Package ‘repfdr’

April 6, 2015

Type Package
Title Replicability Analysis for Multiple Studies of High Dimension
Version 1.1-3
Description Estimation of Bayes and local Bayes false discovery rates for replicability analysis.
License GPL (>= 2)
Depends R (>= 2.10)
Imports splines
NeedsCompilation yes
Repository CRAN
URL http://github.com/shay-y/repfdr
BugReports http://github.com/shay-y/repfdr/issues
Author Ruth Heller [cre, aut],
      Shachar Kaufman [aut],
      Shay Yaacoby [aut],
      Daniel Yekutieli [aut]
Maintainer Ruth Heller <ruheller@gmail.com>
Date/Publication 2015-04-06 07:51:22

R topics documented:

  binned_zmat ........................................ 2
  binned_zmat_sim .................................. 2
  em.control ....................................... 3
  hconfigs ......................................... 4
  hmat_sim ......................................... 5
  ldr ................................................ 6
  piem ............................................... 7
  repfdr ............................................ 9
  SNPlocations .................................... 12
  zmat_sim ......................................... 13
  ztobins ........................................... 14
**binned_zmat_sim**

**Description**

This data was created from the zmat_sim matrix using ztobins function. It contains two objects to be input to the main function repfdr.

**Usage**

```r
data(binned_zmat_sim)
```

**Format**

The file includes two objects - a matrix and 3d array:

bz_sim is a matrix of binned 10000 z-scores (in rows) in each of the 3 studies (columns).
pbz_sim is a 3-dimensional array which contains for each study (first dimension), the probabilities of a z-score to fall in the bin (second dimension), under each hypothesis status (third dimension).

**Examples**

```r
data(binned_zmat_sim)
bz[1:5,]
pbz[,1:5,]
```
**em.control**

**Examples**

```r
data(binned_zmat_sim)
bz_sim[1:5,]
pbz_sim[1:5,]
```

**em.control**  
*Control Parameters for the EM algorithm*

**Description**

Input parameters for the EM algorithm.

**Usage**

```r
em.control(pi.initial = NULL, max.iter = 10000, tol = 1e-12,
            nr.threads = 0, verbose = TRUE)
```

**Arguments**

- `pi.initial`: Initial guess for the probabilities of the vectors of associations status. If NULL then 0.9 is assigned for the c(0,...,0) configuration and 0.1 is distributed uniformly for all other configurations.
- `max.iter`: Maximum number of EM iterations.
- `tol`: Tolerance (in maximum absolute difference between two EM iterations in estimated probabilities) before declaring convergence and stopping.
- `nr.threads`: Number of processing threads to use. If zero (the default), will automatically detect the number of compute cores available and spawn one thread per core.
- `verbose`: An indicator of whether to report progress (running iteration number) during computation.

**Details**

The function is used inside the control argument in *repfdr* and *piem*.

**Value**

A list with the input values.

**See Also**

*repfdr* *piem*
## Examples

```r
## Not run:
data(binned_zmat)
out <- repfdr(pbz, bz, "replication",
            control = em.control(pi.initia = c(0.48, rep(0.02, 26)),
                        verbose = TRUE, nr.threads = 1))
# iterations are printed; run bit slower (1 thread)

## End(Not run)
```

---

### hconfigs

**Enumeration of all possible vectors of association status.**

### Description

The function generates a matrix with all possible vectors of association status (in rows), given the number of studies and number of possible association status states in each study (2 or 3).

### Usage

```r
hconfigs(n.studies, n.association.status = 3, studies.names = NULL)
```

#### Arguments

- `n.studies` Number of studies in the analysis.
- `n.association.status` either 2 for no-association\association or 3 for no-association\negative-association\positive-association.
- `studies.names` Optional study names to display.

### Details

This matrix should be used when selecting the rows indices for the association status vectors that are in the non-null set, specified by the used in `non.null.rows` in the function `repfdr`.

