Package 'POCRE'

July 20, 2018

Title Penalized Orthogonal-Components Regression

Version 0.5.0

Date 2018-07-15

Author Dabao Zhang, Zhongli Jiang, Zeyu Zhang

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 (>= 2.2.0),pracma,EbayesThresh

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2018-07-20 13:20:03 UTC

R topics documented:

cvpocre 2
plot.pocre 3
plot.pocrepath 4
pocre 6
pocrepath 8
pocrescreen 10
selectmodel 11
sim5ydata 12
simdata 13
sipocre 14

Index **16**

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' (emprical Bayes thresholding after Fisher's z-transformation, by default), 'ebt' (emprical 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 optinal 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

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348-1360

Johnstone IM and Silverman BW (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, 32: 1594-1649.

Zhang C-H (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38: 894-942.

Zhang D, Lin Y, and Zhang M (2009). Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics*, 3: 781-796.

See Also

pocrescreen, pocrepath, pocre.

Examples

```
## 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)
```

---

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

Zhang D (2018). R package POCRE: Exploring high-dimensional data via supervised dimension reduction. Manuscript.

Zhang D, Lin Y, and Zhang M (2009). Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics*, 3: 781-796.

See Also

pocre, plot.pocrepath, pocrepath.

Examples

```
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 paramter values, it provides three types of plots to help select an appropriate tuning parameter value.

Usage

```
## 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

Zhang D (2018). R package POCRE: Exploring high-dimensional data via supervised dimension reduction. Manuscript.

Zhang D, Lin Y, and Zhang M (2009). Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics*, 3: 781-796.

See Also

pocrepath, plot.pocre, pocre.

Examples

```
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 $X\vartheta_1, X\vartheta_2, \ldots$,

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

where $var[\epsilon] = \sigma^2$ and $\beta = \sum_j \varpi_j\vartheta_j$. 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 μ and regression coefficients in β 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

```
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' (emprical Bayes thresholding after Fisher's z-transformation, by default), 'ebt' (emprical 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 paramete.
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 σ^2.
rsq	R^2 value of the fitted regression model.
nzBeta	number of non-zero regression coefficients in β.
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

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348-1360

Johnstone IM and Silverman BW (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, 32: 1594-1649.

Zhang C-H (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38: 894-942.

Zhang D, Lin Y, and Zhang M (2009). Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics*, 3: 781-796.

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
```

---

pocrepath *Build a POCRE Path for Different Values of Tuning Parameters*

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 β 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 decrase 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' (emprical Bayes thresholding after Fisher's z-transformation, by default), 'ebt' (emprical 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

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348-1360

Johnstone IM and Silverman BW (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, 32: 1594-1649.

Zhang C-H (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38: 894-942.

Zhang D (2018). R package POCRE: Exploring high-dimensional data via supervised dimension reduction. Manuscript.

Zhang D, Lin Y, and Zhang M (2009). Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics*, 3: 781-796.

See Also

plot.pocrepath, selectmodel, pocre.

Examples

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

ppres <- pocrepath(yy,xx)
```

pocrescreen *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.

Usage

```
pocrescreen(y, x, maxvar=nrow(x), maxcmp=5, x.include=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).
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

Zhang D (2018). R package POCRE: Exploring high-dimensional data via supervised dimension reduction. Manuscript.

Zhang D, Lin Y, and Zhang M (2009). Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics*, 3: 781-796.

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

```
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

Chen J and Chen Z (2008) Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95: 759-771.

Zhang D, Lin Y, and Zhang M (2009). Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics*, 3: 781-796.

See Also

pocrepath, plot.pocrepath.

Examples

```
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)
```

---

sim5ydata *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

Zhang D, Lin Y, and Zhang M (2009). Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics*, 3: 781-796.

See Also

pocrescreen, pocrepath, pocre, cvpocre.

Examples

```
data(sim5ydata)
```

`simdata` *A Set of Simulated Data with Single Response Variable*

Description

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

Usage

```
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

Zhang D, Lin Y, and Zhang M (2009). Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics*, 3: 781-796.

See Also

`pocrescreen`, `pocrepath`, `pocre`.

Examples

```
data(simdata)
```

`sipocre` *Penalized Orthogonal-Components Regression (POCRE) with Significance Inference*

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

```
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 decrase 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-valdiation) or `'fixed'` (a pre-specified value).
`ptype`	a character to indicate the type of penalty: `'ebtz'` (emprical Bayes thresholding after Fisher's z-transformation, by default), `'ebt'` (emprical 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-sepcified 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,

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

Author(s)

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

References

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348-1360

Johnstone IM and Silverman BW (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, 32: 1594-1649.

Meinshausen N, Meier L, and Buhlmann P (2009) p-Values for High-Dimensional Regression. *Journal of the American Statistical Association*, 104: 1671-1681.

Zhang C-H (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38: 894-942.

Zhang D, Lin Y, and Zhang M (2009). Penalized orthogonal-components regression for large p small n data. *Electronic Journal of Statistics*, 3: 781-796.

See Also

pocre.

Examples

```
## Not run:
data(simdata)
xx <- simdata[,-1]
yy <- simdata[,1]

sipres <- sipocre(yy,xx)

## End(Not run)
```

Index

∗Topic **datasets**
sim5ydata, 12
simdata, 13

cvpocre, 2, *8*, *10*, *12*

ggplot, *4, 5*

par, *4, 5*
plot.pocre, 3, *5*, *8*
plot.pocrepath, *4*, 4, *9*, *12*
pocre, *3–5*, 6, *9–13*, *15*
pocrepath, *3–5*, *8*, 8, *10–13*
pocrescreen, *3*, *8*, 10, *12, 13*

selectmodel, *9*, 11
sim5ydata, 12
simdata, 13
sipocre, 14