Package ‘iClusterPlus’

April 25, 2017

Title Integrative clustering of multi-type genomic data
Version 1.12.0
Date 2016-10-03
Depends R (>= 2.15.0), parallel
Suggests RUnit, BiocGenerics
Author Qianxing Mo, Ronglai Shen
Maintainer Qianxing Mo <qmo@bcm.edu>, Ronglai Shen <shenr@mskcc.org>
Description Integrative clustering of multiple genomic data using a joint latent variable model
LazyData yes
License GPL (>= 2)
biocViews Microarray, Clustering
NeedsCompilation yes

R topics documented:

breast.chr17 ................................................................. 2
CNregions ................................................................. 2
compute.pod ............................................................... 3
coord ................................................................. 4
gbm ................................................................. 5
glp ................................................................. 5
iCluster ................................................................. 6
iCluster2 ................................................................. 7
iClusterPlus ........................................................... 9
plotHeatmap ............................................................ 11
plotiCluster .............................................................. 12
plotRI ................................................................. 13
simuResult ............................................................. 14
tune.iCluster2 .......................................................... 14
tune.iClusterPlus ...................................................... 15
utility ................................................................. 17
variation.hg18.v10.nov.2010 .......................................... 18

Index 19
breast.chr17  

*Breast cancer data set DNA copy number and mRNA expression measure on chromosome 17*

**Description**

This is a subset of the breast cancer data from Pollack et al. (2002).

**Usage**

```r
data(breast.chr17)
```

**Format**

A list object containing two data matrices: DNA and mRNA. They consist chromosome 17 data in 41 samples (4 cell lines and 37 primary tumors).

**Source**

This data can be downloaded at [http://www.pnas.org/content/99/20/12963/suppl/DC1](http://www.pnas.org/content/99/20/12963/suppl/DC1)

**References**


---

**CNregions**  

*A function to remove redundant copy number regions*

**Description**

This function is used to reduce copy number regions.

**Usage**

```r
CNregions(seg, epsilon=0.005, adaptive=FALSE, rmCNV=FALSE, cnv=NULL, frac.overlap=0.5, rmSmallseg=TRUE, nProbes=15)
```

**Arguments**

- `seg` DNAcopy CBS segmentation output.
- `epsilon` the maximum Euclidean distance between adjacent probes tolerated for denying a nonredundant region. epsilon=0 is equivalent to taking the union of all unique break points across the n samples.
- `adaptive` Vector of length-m lasso penalty terms.
- `rmCNV` If TRUE, remove germline CNV.
- `cnv` A data frame containing germline CNV data.
- `frac.overlap` A parameter needed to be explain.
- `rmSmallseg` If TRUE, remove small segment.
- `nProbes` The segment length threshold below which the segment will be removed if rmSmallseg = TRUE.
compute.pod

Value
A matrix with reduced copy number regions.

Author(s)
Ronglai Shen <shenr@mskcc.org>

References

See Also
breast.chr17,plotiCluster,compute.pod,iCluster,iClusterPlus

Examples

```r
#data(gbm)
#library(GenomicRanges)
#library(cluster)
#reducedM=CNregions(seg,epsilon=0,adaptive=FALSE,rmCNV=TRUE,cnv=NULL,
# frac.overlap=0.5, rmSmallseg=TRUE,nProbes=5)
```

compute.pod A function to compute the proportion of deviation from perfect block diagonal matrix

Description
A function to compute the proportion of deviation from perfect block diagonal matrix.

Usage
compute.pod(fit)

Arguments

fit A iCluster object

Value
pod proportion of deviation from perfect block diagonal matrix

Author(s)
Ronglai Shen <shenr@mskcc.org>
coord
genomic coordinates

description

genomic coordinates for the copy number data in gbm

Usage

data(coord)

Format

A data matrix consists of chr number, start and end position for the genes included in the gbm copy number data.

References

gbm

**GBM data**

**Description**

This is a subset of the glioblastoma dataset from the cancer genome atlas (TCGA) GBM study (2009) used in Shen et al. (2012).

**Usage**

data(gbm)

**Format**

A list object containing three data matrices: copy number, methylation and mRNA expression in 84 samples.

**Value**

- gbm.seg: GBM copy number segmentation results generated by DNAcopy package.
- gbm.exp: GBM gene expression data.
- gbm.mut: GBM mutation data.

