- pocre(yy,xx,lambda=tp,maxvar=n/log(n))
## End(Not run)
```

Description

A pre-specified number (i.e., maxvar) of covariates will be selected for generalized linear models by constructing maxcmp components with generalized POCRE. Each component will be constructed by selecting maxvar/macmp covariates which are most relevant to the response variable(s). Similar to pocrescreen, gps selects covariates for their top relevance to the response variable(s) without penalization.
Usage

gps(y, x, family="binomial", bc.method="optimal", x.include=NULL, weights=NULL, maxcmp=10, maxvar=NULL, tol = 1e-6, maxit = 100)

Arguments

- **y**: n*q matrix, values of q response variables (allow for multiple response variables).
- **x**: n*p matrix, values of p predicting variables (excluding the intercept).
- **family**: Family objects as `family`. Currently only support "gaussian", "binomial" (by default), and "poisson".
- **bc.method**: Bias correction method.
- **x.include**: a vector of indices indicating covariates which should always be included in the model (so not counted into selected maxvar covariates).
- **weights**: A vector, including a prespecified weight for each observation (set as 1/n by default).
- **maxcmp**: maximum number of components to be constructed.
- **maxvar**: maximum number of selected variables.
- **tol**: tolerance of precision in iterations.
- **maxit**: maximum number of iterations to be allowed.

Value

- A vector of indices of selected covariates (excluding those in x.include).

Author(s)

Dabao Zhang, Zhongli Jiang, Yu-ting Chen, Department of Statistics, Purdue University

References


See Also

- `pocrescreen`

Examples

```r
# Binomial Data
data(simbin)
gps(simbin[,1], simbin[-1], maxcmp=3, maxvar=9)
gps(simbin[,1], simbin[-1], x.include=103:104, maxcmp=3, maxvar=9)

# Count Data
data(simpoi)
gps(simpoi[,1], simpoi[-1], family='poisson', maxcmp=5, maxvar=10)
```
plot.pocre

Visualization of a pocre Object

Description
Plot the regression coefficients, and the loadings of all components for a fitted model by POCRE.

Usage
## S3 method for class 'pocre'
plot(x, x.id = NA, which=1:2, cex=.5, ...)

Arguments
- `x`: a pocre object, i.e., the result from `pocre`.
- `x.id`: a vector indicating the indices or positions of the covariates in the original data.
- `which`: 1 for plotting the regression coefficients, 2 for plotting the loadings of all components.
- `cex`: A numerical value giving the amount by which plotting text and symbols should be magnified relative to the default, see `par`.
- `...`: additional arguments accepted by `ggplot`.

Author(s)
Dabao Zhang, Zhongli Jiang, Zeyu Zhang, Department of Statistics, Purdue University

References

See Also
- `pocre`, `plot.pocrepath`, `pocrepath`.

Examples
```r
data(simdata)
xx <- scale(as.matrix(simdata[, -1]))
yy <- scale(as.matrix(simdata[, 1]))

## Fit with pocre()
pres <- pocre(yy, xx, lambda=0.9)

# plot(pres, which=1)
plot(pres)
```
plot.pocrepath

Visulaization of a POCRE Path

Description

For a series models built by POCRE for different tuning parameter values, it provides three types of plots to help select an appropriate tuning parameter value.

Usage

```r
## S3 method for class 'pocrepath'
plot(x, which=1:3, cex=.5, lwd=1, ...)
```

Arguments

- `x` a `pocrepath` object, i.e., the result from `pocrepath`.
- `which` 1 for plotting the tuning parameter vs. (beta, ![beta!=0]), 2 for plotting the tuning parameter vs. (beta, R^2), 3 for plotting the tuning parameter vs. (R^2, ![beta!=0]).
- `cex` A numerical value giving the amount by which plotting text and symbols should be magnified relative to the default, see `par`.
- `lwd` line width, see `par`.
- `...` additional arguments accepted by `ggplot`.

Author(s)

Dabao Zhang, Zhongli Jiang, Zeyu Zhang, Department of Statistics, Purdue University

References


See Also

`pocrepath, plot.pocre, pocre`.

Examples

