Package ‘iCluster’

February 20, 2015

Title Integrative clustering of multiple genomic data types
Version 2.1.0
Date 2012-05-01
Depends R (>= 2.15.0), lattice, caTools, gdata, gtools, gplots, parallel
Author Ronglai Shen
Maintainer Ronglai Shen <shenr@mskcc.org>
Description Integrative clustering of multiple genomic data types using a joint latent variable model.
LazyData yes
License GPL (>= 2)
biocViews Integrated omic data, Bioinformatics
Repository CRAN
Date/Publication 2012-05-08 04:07:00
NeedsCompilation no

R topics documented:

<table>
<thead>
<tr>
<th>Topic</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>breast.chr17</td>
<td>2</td>
</tr>
<tr>
<td>compute.pod</td>
<td>2</td>
</tr>
<tr>
<td>coord</td>
<td>3</td>
</tr>
<tr>
<td>gbm</td>
<td>4</td>
</tr>
<tr>
<td>glp</td>
<td>4</td>
</tr>
<tr>
<td>iCluster</td>
<td>5</td>
</tr>
<tr>
<td>iCluster2</td>
<td>6</td>
</tr>
<tr>
<td>plotHeatmap</td>
<td>7</td>
</tr>
<tr>
<td>plotiCluster</td>
<td>9</td>
</tr>
<tr>
<td>plotRI</td>
<td>10</td>
</tr>
<tr>
<td>simu.datasets</td>
<td>11</td>
</tr>
<tr>
<td>tune.iCluster2</td>
<td>11</td>
</tr>
</tbody>
</table>

Index 14
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
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

References

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>

References


See Also

`iCluster`, `iCluster2`, `plotICluster`

Examples

```r
# library(iCluster)
# 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)
```

coord

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

References


---

coord

<table>
<thead>
<tr>
<th>coord</th>
<th>genomic coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.
**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 55 samples.

**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:

expZ : Relaxed cluster indicator matrix.

W : Coefficient matrix.

clusters : Cluster assignment.

conv.rate : Convergence history.

Author(s)

Ronglai Shen <shenr@mskcc.org>

References


See Also

breast.chr17.plotiCluster, compute.pod
iCluster2

A variant of the iCluster method with variance weighted shrinkage

Description

iCluster function with variance-weighted shrinkage (see Shen et al. PLoS ONE, 2012)

Usage

icluster2(datasets, k, lambda=NULL, scale=T, scalar=F, max.iter=10, verbose=T)

Arguments

datasets: A list containing data matrices. For each data matrix, the rows represent samples, and the columns represent genomic features.
k: Number of classes for the samples.
lambda: Penalty term for the coefficient matrix of the iCluster model.
scalar: Logical value. If true, a degenerate version assuming scalar covariance matrix is used.
max.iter: maximum iteration for the EM algorithm
scale: Logical value. If true, data matrix is column centered
verbose: Logical value. If true, print message.

Value

A list with the following elements.

expZ: Latent variable matrix
\( \hat{\psi} \): The iCluster model coefficient matrix
\( \hat{\Sigma} \): The estimated covariance matrix
clusters: Cluster indicator for samples

Author(s)

Ronglai Shen <shenr@mskcc.org>
References


See Also

tune.iCluster2, plot.iCluster, compute.pod, plotHeatmap

Examples

```r
library(iCluster)
library(caTools, lib.loc="/apps/Rlib64/")
library(gdata, lib.loc="/apps/Rlib64/"
library(gtools, lib.loc="/apps/Rlib64/")
library(gplots, lib.loc="/apps/Rlib64/"
library(lattice, lib.loc="/apps/Rlib64/")
data(gbm)

#setting the penalty parameter lambda=0 returns non-sparse fit
#fit=iCluster2(datasets=gbm, k=3, lambda=list(0.44,0.33,0.28))

#plot.iCluster(fit=fit, label=rownames(gbm[[1]]))

#compute.pod(fit)

#data(coord)
#chr=coord[,1]
#plotHeatmap(fit=fit, data=gbm, feature.order=c(FALSE,TRUE,TRUE),
#sparse=c(FALSE,TRUE,TRUE),plot.chr=c(TRUE,FALSE,FALSE), chr=chr)
```