### Value

Matrix with rows indicating all the possible vectors of association status.

### See Also

`repfdr`
**hmat_sim**

**Examples**

\[(H \leftarrow \text{hconfigs(n.studies = 3)})\]

# in replication analysis the non-null vectors are:
\[H[\text{apply}(H,1,\text{function}(y)\{(\text{sum}(y==1)>1 \mid \text{sum}(y==-1)>1)\}),,]\]

# in meta-analysis there is only one null vector \(c(0,0,0)):
\[H[\text{rowSums}(\text{abs}(H))]!=0,]\]

\[\text{hconfigs(n.studies = 3, n.association.status= 2)}\]

---

**hmat_sim**

*Simulated data set - indicators of association status matrix*

**Description**

A matrix of size 10000x3 of indicators of whether each z-score from `zmat_sim` belongs to a non-null hypothesis for the feature in the study (1) or to a null hypothesis for the feature in the study (0).

**Usage**

`data(hmat_sim)`

**Format**

`hmat_sim` is a matrix of 10000 rows, each row a vector of the true association status from which the z-scores in the same row in `zmat_sim` was generated. Specifically, for a zero entry in `hmat_sim` the corresponding z-score in `zmat_sim` was generated from the standard normal distribution, and for a unit entry in `hmat_sim` the corresponding z-score in `zmat_sim` was generated from the normal distribution with mean 3 and variance one.

**Examples**

#### use `hmat_sim` to generate the simulated z-scores:

```r
data(hmat_sim)
m <- nrow(hmat_sim)
set.seed(12)
zm\_sim1 <- matrix(rnorm(n=3\*m, mean=hmat\_sim\*3), nrow=m, ncol=3)
rm(m, H)
data(zmat_sim)
stopifnot(all.equal(zmat\_sim1, zmat\_sim))
```

#### `hmat_sim` was generated by the following code:

```r
H <- hconfigs(n.studies= 3, n.association.status=2)
f <- c(0.895,0.005,0.005,0.02,0.005,0.02,0.02,0.02,0.03) \# frequencies for the association status vectors
m = 10000 \# number of tests in each study
```
Estimation of posterior probabilities for the vectors of association status

Description

The function finds the posterior probabilities of each vector of association status for each feature, given the feature’s vector of binned z-scores.

Usage

```r
ldr(pdf.binned.z, binned.z.mat, Pi, h.vecs = NULL)
```

Arguments

- `pdf.binned.z`: Same input as in `repfdr`. A 3-dimensional array which contains for each study (first dimension), the probability of a z-score to fall in the bin (second dimension), under each hypothesis status (third dimension). The third dimension can be of size 2 or 3, depending on the number of association states: if the association can be either null or only in one direction, the dimension is 2; if the association can be either null, or positive, or negative, the dimension is 3. Element `[[1]]` in the output of `ztobins`.
- `binned.z.mat`: Same input as in `repfdr`. A matrix of the bin numbers for each of the z-scores (rows) in each study (columns). Element `[[2]]` in the output of `ztobins`.
- `Pi`: The estimated prior probabilities for each association status vector. Can be extracted from the output of `repfdr` or `piem`, see Example section.
- `h.vecs`: The row indices in `H` (see `hconfigs`), corresponding to the association status vectors. By default the posterior probabilities of all possible vectors of association status are computed.

Details

A subset of features (e.g. most significant) can be specified as the rows in `binned.z.mat`, so the posterior probabilities of the vectors of association status are computed for this subset of features. See Example section.

