Package ‘fastcmh’

October 13, 2022

Title Significant Interval Discovery with Categorical Covariates

Version 0.2.7

Author Felipe Llinares Lopez, Dean Bodenham

Maintainer Dean Bodenham <deanbodenhambsse@gmail.com>

Description A method which uses the Cochran-Mantel-Haenszel test with significant pattern mining to detect intervals in binary genotype data which are significantly associated with a particular phenotype, while accounting for categorical covariates.

Depends R (>= 3.3.0), bindata

License GPL-2 | GPL-3

LinkingTo Rcpp

Imports Rcpp

Suggests testthat

RoxygenNote 5.0.1

SystemRequirements C++11

NeedsCompilation yes

Repository CRAN

Date/Publication 2016-09-13 21:19:17

R topics documented:

demofastcmh 2
makefastcmhdata 3
runfastcmh 4

Index **8**

demofastcmh *Demo of fastcmh*

Description

This function runs a demo for fastcmh, by first creating a sample data set and then running fastcmh on this data set.

Usage

```
demofastcmh(saveToFolder = FALSE, folder = NULL)
```

Arguments

`saveToFolder`	A flag indicating whether or not the data files created for the demo should be saved to file. The default is `FALSE`, i.e. no files are saved to the folder. The only reason to save demo data to a folder is for the user to be able to have a look at the files after the demo.
`folder`	The folder in which the data for the demo will be saved. Default is the current directory, `"./"`. The demo data will created in `folder/data` and the results will be saved in `folder/results` as an RData file.

Details

This function will first create a sample data set in `folder/data`, and will then run `runfastcmh` on this data set, before saving the each step showing the R code that can be used to do the step, then running that R code, and then waiting for the user to press enter before moving onto the next step. If `saveToFolder=FALSE`, (default) then no files are saved and all the results are kept in memory.

See Also

`runfastcmh`

Examples

```
demofastcmh()
```

makefastcmhdata *Create sample data for fastcmh*

Description

This function creates sample data for use with the runfastcmh method.

Usage

```
makefastcmhdata(folder = "./", xfilename = "data.txt",
  yfilename = "label.txt", covfilename = "cov.txt", K = 2, L = 1000,
  n = 200, noiseP = 0.3, corruptP = 0.05, rho = 0.8, tau1 = 100,
  taulength1 = 4, tau2 = 200, taulength2 = 4, seednum = 2,
  truetaufilename = "truetau.txt", showOutput = FALSE, saveToList = FALSE)
```

Arguments

folder	The folder in which the data will be saved. Default is current directory "./".
xfilename	The name of the data file. Default is "data.txt"
yfilename	The name of the label file. Default is "label.txt"
covfilename	The name of the file containing the covariate categories . This file actually just contains K numbers, where K is the number of covariates. Default is "cov.txt"
K	The number of covariates (a positive integer). Default is K=2.
L	The number of features (length of each sequence). Default is L=1000.
n	The number of samples (cases and controls combined). Default is n=200, i.e. 100 cases and 100 controls.
noiseP	The background noise in the data (as a probability of 0/1 being flipped). Default is noiseP=0.3
corruptP	The probability of data corruption: each bit has probability corruptP of being flipped. Default is corruptP=0.05.
rho	The strength of the confounding in the confounded interval (as a probability). Default is rho=0.8 (i.e. a very strong signal).
tau1	The location of the significant interval (starting point). Default value is tau1=100.
taulength1	The length of the significant interval. Default value is taulength1=4, so default significant interval is [100, 103].
tau2	The location of the confounded significant interval (starting point). Default value is tau2=200.
taulength2	The length of the confounded significant interval. Default value is taulength2=4, so default significant interval is [200, 203].
seednum	The seed used for generating the data. Default value is seednum=2.
truetaufilename	The file where the location of the true significant intervals are saved (as opposed to the detected significant intervals). Default is "truetau.txt".

showOutput — Flag to decide whether or not to show output, where files are created, their names, etc. Default is FALSE, so will save to folder by default. However, all of the examples use saveToList=TRUE in order to avoid writing to file. The list will consist of data, label and cov data frames, when saveToList=TRUE.

saveToList — Flag to decide whether or not to save data to the folder, or to return (output) the data as a list. By default, saveToList=FALSE.

See Also

runfastcmh

Examples

```
#make a small sample data set, using the default parameters
mylist <- makefastcmhdata(showOutput=TRUE, saveToList=TRUE)

#make a very small sample data set
mylist <- makefastcmhdata(n=20, L=10, tau1=2, taulength1=2,
       tau2=6, taulength2=2, saveToList=TRUE)
```

runfastcmh *Run the fastcmh algorithm*

Description

This function runs the FastCMH algorithm on a particular data set.

Usage

```
runfastcmh(folder = NULL, data = NULL, label = NULL, cov = NULL,
  alpha = 0.05, Lmax = 0, showProcessing = FALSE, saveAllPvals = FALSE,
  doFDR = FALSE, useDependenceFDR = FALSE, saveToFile = FALSE,
  saveFilename = "fastcmhresults.RData", saveFolder = NULL)
```

Arguments

folder — The folder in which the data is saved. If the any of data, label and pvalue arguments are not specified, then filenames must have following a naming convention inside the folder: the data file is "data.txt" (i.e. the full path is "folder/data.txt"), the phenotype label file is label.txt, and covariate label file is cov.txt. More details on the structure of these files is given below, or the user can use the makefastcmhdata function to see an example of the correct data formats. If folder="/data/", the data in fastcmh/inst/extdata is used.

data — The filename for the data file. Default is NULL. The data file must be an L x n txt file containing only 0s and 1s, which are space-separated in each row, while each row is on a separate newline.

