Package 'fdrDiscreteNull'

October 13, 2022

Type Package

Title False Discovery Rate Procedures Under Discrete and Heterogeneous Null Distributions

Version 1.4

Date 2020-05-01

Author Xiongzhi Chen and Rebecca W. Doerge <rwdoerge@andrew.cmu.edu>

Maintainer Xiongzhi Chen <xiongzhi.chen@wsu.edu>

Description It is known that current false discovery rate (FDR) procedures can be very conservative when applied to multiple testing in the discrete paradigm where p-values (and test statistics) have discrete and heterogeneous null distributions. This package implements more powerful weighted or adaptive FDR procedures for FDR control and estimation in the discrete paradigm. The package takes in the original data set rather than just the p-values in order to carry out the adjustments for discreteness and heterogeneity of p-value distributions. The package implements methods for two types of test statistics and their p-values: (a) binomial test on if two independent Poisson distributions have the same means, (b) Fisher's exact test on if the conditional distribution is the same as the marginal distribution for two binomial distributions, or on if two independent binomial distributions have the same probabilities of success.

Imports MCMCpack, qvalue

Depends R(>= 3.2.0)

URL http://math.wsu.edu/faculty/xchen/welcome.php

License LGPL

Encoding UTF-8

NeedsCompilation no

Repository CRAN

Date/Publication 2020-04-25 00:40:03 UTC

R topics documented:

`BHPlusTwoSide` *FDR control for multiple testing based on p-values with cadlag distribution functions.*

Description

Implement the BH+ procedures of Chen, X. (2019) for FDR control for multiple testing based on p-values whose distributions are cadlag, i.e., right-continuous with left-limits. This includes conventional p-values and mid p-values. Currently, the methods are implemented for two-sided p-values.

Usage

```
BHPlusTwoSide(data=NULL, Test=c("Binomial Test", "Fisher's Exact Test"),
                          FET_via = c("PulledMarginals","IndividualMarginals"),
                                  FDRlevel=NULL, espilon=NULL)
```

Arguments

`data`	Data to be analyzed in the form of a matrix for which observations for a single entity are in a row. Format of data will be checked by this function automatically and the functions stops execution if the format is wrong.
`Test`	The type of test to be conducted. It should be exactly one entry from the string c("Binomial Test", "Fisher's Exact Test"). Currently no other type of test is supported by the package.
`FET_via`	When the type of test is the Fisher's exact test, how the marginal counts are formed should be specified to be exactly one entry from the string c("PulledMarginals", "IndividualMarginals"). When "PulledMarginals" is used, the data matrix should have only two clumns, each row of which contains the observed counts for the two binomial distributions, whereas when "IndividualMarginals" is used the data matrix should have four columns, each row of which has the first and third entries as the observed count and total number of trials of one binomial distribution, and the second and fourth entries as the observed count and total number of trials of the other binomial distribution. For other types of test, this argument need not to be specified.
`FDRlevel`	The nominal false discovery rate (FDR) no larger than which the method to be applied is to have.
`espilon`	A scalar used to determine the guiding value for the estimator of the proportion of true null hypotheses. It is usually set to be 0.01.

Value

It returns the following lists:

`BH`	Restuls obtained by the Benjamini-Hochberg (BH) procedure when applied to conventional p-values.
`BHplus`	Results obtained by the BH+ procedure when applied to conventional p-values.
`aBHplus`	Results obtained by the adaptive BH+ procedure when applied to conventional p-values.
`MidpBHplus`	Results obtained by the BH+ procedure when applied to mid p-values.
`aMidpBHplus`	Results obtained by the adaptive BH+ procedure when applied to mid p-values.

Each of the above contains:

`pi0Est`	The estimated proprtion of true nulls, where for non-adaptive procedure, it is set to be 1.
`Threshold`	The threshold below which p-values and their associated hypotheses are rejected.
`IndicesOfDiscoveries`	The row indices of the data matrix whose corresponding hypotheses are rejected.

For an adaptive procedure, each of the above lists contains:

`Tuning`	The guiding value for the estimator of the proportion.
`TuningVon`	If TuningVon=0, the guiding value is chosen by theory, meaning that the adaptive procedure is conservative; if TuningVon=1, the guiding value is chosen approximately, meaning that the adaptive procedure may not be conservative. A user may get the "Warning message: In max(DevCount) : no non-missing arguments to max; returning -Inf", which can be safely ignored.

