mvtests Software Documentation

June 9, 2020

Version 0.3
Date June 9, 2020
Title Multivariate Tests of Association for Multiple Phenotypes
Correspondence Debashree Ray, Ph.D. `<dray@jhu.edu>`
Description mvtests package is meant to be a suite of functions implementing single-variant genetic association test of multiple phenotypes. Currently, it implements a cross-phenotype test using proportional odds regression model of genotype on phenotypes, and a minimum p-value test based on Nyholt-Šidák correction for multiple traits.
Depends MASS, lmtest, R (>= 3.0.1)

`grabWarnings` *To grab warnings as objects from function implementations.*

Description

The R function `grabWarnings` takes all the warning messages from a function implementation and outputs a list with elements `output` (containing the output from the function implementation), and `warnings` (containing a list of warning messages). When implementing the multi-trait analyses in this package for millions of variants, it is not possible to keep track of the warning messages that are printed in the log file. This function allows the user to collect these warning messages as an object, which can be saved in a separate file for later use.

Usage

```
grabWarnings(expr)
```

Arguments

`expr` Any R function implementation from this package or otherwise.

Value

`output` The output that one would obtain from the function implementation.

`warnings` A list of all warning messages that arise during the function implementation.

Source

A user-suggested function in `https://stackoverflow.com/questions/3903157/`.

`minP.sidak` *Minimum p-value corrected for multiple tests using Šidák's approach.*

`level.sidak` *Šidák corrected significance threshold for multiple tests.*

Description

The R function `minP.sidak` implements the minimum p-value approach based on p-values from multiple tests while `level.sidak` provides multiple-test corrected significance threshold. Šidák multiple-testing correction is used, which requires the number of independent tests. Nyholt's approach for estimating the approximate number of independent tests is used for correlated tests.

Usage

```
minP.sidak(P, R=NULL, method="Nyholt")
level.sidak(level=5e-8, R=NULL, method="Nyholt")
```

Arguments

`P` A vector of p-values from $K(> 1)$ multiple tests.

`level` Uncorrected significance threshold to be used for the tests. Default value is 5×10^{-8}, the genome-wide significance threshold.

`R` A $K \times K$ correlation matrix of the variables whose association tests gave the p-values in `P`. Default value is `NULL`.

`method` Either the Nyholt method for correlated tests (`Nyholt`) or the usual Šidák correction for independent tests (`independent`) is used for testing genetic association of the traits. Default value is `Nyholt`.

Value

`minP` The minimum of the Šidák corrected p-values.

`level.corrected` The Šidák corrected significance threshold `level`.

`K` Total number of traits/tests.

`K.indep` (If `method="Nyholt"`) Approximate number of independent tests based on Nyholt's approach.

`method` The user-input choice of method used for the Šidák multiple-testing correction.

References

Nyholt, D.R. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet*, 74(4):765-769, 2004.

Šidák, Z. K. Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc*, 62(318):626-633, 1967.

`pom` *Genetic association test of multiple phenotypes using proportional odds model*

Description

POM uses a proportional odds model for testing association between a genetic marker and multiple traits using individual-level genotype phenotype data. The traits may be continuous and/or binary, correlated and/or independent. One may use likelihood ratio test (LRT) or the Wald test. The R function `pom` implements this association test.

Usage

```
pom(Y, X, COV=NULL, test.method="LRT",
    msg.mute=FALSE, no.format.check=FALSE, ...)
```

Arguments

`Y` The $n \times K$ phenotype matrix, where n is the number of individuals and K is the number of phenotypes. The joint association of all K phenotypes with the single marker will be tested. `Y` needs to be in R data frame format.

`X` The $n \times 1$ column matrix for the single genetic marker, where n is the number of individuals. `X` needs to be in R data frame format.

`COV` The $n \times q$ matrix of covariates that need to be adjusted in the model. q is the number of such covariates. `COV` needs to be in R data frame format. The default value is `NULL` assuming there is no covariate in the model.