Value

Matrix with rows that contain for each of the vectors of association status the posterior probabilities. The columns are the different feature.
piem

See Also
   repfdr, piem, hconfigs

Examples

```r
## Not run:
data(binned_zmat)
data(Pi)

# Fdr calculation:
output3 <- repfdr(pbz, bz, "replication", Pi.previous.result = Pi)
BayesFdr <- output3$mat[,"Fdr"]
sum(BayesFdr <= 0.05)

# The posterior probabilities for the first five features with Bayes FDR at most 0.05:
post <- ldr(pbz,bz[which(BayesFdr <= 0.05)[1:5],],Pi)
round(post,4)

# posteriors for a subset of the association status vectors can also be reported,
# here the subset is the four first association status vectors:
post <- ldr(pbz,bz[which(BayesFdr <= 0.05)[1:5],],Pi,h.vecs=1:4)
round(post,4)

## End(Not run)
```

---

### Description

The function calls an expectation-maximization (EM) algorithm to estimate the prior probabilities of each association status vector. It is also used internally in repfdr.

### Usage

```r
piem(pdf.binned.z, binned.z.mat, control = em.control())
```

### Arguments

- **pdf.binned.z**  
  Same input as in repfdr. A 3-dimensional array which contains for each study (first dimension), the probabilities of a z-score to fall in the bin (second dimension), under each hypothesis status (third dimension). The third dimension can be of size 2 or 3, depending on the number of association states: if the association can be either null or only in one direction, the dimension is 2; if the association can be either null, or positive, or negative, the dimension is 3. Element [[]] in the output of ztobins.
binned.z.mat | Same input as in `repfdr`. A matrix of the bin numbers for each the z-scores (rows) in each study (columns). Element [[2]] in the output of `ztobins`.

control | List of control parameters to pass to the EM algorithm. See `em.control`.

Details

The implementation of the EM algorithm is in C, and allows parallel processing. By default, the software automatically detects the number of available processing threads. See `em.control` for the option of providing the number of threads to use, as well as for the additional control parameters.

Value

- `all.iterations` Matrix with number of columns equal to the number of EM iterations, and each column is the estimated probability distribution of the vector of association status.
- `last.iteration` Matrix of the vectors of association status along with the column vector of the last EM iteration, which contains the estimated probabilities of the vectors of association status.

Author(s)

C implementation by Shachar Kaufman.

References


See Also

- `repfdr`

Examples

```r
## Not run:
data(binned_zmat)
output_piem <- piem(pbz, bz)

# extract the last iteration to use it in repfdr (see help(repfdr)):
Pi1 <- output_piem$last.iteration
data(Pi)
stopifnot(all.equal(Pi, Pi1))

# simulation data:
data(binned_zmat_sim)
output_piem_sim <- piem(pbz_sim, bz_sim)
Pi_sim <- output_piem_sim$last.iteration

# following are the true proportions in the data: (see help(hmat_sim) for data generation details.)
f <- c(0.895, 0.005, 0.005, 0.02, 0.005, 0.02, 0.02, 0.03)
```
repfdr

Bayes and local Bayes false discovery rate estimation for replicability analysis

Description

Estimate Bayes and local Bayes false discovery rates (FDRs) from multiple studies, for replicability analysis and for meta-analysis, as presented in Heller and Yekutieli (see reference below).

Usage

repfdr(pdf.binned.z, binned.z.mat, non.null = c("replication", "meta-analysis", "user-defined"),
non.null.rows = NULL, Pi.previous.result = NULL, control = em.control())

Arguments

pdf.binned.z A 3-dimensional array which contains for each study (first dimension), the probabilities of a z-score to fall in the bin (second dimension), under each hypothesis status (third dimension). The third dimension can be of size 2 or 3, depending on the number of association states: if the association can be either null or non-null (e.g. only in one direction), the dimension is 2; if the association can be either null, or positive, or negative, the dimension is 3. Element [11] in the output of ztobins.

binned.z.mat A matrix of the bin numbers for each of the z-scores (rows) in each study (columns). Element [2] in the output of ztobins.

non.null Indicates the desired analysis: replication, meta-analysis or user-defined. When user-defined is selected non.null.rows must be specified.

non.null.rows Vector of row indices in H (see hconfigs), indicating which vectors of association status should be considered as non-null in the analysis. H is the output of hconfigs(dim(pdf.binned.z)[1], dim(pdf.binned.z)[3]), i.e. the matrix with rows indicating the possible vectors of association status, where dim(pdf.binned.z)[1] is the number of studies and dim(pdf.binned.z)[3] is the number of association states in each study (2 or 3).

