Package 'eLNNpairedCov'

January 11, 2024

Type Package

Title Model-Based Gene Selection for Paired Data

Version 0.3.2

Date 2023-12-22

Depends R (>= 4.0.0), Biobase

Imports MASS, graphics, stats, limma, methods, parallel

biocViews Bioinformatics, DifferentialExpression

Description
Model-based clustering for paired data based on the regression of a mixture of Bayesian hierarchical models on covariates. Zhang et al. (2023) <doi:10.1186/s12859-023-05556-x>.

License GPL (>= 2)

NeedsCompilation no

Author Yixin Zhang [aut, cre],
Wei Liu [aut, ctb],
Weiliang Qiu [aut, ctb]

Maintainer Yixin Zhang <zhyl133@gmail.com>

Repository CRAN

Date/Publication 2024-01-11 09:30:07 UTC

R topics documented:

eLNNpairedCov *Model-Based Clustering for Paired Data Adjusting for Covariates*

Description

Model-based clustering based on extended log-normal normal model for paired data adjusting for covariates.

Usage

```
eLNNpairedCov(
  EsetDiff,
  fmla = ~Age + Sex,
  probeID.var = "probeid",
  gene.var = "gene",
  chr.var = "chr",
  scaleFlag = TRUE,
  Maxiter =10,
  maxIT = 10,
  b=c(2,2,2),
  converge_threshold = 1e-3,
  optimMethod = "L-BFGS-B",
  bound.alpha = c(0.001, 6),
  bound.beta = c(0.001, 6),
  bound.k = c(0.001, 0.9999),
  bound.eta = c(-10, 10),
  mc.cores = 1,
  verbose=FALSE)
```

Arguments

EsetDiff	An ExpressionSet object storing the log2 difference between post-treatment and pre-treatment.
fmla	A formula without outcome variable.
probeID.var	character. Indicates the probe id.
gene.var	character. Indicates the gene symbol.
chr.var	character. Indicates the chromosome.
scaleFlag	logical. Indicating if rows (probes) need to be scaled (but not centered).
Maxiter	integer. The max allowed number of iterations for EM algorithm. Default value is maxRT = 100.
maxIT	integer. The max allowed number of iterations in R built-in function optim. Default value is maxIT = 100. maxIT.
b	numeric. A vector of concentration parameters used in Dirichlet distribution. Default value is b = c(2,2,2).

converge_threshold
: numeric. One of the two termination criteria of iteration. The smaller this value is set, the harder the optimization procedure in eLNNpaired will be considered to be converged. Default value is converge_threshold = 1e-6.

optimMethod
: character. Indicates the method for optimization. optim.

bound.alpha
: numeric. A vector of 2 positive numbers to specify lower and upper bound of estimate of α_c, c="OE", "UE", or "NE".

bound.beta
: numeric. A vector of 2 positive numbers to specify lower and upper bound of estimate of β_c, c="OE", "UE", or "NE".

bound.k
: numeric. A vector of 2 positive numbers to specify lower and upper bound of estimate of k_c, c="OE", "UE", or "NE".

bound.eta
: numeric. A vector of p+1 positive numbers to specify lower and upper bound of estimate of η_c, c="OE", "UE", or "NE", where p is the number of covariates.

mc.cores
: integer. A positive integer specifying number of computer cores to be used by parallel computing.

verbose
: logic. An indicator variable telling if print out intermediate results: FALSE for not printing out, TRUE for printing out. Default value is verbose = False.

Details

A gene will be assigned to cluster "NE" if its posterior probability for non-differentially expressed gene cluster is the largest. A gene will be assigned to cluster "OE" if its posterior probability for over-expressed gene cluster is the largest. A gene will be assigned to cluster "UE" if its responsibility for under-expressed gene cluster is the largest.

Value

A list of 9 elementes:

par.ini
: initial estimate of parameter

par.final
: A vector of the estimated model parameters in original scale.

memGenes
: probe cluster membership based on eLNNpairedCov algorithm.

memGenes2
: probe cluster membership based on eLNNpairedCov algorithm. 2-categories: "DE" indicates differentially expressed; "NE" indicates non-differentially expressed.

memGenes.limma
: probe cluster membership based on limma.

res.ini
: results of limma analysis

update_info
: object returned by optim function

wmat
: matrix of responsibilities

iter
: number of EM iterations.