**References**


---

glp

**good lattice points using the uniform design**

**Description**

good lattice points using the uniform design (Fang and Wang 1995)

**Usage**

data(glp)

**Format**

A list object containing sampling design for s=2-5 where s is the number of tuning parameters.

**References**


**Description**

Given multiple genomic data types (e.g., copy number, gene expression, DNA methylation) measured in the same set of samples, iCluster fits a regularized latent variable model based clustering that generates an integrated cluster assignment based on joint inference across data types.

**Usage**

iCluster(datasets, k, lambda, scalar=FALSE, max.iter=50, epsilon=1e-3)

**Arguments**

datasets A list object containing m data matrices representing m different genomic data types measured in a set of n samples. For each matrix, the rows represent samples, and the columns represent genomic features.

k Number of subtypes.

lambda Vector of length-m lasso penalty terms.

scalar If TRUE, assumes scalar covariance matrix Psi. Default is FALSE.

max.iter Maximum iteration for the EM algorithm.

epsilon EM algorithm convergence criterion.

**Value**

A list with the following elements.

meanZ Relaxed cluster indicator matrix.

beta Coefficient matrix.

clusters Cluster assignment.

conv.rate Convergence history.

**Author(s)**

Ronglai Shen <shenr@mskcc.org>

**References**


**See Also**

breast.chr17, plotiCluster, compute.pod
Examples

```r
data(breast.chr17)
fit=iCluster(breast.chr17, k=4, lambda=c(0.2,0.2))
plotICluster(fit=fit, label=rownames(breast.chr17$[2]))
compute.pod(fit)

#library(gplots)
#library(lattice)
#col.scheme = alist()
#col.scheme[[1]] = bluered(256)
#col.scheme[[2]] = greenred(256)
#cn.image=breast.chr17$[2]
#cn.image[cn.image>1.5]=1.5
#cn.image[cn.image<-1.5]=-1.5
#exp.image=breast.chr17$[1]
#exp.image[exp.image>3]=3
#exp.image[exp.image<-3]=3
#plotHeatmap(fit, datasets=list(cn.image,exp.image), type=c("gaussian","gaussian"),
# row.order=c(FALSE,FALSE), width=5, col.scheme=col.scheme)
```

---

**iCluster2**

**Integrative clustering of multiple genomic data types**

**Description**

Given multiple genomic data types (e.g., copy number, gene expression, DNA methylation) measured in the same set of samples, iCluster fits a regularized latent variable model based clustering that generates an integrated cluster assignment based on joint inference across data types.

**Usage**

```r
iCluster2(x, K, lambda, method=c("lasso","enet","flasso","glasso","gflasso"),
    chr=NULL, maxiter=50, eps=1e-4, eps2=1e-8)
```

**Arguments**

- **x**: A list object containing m data matrices representing m different genomic data types measured in a set of n samples. For each matrix, the rows represent samples, and the columns represent genomic features.
- **K**: Number of subtypes.
- **lambda**: A list with m elements; each element is a vector with one or two elements depending on the methods used.
- **method**: Method used for clustering and variable selection.
- **chr**: Chromosome labels
- **maxiter**: Maximum iteration for the EM algorithm.
- **eps**: EM algorithm convergence criterion 1.
- **eps2**: EM algorithm convergence criterion 2.
Value

A list with the following elements.

- `cluster`: Cluster assignment.
- `centers`: Cluster centers.
- `Phivec`: Parameter $\phi$; a vector.
- `beta`: Parameter $B$; a matrix.
- `meanZ`: $\text{meanZ}$
- `EZZt`: $\text{EZZt}$
- `dif`: Difference
- `iter`: Iteration

Author(s)

