Package: mrclust (via r-universe)

January 24, 2025

Type Package

Title Identifying Clustered Heterogeneity in Mendelian Randomization Analyses

Version 0.1.0

Author Christopher Neal Foley

Maintainer The package maintainer <chris.neal.foley@gmail.com>

Description Performs likelihood based clustering on univariate observations with known uncertainty (via standard error data), whilst accounting for possible null and junk components in the sample.

License GPL-3

Encoding UTF-8

LazyData true

Depends R (>= 3.5)

Imports ggplot2, RColorBrewer, MendelianRandomization

RoxygenNote 7.1.1

Suggests knitr, data.table, rmarkdown

VignetteBuilder knitr

Config/pak/sysreqs libgmp3-dev make libicu-dev libssl-dev

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

RemoteUrl https://github.com/cnfoley/mrclust

RemoteRef HEAD

RemoteSha 039ed855bda28b76195222a8ddd2b7b5041b775b

Contents

DBP_CAD *Genetic association data from diastolic blood pressure (DBP) and coronary artery disease (CAD) GWAS.*

Description

A dataset containing chromosome position, rsid and allele information as well as estimates of the regression coefficients and associated standards errors from the SBP and CAD GWAS.

Usage

```
DBP_CAD
```

Format

A data frame with 119 rows and 8 variables:

chr.pos chromosome position

rsid RSID

bx estimated regression coefficient with risk-factor, SBP

bxse standard error of estimated regression coefficient with risk-factor, SBP

by estimated regression coefficient with outcome, CAD

byse standard error of estimated regression coefficient with outcome, CAD

a1 a1 allele

a2 a2 allele

Source

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6284793/

http://www.phenoscanner.medschl.cam.ac.uk

`mr_clust_em` *MR-Clust mixture model fitting*

Description

Assessment of clustered heterogeneity in Mendelian randomization analyses using expectation-maximisation (EM) based model fitting of the MR-Clust mixture model. Function output includes both data-tables and a visualisation of the assingment of variants to clusters.

Usage

```
mr_clust_em(
  theta,
  theta_se,
  bx,
  by,
  bxse,
  byse,
  obs_names = NULL,
  max_iter = 5000,
  tol = 1e-05,
  junk_sd = NULL,
  junk_mean = 0,
  stop_bic_iter = 5,
  min_clust_search = 10,
  results_list = list("all", "best"),
  cluster_membership = list(by_prob = 0.1, bound = 0),
  plot_results = list("best", min_pr = 0.5),
  trait_search = FALSE,
  trait_pvalue = 1e-05,
  proxy_r2 = 0.8,
  catalogue = "GWAS",
  proxies = "None",
  build = 37
)
```

Arguments

`theta`	numeric vector of length the number of variants, the i-th element is a ratio-estimate for the i-th genetic variant.
`theta_se`	numeric vector of length the number of variants, the i-th element is the standard error of the ratio-estimate for the i-th genetic variant.
`bx`	numeric vector of length the number of variants, the i-th element is the estimated regression coefficient - i.e. beta-x value - relating the i-th genetic variant to the risk-factor.

by	numeric vector of length the number of variants, the i-th element is the estimated regression coefficient - i.e. beta-y value - relating the i-th genetic variant to the outcome.
bxse	numeric vector of length the number of variants, the i-th element is the standard error of the estimated regression coefficient relating the i-th genetic variant to the risk-factor.
byse	numeric vector of length the number of variants, the i-th element is the standard error of the estimated regression coefficient relating the i-th genetic variant to the outcome.
obs_names	character vector of length the number of variants, the i-th element is the name of the i-th genetic variants - e.g. the rsID.
max_iter	numeric integer denoting the maximum number of iterations to take before stopping the EM-algorithm's search for a maxima in the log-likelihood.
tol	numeric scalar denoting the maximum absolute difference between two computations of the log-likelihood with which we accept that a maxima in the log-likelihood has been computed.
junk_sd	numeric scalar denoting the scale parameter in the generalised t-distribution
junk_mean	numeric scalar denoting the mean of the generalised t-distribution. By default mean is set to zero.
stop_bic_iter	numeric integer I, for computational efficiency - particularly when analysing large numbers of variants - we can stop the EM-algorithm if the BIC is monotonic increasing over the previous I increases in the number of clusters K. By default evidence supporting at least 10 clusters in the data is computed and so, for example, if the BIC from models which assume 6 clusters; 7 clusters; ... or; 10 clusters is monotonic increasing - in the number of clusters K -then the EM-algorithm is stopped and the model whose K minimises the BIC is returned.
min_clust_search	numeric integer which denotes the minimum number of clusters searched for in the data - default computes evidence supporting up to K=10 clusters which might explain any clustered heterogeneity in the data.
results_list	character list allowing users to choose whether to return a table with the variants assigned to: "all" of the clusters; a single "best" cluster or; both. By default we return both, i.e. results_list = list("all", "best").
cluster_membership	numeric list which allows users to output a list which, for each cluster, returns the variants assigned to the cluster by stratified by the probability of belonging to the cluster. By default, cluster_membership = list(by_prob = 0.1, bound = 0); so that MRClust returns a list, which for each cluster, outputs the variants assigned to the cluster with probability between (0.9,1); (0.8,0.9);... and finally; (0.1,0), i.e. by probability increments 0.1 from 1 to a lower bound of 0.
plot_results	numeric list which allows users to plot the output of MRClust. By default, plot_results = list("best", min_pr = 0.5); so that the best clustering is plotted with variants assigned to a cluster with probability above 0.5.
trait_search	logical, for each of the non-null and non-junk clusters search phenoscanner for traits associated with the variants.

