Package 'pETM'

May 22, 2016

Type Package

Title Penalized Exponential Tilt Model

Version 0.1.6

Date 2016-05-22

Author Hokeun Sun

Maintainer Hokeun Sun <hsun@pusan.ac.kr>

Description In analysis of high-dimensional DNA methylation data, a penalized exponential tilt model can identify differentially methylated loci between cases and controls, using network based regularization. It is able to detect any differences in means only, in variances only or in both means and variances.

License GPL-2

Repository CRAN

NeedsCompilation yes

R topics documented:

---

pETM *Penalized Exponential Tilt Model*

Description

Fit a penalized exponential tilt model (ETM) to identify differentially methylated loci between cases and controls. ETM is able to detect any differences in means only, in variances only or in both means and variances.

A penalized exponential tilt model using combined lasso and Laplacian penalties is applied to high-dimensional DNA methylation data with case-control association studies. When CpG sites are correlated with each other within the same gene or the same genetic region, Laplacian matrix can be imposed into the penalty function to encourage grouping effects among linked CpG sites. The selection probability of an individual CpG site is computed based on a finite number of resamplings.

Usage

```
pETM(x,y,cx=NULL,alpha=0.1,maxit=100000,thre=1e-6,group=NULL,lambda=NULL,
      type=c("ring","fcon"),etm=c("none","normal","beta"),psub=0.5,nlam=10,
      kb=10,K=100)
```

Arguments

x	Observed DNA methylation beta values consisting of n samples and p CpG sites. It should be (n x p) design matrix without an intercept.
y	The phenotype outcome coded as 1 for cases and 0 for the controls.
cx	The covariates such as age and gender. It should be (n x m) matrix, where m is the number of the covariates.
alpha	The penalty mixing parameter with $0 \leq \alpha \leq 1$ and default is 0.1. See details.
maxit	Maximum number of passes over the data for all regularization values, and default is 10^5. For fast computation, use a smaller value than the default value.
thre	Convergence threshold for coordinate descent algorithm. The default value is 1E-6. For fast computation, use a larger value than the default value.
group	The integer vector describing the size of genes or genetic regions. The length of group should be equivalent to the total number of genes or genetic regions, and the sum of group should be the same as the total number of CpG sites. If no group information is available, i.e., not specified, the pETM performs an elastic-net regularization procedure of a logistic regression. See details.
lambda	A sequence of regularization tuning parameter can be specified. Typical usage is to have the program compute its own lambda sequence based on nlam and kb.
type	A type of network within each group when group is specified. "ring" and "fcon" represent a ring and fully connected network, respectively. Default is "ring". See details.
etm	A type of an exponential tilt model. none does not perform an exponential tilt model, instead an ordinary penalized logistic regression model is applied. normal performs a penalized exponential tilt model based on a Gaussian distribution, and beta performs a penalized exponential tilt model based on a Beta distribution. See details.
psub	The proportion of subsamples used for resamplings, and psub$\in [0.5, 1)$. The default is 0.5.
nlam	The number of lambda values used for resamplings, and default is 10. For fast computation, use a smaller value than the default value.
kb	The number of burn-out replications before resamplings to properly adjust a sequence of lambda values and default is 10.
K	The number of resamplings, and default is 100.

Details

The exponential tilt model based on a logistic regression is defined as

$$\log \frac{p(x_i)}{1-p(x_i)} = \beta_0 + h_1(x_i)^T \beta_1 + h_2(x_i)^T \beta_2,$$

where $h_1(\cdot)$ and $h_2(\cdot)$ are pre-specified functions. For example $h_1(x) = x$ and $h_2(x) = x^2$ if etm is normal and $h_1(x) = -\log(x)$ and $h_2(x) = -\log(1-x)$ if etm is beta.

The penalty function of pETM is defined as

$$\alpha||\beta||_1 + (1-\alpha)(\beta^T L\beta)/2,$$

where L is a Laplacian matrix describing a group structure of CpG sites. This penalty is equivalent to the Lasso penalty if alpha=1. When group is not defined, L is replaced by an identity matrix. In this case, pETM performs an elastic-net regularization procedure since the second term of the penalty simply reduces to the squared l_2 norm of β.

If group sizes of CpG sites are listed in group, it is assumed that CpG sites within the same genes are linked with each other like a ring or a fully connected network. In this case, the Laplacian matrix forms a block-wise diagonal matrix. The ring network assumes only adjacent CpG sites within the same genes are linked with each other, while every CpG sites within the same genes are linked with each other for fully connected network. For a big gene, ring network is recommended for computational speed-up.

The selection result is summarized as the selection probability of individual CpG sites. The psub portions of n samples are randomly selected without replacement K times. For each subsample of (x,cx,y), pETM is applied to find non-zero coefficients of CpG sites along with nlam lambda values. The selection probability of each CpG site is then computed based on the maximum proportion of non-zero regression coefficients among K replications.

Value

selprob	The selection probabilities of p CpG sites
topsp	The selection probability of each CpG site is listed in descending order along with the name of CpG sites.
lambda	The actual sequence of lambda values used
valid.K	The actual number of resamplings used

Author(s)

Hokeun Sun <hsun@pusan.ac.kr>

References

H. Sun and S. Wang (2012) *Penalized Logistic Regression for High-dimensional DNA Methylation Data with Case-Control Studies*, Bioinformatics 28(10), 1368–1375

H. Sun and S. Wang (2013) *Network-based Regularization for Matched Case-Control Analysis of High-dimensional DNA Methylation Data*, Statistics in Medicine 32(12), 2127–2139

H. Sun and S. Wang (2016) *Penalized Exponential Tilt Model for Analysis of High-dimensional DNA Methylation Data*, Manuscript

Examples

```
n <- 100
p <- 500
x <- matrix(rnorm(n*p), n, p)
y <- rep(0:1, c(50,50))
```

```
# a total of 200 genes each of which consists of 1, 2, or 5 CpG sites
gr <- rep(c(1,2,5), c(50,100,50))

# ordinary penalized logistic regression
g1 <- pETM(x, y, group=gr, K=10)

# penalized exponential tilt model based on Gaussian distribution
g2 <- pETM(x, y, group=gr, etm = "normal", K=10)

# penalized exponential tilt model based on Beta distribution
x2 <- matrix(runif(n*p), n, p)
g3 <- pETM(x2, y, group=gr, etm = "beta", K=10)
```

Index