It also returns the following:

`pval`	Vector of conventional p-values, whose indices matches those of the hypotheses.
`pvalSupp`	It is a list, each of whose element is the support of a conventional p-value. The indice of pvalSupp match those of pval.
`Midpvals`	Vector of mid p-values, whose indices matches those of the hypotheses.
`pvalMidSupp`	It is a list, each of whose element is the support of a mid p-value. The indice of pvalSupp match those of Midpvals.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. Ser. B 57(1): 289-300.

Hwang, J. T. G. and Yang, M.-C. (2001). An optimality theory for mid pcvalues in 2 x 2 contingency tables. Statistica Sinica 11(3): 807-826.

Chen, X. (2019). False discovery rate control for multiple testing based on discrete p-values. https://arxiv.org/abs/1803.06040; Biometrical Journal (in press).

Gilbert, P. B. (2005). A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics, J. R. Statist. Soc. Ser. C 54(1): 143-158.

Examples

```
library(fdrDiscreteNull)
data(hivdata)
m = dim(hivdata)[1]
hivdata = cbind(hivdata,rep(73,m),rep(73,m))
hivdataA = hivdata[rowSums(hivdata[,1:2])>=5,]
ResTmp = BHPlusTwoSide(data = hivdataA, Test = "Fisher's Exact Test",
 FET_via = "IndividualMarginals",FDRlevel = 0.05)
```

`GeneralizedEstimatorsGrouped` *Weighted multiple hypothesis testing under discrete and heterogeneous null distributions.*

Description

Implement weighted multiple testing procedure of Chen, X., Doerge, R. and Sanat, S. K. (2019) for independent p-values whose null distributions are super-uniform but not necessarily identical or continuous, where groups are formed by the infinity norm for functions, p-values weighted by data-adaptive weights, and multiple testing conducted. The weights are not constructed using plug-in null proportion estimators. For multiple testing based on p-values of Binomial tests or Fisher's exact tests, grouping using quantiles of observed counts is recommended both for fast implementation and excellent power performance of the weighted FDR procedure.

Usage

```
GeneralizedEstimatorsGrouped(data_in = NULL,
 grpby = c("quantileOfRowTotal","kmeans","InfNorm"),ngrp_in = NULL,
 GroupMergeSize = 50,test_in = NULL,FET_via_in = NULL,OneSide_in = NULL,
 FDRlevel_in = NULL, eNetSize = NULL, unif_tol= 10^-3,
 TuningPar = c(0.5,100),lambda = 0.5)
```

Arguments

`data_in`	Data to be analyzed in the form of a matrix for which observations for a single entity are in a row. Format of data will be checked by this function automatically and the functions stops execution if the format is wrong.
`grpby`	The method to be used to form the groups. It should be exactly one entry from the string c("quantileOfRowTotal","kmeans","InfNorm"). Grouping by "quantileOfRowTotal" is a good choice as demonstrated by simulation stuides and it is very fast.
`ngrp_in`	The number of groups to be formed from the orginal data. It refers to the number of groups that the rows of the data matrix will be formed, and also to the number of groups that the discrete null distributions and their associated p-values will be formed.
`GroupMergeSize`	When the grouping method is "InfNorm", the default minimal group size is "GroupMergeSize", which is 50 by default.

`test_in` The type of test to be conducted. It should be exactly one entry from the string c("Binomial Test", "Fisher's Exact Test"). Currently no other type of test is supported by the package.

`FET_via_in` When the type of test is the Fisher's exact test, how the marginal counts are formed should be specified to be exactly one entry from the string 'c("PulledMarginals", "IndividualMarginals")'. When "PulledMarginals" is used, the data matrix should have only two clumns, each row of which contains the observed counts for the two binomial distributions, whereas when "IndividualMarginals" is used the data matrix should have four columns, each row of which has the first and third entries as the observed count and total number of trials of one binomial distribution, and the second and fourth entries as the observed count and total number of trials of the other binomial distribution. For other types of test, this argument need not to be specified.

`OneSide_in` Specify if one-sided p-value is to be computed from the test. If "OneSide_in= NULL", then two-sided p-value will be computed; if 'OneSide_in="Left"', then the p-value is computed using the left tail of the CDF of the test statistics; if 'OneSide_in="Right" ', then the p-value is computed using the right tail of the CDF of the test statistics.

