Package 'MRBEEX'

August 23, 2024

Type Package

Title Mendelian Randomization using Bias-correction Estimating Equation.

Version 0.1.0

Author Yihe Yang

Maintainer Yihe Yang <yxy1234@case.edu>

Description MRBEEX extends the existing MRBEE package by introducing several advanced functions. It includes the MRBEE_IMRP function, which uses the IMRP algorithm to iteratively estimate causal effects and test for horizontal pleiotropy. The MRBEE_IPOD function employs the IPOD algorithm to select horizontal pleiotropy and allows for the inclusion of correlated instrumental variables with the corresponding LD matrix. MRBEE_SuSiE further extends MRBEE_IPOD by using the SuSiE method for selecting exposures while continuing to use the IPOD algorithm for horizontal pleiotropy selection.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 7.2.3

Imports MASS,
Rcpp (>= 1.0.0),
RcppArmadillo,
FDRestimation,
mvtnorm,
utils,
varbvs,
susieR,
CppMatrix,
Matrix

Remotes harryyiheyang/CppMatrix,
stephenslab/susieR,
pcarbo/varbvs

ORCID 0000-0001-6563-3579

R topics documented:

`clump_cluster` *Clustering second data frame based on closest SNP centers from first data frame*

Description

This function performs clustering of SNPs in a second data.frame based on the closest SNP centers defined in a first data.frame. Both data.frames should include SNP, BP, and CHR columns. This function scales CHR and BP to ensure distinctiveness across chromosomes and employs Euclidean distance to find the nearest cluster centers from the first data.frame for each SNP in the second data.frame.

Usage

```
clump_cluster(df1, df2)
```

Arguments

`df1`	A data.frame representing the output of a plink clump with parameters r2=0.01. It contains columns for SNP, BP (base pair position), CHR (chromosome), and P (p-value).
`df2`	A data.frame similar to df1, representing a plink output with a less stringent r2 value, typically r2=0.5, including columns for SNP, BP, CHR, and P.

Details

The function first standardizes the CHR and BP columns by multiplying CHR by 10000 and dividing BP by 1e6. This standardization helps to manage the scale differences between chromosome numbers and base pair positions. After standardization, it calculates the Euclidean distances between each SNP in df2 to all SNP centers in df1, assigns each SNP in df2 to the nearest center from df1, and adds a new column 'cluster' to df2 to reflect this assignment.

Value

A modified version of df2 where each SNP is annotated with a 'cluster' index corresponding to the closest SNP center from df1 based on scaled CHR and BP values.

Examples

```
df1 <- data.frame(SNP=c("rs1", "rs2"), CHR=c(1, 1), BP=c(150000, 250000), P=c(0.001, 0.002))
df2 <- data.frame(SNP=c("rs1", "rs3", "rs2", "rs4"), CHR=c(1,1,1,1),
                  BP=c(150000,160000,250000,260000),
                  P=c(0.001,0.003,0.002, 0.004))
clustered_df2 <- clump_cluster(df1, df2)
```

`cluster_snps` *Clustering SNPs based on p-value and proximity with a PLINK C+T file.*

Description

This function clusters SNPs within a given window size based on their P-value and proximity. It iterates through each chromosome, finds the SNP with the smallest P-value, and groups all SNPs within the specified window size around this SNP into a cluster.

Usage

```
cluster_snps(df, window_size = 1e+06)
```

Arguments

`df`	A data.frame containing SNP data with columns for SNP (SNP ID), CHR (chromosome), BP (base pair position), and P (p-value).
`window_size`	An integer specifying the window size around each SNP (in base pairs) within which other SNPs are considered for clustering. Default to 1e6.

Details

The function processes each chromosome independently. It orders the SNPs by their base pair positions, identifies the SNP with the smallest P-value, and clusters all SNPs within the specified window size around this SNP. The process is repeated until all SNPs are assigned to a cluster.

Value

A data.frame containing the clustered SNPs with an additional column 'ClusterSize' indicating the number of SNPs in each cluster.