`test.method` Either the likelihood ratio test (`LRT`) or the Wald test (`Wald`) is used for testing genetic association of the traits. Default value is `LRT`.

`msg.mute` Default value is `FALSE`, which allows messages to print when the analysis is in progress.

`no.format.check` Default value is `FALSE`, which allows the code to check if the input parameters conform to the required format.

`...` Additional arguments to be passed to `polr()` (used when proportional odds model framework is used) or `glm()` (used when logistic model framework is used).

Details

For testing joint association of multiple phenotypes $\boldsymbol{Y}$ ($n \times K$ matrix) and a single genetic marker $\boldsymbol{X}$ ($n \times 1$ matrix), one can consider a reverse regression approach: regressing $\boldsymbol{X}$ on $\boldsymbol{Y}$ assuming a proportional odds model (POM). For a given individual with genotype X (taking values $0, 1$ or 2) and phenotypes $\boldsymbol{y}$ (binary and/or continuous) and no covariate (for simplicity), the assumed model is

$$\text{logit}\left(\mathbb{P}(X \leq j|\boldsymbol{y})\right) = \alpha_j + \boldsymbol{\beta}'\boldsymbol{y},\ \ j = 0, 1$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_K)'$ are the genetic effects of interest. This model can also be written as

$$\begin{aligned}
\mathbb{P}\left(X = 0|\boldsymbol{y}\right) &= \frac{1}{1 + e^{-\alpha_0 - \sum_{k=1}^{K}\beta_k y_k}} \\
\mathbb{P}\left(X = 1|\boldsymbol{y}\right) &= \frac{1}{1 + e^{-\alpha_1 - \sum_{k=1}^{K}\beta_k y_k}} - \frac{1}{1 + e^{-\alpha_0 - \sum_{k=1}^{K}\beta_k y_k}} \\
\mathbb{P}\left(X = 2|\boldsymbol{y}\right) &= \frac{1}{1 + e^{\alpha_1 + \sum_{k=1}^{K}\beta_k y_k}}
\end{aligned}$$

For a genetic variant with low allele frequency, it is possible for X to take only 2 values (instead of 3). In that case, a logistic regression is used.

The null hypothesis of no association of the genetic marker (say, single nucleotide polymorphism or SNP) with the phenotypes is $H_0 : \beta_1 = ... = \beta_K = 0$. The `test.method="LRT"` (default) implements likelihood ratio test to test H_0 against the alternative hypothesis $H_a : \beta_k \neq 0$ for at least one k. Under the null H_0, the LRT statistic has an asymptotic chi-squared distribution with K degrees of freedom.

When `test.method="Wald"`, the test statistic is $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\Sigma}}_\beta\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\Sigma}}_\beta$ is the estimated variance-covariance matrix of estimates $\hat{\boldsymbol{\beta}}$. Under the null H_0, the Wald test statistic, too, has an asymptotic chi-squared distribution with K degrees of freedom. Additionally, a version of the Wald test using F distribution is implemented using the `lmtest` R package, which requires fitting the full model as well as the null model (model under H_0).

We request that the reference for Ray and Chatterjee (2020) be cited if this software is used in any publication. In the Ray and Chatterjee (2020) article, we explored advantages and pitfalls of some of the currently used single-variant cross-phenotype methods in genome-wide analysis of rare, low-frequency and common variants when the basic assumption of multivariate normality is not satisfied. Based on our findings, we recommend extra caution when applying cross-phenotype association tests in GWAS with low-frequency or rare variants due to possible violation of multivariate normality assumption. However, we found that robust association testing is still possible for variants with minor allele count (MAC) > 30 by applying the POM-LRT approach (`pom()` with `test.method="LRT"`) when individual-level data are available. Our recommendation is based on an MAC threshold (instead of a minor allele frequency - MAF - threshold as is commonly used) because we found consistent type I error calibration of methods when the MAC is kept constant.