Qianxing Mo <qmo@bcm.edu>, Ronglai Shen, Sijian Wang

References


See Also

`plotiCluster`, `compute.pod`, `iClusterPlus`

Examples

```r
## clustering
n1 = 20
d2 = 20
n3 = 20
n = n1+n2+n3
p = 5
q = 100

x = NULL
x1a = matrix(rnorm(n1*p), ncol=p)
x2a = matrix(rnorm(n1*p, -1.5,1), ncol=p)
x3a = matrix(rnorm(n1*p, 1.5, 1), ncol=p)
xa = rbind(x1a,x2a,x3a)
xb = matrix(rnorm(n*q), ncol=q)
x[[1]] = cbind(xa,xb)

x1a = matrix(rnorm(n1*p), ncol=p)
x2a = matrix(rnorm(n1*p, -1.5,1), ncol=p)
x3a = matrix(rnorm(n1*p, 1.5, 1), ncol=p)
xa = rbind(x1a,x2a,x3a)
xb = matrix(rnorm(n*q), ncol=q)
x[[2]] = cbind(xa,xb)

x1a = matrix(rnorm(n1*p), ncol=p)
x2a = matrix(rnorm(n1*p, -1.5,1), ncol=p)
```
iClusterPlus

Description

Given multiple genomic data types (e.g., copy number, gene expression, DNA methylation) measured in the same set of samples, iClusterPlus fits a regularized latent variable model based clustering that generates an integrated cluster assignment based on joint inference across data types.
iClusterPlus

Usage

iClusterPlus(dt1, dt2=NULL, dt3=NULL, dt4=NULL, type=c("gaussian", "binomial", "poisson", "multinomial"), K=2, alpha=c(1,1,1,1), lambda=c(0.03, 0.03, 0.03, 0.03), n.burnin=100, n.draw=200, maxiter=20, sdev=0.05, eps=1.0e-4)

Arguments

dt1 A data matrix. The rows represent samples, and the columns represent genomic features.
dt2 A data matrix. The rows represent samples, and the columns represent genomic features.
dt3 A data matrix. The rows represent samples, and the columns represent genomic features.
dt4 A data matrix. The rows represent samples, and the columns represent genomic features.
type Data type, which can be gaussian, binomial, poisson, multinomial.
K The number of eigen features. Given K, the number of clusters is K+1.
alpha Vector of elasticnet penalty terms. At this version of iClusterPlus, elasticnet is not used. Therefore, all the elements of alpha are set to 1.
lambda Vector of lasso penalty terms.
n.burnin Number of MCMC burnin.
n.draw Number of MCMC draw.
maxiter Maximum iteration for the EM algorithm.
sdev standard deviation of random walk proposal.
eps Algorithm convergence criterion.

Value

A list with the following elements.

alpha Intercept parameter.
beta Information parameter.
clusters Cluster assignment.
centers Cluster center.
meanZ Latent variable.
BIC Bayesian information criterion.
dev.ratio see dev.ratio defined in glmnet package.
dif absolute difference for the parameters in the last and next-to-last iterations.

Author(s)

Qianxing Mo <qmo@bcm.edu>, Ronglai Shen, Sijian Wang
plotHeatmap

References

See Also
plotiCluster,iCluster,compute.pod

Examples
# see iManual.pdf

plotHeatmap A function to generate heatmap panels sorted by integrated cluster assignment.

Description
A function to generate heatmap panels sorted by integrated cluster assignment.

Usage
plotHeatmap(fit,datasets,type=c("gaussian","binomial","poisson","multinomial"), sample.order=NULL,row.order=NULL,sparse=NULL,threshold=rep(0.25,length(datasets)), width=5,scale=rep("none",length(datasets)),col.scheme=rep(list(bluered(256)), length(datasets)), chr=NULL, plot.chr=NULL, cap=NULL)

Arguments
fit A iCluster object.
datasets A list object of data matrices.
type Types of data in the datasets.
sample.order User supplied cluster assignment.
row.order A vector of logical values each specify whether the genomic features in the corresponding data matrix should be reordered by similarity. Default is TRUE.
sparse A vector of logical values each specify whether to plot the top cluster-discriminant features. Default is FALSE.
threshold When sparse is TRUE, a vector of threshold values to include the genomic features for which the absolute value of the associated coefficient estimates fall in the top quantile. threshold=c(0.25,0.25) takes the top quartile most discriminant features in data type 1 and data type 2 for plot.
width Width of the figure in inches
scale A vector of logical values each specify whether data should be scaled. Default is FALSE.
col.scheme Color scheme. Can use bluered(n) in gplots R package.
chr A vector of chromosome number.
plotiCluster

A function to generate cluster separability matrix plot.

Description

A function to generate cluster separability matrix plot.

Usage

plotiCluster(fit, label=NULL)

Arguments

fit A iCluster object

label Sample labels

Value

no value returned.

