Package ‘EMMIXgene’

March 23, 2020

Type Package
Version 0.1.3
Title A Mixture Model-Based Approach to the Clustering of Microarray Expression Data
Description Provides unsupervised selection and clustering of microarray data using mixture models. Following the methods described in McLachlan, Bean and Peel (2002) <doi:10.1093/bioinformatics/18.3.413> a subset of genes are selected based one the likelihood ratio statistic for the test of one versus two components when fitting mixtures of t-distributions to the expression data for each gene. The dimensionality of this gene subset is further reduced through the use of mixtures of factor analyzers, allowing the tissue samples to be clustered by fitting mixtures of normal distributions.
Encoding UTF-8
Author Andrew Thomas Jones
Maintainer Andrew Thomas Jones <andrewthomasjones@gmail.com>
License GPL (>= 3)
LazyData TRUE
Suggests R.rsp
LinkingTo Rcpp, RcppArmadillo, BH
Depends R (>= 3.3.0)
Imports Rcpp (>= 0.12.5), stats, mclust, reshape, ggplot2, scales, tools
SystemRequirements C++11
RoxygenNote 6.1.1
VignetteBuilder R.rsp
NeedsCompilation yes
Repository CRAN
Date/Publication 2020-03-23 15:50:14 UTC
all_cluster_tissues

Clusters tissues using all group means

Description

Clusters tissues using all group means

Usage

all_cluster_tissues(gen, clusters, q = 6, G = 2)

Arguments

- **gen**: EMMIXgene object
- **clusters**: mclust object
- **q**: number of factors if using mfa
- **G**: number of components if using mfa

Value

A clustering for each sample (columns) by each group (rows)

Examples

```r
example <- plot_single_gene(alon_data, 1)
# only run on first 100 genes for speed
alon_sel <- select_genes(alon_data[seq_len(100), 1])
alon_clust <- cluster_genes(alon_sel, 2)
alon_tissue_all <- all_cluster_tissues(alon_sel, alon_clust, q=1, G=2)
```
Normalized gene expression values from Alon et al. (1999).

Description

Usage
data(alon_data)

Format
A data frame with 2000 rows (genes) and 62 variables (samples).

Examples
dim(alon_data)

cluster_genes
Clustering genes using mixtures of normal distributions

Description
Sorts genes into clusters using mixtures of normal distributions with covariance matrices restricted to be multiples of the identity matrix.

Usage
cluster_genes(gen, g = NULL)

Arguments
- **gen**: an EMMIXgene object produced by select_genes().
- **g**: The desired number of gene clusters. If not specified will be selected automatically on the basis of BIC.

Value
An array containing the clustering.
cluster_tissues

Examples

#only run on first 100 genes for speed
alon_sel <- select_genes(alon_data[seq_len(100), ])
alon_clust<- cluster_genes(alon_sel , 2)

cluster_tissues(gen, clusters, method = "t", q = 6, G = 2)

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>gen</td>
<td>EMMIXgene object</td>
</tr>
<tr>
<td>clusters</td>
<td>mclust object</td>
</tr>
<tr>
<td>method</td>
<td>Method for separating tissue classes. Can be either ‘t’ for a univariate mixture of t-distributions on gene cluster means, or ‘mfa’ for a mixture of factor analyzers.</td>
</tr>
<tr>
<td>q</td>
<td>number of factors if using mfa</td>
</tr>
<tr>
<td>G</td>
<td>number of components if using mfa</td>
</tr>
</tbody>
</table>

Value

a clustering for each sample (columns) by each group(rows)

Examples

#only run on first 100 genes for speed
alon_sel <- select_genes(alon_data[seq_len(100), ])
alon_clust<- cluster_genes(alon_sel,2)
alon_tissue_t<-
  cluster_tissues(alon_sel,alon_clust,method='t')
alon_tissue_mfa<-
  cluster_tissues(alon_sel, alon_clust,method='mfa',q=2,G=2)
Description


Functions

select_genes: Selects the most differentially expressed genes.
cluster_genes: Clusters the genes using a mixture model approach.
cluster_tissues: Clusters the tissues based on the differences between the tissue samples among the gene groups.

See vignette('The-EMMIXgene-Workflow') for more details.

golub_data

Normalized gene expression values from Golub et al. (1999).

Description


Usage

data(golub_data)

Format

A data frame with 3731 rows (genes) and 72 variables (samples). # @examples dim(golub_data)
heat_maps

Description
Plot heat maps of gene expression data. Optionally sort the x-axis according to a predetermined clustering.

Usage
heat_maps(dat, clustering = NULL, y_lab = NULL)

Arguments
- dat: matrix of gene expression data.
- clustering: a vector of sample classifications. Must be same length as the number of columns in dat.
- y_lab: optional label for y-axis.

Value
A ggplot2 heat map.

