metaUSAT Software Documentation

December 11, 2017

Version 1.17

Date December 11, 2017

Title Meta-analysis using Unified Score-based Association Test

Correspondence Debashree Ray, Ph.D. <dray@jhu.edu>

Description metaUSAT uses a data-adaptive weighted score-based test statistic for testing association of multiple phenotypes with a single marker using summary statistics. The association test can be done for one or more studies with or without overlapping samples.

Depends CompQuadForm, minqa, psych, survey, R (>= 3.0.1)

`cor.pearson` *Estimating Pearson correlation matrix of summary statistics*

Description

This function is used to estimate the Pearson correlation matrix of summary statistics to be used in function `metausat`. When two or more studies are considered, the between-study correlations reflect the overlap between studies. This is the recommended approach for estimating `R`.

Usage

```
cor.pearson(Z.matrix, P.matrix, p.threshold=1e-5)
```

Arguments

`Z.matrix` The $K \times p$ matrix of summary statistics of all p genetic variants. If there is a single study, K is the number of traits to be tested from that study. If there are multiple studies, K is the total number of traits across all studies (i.e., $K =$ number of traits per study $\times$ number of studies).

`P.matrix` The $K \times p$ matrix of p-values of all K traits across p genetic variants. Like the summary statistics in `Z.matrix`, the p-values of `P.matrix` correspond to individual test of each trait against each genetic variant. The order in which traits/studies and the genetic variants are arranged in `Z.matrix` and `P.matrix` must be same.

`p.threshold` The p-value threshold used to determine which genetic variants are likely not associated. A liberal threshold needs to be used to screen out any signal that may affect the estimate of the correlation matrix `R`. Genetic variants (here, columns) with p-values smaller than this threshold for any trait are removed before estimating `R`. Default value is 10^{-5}.

Value

`R` The estimated $K \times K$ correlation matrix of the (univariate) summary statistics under the null hypothesis of no association. The `R` needs to be calculated only once.

`cor.tetrachor` *Estimating tetrachoric correlation matrix of summary statistics*

Description

This function provides another approach for estimating `R` using the tetrachoric correlation matrix of summary statistics. Tetrachoric correlation coefficient is more robust to outliers compared to Pearson correlation coefficient.

Usage

```
cor.tetrachor(Z.matrix)
```

Arguments

`Z.matrix` The $K \times p$ matrix of summary statistics of all p genetic variants. If there is a single study, K is the number of traits to be tested from that study. If there are multiple studies, K is the total number of traits across all studies (i.e., $K =$ number of traits per study $\times$ number of studies).

Value

`R` The estimated $K \times K$ correlation matrix of the (univariate) summary statistics. The `R` needs to be calculated only once.

`metausat` *Meta-analysis using Unified Score-based Association Test*

Description

metaUSAT uses a weighted score-based test statistic for testing association between a genetic marker and multivariate trait using summary statistics from individual trait analysis. The traits may be continuous and/or binary, correlated and/or independent, and may be from one or more studies. One can perform meta-analysis of a single trait over multiple studies, or multiple traits over one or more studies (which may include overlapping samples). The R function `metausat` implements this association test.

Usage

```
metausat(Z, R, weights=1, metamanova=FALSE,
    AbsTol=.Machine$double.eps^0.8)
```

Arguments

`Z` The vector of K summary statistics for a given genetic variant. If there is a single study, K is the number of traits to be tested from that study. If there are multiple studies, K is the number of traits $\times$ number of studies (assuming no missing trait). If there is any missing (`NA`) summary statistic, it needs to be removed before applying `metausat`. The joint association of all K phenotypes (across studies) with the single marker will be tested.

`R` The $K \times K$ estimated correlation matrix of the summary statistics. It can be estimated using function `cor.pearson` (based on Pearson's correlation, the recommended method described in Ray and Boehnke, 2017) or using `cor.tetrachor` (based on tetrachoric correlation). `R` needs to be in R matrix format.

`weights` The K vector of weights for the summary statistics `Z`. When sample sizes differ, suitable weights include square-root of sample size of each trait from each study. `weights` needs to be in R vector format. The default value is 1, i.e., it is assumed that all traits from all studies have the same weight of 1. Note that `weights` do not affect metaMANOVA.

`metamanova` If value is `FALSE` (default), metaMANOVA statistic and p-value will not be included in the final output.

`AbsTol` The user can specify the absolute tolerance value used in the numerical integration for evaluating metaUSAT p-value. Default value is 3×10^{-13}. `integrate()` function is used for numerical integration.