Author(s)

Yixin Zhang <zhyl133@gmail.com>, Wei Liu <liuwei@mathstat.yorku.ca>, Weiliang Qiu <weiliang.qiu@sanofi.com>

References

Zhang Y, Liu W, Qiu W. A model-based clustering via mixture of hierarchical models with covariate adjustment for detecting differentially expressed genes from paired design. *BMC Bioinformatics* 24, 423 (2023)

Examples

```
data(esDiff)

res = eLNNpairedCov(EsetDiff = esDiff,
    fmla = ~Age + Sex,
    probeID.var = "probeid",
    gene.var = "gene",
    chr.var = "chr",
    scaleFlag = FALSE,
    mc.cores = 1,
        verbose = TRUE)

# true probe cluster membership
memGenes.true = fData(esDiff)$memGenes.true
print(table(memGenes.true))

# probe cluster membership
memGenes.limma = res$memGenes.limma
print(table(memGenes.limma))

# final probe cluster membership
memGenes = res$memGenes
print(table(memGenes))

# cross tables
print(table(memGenes.true, memGenes.limma))
print(table(memGenes.true, memGenes))

# accuracies
print(mean(memGenes.true == memGenes.limma))
print(mean(memGenes.true == memGenes))
```

`eLNNpairedCovSEM` *Model-Based Clustering for Paired Data Adjusting for Covariates Using Simulated Annealing Modified EM*

Description

Model-based clustering based on extended log-normal normal model for paired data adjusting for covariates.

Usage

```
eLNNpairedCovSEM(
  EsetDiff,
  fmla = ~Age + Sex,
  probeID.var = "probeid",
  gene.var = "gene",
  chr.var = "chr",
  scaleFlag = TRUE,
  Maxiter =10,
  maxIT = 10,
  b=c(2,2,2),
  converge_threshold = 1e-3,
  optimMethod = "L-BFGS-B",
  bound.alpha = c(0.001, 6),
  bound.beta = c(0.001, 6),
  bound.k = c(0.001, 0.9999),
  bound.eta = c(-10, 10),
  mc.cores = 1,
  temp0 = 2,
  r_cool=0.9,
  verbose=FALSE)
```

Arguments

EsetDiff	An ExpressionSet object storing the log2 difference between post-treatment and pre-treatment.
fmla	A formula without outcome variable.
probeID.var	character. Indicates the probe id.
gene.var	character. Indicates the gene symbol.
chr.var	character. Indicates the chromosome.
scaleFlag	logical. Indicating if rows (probes) need to be scaled (but not centered).
Maxiter	integer. The max allowed number of iterations for EM algorithm. Default value is maxRT = 100.
maxIT	integer. The max allowed number of iterations in R built-in function optim. Default value is maxIT = 100. maxIT.
b	numeric. A vector of concentration parameters used in Dirichlet distribution. Default value is b = c(2,2,2).
converge_threshold	numeric. One of the two termination criteria of iteration. The smaller this value is set, the harder the optimization procedure in eLNNpaired will be considered to be converged. Default value is converge_threshold = 1e-6.
optimMethod	character. Indicates the method for optimization. optim.
bound.alpha	numeric. A vector of 2 positive numbers to specify lower and upper bound of estimate of α_c, c="OE", "UE", or "NE".

bound.beta	numeric. A vector of 2 positive numbers to specify lower and upper bound of estimate of β_c, c="OE", "UE", or "NE".
bound.k	numeric. A vector of 2 positive numbers to specify lower and upper bound of estimate of k_c, c="OE", "UE", or "NE".
bound.eta	numeric. A vector of p+1 positive numbers to specify lower and upper bound of estimate of η_c, c="OE", "UE", or "NE", where p is the number of covariates.
mc.cores	integer. A positive integer specifying number of computer cores to be used by parallel computing.
temp0	numeric. Initial temperature in simulated-annealing modified EM.
r_cool	numeric. Cooling rate in simulated-annealing modified EM, which is inside interval $(0, 1)$.
verbose	logic. An indicator variable telling if print out intermediate results: FALSE for not printing out, TRUE for printing out. Default value is verbose = False.

Details

A gene will be assigned to cluster "NE" if its posterior probability for non-differentially expressed gene cluster is the largest. A gene will be assigned to cluster "OE" if its posterior probability for over-expressed gene cluster is the largest. A gene will be assigned to cluster "UE" if its responsibility for under-expressed gene cluster is the largest.