Pi.previous.result An optional Vector of probabilities for each association status. If NULL, then the probabilities are estimated with the EM algorithm. An estimation result from a previous run of repfdr or piem can be supplied to shorten the run-time of the function, see Example section.

control List of control parameters to pass to the EM algorithm. See em.control.
Details

For \( N \) studies, each examining the same \( M \) features, the binned z-scores and the (estimated) probabilities under the null and non-null states in each study are given as input. These inputs can be produced from the z-scores using the function `ztobins`.

The function calls `piem` for the computation of the probabilities for each vector of association status. The number of probabilities estimated is \( x^N \), where \( x=2,3 \) is the number of possible association states in each study.

The function calls `ldr` for the computation of the conditional probability of each of the vectors of association status in the null set given the binned z-scores. The null set contains the rows in `hconfigs(N,x)` that: are excluded from `non.null.rows` if `non.null` is `user_defined`; that are non-zero if `non.null` is `meta_analysis`; that contain at most one 1 if `non.null` is replication and \( x=2 \); that contain at most one 1 or one -1 if `non.null` is replication and \( x=3 \).

The local Bayes FDR is estimated to be the sum of conditional probabilities in the null set for each feature. The empirical Bayes FDR is the average of all local Bayes FDRs that are at most the value of the local Bayes FDR for each feature. The list of discoveries at level \( q \) are all features with empirical Bayes FDR at most \( q \).

Value

- mat
  
  An \( M \times 2 \) Matrix with a row for each feature (\( M \) rows) and two columns, the estimated local Bayes FDR (fdr) and the estimated Bayes FDR (Fdr).

- Pi

  Vector of the estimated probabilities for each of the \( x^N \) possible vectors of association status.

Author(s)

Ruth Heller, Shachar Kaufman, Shay Yaacoby, Daniel Yekutieli.

References


Examples

```r
### Example 1: a simulation; each feature in each study has two association states,
###
# a) Replicability analysis:
data(binned_zmat_sim) # this loads the binned z-scores as well as the (estimated) probabilities
# in each bin for each state
output.rep <- repfdr(pbz_sim, bz_sim, "replication")
BayesFdr.rep <- output.rep$mat[, "Fdr"]
Rej <- (BayesFdr.rep <= 0.05)
sum(Rej)
```
# which of the tests are true replicability findings? (we know this since the data was simulated)
data(hmat_sim)
true.rep <- apply(hmat_sim,1,function(y){ sum(y>1) })

# Compute the false discovery proportion (FDP) for replicability:
sum(Rej * !true.rep) / sum(true.rep)

# we can use the previously calculated Pi for further computations (e.g, meta-analysis):
Pi_sim <- output.rep$pi

# b) meta-analysis:
output.meta <- repfdr(pbz_sim, bz_sim, "meta-analysis", Pi.previous.result = Pi_sim)
BayesFdr.meta <- output.meta$mat[,"Fdr"]
Rej <- (BayesFdr.meta <= 0.05)
sum(Rej)

# which of the tests are true association findings? (we know this since the data was simulated)
ture.assoc <- rowSums(hmat_sim) >= 1

# Compute the false discovery proportion (FDP) for association:
sum(Rej * !true.assoc) / sum(true.assoc)

## Not run:
#### Example 2: SNPs data; each SNP in each study has three association states, negative, null, or positive:

# load the bins of the z-scores and their probabilities.
data(binned_zmat)

# load the prior probabilities for each association status vector.
data(Pi)
Pi # the proportions vector was computed using piem()

