Package ‘dasev’

June 17, 2019

Type Package

Title Differential abundance analysis and cluster analysis with empirical Bayes shrinkage estimation of variance for zero-inflated data

Version 0.1.0

Author Zhengyan Huang
Chi Wang, PhD

Maintainer Zhengyan Huang <`Zhengyan.Huang@uky.edu`>

Description Perform differential abundance analysis and cluster analysis for zero-inflated data. A inverse-gamma distribution is applied as the prior distribution of the variance to more robustly estimate the variances.

Depends R (>= 3.1)

License GPL (>= 2)

Encoding UTF-8

LazyData true

RoxygenNote 6.1.1

NeedsCompilation no

R topics documented:

`DASEV` *Differential abundance analysis for zero-inflated data.*

Description

This function processes the two group differential abundance (DA) analysis for zero-inflated data. It has the flexibility to let users to choose various detection limits (DL). The function returns estimates on group means, group biological point mass value (BPMV) proportions, standard deviation, and p-values.

Usage

```
DASEV(indata, cov.matrix, test_cov, test_cov_conti = NULL,
  min.non0n = 3, requiredn = 10, requiredn2 = 30,
  DL_method = "Fixed Difference", DL_value = 0.1, maxit_MLE = 100,
  maxit = 10000, test_model = c("Both", "Mean", "Pzero"))
```

Arguments

`indata`	Specifies the input data matrix. Rows are features and columns are subjects.
`cov.matrix`	A matrix for variables. Categorical data should be represented by dummy variables contain only 0 and 1.
`test_cov`	A vector corresponding to the coloums in `cov.matrix` which specify the variables to be tested.
`test_cov_conti`	A vector corresponding to the coloums in `cov.matrix` which specify the continuous variables.
`min.non0n`	The minimum number of nonzero observations for features to be included in the analysis. Default and the minimum value is 3. Please input a number equal or larger than 3.
`requiredn`	The minimum number of nonzero obs while getting prior distribution of variance. Default value is 10.
`requiredn2`	The minimum number of features while getting prior distribution of variance. Default value is 30. If requiredn is specified, it will retrun features used in getting prior distribution of variance. If the number of features used is less than requiredn2, the function returns top requiredn2 features with the smallest all point mass value (PMV) proportions.
`DL_method`	Specifies the detection limit method. The options are: Fixed Difference (default): for each feature, DL is the minimun value for all nonzero observation minus a number sepcified by DL_value Fixed Rate: for each feature, DL is the minimun value for all nonzero observation devided by a number sepcified by DL_value Fixed Value: DL is the same value (a number sepcified by DL_value) for all features.
`DL_value`	Custom specified number. Default is 0.1 for "Fixed Difference".
`maxit_MLE`	The maximum number of iterations while re-estimating the model parameters. The default is 100.
`maxit`	The maximum number of iterations applied to the optimization function.
`test_model`	A vector of models to be tested. The default value is c("Both","Mean","Pzero"). The default is to achieve all three comparsions listed as follow: If test_model contains "Both", the function compares both difference in group means and BPMV proportions. If test_model contains "Mean", the function compares difference in group means. If test_model contains "Pzero", the function compares difference in BPMV proportions.

Details

The optimization function used here is `optim` with the method option as `"BFGS"`.

If using "Fixed Difference" as the method to get detection limit, a small positive value is recommended for `DL_value`.
If using "Fixed Rate" as the method to get detection limit, a positive value larger than 1 is recommended for `DL_value`.
If using "Fixed Value" as the method to get detection limit, a value smaller than the log value of the minimum observation in the dataset is recommended for `DL_value`.

This function requires at least one zero and one non-zero observations in each group.

Value

`feature_names`	A vector of feature names substracted from the input data.
`pvalue_both`	A vector of estimated p-values for comparing both difference in group means and BPMV proportions. pvalue_both will be NULL if test_model doesn't contain "Both".
`pvalue_mean`	A vector of estimated p-values for comparing only difference in group means. pvalue_mean will be NULL if test_model doesn't contain "Mean".
`pvalue_zero`	A vector of estimated p-values for comparing only difference in group BPMV proportions.pvalue_zero will be NULL if test_model doesn't contain "Pzero".
`DL`	A vector of detection limits used.
`estimates`	A matrix of estimates on optimization parameters, which can be used to calculate group means, group BPMV proportions, and standard deviation.

See Also

`Getsample`

Examples

```
#Get simulation samples#
data(simpool)
sim <- Getsample(numc=1000,numobs=100,pdiff=0.2,lfc=c(log(2),-log(2)),pzerodiff=NULL, simpool)

data<- sim$simdata
paradata <- sim$Parameters
indall <- c()
for(i in 1:nrow(data)){
  ind <- (!all(data[i,][1:100]!=0)
          &!all(data[i,][1:100]==0)
          &!all(data[i,][101:200]!=0)
          &!all(data[i,][101:200]==0))
  indall<- c(indall, ind)
}
data_used <- data[indall,]

example_result <- DASEV(indata=data_used,cov.matrix=cbind(rep(1,200),c(rep(0,100),rep(1,100))),
test_cov= 2,test_cov_conti=NULL, min.non0n=3, requiredn=10, requiredn2=30,
DL_method= "Fixed Difference",DL_value=0.1,maxit_MLE=100,maxit=1000,
test_model=c("Both","Mean","Pzero"))
```

`Getsample` *Get simulated zero-infated dataset.*

Description

This function returns a simulated dataset contains two groups with key parameters used in the simulation.

Usage

```
Getsample(numc = 1000, numobs = 100, pdiff = 0.2, lfc = NULL,
  pzerodiff = NULL, simpool)
```

Arguments

`numc`	Specifies the number of features. Defaults to 1000.
`numobs`	Specifies the number of observations per group. Defaults to 100.
`pdiff`	Specifies the percentage of differentially abundent features. Defaults to 0.2.
`lfc`	The log fold change value as the difference in group means. Default to NULL. The value can be a single number or a numeric vector. For example lfc=c(log(2),-log(2)).
`pzerodiff`	The ratio for odds of BPMVs in two groups. Defaults to NULL. The value can be a single number or a numeric vector. For example pzerodiff=c(0.5,2). pzerodiff equals to (p1/(1-p1))/(p0(/(1-p0))).
`simpool`	A data frame contains detection limit, mean, BPMV proportion, and standard deviation. Simulation parameters are randomly sampled from this data frame.

Details

This function assigns same number of observations for both groups.

Value

`simdata`	A data matrix for the simulated dataset.
`Parameters`	A data matrix contains parameters used in the simulation. DL is the detection limit. mu0 is the group mean for group 0. mu1 is the group mean for group 1. p0 is the BPMV proportion for group 0. p1 is the BPMV proportion for group 1. sd is the standard deviation.

Examples

```
#Get simulation samples#
data(simpool)
simdata1 <- Getsample(numc=1000,numobs=100,pdiff=0.2,lfc=c(log(2),-log(2)),pzerodiff=NULL, simpool)
simdata2 <- Getsample(numc=1000,numobs=100,pdiff=0.2,lfc=NULL,pzerodiff=c(0.5,2), simpool)
```

`simpool` *Parameters data matrix for simulation example*

Description

This data matrix contains four variables. DL is the detection limit, mu is the group mean, p0 is the BPMV proportion, and sd is the standard deviation.

Usage

```
data(simpool)
```

Format

An object of class `matrix` with 5253 rows and 4 columns.

Examples

```
data(simpool)
```

Index