# Package 'wOGL'

November 21, 2023

**Type** Package

**Title** wOGL: Weighted Overlapping Group Lasso for the selection of gene sets as groups

**Version** 0.1.0

**Author** Dan Huang, Geunsu Jo

**Maintainer** Dan Huang <dhuang1221@gmail.com>

**Description** Gene set analysis aims to identify differentially expressed gene sets, often ignoring genetic network structure, which is less effective for sparse signals. Weighted Overlapping Group Lasso leverages network knowledge to identify interconnected genes, combining network-based regularization with overlapping group lasso, using l2-norm of regression coefficients for individual genes can play a role of the weight of gene sets for group selection.

**License** GPL-2

**Encoding** UTF-8

**Roxygen** list(markdown = TRUE)

**RoxygenNote** 7.2.3

**LazyData** true

**Depends** Matrix, mnormt, glmnet, pclogit, gglasso

**Suggests** stats

## R topics documented:

---

adjacency                          *adjacency matrix function*

---

#### Description

Definition of the incidence matrix $A$: `A[i, j]` equals 1 when group i includes variable j.

## Usage

```
adjacency(x, group)
```

## Arguments

| | |
|---|---|
| x | description x an input matrix of dimension nobs x nvars. Each row represents an observation, each column corresponds to a covariate. |
| group | description groups an integer vector indicating group sizes or as a symmetric adjacency matrix, which characterizes the grouping or graph structure of the predictors in 'x'. |

---

| per.gglasso | *Permutation test on selection probability with 'gglasso'* |
|---|---|

---

## Description

The permutation test on selection probability with 'gglasso' is computed based on resamplings.

## Usage

```
per.gglasso(x, y, group, nperm = nperm, maxit = 1e+07)
```

## Arguments

| | |
|---|---|
| x | description x an input matrix of dimension nobs x nvars. Each row represents an observation, each column corresponds to a covariate. |
| y | description y a response variable, where 'y = 1' corresponds to the case and 'y = 0' corresponds to the control. |
| group | description groups an integer vector indicating group sizes or as a symmetric adjacency matrix, which characterizes the grouping or graph structure of the predictors in 'x'. |
| nperm | the number of permutation is a user-defined parameter, with the default value set to 100. |
| maxit | the maximum number of iterations is $1e + 07$. |

---

| sel.gglasso | *selection probability with 'gglasso'* |
|---|---|

---

## Description

The selection probability with 'gglasso' is computed based on resamplings.

## Usage

```
sel.gglasso(
  x,
  y,
  group,
  ...,
  alpha = 0.1,
  psub = 0.5,
  N.lam = 15,
  K = 100,
  eps = 1e-04,
  maxit = 1e+07
)
```

## Arguments

| | |
|---|---|
| x | description x an input matrix of dimension nobs x nvars. Each row represents an observation, each column corresponds to a covariate. |
| y | description y a response variable, where 'y = 1' corresponds to the case and 'y = 0' corresponds to the control. |
| group | description groups an integer vector indicating group sizes or as a symmetric adjacency matrix, which characterizes the grouping or graph structure of the predictors in 'x'. |
| ... | additional arguments that can be supplied to gglasso. |
| alpha | the penalty mixing parameter ranges from 0 to 1, with the default value set to 0.1. |
| psub | the proportion of subsamples used for resampling is denoted as psub, the default value is 0.5. |
| N.lam | the number of lambda values used for resamplings is specified, with the default value set to 15. |
| K | the number of resamplings is a user-defined parameter, with the default value set to 100. |
| eps | the numerical computations, the tolerance for small numerical values is $1e - 04$. |
| maxit | the maximum number of iterations is $1e + 07$. |

---

| | |
|---|---|
| wOGL | *Weighted Overlapping Group Lasso for the selection of gene sets as groups* |

---

## Description

Gene set analysis aims to identify differentially expressed gene sets, often ignoring genetic network structure, which is less effective for sparse signals. Weighted Overlapping Group Lasso leverages network knowledge to identify interconnected genes, combining network-based regularization with overlapping group lasso, using l2-norm of regression coefficients for individual genes can play a role of the weight of gene sets for group selection.

## Usage

```
wOGL(
  x,
  y,
  group,
  adjm,
  nperm = nperm,
  stra = NULL,
  nfold = 5,
  alpha = 0.1,
  nlam = 100,
  N.lam = 15,
  K = 100,
  sgnc = NULL,
  maxit = 1e+07
)
```

## Arguments

| | |
|---|---|
| x | description x an input matrix of dimension nobs x nvars. Each row represents an observation, each column corresponds to a covariate. |
| y | description y a response variable, where 'y = 1' corresponds to the case and 'y = 0' corresponds to the control. |
| group | description groups an integer vector indicating group sizes or as a symmetric adjacency matrix, which characterizes the grouping or graph structure of the predictors in 'x'. |
| adjm | the incidence matrix A: A[i, j] equals 1 when group i includes variable j. |
| nperm | the number of permutation tests. |
| stra | a vector of consecutive integers is used to indicate the stratum of each observation. Each stratum must contain exactly one case and at least one control. If not specified, 'pclogit' will perform a standard logistic regression. |
| nfold | the default number of folds is 5. While 'nfold' can go up to the sample size, it is not advisable for large datasets. |
| alpha | the penalty mixing parameter ranges from 0 to 1, with the default value set to 0.1. |
| nlam | the number of lambda values, with the default value set to 100. |
| N.lam | the number of lambda values used for resamplings is specified, with the default value set to 15. |
| K | the number of resamplings is a user-defined parameter, with the default value set to 100. |
| sgnc | regression coefficients' signs can be provided if the 'group' is specified as either a list of group sizes or an adjacency matrix. |
| maxit | the maximum number of iterations is $1e + 07$. |

## Details

More detailed information, please refer to the provided reference below.

## Value

A matrix includes both the order of weighted selection probabilities arranged from the highest to the lowest and the corresponding weighted selection probabilities.

## 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 (2012) Network-based Regularization for Matched Case-Control Analysis of High-dimensional DNA Methylation Data, manuscript

Yang Yi and Hui Zou (2015) A fast unified algorithm for solving group-lasso penalize learning problems, Statistics and Computing 25, 1129-1141.

## Examples

```
n <- 200
p <- 1000
x <- matrix(rnorm(n*p), nrow=n, ncol=p)
y <- c(rep(0, n/2), rep(1, n/2))

# a total of 10 groups, each consisting of different number of and overlapping members
group <- list(gr1 = c(1:31), gr2 = c(1, 17:54),
              gr3 = c(1, 42:61), gr4 = c(1, 47:76),
              gr5 = c(1, 65:92), gr6 = c(1, 78:108),
              gr7 = c(1, 82:125), gr8 = c(1, 94:140),
              gr9 = c(1, 106:143), gr10 = c(1, 118:160))

# an adjacency matrix
adjm <- adjacency(x=x, group=group)

# weighted overlapping group lasso
wOGL <- wOGL(x=x, y=y, group=group, adjm=adjm, nperm=10, stra=NULL,
             nfold=5, alpha=0.1, nlam=100, N.lam=15, K=100, sgnc=NULL, maxit=1e+07)
```

# Index