Details

Summary statistics for individual traits are often publicly available for various genome-wide association studies (GWAS). For a single GWAS, let $\boldsymbol{Y}_k$ be the $n \times 1$ vector of values for the k-th trait ($k = 1, 2, ..., K$), and $\boldsymbol{X}$ be the $n \times 1$ vector of genotypes. When each trait is tested individually, the assumed model is

$$\boldsymbol{Y}_k = \boldsymbol{\alpha}_k + \beta_k \boldsymbol{X} + \boldsymbol{\epsilon}_k, \;\; \boldsymbol{\epsilon}_k \sim N_n(\boldsymbol{0}, \sigma_k^2 \boldsymbol{I}_n) \;\; \text{for all } k = 1, 2, ..., K$$

for continuous traits, or

$$\text{logit}\big(P(\boldsymbol{Y}_k = 1 | \boldsymbol{X})\big) = \boldsymbol{\alpha}_k + \beta_k \boldsymbol{X} \;\; \text{for all } k = 1, 2, ..., K$$

for binary traits. For the k-th trait, β_k is the genetic effect and the corresponding summary statistic for testing no association is $Z_k = \hat{\beta}_k / \text{se}(\hat{\beta}_k)$, where $\hat{\beta}_k$ is the maximum likelihood estimate (MLE) of β_k and $\text{se}(\hat{\beta}_k)$ is its standard error. For more than one GWAS, we similarly have univariate summary statistic for each trait from each study.

For testing joint association of K traits (the overall number of traits across all studies) and a single genetic marker, we use only the summary statistics $\boldsymbol{Z} = (Z_1, ..., Z_k, ..., Z_K)'$. Under

the null hypothesis of no association (i.e., none of the K traits is associated with the genetic variant), $\boldsymbol{Z}$ has an asymptotic multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix R. To estimate $\boldsymbol{R}$, we use the Z-statistics of all the variants across the genome that are not associated with any of the K traits (i.e., genetic variants with p-values greater than a pre-defined significance threshold, say 10^{-5}, for any trait), and calculate the Pearson correlation.

A test that has been recently used in a few publications is metaMANOVA (meta-analysis version of Multivariate Analysis of Variance), which has a test statistic of the form $\boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}$. This is usually very powerful in detecting association. However, as shown in Ray et al. (2016), Ray and Boehnke (2017), metaMANOVA may lose substantial power in certain situations of association (and phenotypic correlation structure), which is not known apriori. In such situations, marginal association tests perform better than such joint association tests. metaUSAT aims to maximize power by adaptively using the data to combine the metaMANOVA test and a marginal association test SSU[1].

If T_M and T_S are the metaMANOVA and SSU test statistics respectively, we consider a weighted test statistic $T_\omega = \omega T_M + (1-\omega)T_S$, which can be expressed as a linear combination of chi-square variables. Apriori the optimal weight ω is not known. We propose metaUSAT as

$$T_{\text{metaUSAT}} = \min_{0 \leq \omega \leq 1} p_\omega$$

where, for a given $\omega \in [0, 1]$, p_ω is the p-value of the weighted statistic T_ω. For practical purposes, a grid of 11 equispaced ω values are considered: $\{\omega_1 = 0, \omega_2 = 0.1, ..., \omega_{10} = 0.9, \omega_{11} = 1\}$. To find the p-value p_{metaUSAT} of our test statistic T_{metaUSAT}, we need the null distribution of metaUSAT. We propose an approximate p-value calculation using a one-dimensional numerical integration, which makes metaUSAT suitable for application on a genome-wide scale.

For more details on how metaUSAT may be used, please refer Ray and Boehnke (2017). We request that the reference for Ray and Boehnke (2017) be cited if this software is used in any publication.

[1] Kim, J., Bai, Y., and Pan, W. (2015). An adaptive association test for multiple phenotypes with GWAS summary statistics. *Genetic Epidemiology*, 39, 651-663

Value

`T.metausat`	The value of the metaUSAT test statistic (scalar).
`omg.opt`	The optimal weight ω based on a grid search over $[0, 1]$.
`p.metausat`	The p-value of association based on the metaUSAT statistic.
`AbsTol`	The absolute tolerance for the numerical integral used in evaluating metaUSAT p-value.
`error.msg`	The error message from the numerical integral used in evaluating metaUSAT p-value. When p-value if calculated without error, `OK` message is displayed. If p-value is `NA`, an error message concerning absolute tolerance is displayed. In that case, changing `AbsTol` may help.
`T.metamanova`	The value of the metaMANOVA test statistic (scalar). Provided if `metamanova=TRUE`.
`p.metamanova`	The p-value of association based on metaMANOVA statistic. Provided if `metamanova=TRUE`.

References

Ray, D., Boehnke, M. Methods for meta-analysis of multiple traits using GWAS summary statistics. *Genetic Epidemiology*, DOI: 10.1002/gepi.22105, 2017.

Ray, D., Pankow, J.S., Basu, S. USAT: A Unified Score-based Association Test for Multiple Phenotype-Genotype Analysis. *Genetic Epidemiology*, 40(1):20-34, 2016.

Example

```
source("metausat_v1.17.R")
# simulate summary statistics on 2 phenotypes on 1e+6 genetic variants
library(MASS) # needed for multivariate normal simulation
Z.matrix<-mvrnorm(n=1e+6, mu=c(0,0), Sigma=matrix(c(1,0.2,0.2,1),2,2))
# estimate correlation matrix R
# since the p-value matrix is not available, cor.tetrachor is used here
R<-cor.tetrachor(Z.matrix)
## apply metaUSAT to test association with the 1st genetic variant
Z<-Z.matrix[1,]
out<-metausat(Z=Z, R=R, weights=1)
# USAT test statistic and p-value
t<-out$T.metausat
p<-out$p.metausat
```