trait_pvalue	numeric scalar for use with trait_search, representing the maximum p-value with with at least one variant in the cluster must be associated with a trait for it to be returned in the phenoscanner search. Default value is GWA significance, i.e. 5*10^-8.
proxy_r2	numeric scalar for use with trait search, allowing variants whose r2>=proxy_r2 to be included in the trait search. Default r2=0.8.
catalogue	character, for use with trait search. From Phenoscanner (http://www.phenoscanner.medschl.cam.ac.uk/inf "the catalogue to be searched (options: None, GWAS, eQTL, pQTl, mQTL, methQTL)". Default setting is catalogue = "GWAS".
proxies	character, for use with trait search. From Phenoscanner (http://www.phenoscanner.medschl.cam.ac.uk/inf "the proxies database to be searched (options: None, AFR, AMR, EAS, EUR, SAS)". Default setting is proxies = "None"
build	integer, for use with trait search. From Phenoscanner (http://www.phenoscanner.medschl.cam.ac.uk/infor "Human genome build numbers (options: 37, 38; default: 37)". Default setting is build = 37.

Value

Returned are: estimates of the putative number of clusters in the sample, complete with allocation probabilities and summaries of the association estimates for each variant; plots which visualise the allocation of variants to clusters and; several summaries of the fitting process, i.e. the BIC and likelihood estimates.

PP_CAD *Genetic association data from pulse pressure (PP) and coronary artery disease (CAD) GWAS.*

Description

A dataset containing chromosome position, rsid and allele information as well as estimates of the regression coefficients and associated standards errors from the SBP and CAD GWAS.

Usage

```
PP_CAD
```

Format

A data frame with 121 rows and 8 variables:

chr.pos chromosome position

rsid RSID

bx estimated regression coefficient with risk-factor, SBP

bxse standard error of estimated regression coefficient with risk-factor, SBP

by estimated regression coefficient with outcome, CAD

byse standard error of estimated regression coefficient with outcome, CAD

a1 a1 allele

a2 a2 allele

Source

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6284793/

http://www.phenoscanner.medschl.cam.ac.uk

`pr_clust` *Cluster size and assignemnt probabilities*

Description

Keep results based on a minimum allocation probability and number of observations in a cluster.

Usage

```
pr_clust(dta, prob = 0.5, min_obs = 1)
```

Arguments

`dta`	table of results from mr_clust_em$results$best.
`prob`	numeric scalar, keep only variants assigned to clusters above this allocation probability.
`min_obs`	integer, keep only variants assinged to clusters with more than or equal to min_obs members.

Value

The results

`SBP_CAD` *Genetic association data from systolic blood pressure (SBP) and coronary artery disease (CAD) GWAS.*

Description

A dataset containing chromosome position, rsid and allele information as well as estimates of the regression coefficients and associated standards errors from the SBP and CAD GWAS.

Usage

```
SBP_CAD
```

Format

A data frame with 121 rows and 8 variables:

chr.pos chromosome position

rsid RSID

bx estimated regression coefficient with risk-factor, SBP

bxse standard error of estimated regression coefficient with risk-factor, SBP

by estimated regression coefficient with outcome, CAD

byse standard error of estimated regression coefficient with outcome, CAD

a1 a1 allele

a2 a2 allele

Source

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6284793/

http://www.phenoscanner.medschl.cam.ac.uk

---

two_stage_plot *Plotting clustered ratio-estimates*

Description

Plot of the two-stage regression estimates, i.e. G-X and G-Y associations, annotated with cluster allocation labels and cluster mean estimates.

Usage

```
two_stage_plot(res, bx, by, bxse, byse, obs_names)
```

Arguments

res	table of results from mr_clust_em$results$best.
bx	numeric vector of length the number of variants, the i-th element is the estimated regression coefficient - i.e. beta-x value - relating the i-th genetic variant to the risk-factor.
by	numeric vector of length the number of variants, the i-th element is the estimated regression coefficient - i.e. beta-y value - relating the -th genetic variant to the outcome.
bxse	numeric vector of length the number of variants, the i-th element is the standard error of the estimated regression coefficient relating the i-th genetic variant to the outcome.
byse	numeric vector of length the number of variants, the i-th element is the standard error of the estimated regression coefficient relating the i-th genetic variant to the risk-factor.
obs_names	character vector of length the number of variants, the i-th element is the name of the i-th genetic variants - e.g. the rsID.

Value

Returned is a scatter plot of the two-stage association estimates for each variant in which: clusters are colour coded and variants with larger assignement/inclusion probabilities appear larger.

Index