Package 'MethylCapSig'

August 11, 2015

Title Detection of differentially methylated regions using MethylCap-seq data

Version 1.0.0

Date 2015-08-11

Author Deepak N. Ayyala, David E. Frankhouser, Javkhlan-Ochir Ganbat, Guido Marcucci, Ralf Bundschuh, Pearlly Yan and Shili Lin.

Maintainer Deepak N. Ayyala <ayyala.1@osu.edu>

Description Provides a univariate and several high dimensional multivariate test statistics for detecting differentially methylated regions based on MethylCap-seq data.

Depends R (>= 3.0.0)

Imports geepack

LazyLoad YES

License LGPL-3

NeedsCompilation no

R topics documented:

`MethylCapSig-package` *Detection of differentially methylated regions using MethylCap-seq data.*

Description

The **MethylCapSig** package provides several test statistics useful in detecting differential methylation in genomic regions. While all the functions are illustrated using differential methylation as example, the tests are much generic and are applicable to a wide range of high dimensional problems.

Details

High dimensional data collected on small sample sizes cannot be analyzed using traditional multivariate statistical techniques owing to the *curse of dimensionality*. One such type of data is nucleotide-resolution methylation values obtained from MethylCap-seq experiments. To overcome the small sample issue in two sample mean vector testing problem, several test statistics have been developed by studying the asymptotic properties of functions of the random variables being considered.

MethylCapSig provides five such test statistics to test equality of mean vectors in the two-sample case under high dimensional setting. The four multivariate tests and one univariate test all provide test statistics and p-values based on asymptotic distributions.

Author(s)

Deepak N. Ayyala, David E. Frankhouser, Javkhlan-Ochir Ganbat, Guido Marcucci, Ralf Bundschuh, Pearlly Yan and Shili Lin.

References

Ayyala, D. N., et al. (2015) Statistical methods for detecting differentially methylated regions based on MethylCap-seq data, Manuscript.

---

cqtest *Chen-Qin test statistic*

Description

Calculates the two sample Chen-Qin test statistic and p-value.

Usage

```
cqtest(X, Y)
```

Arguments

X	A matrix of dimension $n \times k$ whose rows represent the samples collected from n (≥ 3) individuals from the first group on k variates.
Y	A matrix of dimension $m \times k$ whose rows correspond to samples collected from m (≥ 3) individuals from the second group on k variates. Default value is null. If not specified, the function performs a one-sample test using X.

Details

The Chen-Qin test statistic is used to test equality of mean vectors for two groups of multivariate observations, where the dimension is greater than the sample size. `cqtest` takes matrices `X` and `Y` as arguments whose rows correspond to samples from the two groups respectively. Depending on the values in `X` and `Y`, the function initially determines whether to perform a one-sample test ($\sum_{i,j} X_{ij}^2 = 0$ or $\sum_{i,j} Y_{ij}^2 = 0$) or a two-sample test. The appropriate test statistic is then calculated and is returned along with the p-value which is calculated using right-tailed normal distribution.

Note: The Chen-Qin test involves calculations on the data which require at least three samples in both the groups to evaluate the test statistic. See Chen and Qin (2010) for further details.

Value

A 2×1 vector consisting of the test statistic and the p-value.

Author(s)

Deepak N. Ayyala, Javkhlan-Ochir Ganbat.

References

Chen, S. X. and Qin, Y. (2010) A two-sample test for high-dimensional data with applications in gene-set testing, Annals of Statistics, 38, 808 – 835.

Examples

```
data(diffMethylData)
cqtest(diffMethylData$region1.x, diffMethylData$region1.y)
# cqtest(diffMethylData$region2.x, diffMethylData$region2.y)
```

---

diffMethylData *Randomly generated nucleotide-resolution methylation signal data*

Description

Nucleotide resolution methylation-signal data for two groups of samples. The signals are randomly generated and to mimic acute myeloid cancer data set studied by Frankhouser et al. (2014). Signals are reported for two regions - region1 with 92 *CpG* sites and region2 with 122 *CpG* sites. While region1 is known to be non-differentially methylated, region2 is differentially methylated. Sample sizes for the two groups are 20 and 10 respectively.