Value

`coef` The estimated coefficients in the model (e.g., α_0, α_1, β_1,...,β_K). If `COV` is not `NULL`, estimated coefficients for covariates are also included.

`SE.coef` The estimated standard errors of the coefficient estimates.

`stat.lrt` (If `test.method="LRT"`) The LRT statistic for testing the null hypothesis H_0.

`df.lrt` (If `test.method="LRT"`) The degrees of freedom of the asymptotic null distribution of the LRT statistic.

`pvalue.lrt` (If `test.method="LRT"`) The p-value from testing H_0 using LRT.

`stat.wald.chisq` (If `test.method="Wald"`) The Wald test statistic for testing the null hypothesis H_0.

`df.wald.chisq` (If `test.method="Wald"`) The degrees of freedom of the asymptotic chi-squared distribution of the Wald test statistic under H_0.

`pvalue.wald.chisq` (If `test.method="Wald"`) The p-value from testing H_0 using Wald test assuming chi-squared distribution.

`stat.wald.F` (If `test.method="Wald"`) The Wald test statistic for testing the null hypothesis H_0 using the F distribution approximation.

`pvalue.wald.F` (If `test.method="Wald"`) The p-value from testing H_0 using Wald test assuming F distribution.

`n.obs` Number of individuals used for testing association. Individuals with missing observations in `Y`, `X` or `COV` are removed.

`geno.dist` The distribution of 0, 1 or 2 minor alleles after removing individuals with missing observations in `Y`, `X` or `COV`.

`error.msg` Saves any error message that arises during the analysis. If no error is encountered, message `"OK"` is returned.

Reference

Ray, D. and Chatterjee, N. Effect of Non-Normality and Low Count Variants on Cross-Phenotype Association Tests in GWAS. *Eur J Hum Genet*, 28(3):300-312, 2020.

Example

```
source("mvtests_v0.3.R")
set.seed(1)
# simulate 2 phenotypes on 1000 individuals
Y<-mvrnorm(n=1000, mu=c(0,0), Sigma=matrix(c(1,0.3,0.3,1),2,2))
# simulate a single marker for 1000 individuals
X<-matrix(rbinom(n=1000, size=2, prob=0.2), ncol=1)    # additive model
# required data-frame formats
Y<-as.data.frame(Y)
X<-as.data.frame(X)
# unique column names for the data-frames
colnames(Y)<-paste("Y",1:2,sep="")
colnames(X)<-"X"
## apply POM to test association of X with Y1 and Y2
out1<-pom(Y=Y, X=X, COV=NULL, test.method="LRT")
out1
out2<-pom(Y=Y, X=X, COV=NULL, test.method="Wald")
out2
## optimization parameters may be changed
## (e.g., when convergence issues come up)
out3<-pom(X=X, Y=Y, test.method="Wald", control=list(maxit=1000))
out3

## Nyholt-Sidak correction of single-trait p-values
P<-sapply(1:2, function(i){
                lmout <- lm(as.matrix(Y[,i])~as.matrix(X))
                summary(lmout)$"coefficients"[2,4]
                })    # vector of single-trait p-values
R<-cor(Y) # trait-correlation matrix
minP.sidak(P=P, R=R, method="Nyholt")    # Nyholt-Sidak correction
minP.sidak(P=P, method="independent")    # Sidak correction
# Obtaining Nyholt-Sidak corrected threshold
level.sidak(level=5e-8, R=R, method="Nyholt")
# Obtaining Sidak corrected threshold assuming 2 uncorrelated traits
```

```
level.sidak(level=5e-8, R=diag(1,2), method="independent")

## using grabWarnings()
X[which(X==2),1]<-1    # Say, X is such that it only takes 2 values
out1.g <- grabWarnings(pom(Y=Y, X=X, COV=NULL, test.method="LRT"))
out1.g$output    # the output from pom()
identical(out1, out1.g$output)
out1.g$warnings  # the warnings, if any, from implementing pom()
```