```r
data(simdata)
xx <- scale(as.matrix(simdata[, -1]))
yy <- scale(as.matrix(simdata[, 1]))

# ppres <- pocrepath(yy, xx, delta=0.01)
ppres <- pocrepath(yy, xx)
```
# plot(ppres)
plot(ppres,which=3)

## pocre

### Penalized Orthogonal-Components Regression (POCRE)

#### Description

Apply POCRE with a pre-specified tuning parameter to build a linear regression model with orthogonal components $X\vartheta_1, X\vartheta_2, \ldots$, 

$$Y = \mu + \sum_j (X\varphi_j)\vartheta_j + \epsilon = \mu + X\beta + \epsilon,$$

where $\text{var}[\epsilon] = \sigma^2$ and $\beta = \sum_j \varphi_j\vartheta_j$. These orthogonal components are sequentially constructed according to supervised dimension reduction under penalty set by the pre-specified tuning parameter.

While the orthogonal components are constructed using the centralized covariates, the intercept $\mu$ and regression coefficients in $\beta$ are estimated for original covariates. The sequential construction stops when no new component can be constructed (returning bSparse=1), or the new component is constructed with more than maxvar covariates (returning bSparse=0).

#### Usage

```r
pocre(y, x, lambda=1, x.nop=NA, maxvar=dim(x)[1]/2,
      maxcmp=10, ptype=c('ebtz','ebt','l1','scad','mcp'),
      maxit=100, tol=1e-6, gamma=3.7, pval=FALSE)
```

#### Arguments

- **y**: n*q matrix, values of q response variables (allow for multiple response variables).
- **x**: n*p matrix, values of p predicting variables (excluding the intercept).
- **lambda**: the tuning parameter (=1 by default).
- **x.nop**: a vector indicating indices of covariates which are excluded only when evaluating the significance of components.
- **maxvar**: maximum number of selected variables.
- **maxcmp**: maximum number of components to be constructed.
- **ptype**: a character to indicate the type of penalty: 'ebtz' (empirical Bayes thresholding after Fisher’s z-transformation, by default), 'ebt' (empirical Bayes thresholding by Johnstone & Silverman (2004)), 'l1' (L_1 penalty), 'scad' (SCAD by Fan & Li (2001)), 'mcp' (MCP by Zhang (2010)).
- **maxit**: maximum number of iterations to be allowed.
- **tol**: tolerance of precision in iterations.
- **gamma**: a parameter used by SCAD and MCP (=3.7 by default).
- **pval**: a logical value indicating whether to calculate the p-values of components.
Value

- mu: estimated intercept of the linear regression.
- beta: estimated coefficients of the linear regression.
- varpi: loadings of the constructed components.
- vartheta: the regression coefficients of the constructed components.
- bSparse: a logical value indicating whether estimated beta has less than maxvar nonzero values.
- lambda: value of the tuning parameter.
- nCmp: number of constructed components.
- n: sample size.
- p: number of covariates.
- xShift: the column means of x.
- yShift: the column means of y.
- sigmae2: estimated error variance $\sigma^2$.
- rsq: $R^2$ value of the fitted regression model.
- nzBeta: number of non-zero regression coefficients in $\beta$.
- omega: internal matrix.
- theta: internal matrix.
- pvalue: p-values of constructed components, available when pval=TRUE.
- seqpv: Type I p-values of components when sequentially including them into the model, available when pval=TRUE.
- indpv: p-values of components when marginally testing each component, available when pval=TRUE.
- loglik: the loglikelihood function, available when pval=TRUE.
- effp: the effective number of predictors, excluding redundant ones, available when pval=TRUE.

Author(s)

Dabao Zhang, Zhongli Jiang, Zeyu Zhang, Department of Statistics, Purdue University

References


pocrepath

See Also

plot.pocre, pocrescreen, pocrepath, cvpocre.

Examples

data(simdata)
xx <- simdata[,,-1]
yy <- simdata[,1]

#pres <- pocre(yy,xx,lambda=0.9)
pres <- pocre(yy,xx)  # lambda=1 by default