Usage

```
data(diffMethylData)
```

Format

A data frame with signal matrices for two groups recorded on two regions.

`region1.x` a 20×92 matrix

`region1.y` a 10×92 matrix

`region1.x` a 20×122 matrix

`region1.y` a 10×122 matrix

References

Frankhouser, D. E., et al. (2014) PrEMeR-CG: inferring nucleotide leve DNA methylation values from MethylCap-seq data, Bioinformatics, 30 (24), 3567 – 3574.

Ayyala, D. N., et al. (2015) Statistical methods for detecting differentially methylated regions based on MethylCap-Seq data, Manuscript.

`methmage` *MethMAGE test*

Description

Calculates a generalized estimating equation (GEE) based test statistic as used in MethMAGE package.

Usage

```
methmage(X, Y)
```

Arguments

`X`	A matrix of dimension $n \times k$ whose rows represent the samples collected from n individuals from the first group on k variates.
`Y`	A matrix of dimension $m \times k$ whose rows correspond to samples collected from m individuals from the second group on k variates.

Details

`methmage` uses a generalized estimating equations (GEE) approach to test for equality of mean vectors for two groups of multivariate observations. Using a first order autoregressive (AR(1)) structure as the working correlation matrix, `methmage` uses `geeglm` function from the **geepack** package to estimate the coefficients and construct the test statistic. To ensure convergence in modest time, maximum number of iterations and convergence criterion (epsilon) are set at 100 and 10^{-8} respectively.

Value

A 2×1 vector consisting of the test statistic and the p-value.

Author(s)

Deepak N. Ayyala, David E. Frankhouser

References

Frankhouser, D. E., et al. (2014) PrEMeR-CG: inferring nucleotide level DNA methylation values from MethylCap-seq data, Bioinformatics, 30 (24), 3567 – 3574.

Examples

```
data(diffMethylData)
methmage(diffMethylData$region1.x, diffMethylData$region1.y)
# methmage(diffMethylData$region2.x, diffMethylData$region2.y)
```

`mvlognormal` *Multivariate lognormal random variable generator.*

Description

Given mean (Mu), variances (Sigma) and correlation structure (R) of the distribution, `mvlognormal` generates multivariate lognormal random variables.

Usage

```
mvlognormal(n, Mu, Sigma, R)
```

Arguments

`n`	Sample size (default value is 1).
`Mu`	Mean vector of length k.
`Sigma`	Vector of length k containing the diagonal of covariances.
`R`	A $k \times k$ matrix comprising the correlation structure of the variables on the log-scale, i.e. $R = cor(log(X))$.

Details

The multivariate lognormal distribution is characterized by its associated normal distribution on the log-scale - if X is lognormal, then $log(X)$ is normal. `mvlognormal` uses this relationship to generate lognormal random variables. Specifying the correlation structure of the actual variable does not guarantee validity of the associated normal distribution. Hence, the function takes correlation matrix of the log-transformed normal variable to ensure existence.

Value

Matrix of dimension $n \times k$, where k is the length of the mean vector.

Author(s)

Deepak N. Ayyala

Examples

```
## Generate 10 samples with dimension 20.
X <- mvlognormal(n = 10, Mu = runif(20, 0, 1),
      Sigma = rep(2, 20), R = toeplitz(0.5^(0:19)));
```

`patest` *Park-Ayyala test statistic*

Description

Calculates the two sample Park-Ayyala test statistic and p-value.

Usage

```
patest(X, Y)
```

Arguments

`X`	A matrix of dimension $n \times k$ whose rows represent the samples collected from n (≥ 4) individuals from the first group on k variates.
`Y`	A matrix of dimension $m \times k$ whose rows correspond to samples collected from m (≥ 4) individuals from the second group on k variates. Default value is null. If not specified, the function performs a one-sample test using X.