`FDRlevel_in` The nominal false discovery rate (FDR) no larger than which the method to be applied is to have.

`eNetSize` The argument is needed only when the argument "InfNorm" is used. It specifies the size of the metric balls to be used to partition the set of discrete cdf's to form the groups.

`unif_tol` The argument is needed only when the argument "InfNorm" is used. It specifies the tolerance on the infinity norm under which a discrete cdf of a p-value will be considered approximately uniform on [0,1]. By default, it is set to be 0.001.

`TuningPar` A vector of 2 scalars (a,b), used to implement the generalized proportion estimator. Let rho be the maximum of the minimum of each support whose minimum is smaller than 1. If rho is smaller 0.5, then the smallest guiding value is set as a times (0.5-rho) and the biggest guiding value as 0.5, and b determines the number of equally spaced guiding values. If rho is at least 0.5, then all guiding values are set to be rho and b=1.

`lambda` A scalor in (0,1) that is used as a tuning parameter to construct data-adaptive weights. By default, it is set to be 0.5.

Value

It returns estimated proportion of true nulls:

`pi0estAll` Estimated proportion of true nulls.

The above quantity is a vector and contains the following:

`pi0E_GE` Estimated proportion of true nulls, obtained by the generalized estimator.

`pi0E_gGE` Estimated proportion of true nulls, obtained by grouping and weighting and the generalized estimator.

`pi0Est_gp*` Estimated proportion of true nulls for each group by the generalized estimator, where * is a group number.

It returns the results on multiple testing that are returned by GeneralizedFDREstimators, plus the following list:

wFDR — Results from the weighted false discovery rate procedure; these results are stored using the same list structure as multiple testing results returned by GeneralizedFDREstimators.

References

Chen, X., Doerge, R. and Sanat, S. K. (2020). A weighted FDR procedure under discrete and heterogeneous null distributions. https://arxiv.org/abs/1502.00973v5; Biometrical Journal (in press).

Lister, R., O'Malley, R., Tonti-Filippini, J., Gregory, B. D., Berry, Charles C. Millar, A. H. and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in arabidopsis. Cell 133(3): 523-536.

See Also

GeneralizedFDREstimators

Examples

```
library(fdrDiscreteNull)
library(qvalue)
data(listerdata)
ResTmp = GeneralizedEstimatorsGrouped(listerdata[1:500,],
  grpby= "quantileOfRowTotal", ngrp_in= 3, GroupMergeSize = 50,
  test_in= "Fisher's Exact Test", FET_via_in = "PulledMarginals",OneSide_in = NULL,
  FDRlevel= 0.05,TuningPar = c(0.5,20),lambda = 0.5)
```

GeneralizedFDREstimators

Adaptive false discovery rate procedure using generalized estimator of the null proportion.

Description

Implement false discovery rate procedures of Chen, X., Doerge, R. and Heyse, J. F. (2018), the Adaptive Benjamini-Hochberg procedure, and the Adaptive Benjamini-Hochberg-Heyse procedure, using the generalized estimator of the proportion of true nulls, for discrete p-values distributions.

Usage

```
GeneralizedFDREstimators(data=NULL,
 Test=c("Binomial Test", "Fisher's Exact Test"),
 FET_via = c("PulledMarginals","IndividualMarginals"),
 OneSide = NULL,FDRlevel=NULL,TuningRange = c(0.5,100))
```

Arguments

data	Data to be analyzed in the form of a matrix for which observations for a single entity are in a row. Format of data will be checked by this function automatically and the functions stops execution if the format is wrong.
Test	The type of test to be conducted. It should be exactly one entry from the string c("Binomial Test", "Fisher's Exact Test"). Currently no other type of test is supported by the package.
FET_via	When the type of test is the Fisher's exact test, how the marginal counts are formed should be specified to be exactly one entry from the string c("PulledMarginals", "IndividualMarginals"). When "PulledMarginals" is used, the data matrix should have only two clumns, each row of which contains the observed counts for the two binomial distributions, whereas when "IndividualMarginals" is used the data matrix should have four columns, each row of which has the first and third entries as the observed count and total number of trials of one binomial distribution, and the second and fourth entries as the observed count and total number of trials of the other binomial distribution. For other types of test, this argument need not to be specified.
OneSide	Specify if one-sided p-value is to be computed from the test. If "OneSide= NULL", then two-sided p-value will be computed; if 'OneSide="Left"', then the p-value is computed using the left tail of the CDF of the test statistics; if 'OneSide="Right" ', then the p-value is computed using the right tail of the CDF of the test statistics.
FDRlevel	The nominal false discovery rate (FDR) no larger than which the method to be applied is to have.
TuningRange	A vector of 2 scalars (a,b). Let rho be the maximum of the minimum of each support whose minimum is smaller than 1. If rho is smaller 0.5, then the smallest guiding value is set as a times (0.5-rho) and the biggest guiding value as 0.5, and b determines the number of equally spaced guiding values. If rho is at least 0.5, then all guiding values are set to be rho and b=1.

