Package ‘GBJ’

January 31, 2024

Type Package

Title Generalized Berk-Jones Test for Set-Based Inference in Genetic Association Studies

Version 0.5.4

Date 2024-01-28

Description Offers the Generalized Berk-Jones (GBJ) test for set-based inference in genetic association studies. The GBJ is designed as an alternative to tests such as Berk-Jones (BJ), Higher Criticism (HC), Generalized Higher Criticism (GHC), Minimum p-value (minP), and Sequence Kernel Association Test (SKAT). All of these other methods (except for SKAT) are also implemented in this package, and we additionally provide an omnibus test (OMNI) which integrates information from each of the tests.

The GBJ has been shown to outperform other tests in genetic association studies when signals are correlated and moderately sparse. Please see the vignette for a quickstart guide or Sun and Lin (2017) <arXiv:1710.02469> for more details.

Depends R (>= 2.10)

Imports Rcpp (>= 0.12.7), mvtnorm, SKAT, stats, BH

LinkingTo Rcpp, BH

License GPL-3

RoxygenNote 7.3.1

Suggests knitr, rmarkdown, bindata, rje, testthat

VignetteBuilder knitr

NeedsCompilation yes

Author Ryan Sun [aut, cre]

Maintainer Ryan Sun <ryansun.work@gmail.com>

Repository CRAN

Date/Publication 2024-01-31 08:40:06 UTC
Description

Calculate the Berk-Jones test statistic and p-value.

Usage

BJ(test_stats, cor_mat = NULL, pairwise_cors = NULL)

Arguments

test_stats Vector of test statistics for each factor in the set (i.e. marginal test statistic for each SNP in a gene).

cor_mat d*d matrix of the correlations between all the test statistics in the set, where d is the total number of test statistics in the set. You only need to specify EITHER cor_mat OR pairwise_cors.

pairwise_cors A vector of all d(d-1)/2 pairwise correlations between the test statistics. You only need to specify EITHER cor_mat OR pairwise_cors.

Value

A list with the elements:

BJ The observed Berk-Jones test statistic.

BJ_pvalue The p-value of this observed value, given the size of the set and correlation structure.
Description

Starting with individual-level data on p factors, generate score test statistics for each factor for input into GBJ/GHC/HC/BJ/minP. Also get the correlations between these test statistics. Designed to be used with linear or logistic or log-linear regression null models.

Usage

calc_score_stats(null_model, factor_matrix, link_function, P_mat = NULL)

Arguments

null_model An R regression model fitted using glm(). Do not use lm(), even for linear regression!
factor_matrix An n*p matrix with each factor as one column. There should be no missing data.
link_function Either "linear" or "logit" or "log"
P_mat The projection matrix used in calculation may be passed in to speed up the calculation. See paper for details. Default is null.

Value

A list with the elements:

test_stats The p score test statistics.
cor_mat The p*p matrix giving the pairwise correlation of every two test statistics.

Examples

Y <- rbinom(n=100, size=1, prob=0.5)
null_mod <- glm(Y~1, family=binomial(link="logit"))
factor_mat <- matrix(data=rnorm(n=100*5), nrow=100)
calc_score_stats(null_mod, factor_mat, "logit")
estimate_ss_cor

**Description**

Estimate the correlations between GWAS summary statistics using reference panel eigenvectors and reference panel genotypes.

**Usage**

```r
estimate_ss_cor(ref_pcs, ref_genotypes, link_function)
```

**Arguments**

- `ref_pcs`: An n*m matrix containing PCs calculated from the reference panel. Here n is the number of subjects in the reference panel and m is roughly the number of PCs used in the original analysis which produced the summary statistics.
- `ref_genotypes`: An n*d matrix holding the genotypes from the reference panel, where the d columns correspond to the d SNPs for which we have summary statistics. No missing data allowed.
- `link_function`: Either "linear" or "logit" or "log".