Examples

```
df <- data.frame(SNP=c("rs1", "rs2", "rs3", "rs4", "rs5"),
                 CHR=c(1, 1, 1, 1, 2),
                 BP=c(100000, 150000, 200000, 250000, 300000),
                 P=c(0.01, 0.02, 0.03, 0.04, 0.05))
window_size <- 50000
clustered_snps <- cluster_snps(df, window_size)
```

MRBEE_IMRP *Mendelian randomization with bias-correction estimating equation: detecting horizontal pleiotropy via hypothesis test.*

Description

This function estimates the causal effect using a bias-correction estimating equation, considering potential pleiotropy and measurement errors.

Usage

```
MRBEE_IMRP(
  by,
  bX,
  byse,
  bXse,
  Rxy,
  max.iter = 30,
  max.eps = 1e-04,
  pv.thres = 0.05,
  var.est = "variance",
  FDR = T,
  adjust.method = "Sidak",
  maxdiff = 1.5
)
```

Arguments

by	A vector (n x 1) of the GWAS effect size of outcome.
bX	A matrix (n x p) of the GWAS effect sizes of p exposures.
byse	A vector (n x 1) of the GWAS effect size SE of outcome.
bXse	A matrix (n x p) of the GWAS effect size SEs of p exposures.
Rxy	A matrix (p+1 x p+1) of the correlation matrix of the p exposures and outcome. The last one should be the outcome.
max.iter	Maximum number of iterations for causal effect estimation. Defaults to 30.
max.eps	Tolerance for stopping criteria. Defaults to 1e-4.
pv.thres	P-value threshold in pleiotropy detection. Defaults to 0.05.
var.est	Method for estimating the variance of residual in pleiotropy test. Can be "robust", "variance", or "ordinal". Defaults is "variance" that estimates the variance of residual using median absolute deviation (MAD).
FDR	Logical. Whether to apply the FDR to convert the p-value to q-value. Defaults to TRUE.
adjust.method	Method for estimating q-value. Defaults to "Sidak".
maxdiff	The maximum difference between the MRBEE causal estimate and the initial estimator. Defaults to 1.5.

Value

A list containing the estimated causal effect, its covariance, and pleiotropy

MRBEE_IPOD *Mendelian randomization with bias-correction estimating equation: selecting horizontal pleiotropy via IPOD algorithm.*

Description

Detailed description of the function goes here.

Usage

```
MRBEE_IPOD(
  by,
  bX,
  byse,
  bXse,
  LD = "identity",
  Rxy,
  cluster.index = c(1:length(by)),
  Nmin = F,
  tauvec = seq(3, 50, by = 5),
  max.iter = 100,
  max.eps = 0.001,
  ebic.gamma = 1,
  reliability.thres = 0.5,
  rho = 2,
  maxdiff = 1.5,
  sampling.time = 100,
  sampling.frac = 0.5,
  sampling.iter = 5,
  theta.ini = F,
  gamma.ini = F
)
```

Arguments

by	A vector (n x 1) of GWAS effect sizes for the outcome.
bX	A matrix (n x p) of GWAS effect sizes for p exposures.
byse	A vector (n x 1) of standard errors for the GWAS effect sizes of the outcome.
bXse	A matrix (n x p) of standard errors for the GWAS effect sizes of the exposures.
LD	A matrix representing the linkage disequilibrium (LD) among instrumental variables. This matrix should be a sparse matrix with lots of zero entries.
Rxy	A matrix (p+1 x p+1) of the correlation matrix including p exposures and the outcome. Outcome should be the last column.
cluster.index	A vector indicating the cluster membership for each instrumental variable. This is used in standard error estimation.
Nmin	Optional; the minimum sample size for the GWAS if not provided, defaults to the number of instrumental variables.
tauvec	A vector of tuning parameters for penalizing horizontal pleiotropy in the IPOD algorithm.

