Package: MixTwice (via r-universe)

September 15, 2024

Type Package

Title Large-Scale Hypothesis Testing by Variance Mixing

Version 2.0

Imports alabama, ashr, fdrtool, Iso, stats

Date 2022-02-28

Author Zihao Zheng and Michael A.Newton

Maintainer Zihao Zheng <zihao.zheng@wisc.edu>

Description Implements large-scale hypothesis testing by variance mixing. It takes two statistics per testing unit -- an estimated effect and its associated squared standard error -- and fits a nonparametric, shape-constrained mixture separately on two latent parameters. It reports local false discovery rates (lfdr) and local false sign rates (lfsr). Manuscript describing algorithm of MixTwice: Zheng et al(2021) <doi:10.1093/bioinformatics/btab162>.

License GPL-2

NeedsCompilation no

Depends R (>= 3.5.0)

Date/Publication 2022-03-02 16:50:02 UTC

Repository https://zzheng68.r-universe.dev

RemoteUrl https://github.com/cran/MixTwice

RemoteRef HEAD

RemoteSha f31a2993c756c1f83e32b0c303033bccdb37b880

Contents

`MixTwice-package` *Large-Scale Hypothesis Testing by Variance Mixing*

Description

Implements large-scale hypothesis testing by variance mixing. It takes two statistics per testing unit – an estimated effect and its associated squared standard error – and fits a nonparametric, shape-constrained mixture separately on two latent parameters. It reports local false discovery rates (lfdr) and local false sign rates (lfsr). Manuscript describing algorithm of MixTwice: Zheng et al(2021) <doi: 10.1093/bioinformatics/btab162>.

Details

The DESCRIPTION file:

Package:	MixTwice
Type:	Package
Title:	Large-Scale Hypothesis Testing by Variance Mixing
Version:	2.0
Imports:	alabama, ashr, fdrtool, Iso, stats
Date:	2022-02-28
Author:	Zihao Zheng and Michael A.Newton
Maintainer:	Zihao Zheng <zihao.zheng@wisc.edu>
Description:	Implements large-scale hypothesis testing by variance mixing. It takes two statistics per testing unit – an estima
License:	GPL-2

Index of help topics:

Author(s)

Zihao Zheng and Michael A.Newton

Maintainer: Zihao Zheng <zihao.zheng@wisc.edu>

References

Zheng et al. *MixTwice: Large scale hypothesis testing for peptide arrays by variance mixing*. **Bioinformatics**, 2021.

Zheng et al. *Disordered Antigens and Epitope Overlap Between Anti Citrullinated Protein Antibodies and Rheumatoid Factor in Rheumatoid Arthritis.* **Arthritis & Rheumatology** 72.2 (2020): 262-272.

Examples

```
data(peptide_data)
## For more detail and example, use ?peptide_data and ?mixtwice
```

`mixtwice` *Large-scale hypothesis testing by variance mixing*

Description

MixTwice deploys large-scale hypothesis testing in the case when testing units provide estimated effects and estimated standard errors. It produces empirical Bayesian local false discovery and sign rates for tests of zero effect.

Usage

```
mixtwice(thetaHat, s2, Btheta = 15, Bsigma2 = 10, df,
method = c("EM-pava", "AugLag"), maxit = 100, prop = 1)
```

Arguments

`thetaHat`	Estimated effect sizes (vector over testing units))
`s2`	Estimated squared standard errors of thetaHat (vector over testing units)
`Btheta`	Grid size parameter for effect distribution
`Bsigma2`	Grid size parameter for variance distribution
`df`	Degrees of freedom in chisquare associated with estimated standard error
`method`	Method used for solving the non-parametric MLE optimization. `method = "EM-pava"` solves the optimization problem using EM approach with pool adjacent violator algorithm (pava) and `method = "EM-pava"` solves the optimization problem directly using Augmented Lagrangian approach.
`maxit`	Number of iterations in EM if `method = "EM-pava"`, with default, `maxit = 100`
`prop`	Proportion of units randomly selected to fit the distribution, with default, `prop = 1` (use all units to fit the distribution).

Details

`mixtwice` takes estimated effects and standard errors over a set of testing units. To compute local error-rate statistics, it finds nonparametric MLEs of the underlying distributions. It is similar to "ashr", except that mixtwice allows both a mixing distribution of underlying effects theta as well as a mixing distribution over underlying variance parameters. Furthermore, it treats the effect mixing distribution nonparametrically, but enforces the shape constraint that this distribution is unimodal with mode at theta=0. (We do not assume symmetry). The distribution of variance parameters is also treated nonparametrically, but with no shape constraints. The observations are assumed to be emitted from a normal distribution (on estimated effects) and an independent Chi-square distribution (on estimated squared standard errors).