Value

It returns the following lists:

BH	Restuls obtained by the Benjamini-Hochberg (BH) procedure.
aBH	Results obtained by the adaptive BH procedure using the generalized estimator of the proportion.
BHH	Results obtained by the Benjamini-Hochberg-Heyse (BHH) procedure.
aBHH	Results obtained by the adaptive BHH (aBHH) procedure using the generalized estimator of the proportion.

Each of the above contains:

pi0Est	The estimated proprtion of true nulls, where for the BH procedure, it is set to be 1.
Threshold	The threshold below which p-values and their associated hypotheses are rejected.

`NumberOfDiscoveries`
: The number of rejections.

`IndicesOfDiscoveries`
: The row indices of the data matrix for the rejections.

It also returns the following:

`pvalues`
: Vector of p-values of the individual tests without grouping.

`pvalSupp`
: It is a list. For binomial test, each entry of the list is a vector, whose first element is the mean of the p-value under the null, second element the p-value itself, and the rest the values at the support of the discrete cdf of the p-value without grouping; for Fisher's exact test, the structure of the list is the same except that in the vector the element denoting the p-value itself is removed.

Finally, it also returns randomized p-values (as "RndPval") and results (as "SARP") of the procedure in Habiger (2015) that is exactly the procedure of Storey et al. (2004) applied to the randomized p-values, and mid p-values (as "MidPval") and "aBHmidP" as the adaptive BH procedure of Benjamini and Hochberg (1995) applied to these mid p-values together with the estimated proportion of true null hypotheses obtained by Storey's estimator in Storey et al. (2004) applied to these mid p-values.

References

Chen, X., Doerge, R. and Heyse, J. F. (2018). Multiple testing with discrete data: proportion of true null hypotheses and two adaptive FDR procedures. Biometrial Journal 60(4): 761-779.

Habiger, J. D. (2015). Multiple test functions and adjusted p-values for test statistics with discrete distributions. J. Stat. Plan. Inference 167: 1-13.

Heyse, J. F. (2011). A false discovery rate procedure for categorical data, in M. Bhattacharjee, S. K. Dhar and S. Subramanian (eds), Recent Advances in Biostatistics: False Discovery Rates, Survival Analysis, and Related Topics, chapter 3.

Lister, R., O'Malley, R., Tonti-Filippini, J., Gregory, B. D., Berry, Charles C. Millar, A. H. and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in arabidopsis. Cell 133(3): 523-536.

See Also

`GeneralizedEstimatorsGrouped`

Examples

```
library(qvalue)
library(fdrDiscreteNull)
data(listerdata)
ResG = GeneralizedFDREstimators(listerdata[1:100,],
  Test= "Fisher's Exact Test", FET_via = "PulledMarginals",
  OneSide = NULL,FDRlevel=0.05,TuningRange = c(0.5,20))
```

hivdata *HIV data*

Description

This data set has been analyzed and provided by the listed reference. There are 118 positions, each of which is under two types of HIV. Each type has 73 subjects.

References

Gilbert, P. B. (2005). A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics, J. R. Statist. Soc. Ser. C 54(1): 143-158.

listerdata *Methylation data for Arabidopsis thaliana*

Description

This data set has been analyzed and provided by the listed reference. There are around 22000 cytosines, each of which is under two conditions. For each cytosine under each condition, there is only one replicate. The discrete count for each replicate can be modelled by binomial distribution, and Fisher'e exact test can be applied to assess if a cytosine is differentially methylated. The filtered data "listerdata.RData" contains cytosines whose total counts for both lines are greater than 5 and whose count for each line does not exceed 25.

References

Lister, R., O'Malley, R., Tonti-Filippini, J., Gregory, B. D., Berry, Charles C. Millar, A. H. and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in arabidopsis, Cell 133(3): 523-536.

Index