max.iter	The maximum number of iterations allowed for convergence of the causal effect estimates.
max.eps	The tolerance level for convergence; iteration stops when changes are below this threshold.
ebic.gamma	The penalty factor for extended Bayesian Information Criterion (eBIC) adjustments on pleiotropy.
reliability.thres	A threshold on bias-correction term, defaults to 0.5.
rho	The penalty multiplier used in the ADMM algorithm within the IPOD framework.
maxdiff	The maximum allowed difference ratio between iterative causal estimates and initial estimations for stabilization.
sampling.time	The number of subsampling iterations used to estimate the standard error of the causal effect estimate. Defaults to 100. When set to 0, a sandwich formula is applied for the estimation.
sampling.frac	The fraction of the data to be used in each subsampling iteration. Defaults to 0.5, meaning that 50% of the data is used in each iteration.
sampling.iter	The number of iteration of MRBCEE.IPOD to be used in each subsampling iteration. Defaults to 5.
theta.ini	Initial estimates for the causal effects; defaults to FALSE, indicating automatic initialization.
gamma.ini	Initial estimates for horizontal pleiotropy effects; also defaults to FALSE for automatic setup.

Value

A list containing detailed results of the analysis, including estimated causal effects, pleiotropy effects, their respective standard errors, and Bayesian Information Criterion (BIC) scores, among other metrics.

MRBEE_SuSiE *Mendelian randomization with bias-correction estimating equation: selecting horizontal pleiotropy via IPOD and selecting exposures via SuSiE.*

Description

Detailed description of the function goes here.

Usage

```
MRBEE_SuSiE(
  by,
  bX,
  byse,
  bXse,
  LD = "identity",
  Rxy,
```

```
  cluster.index = c(1:length(by)),
  Nmin = F,
  Lvec = c(1:min(5, nrow(bX))),
  pip.thres = 0.5,
  tauvec = seq(3, 50, by = 2),
  max.iter = 100,
  max.eps = 0.001,
  susie.iter = 100,
  ebic.theta = 0,
  ebic.gamma = 1,
  empirical.variance.lower = 0.2,
  empirical.variance.upper = 100,
  reliability.thres = 0.5,
  rho = 2,
  maxdiff = 1.5,
  sampling.time = 100,
  sampling.frac = 0.5,
  sampling.iter = 5,
  theta.ini = F,
  gamma.ini = F
)
```

Arguments

by	A vector (n x 1) of GWAS effect sizes for the outcome.
bX	A matrix (n x p) of GWAS effect sizes for p exposures.
byse	A vector (n x 1) of standard errors for the GWAS effect sizes of the outcome.
bXse	A matrix (n x p) of standard errors for the GWAS effect sizes of the exposures.
LD	A matrix representing the linkage disequilibrium (LD) among instrumental variables.
Rxy	A matrix (p+1 x p+1) of the correlation matrix including p exposures and the outcome. Outcome should be the last column.
cluster.index	A vector indicating the cluster membership for each instrumental variable. This is used in standard error estimation.
Nmin	Optional; the minimum sample size for the GWAS if not provided, defaults to the number of instrumental variables.
Lvec	The number of single effects used in SuSiE, defaults to c(1:min(5,nrow(bX))).
pip.thres	The threshold of PIP in SuSiE, below which the causal effect estimate will be set to zero, defaults to 0.5.
tauvec	A vector of tuning parameters for penalizing horizontal pleiotropy in the IPOD algorithm.
max.iter	The maximum number of iterations allowed for convergence of the causal effect estimates.
max.eps	The tolerance level for convergence; iteration stops when changes are below this threshold.
susie.iter	The maximum number of iterations allowed for convergence of SuSiE program, defaults to 50.
ebic.theta	The penalty factor for extended Bayesian Information Criterion (eBIC) adjustments on causal effects

`ebic.gamma` The penalty factor for eBIC adjustments on pleiotropy.

`empirical.variance.lower`
The lower boundary of empirical variance estimate of residual, defaults to 0.5.

`empirical.variance.upper`
The upper boundary of empirical variance estimate of residual, defaults to 2.

`reliability.thres`
A threshold on bias-correction term, defaults to 0.5.