Details

The Park-Ayyala test statistic is used to test equality of mean vectors for two groups of multivariate observations, where the dimension is greater than the sample size. `patest` takes matrices `X` and `Y` as arguments whose rows represent samples from two groups respectively. Depending on the values in `X` and `Y`, the function initially determines whether to perform a one sample test ($\sum_{i,j} X_{i,j}^2 = 0$ or $\sum_{i,j} Y_{i,j}^2 = 0$) or a two-sample test. The appropriate test statistic is then calculated and is returned along with the p-value which is calculated using right-tailed normal distribution.

Note: The Park-Ayyala test statistic involves repeated computation of the covariance matrix, requiring at least four samples in both the groups. See Park and Ayyala (2013) for more details.

Value

A 2×1 vector consisting of the test statistic and the p-value.

Author(s)

Deepak N. Ayyala, Javkhlan-Ochir Ganbat.

References

Park, J. and Ayyala, D. N. (2013) A test for the mean vector in large dimension and small samples, Journal of Statistical Planning and Inference, 143, 929 – 943.

Examples

```
data(diffMethylData)
patest(diffMethylData$region1.x, diffMethylData$region1.y)
# patest(diffMethylData$region2.x, diffMethylData$region2.y)
```

skktest *Srivastava-Katayama-Kano test statistic*

Description

Calculates the two sample test statistic and p-value for the Srivastava-Katayama-Kano test.

Usage

```
skktest(X, Y)
```

Arguments

X	A matrix of dimension $n \times k$ whose rows represent the samples collected from n individuals from the first group on k variates.
Y	A matrix of dimension $m \times k$ whose rows correspond to samples collected from m individuals from the second group on k variates. Default value is null. If not specified, the function performs a one-sample test using X.

Details

The Srivastava-Katayama-Kano test statistic is used to test equality of mean vectors for two groups of multivariate observations, where the dimension is greater than the sample size. `skktest` takes matrices `X` and `Y` as arguments whose rows represent samples from two groups respectively. Depending on the values in `X` and `Y`, the function initially determines whether to perform a one sample test ($\sum_{i,j} X_{i,j}^2 = 0$ or $\sum_{i,j} Y_{i,j}^2 = 0$) or a two-sample test. The appropriate test statistic is then calculated and is returned along with the p-value which is calculated using right-tailed normal distribution.

Value

A 2×1 vector consisting of the test statistic and the p-value.

Author(s)

Deepak N. Ayyala

References

Srivastava, M. S., Katayama, S. and Kano, Y. (2013) A two sample test in high dimensional data, Journal of Multivariate Analysis, 114, 349 – 358.

Examples

```
data(diffMethylData)
skktest(diffMethylData$region1.x, diffMethylData$region1.y)
# skktest(diffMethylData$region2.x, diffMethylData$region2.y)
```

ttest *One sample and two sample t-test using total methylation signal per region*

Description

Performs a two-sample t-test using the total signal observed across all variates in multivariate data.

Usage

```
ttest(X, Y)
```

Arguments

X	A matrix of dimension $n \times k$ whose rows represent the samples collected from n individuals from the first group on k variates.
Y	A matrix of dimension $m \times k$ whose rows correspond to samples collected from m individuals from the second group on k variates. Default value is NULL. If not specified, the function performs a one-sample test using X.

Details

Given multivariate observations collected from two groups, a straightforward univariate test for equality of mean vectors can be constructed by converting the multivariate observations into univariate measures. `ttest` tests equality of means by converting the $k \times 1$ observations into sum, representing each sample by total measurement observed across the variates. Using this total measure as our observed samples, a two-sample t-test is performed. If both groups contain all zero observations, ($\sum_{i,j} X_{i,j}^2 = 0$ and $\sum_{i,j} Y_{i,j}^2 = 0$), then the test statistic is set equal to 0 and a p-value of 1 is returned.

Value

A 2×1 vector consisting of the test statistic and the p-value.

Author(s)

Deepak N. Ayyala

Examples

```
data(diffMethylData)
ttest(diffMethylData$region1.x, diffMethylData$region1.y)
# ttest(diffMethylData$region2.x, diffMethylData$region2.y)
```

Index