Examples
example <- heat_maps(alon_data[seq_len(100), ])

plot_single_gene
Plot a single gene expression histogram with best fitted mixture of t-distributions.

Description
Plot a single gene expression histogram with best fitted mixture of t-distributions according to the EMMIX-gene algorithm.

Usage
plot_single_gene(dat, gene_id, g = NULL, random_starts = 8,
max_it = 100, ll_thresh = 8, min_clust_size = 8, tol = 1e-04,
start_method = "both", three = TRUE, min = -4, max = 2)
**select_genes**

**Description**

Selects genes using the EMMIXgene algorithm.

**Arguments**

- **dat**: matrix of gene expression data.
- **gene_id**: row number of gene to be plotted.
- **g**: force number of components, default = NULL
- **random_starts**: The number of random initializations used per gene when fitting mixtures of t-distributions. Initialization uses k-means by default.
- **max_it**: The maximum number of iterations per mixture fit. Default value is 100.
- **ll_thresh**: The difference in -2 log lambda used as a threshold for selecting between g=1 and g=2 for each gene. Default value is 8, which was chosen arbitrarily in the original paper.
- **min_clust_size**: The minimum number of observations per cluster used when fitting mixtures of t-distributions for each gene. Default value is 8.
- **tol**: Tolerance value used for detecting convergence of EMMIX fits.
- **start_method**: Default value is "both". Can also choose "random" for purely random starts.
- **three**: Also test g=2 vs g=3 where appropriate. Defaults to TRUE.
- **min, max**: Minimum and maximum x-axis values for the plot window.

**Examples**

```r
example <- plot_single_gene(alon_data,1)
#plot(example)
```

**Value**

A ggplot2 histogram with fitted t-distributions overlayed.

**Usage**

```r
select_genes(dat, filename, random_starts = 4, max_it = 100, 
ll_thresh = 8, min_clust_size = 8, tol = 1e-04, 
start_method = "both", three = FALSE)
```
select_genes

Arguments

dat  A matrix or dataframe containing gene expression data. Rows are genes and columns are samples. Must supply one of filename and dat.

filename  Name of file containing gene data. Can be either .csv or space separated .dat. Rows are genes and columns are samples. Must supply one of filename and dat.

random_starts  The number of random initializations used per gene when fitting mixtures of t-distributions. Initialization uses k-means by default.

max_it  The maximum number of iterations per mixture fit. Default value is 100.

ll_thresh  The difference in -2 log lambda used as a threshold for selecting between g=1 and g=2 for each gene. Default value is 8, which was chosen arbitrarily in the original paper.

min_clust_size  The minimum number of observations per cluster used when fitting mixtures of t-distributions for each gene. Default value is 8.

tol  Tolerance value used for detecting convergence of EMMIX fits.

start_method  Default value is "both". Can also choose "random" for purely random starts.

three  Also test g=2 vs g=3 where appropriate. Defaults to FALSE.

Value

An EMMIXgene object containing:

stat  The difference in log-likelihood for g=1 and g=2 for each gene (or for g=2 and g=3 where relevant).

g  The selected number of components for each gene.

it  The number of iterations for each genes selected fit.

selected  An indicator for each genes selected status

ranks  selected gene ids ranked by stat

genes  A dataframe of selected genes.

all_genes  Returns dat or contents of filename.

Examples

#only run on first 100 genes for speed
alon_sel <- select_genes(alon_data[seq_len(100), ])
Description
Cluster tissues

Usage

```
top_genes_cluster_tissues(gen, n_top = 100, method = "mfa", q = 2, g = 2)
```

Arguments

- **gen**: An EMMIXgene object produced by `select_genes()`.
- **n_top**: number of top genes (as ranked by likelihood) to be selected.
- **method**: Method for separating tissue classes. Can be either 't' for a univariate mixture of t-distributions on gene cluster means, or 'mfa' for a mixture of factor analysers.
- **q**: number of factors if using mfa
- **g**: number of components if using mfa

Value
An EMMIXgene object containing:

- **stat**: A matrix containing clustering (0 or 1) for each sample (columns) by each group(rows).
- **top_gene**: The row numbers of the top genes.
- **fit**: The fit object used to determine the clustering.

Examples

```
alon_sel <- select_genes(alon_data[seq_len(100), ])
alon_top_10 <- top_genes_cluster_tissues(alon_sel, 10, method='mfa', q=3, g=2)
```
Index

*Topic datasets
  alon_data, 3
  golub_data, 5

all_cluster_tissues, 2
alon_data, 3

cluster_genes, 3, 5
cluster_tissues, 4, 5

EMMIXgene, 5
EMMIXgene-package (EMMIXgene), 5

golub_data, 5
heat_maps, 6
plot_single_gene, 6
select_genes, 5, 7

top_genes_cluster_tissues, 9