Package 'MORST'

December 9, 2023

Type Package

Title Minimax Optimal Ridge-type Set Test

Version 0.9.0

Author Yaowu Liu

Maintainer Yaowu Liu <yaowuliu615@gmail.com>

Description This package implemented MORST, a fast and powerful test that is designed to be power-robust to the strength of signals. While mainly motivated from the linear regression setting, MORST is a generic test and has several versions. This package implemented the score version of MORST in GLMs and a specific function of MORST tailored to genetic association studies. At the core of MORST, it is the choice of the ridge-type parameter tau_c, which is implemented in this package. With tau_c, one can easily develop other versions of MORST or adapt MORST to other applications. In addition, this package also implemented the Burden, SKAT and ACAT-V tests, and their respective ensemble versions, e.g., the ensemble Burden test.

Depends R (>= 3.2.0)

License GPLv3

Encoding UTF-8

LazyData true

Imports Rcpp (>= 1.0.1), Matrix

LinkingTo Rcpp

RoxygenNote 7.1.2

R topics documented:

`EnsembleSetTests` *Ensemble Set based tests for testing the association between a set of genetic variants and a phenotype*

Description

Calculate the ensemble Burden, ensemble SKAT, and ensemble MORST p-values under one or multiple sets of weights in GLMs or GLMMs.

Usage

```
EnsembleSetTests(
  G,
  obj,
  B = 100,
  tests = c("Burden", "SKAT", "MORST"),
  weights.prior = "none",
  maf.beta.para = c(1, 25),
  is.pvals.path = FALSE,
  alpha = 1e-06,
  tau.type = "approx",
  target_power = 0.5,
  n.points = 50
)
```

Arguments

`G`	a numeric matrix or dgCMatrix with each row as a different individual and each column as a separate gene/snp. Each genotype should be coded as 0, 1, 2.
`obj`	an output from `Null_model_glm` or the function `glmmkin` from the `GMMAT` package. See details.
`B`	the number of base tests.
`tests`	the names of set-based tests. It could only be one or several of "Burden", "SKAT" and "MORST".
`weights.prior`	a character or a numeric vector. It is used to specify the standardized deviations (SDs) of normal variables when generating the random weights. If it is character, it could only be *"none"*, which means the SDs are all equal, or *"beta"*, which means the maf-based beta weights will be used as the SDs and the parameters of beta weights are specified by *maf.beta.para*. If it is a numeric vector, it is the user-specified SDs and hence must have a length equal to ncol(G).
`maf.beta.para`	a vector of parameters for the beta weights for the weighted kernels. It is only used when *weights.prior == "beta"*.
`is.pvals.path`	logical. If *is.pvals.path == TRUE*, the p-value path of the ensemble test as *B* increases will be provided in the output and can be used to draw the ensemble p-value path plot (See examples).
`alpha`	the alpha parameter in MORST. It is suggested to be the significance level.
`tau.type`	either "minimax" or "approx". See `tau_c` for details.
`target_power`	a value that is used when *tau.type == "approx"*. See `tau_c` for details.
`n.points`	number of grid points used when *tau.type == "minimax"*. See `tau_c` for details.

Details

If you want to fit a GLM, please use the `Null_model_glm` function to obtain the null model *obj*. If you have a kinship matrix/GRM and would like to fit a GLMM, please use the the function `glmmkin` from the the `GMMAT` package. Both dense and sparse kinship/GRM can be used.

While the *alpha* parameter is suggested to be the significance level, practically there is no need to set *alpha* less than 1e-08. In most situations, the MORST p-values would only have negligible difference for values of *alpha* less than 1e-06. A super small *alpha* could slow down the computation and might cause some numerical issue. Therefore, the default value for *alpha* is 1e-06.

Value

A list with components:

`pval.ensemble.test` The p-value of the ensemble test

`pval.base.test` The p-values of the individual base tests

`pval.path.ensemble` The p-value path of the ensemble test when *is.pvals.path* == *TRUE*.

Author(s)

Yaowu Liu

References

Liu, Y., Liu, Z., and Lin, X. (2024+) Ensemble methods for testing a global null. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. Accepted.

Examples

```
library(Matrix)
data(Geno)
G<-Geno[,1:100] # Geno is a dgCMatrix of genotypes
Y<-rnorm(nrow(G)); Z<-matrix(rnorm(nrow(G)*4),ncol=4)
obj<-Null_model_glm(Y,Z,family="gaussian")

### The number of base tests B = 10.
EnsembleSetTests(G,obj,B=10)

### Use the beta prior, and only do Burden test
EnsembleSetTests(G,obj,B=20,tests = "Burden",weights.prior = "beta")

### User-specified weights prior
EnsembleSetTests(G,obj,B=20,tests = c("Burden","SKAT"),weights.prior = runif(ncol(G)))

### plot the p-value path of the ensemble tests as the number of base tests B increases
set.seed(123); Y<- rowMeans(G)*4 + rnorm(nrow(G))
Z<-matrix(rnorm(nrow(G)*4),ncol=4);obj<-Null_model_glm(Y,Z,family="gaussian")
res <- EnsembleSetTests(G,obj,B=500,tests = c("Burden","SKAT"),is.pvals.path = TRUE,weights.prior = "beta"
test.name = "Burden"; pvals.path <- res[["pval.path.ensemble"]][,test.name]
plot(c(1:length(pvals.path)), -log10(pvals.path),xlab = "Number of base tests",ylab = "-log10(p-value)",ty
```

`Get_Q_pval` *Get the p-value of a quadratic test*

Description

The survival function (i.e., p-value) of the weighted sum of i.i.d. Chi-squared variables with df=1 (i.e., a quadratic test). A hybrid of davies's method, saddle point method, and liu's method is used.