Value

grid.theta	Support of the estimated mixing distribution on effects
grid.sigma2	Support of the estimated mixing distribution of variances
mix.theta	Estimated distribution of effect size, on grid.theta
mix.sigma2	Estimated distribution of variance, on grid.sigma2
lfdr	Local false discovery rate for each testing unit
lfsr	Local false sign rate for each testing unit

Note

See Zheng et al. 2021 for further detail

Author(s)

Zihao Zheng, Michael A.Newton

References

Zheng et al. *MixTwice: Large scale hypothesis testing for peptide arrays by variance mixing*. Bioinformatics, 2021.

See Also

See Also as ?peptide_data

Examples

```
set.seed(1)

l = 1000 ## number of testing units

neach = 20 ## number of subjects in each group

pi0 = 0.8 ## null proportion

signal1 = rep(0, l)

signal2 = signal1

signal2[1:round((1-pi0)*l)] = rnorm(round((1-pi0)*l), mean = 0, sd = 3)

## I will generate the sigma^2 parameter

sigma2 = rep(1,l)

sigma = sqrt(sigma2)

## Then I can generate data
```

```
data1 <- data2 <- matrix(NA, nrow = l, ncol = neach)

for (i in 1:l) {

  data1[i,] = rnorm(neach, mean = rep(signal1[i], each = neach), sd = sigma[i])
  data2[i,] = rnorm(neach, mean = rep(signal2[i], each = neach), sd = sigma[i])

}

thetaHat = rowMeans(data2) - rowMeans(data1)

sd1 = apply(data1, 1, sd)
sd2 = apply(data2, 1, sd)

s2 = sd1^2/neach + sd2^2/neach

fit.EM = mixtwice(thetaHat, s2, Btheta = 15, Bsigma2 = 10, df = 2*neach - 2,
                  method = "EM-pava", maxit = 100, prop = 1)

## you can try to visualize the result

plot(fit.EM$grid.theta, cumsum(fit.EM$mix.theta), type = "s",
     xlab = "grid.theta", ylab = "ecdf of theta", lwd = 2)

lines(ecdf(signal2 - signal1),cex = 0.1, lwd = 0.5, lty = 2, col = "red")

legend("topleft", legend = c("fit.mix", "true.mix"), col = c("black", "red"),
       lwd = 1, pch = 19)

## calculate false discovery rate and true positive under 0.05

oo = order(fit.EM$lfdr)

# number of discovery

x1 = sum(cumsum(fit.EM$lfdr[oo])/c(1:l) <= 0.05)

# number of true discovery

x2 = sum(cumsum(fit.EM$lfdr[oo])/c(1:l) <= 0.05 & (signal2 != 0)[oo])

# number of real positive

x3 = sum(signal2 != 0)

# number of false discovery

x4 = sum(cumsum(fit.EM$lfdr[oo])/c(1:l) <= 0.05 & (signal2 == 0)[oo])

x4/x1 ## false discovery rate FDR

x2/x3 ## true positive rate
```

```
## null proportion estimation

max(fit.EM$mix.theta)

## you can also try using another method (Augmented Lagrangian), the result would be similar

# fit.AugLag = mixtwice2(thetaHat, s2, Btheta = 15, Bsigma2 = 10, df = 2*neach - 2,
#                        method = "AugLag", prop = 1)
```

---

peptide_data *Peptide array data example*

Description

A high-density peptide microarray example to identify peptides for which antibody binding levels differ between control subjects and rheumatoid arthritis (RA) patients expressing a specific disease marker combination (i.e., CCP+RF+ RA).

Usage

```
data("peptide_data")
```

Format

A data frame with 152603 observations on the following 16 variables.

The first 8 columns are RA patients and the remaining columns are from control subjects.

Details

Each row of the data (`rownames(peptide_data)`) is a probed length-12 peptide and each column of the data (`colnames(peptide_data)`) is a subject with distinct pseudo sample ID. The binding value is doubly-log transformed using natural base to stabilize variance.

Source

Zheng, Zihao, et al. *Disordered Antigens and Epitope Overlap Between Anti Citrullinated Protein Antibodies and Rheumatoid Factor in Rheumatoid Arthritis.* **Arthritis & Rheumatology** 72.2 (2020): 262-272.

References

Zheng et al. *MixTwice: Large scale hypothesis testing for peptide arrays by variance mixing*. Bioinformatics, 2021.

Examples

```
#### load the RA data

data(peptide_data)

#### visualize the data

## each row is a peptide with unique peptide sequence

## each column is a subject with information on group and pseudo ID

colnames(peptide_data)


## z-score for peptide

get_zscore = function(x){

  n = length(x)

  t = t.test(x[1:(n/2)], x[(n/2 + 1):n], var.equal = TRUE)$statistic

  return(qnorm(pt(t, df = n-2)))
}

z = apply(peptide_data, 1, get_zscore)

## visualize the density of z-score

hist(z, probability = TRUE, 100, ylim = c(0,0.4), col = "blue")
lines(density(rnorm(10^5)), lwd =2)
```

Index