Description

Applying POCRE for a series of tuning parameters chosen by a pre-specified step size. The tuning parameter will increase until non-component can be constructed, and then decrease until a non-sparse regression is constructed (i.e., the number of non-zero coefficients in \( \beta \) is more than maxvar).

Usage

pocrepath(y, x, delta=0.1, maxvar=dim(x)[1]/2, x.nop=NA, maxcmp=10,
ptype=c('ebtz','ebt','l1','scad','mcp'), lambda.init=1,
maxit=100, tol=1e-6, maxtps=500, gamma=3.7, pval=(dim(y)[2]==1))

Arguments

y  
n*q matrix, values of q response variables (allow for multiple response variables).

x  
n*p matrix, values of p predicting variables (excluding the intercept).

delta  
step size to increase or decrease from current tuning parameter.

maxvar  
maximum number of selected variables.

x.nop  
a vector indicating indices of covariates which are excluded only when evaluating the significance of components.

maxcmp  
maximum number of components to be constructed.

ptype  
a character to indicate the type of penalty: 'ebtz' (empirical Bayes thresholding after Fisher's z-transformation, by default), 'ebt' (empirical Bayes thresholding by Johnstone & Silverman (2004)), 'l1' (L_1 penalty), 'scad' (SCAD by Fan & Li (2001)), 'mcp' (MCP by Zhang (2010)).

lambda.init  
initial value of the tuning parameter (=1 by default).

maxit  
maximum number of iterations to be allowed.

tol  
tolerance of precision in iterations.
maxtps  maximum number of different values that the tuning parameter is allowed.
gamma   a parameter used by SCAD and MCP (=3.7 by default).
pval    a logical value indicating whether to calculate the p-values of components (not
        implemented for q>1, i.e., multiple response variables).

Value
A list of results from pocre, each for a specific value of the tuning parameter.

Author(s)
Dabao Zhang, Zhongli Jiang, Zeyu Zhang, Department of Statistics, Purdue University

References

See Also
plot.pocrepath, selectmodel, pocre.

Examples
```r
data(simdata)
xx <- simdata[, -1]
yy <- simdata[, 1]
ppres <- pocrepath(yy, xx)
```

Screen Variables Using Penalized Orthogonal-Components Regression (POCRE)

Description
Screen for a pre-specified number (i.e., maxvar) of covariates by constructing maxcmp components with POCRE. Each component will be constructed by selecting maxvar/macmp covariates which are most relevant to the response variable(s). Here POCRE selects covariates for their top relevance to the response variable(s) without penalization.
pocrescreen

Usage

pocrescreen(y, x, maxvar=nrow(x), maxcmp=5, x.include=NULL, tole=1e-6, maxit=100)

Arguments

y  n*q matrix, values of q response variables (allow for multiple response variables).
x  n*p matrix, values of p predicting variables (excluding the intercept).
maxvar  maximum number of selected variables.
maxcmp  maximum number of components to be constructed.
x.include  a vector of indices indicating covariates which should always be included in the model (so not counted into selected maxvar covariates).
tol  tolerance of precision in iterations.
maxit  maximum number of iterations to be allowed.

Value

a vector of indices of selected covariates (excluding those in x.include).

Author(s)

Dabao Zhang, Zhongli Jiang, Zeyu Zhang, Department of Statistics, Purdue University

References


See Also

pocre, pocrepath, cvpocre.

Examples

data(simdata)
xx <- simdata[, -1]
yy <- simdata[, 1]

# Screen for 50 covariates
sidx <- pocrescreen(yy, xx, maxvar=50)

# Screen for 50 additional covariates besides the first 10
xinc <- 1:10
sidx <- pocrescreen(yy, xx, maxvar=50, x.include=xinc)
sidx <- c(xinc, sidx)
selectmodel

Select the Optimal Model

**Description**

Select the optimal model from those fitted by POCRE, on the basis of prespecified criterion, such as EBIC, BIC, AIC, and AICc.

**Usage**