`label` The filename for the phenotype label file. Default is `NULL`. The label file should consist of a single column (i.e. each row is on a separate line) of `0`s and `1`s.

`cov` The filename for the covariate label file. Default is `NULL`. The `cov` file contains a single column of positive integers. The first row, containing value n_1, specifies that the first n_1 columns have covariate value `1`; the second row, containing n_2, specifies that the next n_2 rows have covariate value `2`, etc.

`alpha` The value of the FWER; must be a number between 0 and 1. Default is `alpha=0.05`.

`Lmax` The maximum length of significant intervals which is considered. Must be a non-negative integer. For example, `Lmax=10` searches for significant intervals up to length 10. Setting `Lmax=0` will search for significant intervals up to any length (with algorithm pruning appropriately). Default is `Lmax=0`.

`showProcessing` A flag which will turn printing to screen on/off. Default is `FALSE` (which is “off”).

`saveAllPvals` A flag which controls whether or not all the intervals (less than minimum attainable pvalue) will be returned. Default is `FALSE` (which is “no, do notreturn all intervals”).

`doFDR` A flag which controls whether or not Gilbert’s Tarone FDR procedure (while accounting for positive regression dependence) is performed. Default is `FALSE` (which is “no, do not do FDR”).

`useDependenceFDR`

A flag which controls whether or not Gilbert’s Tarone FDR procedure uses the dependent formulation by Benjamini and Yekutieli (2001), which further adjusts alpha by dividing by the harmonic mean. This flag is only used if `doFDR==TRUE`. Default is `FALSE`.

`saveToFile` A flag which controls whether or not the results are saved to file. By default, `saveToFile=FALSE`, and the data frame is returned in R. See the examples below.

`saveFilename` A string which gives the filename to which the output is saved (needs to have `saveToFile=TRUE`) as an RData file. Default is `"fastcmhresults.RData"`.

`saveFolder` A string which gives the path to which the output will be saved (needs to have `saveToFile=TRUE`). Default is `"./"`.

Details

This function runs the FastCMH algorithm on a particular data set in order to discover intervals that are statistically significantly associated with a particular label, while accounting for categorical covariates.

The user must either supply the folder, which contains files named `"data.txt"`, `"label.txt"` and `"cov.txt"`, or the non-default filenames must be specified individually. See the descriptions of arguments `data`, `label` and `cov` to see the format of the input files, or make a small sample data file using the `makefastcmhdata` function. By default, filtered results are provided. The user also has the option of using an FDR procedure rather than the standard FWER-preserving procedure.

Value

`runfastcmh` will return a list if `saveToFile=FALSE` (default setting), otherwise it will save the list in an .RData file. The fields of the list are:

sig a dataframe listing the significant intervals, after filterting. Columns start, end and pvalue indicate the start and end points of the interval (inclusive), and the *p*-value for that interval.

unfiltered a dataframe listing all the significant intervals before filtering. The filtering compares the overlapping intervals and returns the interval with the smallest p-value in each cluster of overlapping intervals. Dataframe has has structure as sig.

fdr (if doFDR==TRUE) significant intervals using Gilbert's FDR-Tarone procedure, after filtering. Dataframe has same structure as sig.

unfilteredFdr (if doFDR==TRUE) a dataframe listing all the significant intervals before filtering. See description of unfiltered.

allTestablle (if saveAllPvals==TRUE) a dataframe listing all the testable intervals, many of which will not be significant. Dataframe has same structure as sig.

histObs Together with histFreq gives a histogram of maximum attainable CMH statistics.

histFreq Histogram of maximum attainable CMH statistics (only reliable in the testable range).

summary a character string summarising the results. Use cat(...$summary) to print the results with the correct indentation/new lines.

timing a list containing (i) details, a character string summarising the runtime values for the experiment - use cat(...$timing$details) for correct indentation, etc. (ii) exec, the total execution time. (iii) init, the time to initialise the objects. (iv) fileIO, the time to read the input files. (v) compSigThresh, the time to compute the significance threshold. (vi) compSigInt, the time to compute the significant intervals.

Author(s)

Felipe Llinares Lopez, Dean Bodenham

See Also

makefastcmhdata

References

Gilbert, P. B. (2005) *A modified false discovery rate multipl-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics*. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(1), 143-158.

Benjamini, Y., Yekutieli, D. (2001). *The control of the false discovery rate in multiple testing under dependency*. Annals of Statistics, 29(4), 1165-1188.

Examples

```
#Example with default naming convention used for data, label and cov files
# Note: using "/data/" as the argument for folder
#       accesses the data/ directory in the fastcmh package folder
mylist <- runfastcmh("/data/")

#Example where the progress will be shown
mylist <- runfastcmh(folder="/data/", showProcessing=TRUE)
```

```
#Example where many parameters are specified
mylist <- runfastcmh(folder="/data/", data="data2.txt", alpha=0.01, Lmax=7)

#Example where Gilbert's Tarone-FDR procedure is used
mylist <- runfastcmh("/data/", doFDR=TRUE)

#Example where FDR procedure takes some dependence structures into account
mylist <- runfastcmh("/data/", doFDR=TRUE, useDependenceFDR=TRUE)
```

Index

demofastcmh, 2

makefastcmhdata, 3, *4–6*

runfastcmh, *2*, *4*, 4