

Package: mrclust (via r-universe)

September 26, 2024

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

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

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

RemoteRef HEAD

RemoteSha 039ed855bda28b76195222a8ddd2b7b5041b775b

Contents

DBP_CAD	2
mr_clust_em	3
PP_CAD	5
pr_clust	6
SBP_CAD	6
two_stage_plot	7

DBP_CAD	<i>Genetic association data from diastolic blood pressure (DBP) and coronary artery disease (CAD) GWAS.</i>
---------	---

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.phenoscanter.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

<code>theta</code>	numeric vector of length the number of variants, the i-th element is a ratio-estimate for the i-th genetic variant.
<code>theta_se</code>	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.
<code>bx</code>	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.

<code>by</code>	numeric vector of length the number of variants, the <i>i</i> -th element is the estimated regression coefficient - i.e. beta- γ value - relating the <i>i</i> -th genetic variant to the outcome.
<code>bxse</code>	numeric vector of length the number of variants, the <i>i</i> -th element is the standard error of the estimated regression coefficient relating the <i>i</i> -th genetic variant to the risk-factor.
<code>byse</code>	numeric vector of length the number of variants, the <i>i</i> -th element is the standard error of the estimated regression coefficient relating the <i>i</i> -th genetic variant to the outcome.
<code>obs_names</code>	character vector of length the number of variants, the <i>i</i> -th element is the name of the <i>i</i> -th genetic variants - e.g. the rsID.
<code>max_iter</code>	numeric integer denoting the maximum number of iterations to take before stopping the EM-algorithm's search for a maxima in the log-likelihood.
<code>tol</code>	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.
<code>junk_sd</code>	numeric scalar denoting the scale parameter in the generalised t-distribution
<code>junk_mean</code>	numeric scalar denoting the mean of the generalised t-distribution. By default mean is set to zero.
<code>stop_bic_iter</code>	numeric integer <i>I</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>I</i> increases in the number of clusters <i>K</i> . 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 <i>K</i> -then the EM-algorithm is stopped and the model whose <i>K</i> minimises the BIC is returned.
<code>min_clust_search</code>	numeric integer which denotes the minimum number of clusters searched for in the data - default computes evidence supporting up to <i>K</i> =10 clusters which might explain any clustered heterogeneity in the data.
<code>results_list</code>	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. <code>results_list = list("all", "best")</code> .
<code>cluster_membership</code>	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, <code>cluster_membership = list(by_prob = 0.1, bound = 0)</code> ; 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.
<code>plot_results</code>	numeric list which allows users to plot the output of MRClust. By default, <code>plot_results = list("best", min_pr = 0.5)</code> ; so that the best clustering is plotted with variants assigned to a cluster with probability above 0.5.
<code>trait_search</code>	logical, for each of the non-null and non-junk clusters search phenoscanner for traits associated with the variants.

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.phenoscanter.medschl.cam.ac.uk>

pr_clust	<i>Cluster size and assignemnt probabilities</i>
----------	--

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	<i>Genetic association data from systolic blood pressure (SBP) and coronary artery disease (CAD) GWAS.</i>
---------	--

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.phenoscanter.medschl.cam.ac.uk>

two_stage_plot	<i>Plotting clustered ratio-estimates</i>
----------------	---

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 assignment/inclusion probabilities appear larger.

Index

* datasets

DBP_CAD, [2](#)

PP_CAD, [5](#)

SBP_CAD, [6](#)

DBP_CAD, [2](#)

mr_clust_em, [3](#)

PP_CAD, [5](#)

pr_clust, [6](#)

SBP_CAD, [6](#)

two_stage_plot, [7](#)