---

**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, sample.order= NULL, feature.order= NULL,
width= 5, scale= NULL, col.scheme= NULL, sparse= NULL, threshold= NULL,
chr= NULL, plot.chr= NULL, cap= NULL)
```
Arguments

\texttt{fit} \hspace{1cm} \text{A iCluster object}
\texttt{datasets} \hspace{1cm} \text{A list object of data matrices}
\texttt{feature.order} \hspace{1cm} \text{A vector of logical values each specify whether the genomic features in the corresponding data matrix should be reordered by similarity. Default is FALSE.}
\texttt{sparse} \hspace{1cm} \text{A vector of logical values each specify whether to plot the top cluster-discriminant features. Default is FALSE.}
\texttt{threshold} \hspace{1cm} \text{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.}
\texttt{plot.chr} \hspace{1cm} \text{A vector of logical values each specify whether to annotate chromosome number on the left of the panel. Typically used for copy number data type. Default is FALSE.}
\texttt{chr} \hspace{1cm} \text{A vector of chromosome number.}
\texttt{col.scheme} \hspace{1cm} \text{Color scheme. Can use bluered(n) in gplots R package.}
\texttt{sample.order} \hspace{1cm} \text{User supplied cluster assignment.}
\texttt{width} \hspace{1cm} \text{Width of the figure in inches}
\texttt{cap} \hspace{1cm} \text{Image color option}
\texttt{scale} \hspace{1cm} \text{A vector of logical values each specify whether data should be scaled. Default is FALSE.}

Value

no value returned.

Author(s)

Ronglai Shen <shenr@mskcc.org>

References


See Also

\texttt{iCluster}, \texttt{iCluster2}
Examples

```r
#library(iCluster)
data(gbm)
data(coord)
chr=chr[,1]
fit=iCluster2(datasets=gbm, k=3, lambda=list(0.44,0.33,0.28))
plotHeatmap(fit=fit, datasets=datasets, feature.order=c(FALSE,TRUE,TRUE),
sparse=c(FALSE,TRUE,TRUE),plot.chr=c(TRUE,FALSE,FALSE), chr=chr)
```

Description

A function to generate cluster separability matrix plot.

Usage

```r
plotiCluster(fit,label=NULL)
```

Arguments

- `fit`: A iCluster object
- `label`: Sample labels

Value

no value returned.

Author(s)

Ronglai Shen <shenr@mskcc.org>

References


See Also

`icluster`, `computeNPod`
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=allist()
#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)
```

Description

Simulated dataset consists of n=150 samples that fall into three clusters and a total of 200 feature.

Usage

data(simu.datasets)

Format

A list object of two data matrices each is of dimension 150 by 200.

tune.iCluster2 Model tuning function

Description

Model tuning process for choosing the number of clusters k and the lasso penalty parameters.

Usage

tune.iCluster2(datasets, k, n.lambda,nrep, mc.cores,max.iter)

Arguments

datasets A list containing data matrices. For each data matrix, the rows represent samples, and the columns represent genomic features.
k Number of classes for the samples.
nrep Number of training and test data partition for computing the reproducibility index.
n.lambda The number of sampled points for the uniform design. Use the default value by setting n.lambda=NULL.
mc.cores Number of cores to use for parallel computation.
max.iter Number of EM iterations.
Value

A list with the following elements.

- **best.fit**  
  Model fit under the optimal lambda values that give the highest reproducibility index.

- **RI**  
  A vector of reproducibility index associated with each of the sampled lambda combination.

- **ud**  
  Sampled lambda combinations under the uniform design

Author(s)

Ronglai Shen <shenr@mskcc.org>

References


See Also

- **iCluster2.plotiCluster, compute.pod, plotHeatmap**

Examples

```r
library(iCluster)
library(caTools, lib.loc="/apps/Rlib64/")
library(gdata, lib.loc="/apps/Rlib64/")
library(gtools, lib.loc="/apps/Rlib64/")
library(gplots, lib.loc="/apps/Rlib64/")
library(lattice, lib.loc="/apps/Rlib64/")
library(parallel, lib.loc="/apps/Rlib64/")

#data(simu.datasets)
#cv.fit=alist()
#for(k in 2:5){
#  cat(paste("K="",k,sep=""),'\n')
#  cv.fit[[k]]=tune.iCluster2(simu.datasets, k, mc.cores=6)
#}

##Reproducibility index (RI) plot
#plotRI(cv.fit)

##Based on the RI plot, k=3 is the best solution
#best.fit=cv.fit[[3]]$best.fit
```
## Try different color schemes
plotHeatmap(fit=best.fit, datasets=simu.datasets,
# sparse=c(TRUE, TRUE), col.scheme=list(bluered(256), greenred(256)))
Index

*Topic Data integration, subtype discovery, latent variable model
  iCluster2, 6
  tune.iCluster2, 11

*Topic datasets
  breast.chr17, 2
  coord, 3
  gbm, 4
  glp, 4
  simu.datasets, 11

*Topic models
  compute.pod, 2
  iCluster, 5
  plotHeatmap, 7
  plotiCluster, 9
  plotRI, 10

breast.chr17, 2, 5
compute.pod, 2, 5, 7, 9, 12
coord, 3

gbm, 4

glp, 4

iCluster, 3, 5, 8, 9
iCluster2, 3, 6, 8, 12

plotHeatmap, 7, 7, 12
plotiCluster, 3, 5, 7, 9, 12
plotRI, 10

simu.datasets, 11

tune.iCluster2, 7, 10, 11