# a) replicability analysis:
output.rep <- repfdr(pbz, bz, "replication",Pi.previous.result=Pi)
BayesFdr.rep <- output.rep$mat,"Fdr"
Rej <- sum(BayesFdr.rep <= 0.05)
sum(Rej)

# The posterior probabilities for the first five features with Bayes FDR at most 0.05:
post <- ldr(pbz,bz[order(BayesFdr.rep)[1:5],],Pi)
round(post,4)

# posteriors for a subset of the association status vectors can also be reported:
H <- hconfigs( dim(bz)[2], 3)
h.replicability = apply(H, 1, function(y) {sum(y == 1)>1 | sum(y == -1) >1})
post <- ldr(pbz,bz[order(BayesFdr.rep)[1:5],],Pi,h.vecs= which(h.replicability==1))
round(post,4)

# b) meta-analysis:
The code for the manhattan plot is not run. Here is the commented code:

```r
## End(Not run)

## manhattan plot (ploting can take a while):
# code for manhattan plot by Stephen Turner (see copyrights at the source code manhattan.r)

## Not run:
data(snplocations)
par(mfrow=c(2,1))

# Replication
g famously high
manhattan(dataframe=cbind(snplocations,P=BayesFdr.rep),ymax=10.5,pch=20,
  limitchromosomes=1:4,suggestiveline=-log(0.05,10),genomewideline=F,cex=0.25,
  annotate=snplocations$SNP[BayesFdr<=0.05],main="Replication")

# Association
g famously high
manhattan(dataframe=cbind(snplocations,P=BayesFdr.meta),ymax=10.5,cex=0.25,
  limitchromosomes=1:4,suggestiveline=-log(0.05,10),genomewideline=F,pch=20,
  annotate=snplocations$SNP[BayesFdr<=0.05],main="Meta-analysis")
par(mfrow=c(1,1))

## End(Not run)
```

---

**SNPlocations**

*Three GWAS studies SNPs locations and data*

**Description**

SNPlocations includes the locations of SNPs in chromosomes 1 to 4. Data was simulated to the SNPs with HAPGEN2 for three studies and a sample of it was taken (Chromosomes 1 to 4) for the examples. The data is summarized as z-scores (transformed p-values, with inverse standard normal cumulative distribution). The z-scores matrix can be download from the web (see example).

**Usage**

data(SNPlocations)

**Format**

SNPlocations data.frame of 249024 SNPs' names, chromosome number and location on the chromosomes. zmat Matrix of 249024 SNPs' z-scores (in rows) in each of the 3 studies (columns).

**Source**

**Examples**

```r
data(SNPlocations)
head(SNPlocations)
```

```
## Not run:
download.file('http://www.math.tau.ac.il/~ruheller/repfdr_RData/zmat.RData',destfile = "zmat.RData")
load(file = "zmat.RData")

input.to.repfdr <- ztobins(zmat, 3, df= 15)
pbz <- input.to.repfdr$pdf.binned.z
bz  <- input.to.repfdr$binned.z.mat

## End(Not run)
```

---

**Description**

A simulated data set from three studies, with 10000 "features" in each study, each of which yielded a z-score. The data comprises 10000x3 z-scores. See `hmat_sim` for the indicators of association status matrix.

**Usage**

```r
data(zmat_sim)
```

**Format**

`zmat_sim` is a matrix of 10000 z-scores (in rows) in each of the 3 studies (columns).

**Examples**

```r
data(zmat_sim)
head(zmat_sim)
```

```
## Not run:
input.to.repfdr <- ztobins(zmat_sim, 2)
pbz_sim1 <- input.to.repfdr$pdf.binned.z
bz_sim1  <- input.to.repfdr$binned.z.mat

data(binned_zmat_sim)
stopifnot(all.equal(pbz_sim1, pbz_sim))
stopifnot(all.equal(bz_sim1, bz_sim))

## End(Not run)
```

```
#### zmat_sim was generated by the following code:
data(hmat_sim)
```
ztobins

Binning of z-scores and estimation of the probabilities in each bin for the null and non-null states.