**Value**

A list with the elements:

- `cor_mat`: The d*d matrix giving the pairwise correlation of every two test statistics.

**Examples**

```r
ref_pcs <- matrix(data=runif(n=1000, min=-0.2, max=0.2), ncol=5)
ref_genotypes <- matrix(data=rbinom(n=2000, size=2, prob=0.3), ncol=10)
estimate_ss_cor(ref_pcs=ref_pcs, ref_genotypes=ref_genotypes, link_function="linear")
```

---

**FGFR2**

*Genotypes at FGFR2 SNPs for subjects from 'GBR' population in the 1000 Genomes Project.*

**Description**

A dataset containing the genotypes (number of minor alleles) for each of 91 subjects from the 'GBR' population in the 1000 Genomes Projects. There are 64 SNPs documented here, all residing in the FGFR2 gene.

**Usage**

```r
data(FGFR2)
```
Format

A matrix with 91 rows (one for each subject) and 64 columns (one for each SNP)

Source

https://www.internationalgenome.org/data

Description

Calculate the Generalized Berk-Jones test statistic and p-value.

Usage

GBJ(test_stats, cor_mat = NULL, pairwise_cors = NULL)

Arguments

test_stats Vector of test statistics for each factor in the set (i.e. marginal test statistic for each SNP in a gene).
cor_mat d*d matrix of the correlations between all the test statistics in the set, where d is the total number of test statistics in the set. You only need to specify EITHER cor_mat OR pairwise_cors.
pairwise_cors A vector of all d(d-1)/2 pairwise correlations between the test statistics. You only need to specify EITHER cor_mat OR pairwise_cors.

Value

A list with the elements:

GBJ The observed Generalized Higher Criticism test statistic.
GBJ_pvalue The p-value of this observed value, given the size of the set and correlation structure.
err_code Sometimes if your p-value is very small (<10^-12 usually), R/C++ do not have enough precision in their standard routines to calculate the number accurately. In these cases (and very rarely others) we switch to standard Berk-Jones instead (more stable numerically) and let you know with a message here.

Examples

# Should return statistic = 0.9248399 and p_value = 0.2670707
set.seed(100)
Z_vec <- rnorm(5) + rep(1,5)
cor_Z <- matrix(data=0.2, nrow=5, ncol=5)
diag(cor_Z) <- 1
GBJ(test_stats=Z_vec, cor_mat=cor_Z)
### GBJ_pvalue

**Description**

Calculate the p-value for the Generalized Berk-Jones (GBJ) statistic.

**Usage**

```r
GBJ_pvalue(observed_gbj, d, pairwise_cors, times_to_try = 5)
```

**Arguments**

- **observed_gbj**: The observed value of the GBJ statistic.
- **d**: The number of test statistics in the set.
- **pairwise_cors**: A vector of all \( \frac{d(d-1)}{2} \) pairwise correlations between the test statistics, where \( d \) is the total number of test statistics in the set.
- **times_to_try**: Sometimes the numerical root-finder is finnicky, so we have to give it extra chances to try and calculate the p-value if first time is failure. Recommend setting this parameter to 5.

**Value**

The p-value of the GBJ test.

**Examples**

```r
GBJ_pvalue(observed_gbj=2, d=5, pairwise_cors=rep(0.2,10))
```

### gbr_pcs

**Description**

A dataset containing 5 simulated Principal Components (PCs) for each of 91 subjects from the ‘GBR’ population in the 1000 Genomes Projects. These would normally be used as covariates in a regression model to control for population stratification.

**Usage**

```r
data(gbr_pcs)
```

**Format**

A matrix with 91 rows (one for each subject) and 5 columns (one for each PC)
Description

Calculate the Generalized Higher Criticism test statistic and p-value.