`rho` The penalty multiplier used in the ADMM algorithm within the IPOD framework.

`maxdiff` The maximum allowed difference ratio between iterative causal estimates and initial estimations for stabilization.

`sampling.time` The number of subsampling iterations used to estimate the standard error of the causal effect estimate. Defaults to 100. When set to 0, a sandwich formula is applied for the estimation.

`sampling.frac` The fraction of the data to be used in each subsampling iteration. Defaults to 0.5, meaning that 50% of the data is used in each iteration.

`sampling.iter` The number of iteration of MRBCEE.IPOD to be used in each subsampling iteration. Defaults to 5.

`theta.ini` Initial estimates for the causal effects; defaults to FALSE, indicating automatic initialization.

`gamma.ini` Initial estimates for horizontal pleiotropy effects; also defaults to FALSE for automatic setup.

Value

A list containing detailed results of the analysis, including estimated causal effects, pleiotropy effects, their respective standard errors, and Bayesian Information Criterion (BIC) scores, among other metrics.

`summary_generation` *Generating simulated data for Mendelian randomization simulation*

Description

This function generates simulated data for Mendelian Randomization (MR) analysis, considering genetic effects, estimation errors, and horizontal pleiotropy. It allows for different distributions of genetic effects and pleiotropy, and accommodates both independent and correlated instrumental variables (IVs).

Usage

```
summary_generation(
  theta,
  m,
  Rbb,
  Ruv,
  Rnn,
  Nxy,
```

```
  Hxy,
  LD = "identify",
  non.zero.frac,
  pleiotropy.frac = 0,
  pleiotropy.var = 0.5,
  pleiotropy.cor = 0,
  pleiotropy.cor.exposure = c(1:length(theta)),
  pleiotropy.cor.effect = rep(1:length(theta)),
  effect.dis = "normal",
  pleiotropy.dis = "uniform"
)
```

Arguments

`theta`	An (px1) vector of causal effects.
`m`	The number of instrumental variables (IVs).
`Rbb`	An (pxp) correlation matrix of genetic effects.
`Ruv`	An ((p+1)x(p+1)) correlation matrix of residuals in outcome and exposures; the outcome is the first one.
`Rnn`	An ((p+1)x(p+1)) correlation matrix of sample overlap; the outcome is the first one.
`Nxy`	An ((p+1)x1) vector of GWAS sample sizes; the outcome is the first one.
`Hxy`	An ((p+1)x1) vector of heritabilities; the outcome is the first one.
`LD`	An (mxm) correlation matrix of the IVs or "identify" indicating independent IVs.
`pleiotropy.frac`	A number in [0,0.5) indicating the fraction of IVs affected by pleiotropy.
`pleiotropy.var`	A number in $[0, \infty)$ indicating the variance attributed to pleiotropy.
`pleiotropy.cor`	A number indicating the correlation between correlated horizontal pleiotropy and the specified latent variable. If set to 0, uncorrelated horizontal pleiotropy is generated.
`pleiotropy.cor.exposure`	A vector of indices of variables that are correlated with the correlated horizontal pleiotropy.
`pleiotropy.cor.effect`	A vector of effects corresponding to the variables correlated with the correlated horizontal pleiotropy. Specifically, `cor(bX[,pleiotropy.cor.exposure] %*% pleiotropy.cor.effect, CHP)` equals `pleiotropy.cor`.
`effect.dis`	Distribution of genetic effects: "normal" (default), "uniform", or "t" distribution (with degree of freedom 5).
`pleiotropy.dis`	Distribution of pleiotropy effects: "normal" (default), "uniform", or "t" distribution (with degree of freedom 5).
`zero.frac`	An (px1) vector with all entries in (0,1]; each entry is the probability of deltaj such that betaj=betaj'*deltaj.

Value

A list containing simulated GWAS effect sizes for exposures (bX), their standard errors (bXse), the GWAS effect size for the outcome (by), its standard error (byse), the pleiotropy effects (pleiotropy), and the true effects.

Index