Package 'sigclust'

February 20, 2015

Title Statistical Significance of Clustering

Version 1.1.0

Author Hanwen Huang, Yufeng Liu & J. S. Marron

Description SigClust is a statistical method for testing the significance of clustering results. SigClust can be applied to assess the statistical significance of splitting a data set into two clusters. For more than two clusters, SigClust can be used iteratively.

Depends R (>= 2.4.0), methods

Maintainer Hanwen Huang <hanwenh.unc@gmail.com>

License GPL (>= 2)

Repository CRAN

Date/Publication 2014-01-23 12:56:21

NeedsCompilation no

R topics documented:

`plot-methods` *SigClust plot*

Description

Diagnostics and p-value plots from a sigclust object.

Usage

```
## S4 method for signature 'sigclust,missing'
plot(x,y,arg="all",...)
```

Arguments

x	An object of class sigclust.
y	not used
arg	Type of the individual plot: "background": make background standard deviation diagnostic plots. These plots contain the raw data points as well as the corresponding density plots using kernel and robust Gaussian fits; "qq": the QQ plot assessing the quality of robust fit of a Gaussian distribution; "diag": make a null distribution covariance estimation diagnostic plot; "pvalue": make a clustering significance pvalue plot; "all": make all above plots (default).
...	further arguments for plot.

Details

SigClust diagnostic plots are suggested to monitor the performance of the SigClust method for a given dataset.

Author(s)

Hanwen Huang: <hanwenh@email.unc.edu>; Yufeng Liu: <yfliu@email.unc.edu>; J. S. Marron: <marron@email.unc.edu>

References

Liu, Yufeng, Hayes, David Neil, Nobel, Andrew and Marron, J. S, 2008, *Statistical Significance of Clustering for High-Dimension, Low-Sample Size Data*, *Journal of the American Statistical Association* **103**(483) 1281–1293. See also the vignette included with this package.

See Also

sigclust.

Examples

```
## Simulate a dataset from a collection of mixtures of two
## multivariate Gaussian distributions with different means.

mu <- 5
n <- 30
p <- 500
dat <- matrix(rnorm(p*2*n),2*n,p)
dat[1:n,1] <- dat[1:n,1]+mu
dat[(n+1):(2*n),1] <- dat[(n+1):(2*n),1]-mu
```

```
nsim <- 1000
nrep <- 1
icovest <- 3
pvalue <- sigclust(dat,nsim=nsim,nrep=nrep,labflag=0,icovest=icovest)
#sigclust plot
plot(pvalue)
```

---

sigclust *Statistical Significance of Clustering*

Description

Perform a significance analysis of clustering. SigClust studies whether clusters are really there, using the 2-means (k = 2) clustering index as a statistic. It assesses the significance of clustering by simulation from a single null Gaussian distribution. Null Gaussian parameters are estimated from the data.

Usage

```
sigclust(x, nsim, nrep=1, labflag=0, label=0, icovest=1)
```

Arguments

x	A matrix or data.frame of expression data; each row corresponds to a sample and each column to a variable. Data may be properly normalized and may not contain missing values.
nsim	Number of simulated Gaussian samples to estimate the distribution of the clustering index for the main p-value computation.
nrep	Number of steps to use in 2-means clustering computations (default=1, chosen to optimize speed). This has no effect, unless labflag=0.
labflag	An indicator variable specifying if the p-values is for an assigned cluster or for using 2-means; for user assigned clusters labflag=1, otherwise labflag=0.
label	If labflag=0, SigClust uses labels generated by 2-means clustering. If labflag=1, label needs to be set as a numeric, integer vector of 1s and 2s with length nrow(x) which indicates given cluster labels (grouping to be tested for significance).
icovest	Covariance estimation type: 1. Use a soft threshold method as constrained MLE (default); 2. Use sample covariance estimate (recommended when diagnostics fail); 3. Use original background noise thresholded estimate (from Liu, et al, (2008)) ("hard thresholding").

Details

The SigClust method addresses the problem of assessing statistical significance of clustering as a testing procedure. The null hypothesis of `SigClust` is that the data are from a single Gaussian distribution. The signicance of a given clustering is judged by calculating an appropriate p-value. The SigClust method uses a test statistic called the cluster index (CI) which is defined to be the sum of within-class sums of squares about the mean divided by the total sum of squares about the overall mean. The null distribution of the CI can be approximated by simulating from a single Gaussian distribution, fit to the data. Because CI is mean shift invariant, it is enough to take the mean to be 0. Because CI is rotation invariant, we take the covariance to be diagonal. There are three options for estimating the eigenvalues of the covariance matrix: 1. Soft Thresholding (recommended for high dimensions, when the diagnostics indicate assumptions are met). 2. Sample eigenvalues (recommended for low dimensions, and when assumptions, such as Gaussianity fail, but known to be generally conservative). 3. Hard Thresholding.

Value

The function returns an object of class `sigclust`. See help for `sigclust-class` for more details.

Author(s)

Hanwen Huang: <hanwenh@email.unc.edu>; Yufeng Liu: <yfliu@email.unc.edu>; J. S. Marron: <marron@email.unc.edu>

References

Liu, Yufeng, Hayes, David Neil, Nobel, Andrew and Marron, J. S, 2008, *Statistical Significance of Clustering for High-Dimension, Low-Sample Size Data*, *Journal of the American Statistical Association* **103**(483) 1281–1293.

See Also

`plot-methods`.

Examples

```
## Simulate a dataset from a collection of mixtures of two
## multivariate Gaussian distribution with different means.

mu <- 5
n <- 30
p <- 500
dat <- matrix(rnorm(p*2*n),2*n,p)
dat[1:n,1] <- dat[1:n,1]+mu
dat[(n+1):(2*n),1] <- dat[(n+1):(2*n),1]-mu

nsim <- 1000
nrep <- 1
icovest <- 3
pvalue <- sigclust(dat,nsim=nsim,nrep=nrep,labflag=0,icovest=icovest)
#sigclust plot
```

```
plot(pvalue)
```

`sigclust-class` *Class sigclust*

Description

The class sigclust is the output from the function `sigclust`. It is also the input to the plot function `plot-methods`.

Slots

`raw.data`: raw data matrix.

`veigval`: vector of sample eigen values.

`vsimeigval`: vector of eigen values used in simulation

`simbackvar`: background variance fit from the data.

`icovest`: covariance estimation type.

`nsim`: number of simulated Gaussian samples.

`simcindex`: vector of cluster indices based on nsim simulated data sets.

`pval`: simulated sigclust p-value based on empiriccal quantiles.

`pvalnorm`: simulated sigclust p-value based on Gaussian quantiles.

`xcindex`: cluster index based on given data set.

Author(s)

Hanwen Huang: <hanwenh@email.unc.edu>; Yufeng Liu: <yfliu@email.unc.edu>; J. S. Marron: <marron@email.unc.edu>

See Also

`sigclust`, `plot-methods`.

Index