References


See Also

iCluster,iCluster2

Examples

# see iManual.pdf
plotRI

Author(s)
Ronglai Shen <shenr@mskcc.org>

References

See Also
iCluster, compute.pod

Examples
```r
# library(iCluster)
# data(breast.chr17)
# fit=iCluster(datasets=breast.chr17, k=4, lambda=c(0.2,0.2))
# plotICluster(fit=fit, label=rownames(breast.chr17[[2]]))
# compute.pod(fit)
```

---

plotRI

A function to generate reproducibility index plot.

Description
A function to generate reproducibility index plot.

Usage
plotRI(cv.fit)

Arguments
cv.fit
A tune.iCluster2 object

Value
no value returned.

Author(s)
Ronglai Shen <shenr@mskcc.org>

References
See Also
tune.iCluster2

Examples

```r
#data(simu.datasets)
#cv.fit=alist()
#for(k in 2:5){
#  cat(paste("K=",k,sep=""),\n')
#  cv.fit[[k]]=tune.iCluster2(datasets=simu.datasets, k,nrep=2, n.lambda=8)
#}
##Reproducibility index (RI) plot
#plotRI(cv.fit)
```

`simuResult` *The results for the analysis of the simulated data.*

Description

The simulation and analysis are described in iClusterPlus/inst/unitTests/test_iClusterPlus.R.

Usage
data(simuResult)

Format
list

Value
A list of objects returned by the iClusterPlus function.

References
iClusterPlus/inst/unitTests/test_iClusterPlus.R

tune.iCluster2 *Integrative clustering of multiple genomic data types*

Description

Given multiple genomic data types (e.g., copy number, gene expression, DNA methylation) measured in the same set of samples, iCluster fits a regularized latent variable model based clustering that generates an integrated cluster assignment based on joint inference across data types.

Usage
tune.iCluster2(x, K, method=c("lasso","enet","flasso","lasso","glasso"),base=200, chr=NULL,true.class=NULL,lambda=NULL,n.lambda=NULL,save.nonsparse=F,nrep=10,eps=1e-4)
Arguments

- **x**: A list object containing \( m \) data matrices representing \( m \) different genomic data types measured in a set of \( n \) samples. For each matrix, the rows represent samples, and the columns represent genomic features.
- **K**: Number of subtypes.
- **lambda**: User supplied matrix of lambda to tune.
- **method**: Method used for clustering and variable selection.
- **chr**: Chromosome labels
- **n.lambda**: Number of lambda to sample using uniform design.
- **nrep**: Fold of cross-validation.
- **base**: Base.
- **true.class**: True class label if available.
- **save.nonsparse**: Logic argument whether to save the nonsparse fit.
- **eps**: EM algorithm convergence criterion

Value

A list with the following elements.

- **best.fit**: Best fit.
- **best.lambda**: Best lambda.
- **ps**: Rand index
- **ps.adjusted**: Adjusted Rand index.

Author(s)

Qianxing Mo <qmo@bcm.edu>, Ronglai Shen, Sijian Wang

References


See Also

- **iCluster2**

---

**tune.iClusterPlus**

*Integrative clustering of multiple genomic data*

Description

Given multiple genomic data (e.g., copy number, gene expression, DNA methylation) measured in the same set of samples, **tune.iClusterPlus** uses a series of lambda values to fit a regularized latent variable model based clustering that generates an integrated cluster assignment based on joint inference across data.
tune.iClusterPlus

Usage
tune.iClusterPlus(cpus=8, dt1, dt2= NULL, dt3= NULL, dt4= NULL,
type= c("gaussian", "binomial", "poisson", "multinomial"),
K= 2, alpha= c(1, 1, 1, 1), n.lambda= NULL, scale.lambda= c(1, 1, 1, 1),
n.burnin= 200, n.draw= 200, maxiter= 20, sdev= 0.05, eps= 1.0e-4)