Usage

\[ \text{GHC}(\text{test\_stats}, \text{cor\_mat} = \text{NULL}, \text{pairwise\_cors} = \text{NULL}) \]

Arguments

test\_stats Vector of test statistics for each factor in the set (i.e. marginal test statistic for each SNP in a gene).
cor\_mat \(d \times d\) matrix of the correlations between all the test statistics in the set, where \(d\) is the total number of test statistics in the set. You only need to specify EITHER cor\_mat OR pairwise\_cors.
pairwise\_cors A vector of all \(d(d-1)/2\) pairwise correlations between the test statistics. You only need to specify EITHER cor\_mat OR pairwise\_cors.

Value

A list with the elements:

GHC The observed Generalized Higher Criticism test statistic.
GHC\_pvalue The p-value of this observed value, given the size of the set and correlation structure.
err\_code Sometimes if your p-value is very small (>10^(-12) usually), R/C++ do not have enough precision in their standard routines to calculate the number accurately. In these cases (and very rarely others) we switch to standard Higher Criticism instead (more stable numerically) and let you know with a message here.

Examples

\[ \text{set\_seed}(100) \]
\[ \text{Z\_vec} <- \text{rnorm}(5) \]
\[ \text{cor\_Z} <- \text{matrix}(\text{data}=0.2, \text{nrow}=5, \text{ncol}=5) \]
\[ \text{diag(cor\_Z)} <- 1 \]
\[ \text{GHC(test\_stats=Z\_vec, cor\_mat=cor\_Z)} \]
Description

Calculate the Higher Criticism test statistic and p-value.

Usage

\[
\text{HC}(\text{test_stats}, \text{cor_mat} = \text{NULL}, \text{pairwise_cors} = \text{NULL})
\]

Arguments

- **test_stats**: Vector of test statistics for each factor in the set (i.e. marginal test statistic for each SNP in a gene).
- **cor_mat**: \(d \times d\) matrix of the correlations between all the test statistics in the set, where \(d\) is the total number of test statistics in the set. You only need to specify EITHER cor_mat OR pairwise_cors.
- **pairwise_cors**: A vector of all \(d(d-1)/2\) pairwise correlations between the test statistics. You only need to specify EITHER cor_mat OR pairwise_cors.

Value

A list with the elements:

- **HC**: The observed Higher Criticism test statistic.
- **HC_pvalue**: The p-value of this observed value, given the size of the set and correlation structure.

Examples

```r
# Should return statistic = 2.067475 and p.value = 0.2755146
set.seed(100)
Z_vec <- rnorm(5) + rep(1,5)
cor_Z <- matrix(data=0.2, nrow=5, ncol=5)
diag(cor_Z) <- 1
HC(test_stats=Z_vec, cor_mat=cor_Z)
```
Description

Given a vector of individual test statistics and their pairwise correlations, calculate the MinimumP (see Conneely and Boehnke, 2007) second-level test statistic and it’s p-value.

Usage

minP(test_stats, cor_mat = NULL, pairwise_cors = NULL)

Arguments

test_stats    Vector of test statistics for each factor in the set (i.e. marginal test statistic for each SNP in a gene)
cor_mat       d*d matrix of the correlations between all the test statistics in the set, where d is the total number of test statistics in the set. You only need to specify EITHER cor_mat OR pairwise_cors.
pairwise_cors A vector of all d(d-1)/2 pairwise correlations between the test statistics. You only need to specify EITHER cor_mat OR pairwise_cors.

Value

A list with the elements:

minP                  The observed MinimumP test statistic.
minP_pvalue           The p-value of this observed value, given the size of the set and correlation structure.

Examples

# Should return statistic = 0.05918928 and p_value = 0.2525972.
set.seed(100)
Z_vec <- rnorm(5) + rep(1,5)
cor_Z <- matrix(data=0.2, nrow=5, ncol=5)
diag(cor_Z) <- 1
minP(test_stats=Z_vec, cor_mat=cor_Z)
OMNI_individual

Description

Computes the omnibus test statistic combining GBJ, GHC, minP, and SKAT. This version of the function assumes you have the individual factor data (i.e. genotypes) for each subject. If you only have summary statistics, use omni_ss(). You WILL NOT be able to use this function unless you have also loaded the SKAT package (install.packages("SKAT"); library(SKAT)).