```r
selectmodel(ppobj, msc=NULL)
```

**Arguments**

- `ppobj`: output from `pocrepath`.
- `msc`: a value indicating the information criterion: 0 for BIC, (0,1] for EBIC (by default), 2 for AIC, 3 for AICc.

**Value**

output of `pocre` for the optimal model.

**Author(s)**

Dabao Zhang, Zhongli Jiang, Zeyu Zhang, Department of Statistics, Purdue University

**References**


**See Also**

`pocrepath`, `plot.pocrepath`.

**Examples**

```r
data(simdata)
xx <- scale(as.matrix(simdata[, -1]))
yy <- scale(as.matrix(simdata[, 1]))

# ppres <- pocrepath(yy, xx, delta=0.01)
ppres <- pocrepath(yy, xx)
fres <- selectmodel(ppres)
```
A Set of Simulated Data with Multiple Response Variables

Description
A simulated data set with 100 observations, each with five response variable and 1,000 covariates.

Usage
data("sim5ydata")

Format
A data frame with 100 observations on 1005 variables with the first five columns for the response variables, and the rest for the covariates.

Details
The 1,000 covariates are from 10 blocks of independent variables, with each block consisting 100 autoregressively correlated variables. There are a total of 12 covariates affecting the response variables: $x_{50}, x_{51}, x_{150}, x_{153}, x_{250}, x_{256}, x_{350}, x_{359}, x_{450}, x_{467}, x_{550}, x_{583}$.

Author(s)
Dabao Zhang, Zhongli Jiang, Zeyu Zhang, Department of Statistics, Purdue University

References

See Also
pocrescreen, pocrepath, pocre, cvpocre.

Examples
data(sim5ydata)
A Set of Simulated Binomial Data.

Description

A simulated data set with 100 observations, each with one binary response variable and 1,000 covariates.

Usage

data("simbin")

Format

A data frame with 100 observations on 1001 variables with the first column for the response variable, and the rest for the covariates.

Details

The true covariates are 1, 2, 103, 104, 205, and 206.

Author(s)

Dabao Zhang, Zhongli Jiang, Yu-ting Chen, Department of Statistics, Purdue University

References


See Also

gps

Examples

data(simbin)
**simdata**

**A Set of Simulated Data with Single Gaussian Response Variable**

**Description**

A simulated data set with 100 observations, each with one response variable and 1,000 covariates.

**Usage**

```r
data("simdata")
```

**Format**

A data frame with 100 observations on 1001 variables with the first column for the response variable, and the rest for the covariates.

**Details**

The 1,000 covariates are from 10 blocks of independent variables, with each block consisting 100 autoregressively correlated variables. There are a total of 20 covariates affecting the response variables: $x_1, \ldots, x_{10}$, $x_{101}, \ldots, x_{110}$.

**Author(s)**

Dabao Zhang, Zhongli Jiang, Zeyu Zhang, Department of Statistics, Purdue University

**References**


**See Also**

pocrescreen, pocrepath, pocre.

**Examples**

```r
data(simdata)
```
A Set of Simulated Poisson Data.

**Description**

A simulated data set with 100 observations, each with one count response variable and 1,000 covariates.

**Usage**

```r
data("simpoi")
```

**Format**

A data frame with 100 observations on 1001 variables with the first column for the response variable, and the rest for the covariates.

**Details**

The 1,000 covariates are from 10 blocks of independent variables, with each block consisting 100 autoregressively correlated variables. There are a total of 20 covariates affecting the response variables: $x_1, \ldots, x_{10}, x_{101}, \ldots, x_{110}$.

**Author(s)**

Dabao Zhang, Yu-ting Chen, Department of Statistics, Purdue University

**References**


**See Also**

`gps`

**Examples**