Description

For each study, the function discretizes the z-scores into bins and estimates the probabilities in each bin for the null and non-null states.

Usage

ztobins(zmat, n.association.status = 3, n.bins = 120, type = 0, df = 7, central.prop = 0.5)

Arguments

zmat
Matrix of z-scores of the features (in rows) in each study (columns).

n.association.status
either 2 for no-association\association or 3 for no-association\negative-association\positive-association.

n.bins
Number of breaks in the discretization of the z-score axis (the number of bins is n.bins- 1). If the number of z-scores per study is small, we set n.bins to a number lower than the default of 120 (about equals to the square root of the number of z-scores).

type
Type of fitting used for f; 0 is a natural spline, 1 is a polynomial, in either case with degrees of freedom df (so total degrees of freedom including the intercept is df+1).

df
Degrees of freedom for fitting the estimated density f(z).

central.prop
Central proportion of the z-scores used like the area of zero-assumption to estimate pi0.

Details

This utility function outputs the first two arguments to be input in the main function repfd.

```r
set.seed(12)
m <- nrow(hmat_sim)
zmat_sim1 <- matrix(rnorm(n=3*m,mean=hmat_sim*3),nrow=m,ncol=3)
data(zmat_sim)
stopifnot(all.equal(zmat_sim1,zmat_sim))
```
Value

A list with:

- pdf.binned.z: A 3-dimensional array which contains for each study (first dimension), the probabilities of a z-score to fall in the bin (second dimension), under each hypothesis status (third dimension). The third dimension can be of size 2 or 3, depending on the number of association states: if the association can be either null or only in one direction, the dimension is 2; if the association can be either null, or positive, or negative, the dimension is 3.

- binned.z.mat: A matrix of the bin numbers for each the z-scores (rows) in each study (columns).

See Also

repfdr

Examples

```r
## Not run:

# three association states case (H in {-1,0,1}):
download.file('http://www.math.tau.ac.il/~ruheller/repfdr_RData/zmat.RData', destfile = "zmat.RData")
load(file = "zmat.RData")

input.to.repfdr3 <- ztobins(zmat, 3, df = 15)
pbz <- input.to.repfdr3$pdf.binned.z
bz <- input.to.repfdr3$binned.z.mat

# two association states case (H in {0,1}):
data(zmat_sim)

input.to.repfdr <- ztobins(zmat_sim, 2)
pbz_sim <- input.to.repfdr$pdf.binned.z
bz_sim <- input.to.repfdr$binned.z.mat

## End(Not run)
```
Index

*Topic **GWAS**

- binned_zmat, 2
- binned_zmat_sim, 2
- hmat_sim, 5
- repfdr, 9
- SNPlocations, 12
- zmat_sim, 13

*Topic **Replicability**

- binned_zmat, 2
- binned_zmat_sim, 2
- hmat_sim, 5
- repfdr, 9
- SNPlocations, 12
- zmat_sim, 13

*Topic **datasets**

- binned_zmat, 2
- binned_zmat_sim, 2
- hmat_sim, 5
- SNPlocations, 12
- zmat_sim, 13

binned_zmat, 2
binned_zmat_sim, 2
bz (binned_zmat), 2
bz_sim (binned_zmat_sim), 2

em.control, 3, 8, 9
hconfigs, 4, 6, 7, 9, 10
hmat_sim, 5, 13
ldr, 6, 10

manhattan (repfdr), 9

pbz (binned_zmat), 2
pbz_sim (binned_zmat_sim), 2
Pi (piem), 7
piem, 3, 6, 7, 9, 10

repfdr, 3, 4, 6–9, 9, 14, 15

SNPlocations, 2, 12
zmat_sim, 2, 5, 13
ztobins, 2, 6–10, 14