Value

A list of 9 elementes:

par.ini	initial estimate of parameter
par.final	A vector of the estimated model parameters in original scale.
memGenes	probe cluster membership based on eLNNpairedCovSEM algorithm.
memGenes2	probe cluster membership based on eLNNpairedCovSEM algorithm. 2-categories: "DE" indicates differentially expressed; "NE" indicates non-differentially expressed.
memGenes.limma	probe cluster membership based on limma.
res.ini	results of limma analysis
update_info	object returned by optim function
wmat	matrix of responsibilities
iter.EM	number of EM iterations.
tempFinal	final temperature in simulated-annealing modification EM

Author(s)

Yixin Zhang <zhyl133@gmail.com>, Wei Liu <liuwei@mathstat.yorku.ca>, Weiliang Qiu <weiliang.qiu@sanofi.com>

References

Zhang Y, Liu W, Qiu W. A model-based clustering via mixture of hierarchical models with covariate adjustment for detecting differentially expressed genes from paired design. *BMC Bioinformatics* 24, 423 (2023)

Examples

```
data(esDiff)

res.SEM = eLNNpairedCovSEM(EsetDiff = esDiff,
    fmla = ~Age + Sex,
    probeID.var = "probeid",
    gene.var = "gene",
    chr.var = "chr",
    scaleFlag = FALSE,
    mc.cores = 1,
        verbose = TRUE)

# true probe cluster membership
memGenes.true = fData(esDiff)$memGenes.true
print(table(memGenes.true))

# probe cluster membership
memGenes.limma = res.SEM$memGenes.limma
print(table(memGenes.limma))

# final probe cluster membership
memGenes.SEM = res.SEM$memGenes
print(table(memGenes.SEM))

# cross tables
print(table(memGenes.true, memGenes.limma))
print(table(memGenes.true, memGenes.SEM))

# accuracies
print(mean(memGenes.true == memGenes.limma))
print(mean(memGenes.true == memGenes.SEM))
```

`esDiff` *An ExpressionSet Object Storing a Simulated Data*

Description

An ExpressionSet object storing a simulated data of log2 difference of expression levels with 1000 probes, 20 subjects, and 2 covariates.

Usage

```
data("esDiff")
```

Details

This dataset was generated from the mixture of 3-component Bayesian hierarchical models. For true parameters, please refer to the manual for the R function `genSimDat`.

Examples

```
data(esDiff)

print(esDiff)
```

---

genSimDat *Generate Simulated Data*

Description

Generate a simulated dataset from a mixture of Bayesian hierarchical models with two covariates: age and sex.

Usage

```
genSimDat(G, n, psi, t_pi, m.age = 50, sd.age = 5, p.female = 0.5)
```

Arguments

G	integer. Number of probes.
n	integer. Number of samples.
psi	numeric. A vector of model hyper-parameters with elements α_1, β_1, k_1, $\eta_{1,intercept}$, $\eta_{1,age}$, $\eta_{1,sex}$, α_2, β_2, k_2, $\eta_{2,intercept}$, $\eta_{2,age}$, $\eta_{2,sex}$, α_3, β_3, k_3, $\eta_{3,intercept}$, $\eta_{3,age}$, $\eta_{3,sex}$.
t_pi	numeric. A vector of mixture proportions: π_1 (proportion for probes over-expressed in cases); π_2 (proportion for probes under-expressed in cases).
m.age	numeric. mean age.
sd.age	numeric. standard deviation of age.
p.female	numeric. proportion of females.

Value

An ExpressionSet object.

Note

Age will be mean-centered and scaled so that it will have mean zero and variance one.

Author(s)

Yixin Zhang <zhyl133@gmail.com>, Wei Liu <liuwei@mathstat.yorku.ca>, Weiliang Qiu <weiliang.qiu@sanofi.com>

References

Zhang Y, Liu W, Qiu W. A model-based clustering via mixture of hierarchical models with covariate adjustment for detecting differentially expressed genes from paired design. *BMC Bioinformatics* 24, 423 (2023)

Examples

```
set.seed(1234567)

true.psi = c(2, 1, 0.8,
             0.1, -0.01, -0.1,
             2, 1, 0.8,
             -0.1, -0.01, -0.1,
             2, 1, 0.8,
             -0.01, -0.1)
names(true.psi)=c("alpha1", "beta1", "k1",
                  "eta1.intercept", "eta1.Age", "eta1.Sex",
                  "alpha2", "beta2", "k2",
                  "eta2.intercept", "eta2.Age", "eta2.Sex",
                  "alpha3", "beta3", "k3",
                  "eta3.Age", "eta3.Sex")
true.pi=c(0.1, 0.1)
names(true.pi)=c("pi.OE", "pi.UE")
par.true=c(true.pi, true.psi)

esDiff = genSimDat(G = 1000,
       n = 20,
       psi = true.psi,
       t_pi = true.pi,
       m.age = 0, # scaled age
       sd.age = 1, # scaled age
       p.female = 0.5)

print(esDiff)
```

Index