Arguments
cpus Number of CPU used for parallel computation.
dt1 A data matrix. The rows represent samples, and the columns represent genomic features.
dt2 A data matrix. The rows represent samples, and the columns represent genomic features.
dt3 A data matrix. The rows represent samples, and the columns represent genomic features.
dt4 A data matrix. The rows represent samples, and the columns represent genomic features.
type data type, which can be "gaussian", "binomial", "poisson", and "multinomial".
K The number of eigen features. Given K, the number of cluster is K+1.
alpha Vector of elasticnet penalty terms. At this version of iClusterPlus, elasticnet is not used. Therefore, all the elements of alpha are set to 1.
n.lambda Number of lambda are tuned.
scale.lambda A value between (0,1); the actual lambda values will be scale.lambda multiplying the lambda values of the uniform design.
n.burnin Number of MCMC burnin.
n.draw Number of MCMC draw.
maxiter Maximum iteration for the EM algorithm.
sdev standard deviation of random walk proposal.
eps EM algorithm convergence criterion.

Value
A list with the two elements 'fit' and 'lambda', where fit itself is a list and lambda is a matrix. Each row of lambda is the lambda values used to fit iClusterPlus model. Each component of fit is an object return by iClusterPlus, one-to-one corresponding to the row of lambda. Each component of fit has the following objects.

alpha Intercept parameter for the genomic features.
beta Information parameter for the genomic features. The rows and the columns represent the genomic features and the coefficients for the latent variable, respectively.
clusters Cluster assignment.
centers Cluster centers.
meanZ Latent variable.

Author(s)
Qianxing Mo <qmo@bcm.edu>, Ronglai Shen <shenr$mskcc.org>
utility

References


See Also

plotiCluster,iClusterPlus,iCluster2,iCluster,compute.pod

Examples

### see the users’ guide iManul.pdf

utility

Utility functions for iClusterPlus package

Description

Some utility functions for processing the results produced by iClusterPlus methods.

Usage

getBIC(resultList)
getDevR(resultList)
getClusters(resultList)
iManual(view=TRUE)

Arguments

resultList A list object as shown in the following example.
view A logical value TRUE or FALSE

Value

getBIC produce a matrix containing the BIC value for each lambda and K; the rows correspond to the lambda (vector) and the columns correspond to the K latent variables.

getDevR produce a matrix containing the deviance ratio for each lambda and K; the rows correspond to the lambda (vector) and the columns correspond to the K latent variables.

getClusters produce a matrix containing the cluster assignments for the samples under each K; the rows correspond to the samples; the columns correspond to the K latent variables.


Author(s)

Qianxing Mo <qmo@bcm.edu>
References


See Also

tune.iClusterPlus, iClusterPlus, iCluster2

Examples

```r
### see the users' guide iManual.pdf

data(simuResult)
#BIC = getBIC(simuResult)
#devR = getDevR(simuResult)
#clusters = getClusters(simuResult)
```

---

**variation.hg18.v10.nov.2010**

*Human genome variants of the NCBI 36 (hg18) assembly*

Description

Human genome variants of the NCBI 36 (hg18) assembly

Usage

data(variation.hg18.v10.nov.2010)

Format

data frame

Value

variation.hg18.v10.nov.2010

*Human genome variants of the NCBI 36 (hg18) assembly*

References

http://projects.tcag.ca/variation/tableview.asp?table=DGV_Content_Summary.txt
Index

*Topic datasets
  breast.chr17, 2
  coord, 4
  gbm, 5
  glp, 5
  simuResult, 14
  variation.hg18.v10.nov.2010, 18

*Topic models
  CNregions, 2
  compute.pod, 3
  iCluster, 6
  iCluster2, 7
  iClusterPlus, 9
  plotHeatmap, 11
  plotiCluster, 12
  plotRI, 13
  tune.iCluster2, 14
  tune.iClusterPlus, 15
  utility, 17

breast.chr17, 2, 3, 6

CNregions, 2
compute.pod, 3, 3, 6, 8, 11, 13, 17
coord, 4

gbm, 5
getBIC (utility), 17
getClusters (utility), 17
getDevR (utility), 17

iCluster, 3, 4, 6, 11–14, 17
iCluster2, 4, 7, 12, 15, 17, 18
iClusterPlus, 3, 8, 9, 17, 18
iManual (utility), 17

plotHeatmap, 11
plotiCluster, 3, 4, 6, 8, 11, 12, 17
plotRI, 13

simuResult, 14

tune.iCluster2, 14
tune.iClusterPlus, 15, 18

utility, 17
variation.hg18.v10.nov.2010, 18

19