```r
data(simpoi)
```
Description

Applying POCRE to select variables and evaluate the significance of selected variables using the multiple splitting method by Meinshausen et al. (2009). The tuning parameter may be selected based on either an information criterion or k-fold cross-validation. The tuning parameter can also be fixed at a prespecified value.

Usage

```r
sipocre(y, x, n.splits=10, delta=0.1, crit=c('ic','cv','fixed'),
   ptype=c('ebtz','ebt','l1','scad','mcp'), maxvar=dim(x)[1]/2,
   msc=NA, maxit=100, maxcmp=50, gamma=3.7, tol=1e-6,
   n.folds=10, lambda=1, n.train=round(nrow(x)/2))
```

Arguments

- `y`: n*q matrix, values of q response variables (allow for multiple response variables).
- `x`: n*p matrix, values of p predicting variables (excluding the intercept).
- `n.splits`: number of random splits (=10 by default).
- `delta`: step size to increase or decrease from current tuning parameter.
- `crit`: character indicating the criterion to choose the tuning parameter: 'ic' (information criteria such as AIC, AICc, BIC, EBIC), 'cv' (k-folds cross-validation) or 'fixed' (a pre-specified value).
- `ptype`: a character to indicate the type of penalty: 'ebtz' (empirical Bayes thresholding after Fisher's z-transformation, by default), 'ebt' (empirical Bayes thresholding by Johnstone & Silverman (2004)), 'l1' (L_1 penalty), 'scad' (SCAD by Fan & Li (2001)), 'mcp' (MCP by Zhang (2010)).
- `maxvar`: maximum number of selected variables.
- `msc`: value(s) to indicate the penalty related to the information criterion: 0~1 for (E)BIC, 2 for AIC, 3 for AICc, used when `crit='ic'`.
- `maxit`: maximum number of iterations to be allowed.
- `maxcmp`: maximum number of components to be constructed.
- `gamma`: a parameter used by SCAD and MCP (=3.7 by default).
- `tol`: tolerance of precision in iterations.
- `n.folds`: number of folds in k-folds cross-validation, used when `crit='cv'`.
- `lambda`: pre-specified value for the tuning parameter, used when `crit='fixed'`.
- `n.train`: sample size of the training data set.
Value

a list consisting of the following components,

\textbf{cpv} component-based p-values which are calculated by testing the constructed components, either a matrix (when \texttt{crit='ic'}, in this case each column corresponds to one value in msc) or a vector (when \texttt{crit='cv'} or \texttt{crit='fixed'}). 

\textbf{xpv} traditional p-values, either a matrix (when \texttt{crit='ic'}, in this case each column corresponds to one value in msc) or a vector (when \texttt{crit='cv'} or \texttt{crit='fixed'}). 

Author(s)

Dabao Zhang, Zhongli Jiang, Zeyu Zhang, Department of Statistics, Purdue University

References


See Also

\texttt{pocre}.

Examples

```r
## Not run:
data(simdata)
x <- simdata[, -1]
y <- simdata[, 1]
sipres <- sipocre(y, x)
## End(Not run)
```
Index

* datasets
  sim5ydata, 13
  simbin, 14
  simdata, 15
  simpoi, 16

cvpocre, 2, 9, 11, 13

family, 4

ggplot, 5, 6
gps, 3, 14, 16

par, 5, 6
plot.pocre, 5, 6, 9
plot.pocrepath, 5, 6, 10, 12
pocre, 3, 5, 6, 7, 10–13, 15, 18
pocrepath, 3, 5, 6, 9, 9, 11–13, 15
pocrescreen, 3, 4, 9, 10, 13, 15

selectmodel, 10, 12
sim5ydata, 13
simbin, 14
simdata, 15
simpoi, 16
sipocre, 17