Usage

```
Get_Q_pval(Q, w)
```

Arguments

`Q`	The value of the quadratic test statistic. It must be positive.
`w`	a numeric vector of positive weights. See details.

Details

Compute $P[T > Q]$, where $T = \sum_{i=1}^{k} w_i X_i$, X_i's are i.i.d. Chi-squared variables with df=1, and w_i's are the elements of *w*.

A hybrid of several methods is used to acheive both accuracy and computation effeciency. Specifically, when the p-value > 1e-10, the davies' method is used; when 1e-10 < p-value < 1e-15, the saddle point method is used; when p-value < 1e-15, liu's method is used.

Value

The p-value of the quadratic test.

Author(s)

Yaowu Liu

References

Davies, R. B. (1980). Algorithm AS 155: The distribution of a linear combination of X2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3):323-333.

Kuonen, D. (1999). Miscellanea. saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, 86(4):929-935.

Liu, H., Tang, Y., and Zhang, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*,53(4):853-856.

Examples

```
Get_Q_pval(50,seq(1,10,1))
Get_Q_pval(200,seq(1,10,1))
Get_Q_pval(1000,seq(1,10,1))
```

MORST_glm *The score version of MORST for Generalized Linear Models.*

Description

Calculate the score version MORST p-value for GLM. For genetic association study, please use the function SetBasedTests.

Usage

```
MORST_glm(
  X,
  obj,
  alpha = 0.05,
  weights = NULL,
  tau.type = "approx",
  target_power = 0.5,
  n.points = 50
)
```

Arguments

X	a numeric matrix or dgCMatrix of predictors.
obj	an output from Null_model_glm.
alpha	the alpha parameter in MORST. It is suggested to be the significance level.
weights	a numeric vector of nonnegative weights. If NULL, the equal weight is used.
tau.type	either "minimax" or "approx". See tau_c for details.
target_power	a value that is used when *tau.type* == *"approx"*. See tau_c for details.
n.points	number of grid points used when *tau.type* == *"minimax"*. See tau_c for details.

Value

The p-value of MORST.

Author(s)

Yaowu Liu

References

Liu, Y., Li, Z., and Lin, X. (2020+) A Minimax Optimal Ridge-Type Set Test for Global Hypothesis with Applications in Whole Genome Sequencing Association Studies. *Journal of the American Statistical Association*. Accepted.

Examples

```
X<-matrix(rnorm(20000),ncol=20); Z=matrix(rnorm(nrow(X)*4),ncol=4)

### linear regression for continuous outcome
Y<-rnorm(nrow(X));obj<-Null_model_glm(Y,Z,family="gaussian")
MORST_glm(X,obj)

### Logistic regression for binary outcome
Y<-rbinom(nrow(X),1,0.4);obj<-Null_model_glm(Y,Z,family="binomial")
MORST_glm(X,obj,alpha = 1e-04)

### Binomial outcome
Y<-rbinom(nrow(X),5,0.4);Y<-cbind(Y,5-Y);obj<-Null_model_glm(Y,Z,family="binomial")
MORST_glm(X,obj,weights=runif(ncol(X)))

### Poisson outcome
Y<-rpois(nrow(X),5);obj<-Null_model_glm(Y,Z,family="poisson")
MORST_glm(X,obj,alpha = 1e-04,weights=runif(ncol(X)),tau.type = "minimax")
```

`Null_model_glm` *Fit the GLM null model for MORST*

Description

fit a GLM null model

Usage

```
Null_model_glm(Y, Z, family = "gaussian", include_intercept = TRUE)
```

Arguments

`Y`	a numeric vector of outcomes.
`Z`	a numeric matrix of covariates that need to be adjusted.
`include_intercept`	logical. If `TRUE`, the intercept will be included in the null model.
`famlily`	a character. Should be "gaussian", "binomial" or "poisson". For each famliy, the cannonical link is used.

Value

This function returns an object that has model parameters and residuals of the NULL model of no association between outcomes Y and predictors X after adjusting for covariates Z. After obtaining it, please use `MORST_glm` or `SetBasedTests` to conduct the association test.