Usage

OMNI_individual(null_model, factor_matrix, link_function, num_boots = 100)

Arguments

null_model An R regression model fitted using glm(). Do not use lm(), even for linear regression!

factor_matrix An n*d matrix with each factor (i.e. each SNP) as one column. There should be no missing data.

link_function Either "linear" or "logit" or "log".

num_boots Number of bootstrap repetitions to find correlation matrix of set-based statistics.

Value

A list with the elements:

OMNI The observed omnibus test statistic.

OMNI_pvalue The p-value of the OMNI test

err_code Sometimes if your p-value is very small (< 1*10^-10)), R may run into numerical issues. This message will alert you if such a situation occurs.

Examples

factor_matrix <- matrix(data=rbinom(n=1000, size=2, prob=0.3), ncol=5)
Y <- rnorm(n=200)
null_mod <- glm(Y ~ 1)
OMNI_individual(null_model=null_mod, factor_matrix=factor_matrix, link_function='linear', num_boots=5)
Description

Computes the omnibus test statistic combining GBJ, GHC, minP, and SKAT. This version of the function assumes you are using GWAS summary statistics. If you individual-level genotype data, use omni_individual().

Usage

OMNI_ss(test_stats, cor_mat, num_boots = 100)

Arguments

test_stats Vector of test statistics for each factor in the set (i.e. marginal test statistic for each SNP in a gene)
cor_mat d*d matrix of the correlations between all the test statistics in the set, where d is the total number of test statistics in the set.
num_boots Number of bootstrap repetitions to find correlation matrix of set-based statistics.

Value

A list with the elements:

OMNI The observed omnibus test statistic.
OMNI_pvalue The p-value of the OMNI test
err_code Sometimes if your p-value is very small (< 1*10^(-10)), R may run into numerical issues. This message will alert you if such a situation occurs.

Examples

cor_mat <- matrix(data=0.3, nrow=5, ncol=5)
diag(cor_mat) <- 1
test_stats <- as.numeric(mvtnorm::rmvnorm(n=1, sigma=cor_mat))
OMNI_ss(test_stats=test_stats, cor_mat=cor_mat, num_boots=5)
Description

Starting with individual-level data on p factors, generate score test statistics for each factor for input into GBJ/GHC/HC/BJ/minP. DOES NOT get the correlations (assumed known).

Usage

score_stats_only(null_model, factor_matrix, link_function, P_mat = NULL)

Arguments

null_model An R regression model fitted using glm(). Do not use lm(), even for linear regression!
factor_matrix An n*d matrix with each factor as one column. There should be no missing data.
link_function Either "linear" or "logit" or "log".
P_mat The projection matrix used in calculation may be passed in to speed up the calculation. See paper for details. Default is null.

Value

The d score test statistics.

Examples

Y <- rbinom(n=100, size=1, prob=0.5)
null_mod <- glm(Y~1, family=binomial(link="logit"))
factor_matrix <- matrix(data=rnorm(n=100*5), nrow=100)
score_stats_only(null_mod, factor_matrix, "logit")

Description

Survival (1 minus the CDF) function of standard normal random variable.

Usage

surv(x)

Arguments

x Vector of quantiles
Value

Probability that a standard normal random variable is greater than \( x \).

Examples

\[
surv(0) \ # \ Should \ return \ 0.5
\]
Index

* datasets
  FGFR2, 4
  gbr_pcs, 6

BJ, 2

calc_score_stats, 3
estimate_ss_cor, 4

FGFR2, 4

GBJ, 5
GBJ_pvalue, 6
gbr_pcs, 6
GHC, 7

HC, 8

minP, 9

OMNI_individual, 10
OMNI_ss, 11

score_stats_only, 12
surv, 12