Author(s)

Yaowu Liu

Examples

```
X<-matrix(rnorm(20000),ncol=20); Z=matrix(rnorm(nrow(X)*4),ncol=4)

### linear regression for continuous outcome
Y<-rnorm(nrow(X));obj<-Null_model_glm(Y,Z,family="gaussian")

### Logistic regression for binary outcome
Y<-rbinom(nrow(X),1,0.4);obj<-Null_model_glm(Y,Z,family="binomial")

### Binomial outcome
Y<-rbinom(nrow(X),5,0.4);Y<-cbind(Y,5-Y);obj<-Null_model_glm(Y,Z,family="binomial")

### Poisson outcome
Y<-rpois(nrow(X),5);obj<-Null_model_glm(Y,Z,family="poisson")
```

`SetBasedTests` *Set based tests for testing the association between a set of genetic variants and a phenotype*

Description

Calculate the Burden, SKAT, ACAT-V and MORST p-values under one or multiple sets of weights in GLMs or GLMMs.

Usage

```
SetBasedTests(
  G,
  obj,
  alpha = 1e-06,
  weights.beta = matrix(c(1, 25, 1, 1), nrow = 2),
  weights = NULL,
  tau.type = "approx",
  target_power = 0.5,
  n.points = 50,
  mac.thresh = 10
)
```

Arguments

`G`	a numeric matrix or dgCMatrix with each row as a different individual and each column as a separate gene/snp. Each genotype should be coded as 0, 1, 2.
`obj`	an output from `Null_model_glm` or the function `glmmkin` from the `GMMAT` package. See details.
`alpha`	the alpha parameter in MORST. It is suggested to be the significance level.
`weights.beta`	a numeric vector/matrix of parameters for the beta weights for the weighted kernels. If it is a matrix, each column corresponds to one set of the beta-weights parameters. If you want to use your own weights, please use the "weights" parameter. It will be ignored if "weights" parameter is not null.

weights	a numeric vector/matrix of weights for the SNPs. If it is a matrix, each column corresponds to one set of weights. When it is NULL, the beta weight with the "weights.beta" parameter is used.
tau.type	either "minimax" or "approx". See tau_c for details.
target_power	a value that is used when *tau.type* == *"approx"*. See tau_c for details.
n.points	number of grid points used when *tau.type* == *"minimax"*. See tau_c for details.
mac.thresh	a threshold of the minor allele count (MAC) that is used in the ACAT-V test. SNPs with MAC less than this threshold will be first aggregated by the Burden test in ACAT-V.

Details

If you want to fit a GLM, please use the Null_model_glm function to obtain the null model *obj*. If you have a kinship matrix/GRM and would like to fit a GLMM, please use the the function glmmkin from the the GMMAT package. Both dense and sparse kinship/GRM can be used.

The ACAT-V p-value might be slightly different from the result from the ACAT_V function in the ACAT package. This is because the variant-level p-values are calculated using slightly different methods.

While the *alpha* parameter is suggested to be the significance level, practically there is no need to set *alpha* less than 1e-08. In most situations, the MORST p-values would only have negligible difference for values of *alpha* less than 1e-06. A super small *alpha* could slow down the computation and might cause some numerical issue. Therefore, the default value for *alpha* is 1e-06.

Value

The p-values of Burden, SKAT, ACAT-V and MORST under under one or multiple choices of weights.

Author(s)

Yaowu Liu

Examples

```
library(Matrix)
data(Geno)
G<-Geno[,1:100] # Geno is a dgCMatrix of genotypes
Y<-rnorm(nrow(G)); Z<-matrix(rnorm(nrow(G)*4),ncol=4)
obj<-Null_model_glm(Y,Z,family="gaussian")
SetBasedTests(G,obj)
```

---

tau_c *The choice of tau in MORST*

Description

Calculate the parameter tau in MORST based on either the minimax criterion or an approximation of the minimax solution

Usage

```
tau_c(eg.values, alpha, tau.type = "approx", target_power = 0.5, n.points = 50)
```

Arguments

eg.values	a numeric vector of non-negative eigenvalues.
alpha	the alpha parameter in MORST. It is suggested to be the significance level.
tau.type	either "minimax" or "approx". If *tau.type == "minimax"*, tau is calculated based on the minimax criterion; if *tau.type == "approx"*, tau is an approximation to the minimax tau. Default value is "approx".
target_power	a value that is used when *tau.type == "approx"*. See details.
n.points	number of grid points used when *tau.type == "minimax"*. Should be at least 20 to have reasonable accuracy.

Details

The approximation method is substantially faster than the minimax method and will be introduced soon.

Value

the parameter tau in MORST

Author(s)

Yaowu Liu

References

Liu, Y., Li, Z., and Lin, X. (2020+) A Minimax Optimal Ridge-Type Set Test for Global Hypothesis with Applications in Whole Genome Sequencing Association Studies. *Journal of the American Statistical Association*. Accepted.

Examples

```
tau_c(seq(1,100,1),1e-04)
tau_c(seq(1,100,1),1e-04,tau.type = "minimax")
tau_c(c(2,rep(0.2,20)),0.05,tau.type = "minimax",n.points = 200)
```

Index