USAT Software Documentation

December 12, 2016

Version 1.21

Date December 12, 2016

Title Unified Score-based Association Test

Correspondence Debashree Ray, Ph.D. <debashr@umich.edu>;

Saonli Basu, Ph.D. <saonli@umn.edu>

Description USAT uses a data-adaptive weighted score-based test statistic for testing association of multiple continuous phenotypes with a single marker.

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

`usat` *Unified Score-based Association Test*

Description

USAT uses a weighted score-based test statistic for testing association between a genetic marker and multivariate phenotypes. It is designed for unrelated individuals. The R function `usat` implements this association test.

Usage

```
usat(Y, X, COV=NULL, na.check=TRUE, na.check.msg=TRUE,
     manova.out=FALSE, AbsTol=.Machine$double.eps^0.8)
```

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 matrix 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 matrix 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 matrix format. The default value is `NULL`, i.e., it is assumed there is no covariate in the model.

`na.check` If value is `TRUE` (default), the code will check for presence of missing values (coded as `NA`). USAT requires complete observations and any individual with at least one missing value in either `Y`, `X` or `COV` will be removed. Removal of missing observations may substantially reduce the sample size n and hence power to detect association. If a substantial proportion of the individuals have missing data, it is recommended (if possible) to impute the missing values before using USAT.

`na.check.msg` If value is `TRUE` (default), user will receive message updates when presence of missing values (coded as `NA`) is checked.

`manova.out` If value is `FALSE` (default), MANOVA 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 USAT p-values. Default value is 3×10^{-13}. `integrate()` function is used for numerical integration.

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 may use the well-known MANOVA (Multivariate Analysis of Variance) test. MANOVA is usually very powerful in detecting association. However, as shown in Ray et al. (2016), MANOVA 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. USAT maximizes power by adaptively using the data to combine the MANOVA test and a marginal association test SSU[1].

If T_M and T_S are the MANOVA 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 our optimal unified test USAT as

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

where, for a given $\omega \in [0,1]$, p_ω is the p-value of the 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_{USAT} of our test statistic T_{USAT}, we need the null distribution of USAT. We propose an approximate p-value calculation using a one-dimensional numerical integration, which makes USAT suitable for application on a genome-wide scale. USAT took 174 minutes to test $55,775$ single marker associations on a real dataset with $n = 9,964$ individuals, $K = 4$ phenotypes and $q = 3$ covariates using a single core of an Intel(R) Xeon(R) CPU X5660 @2.80GHz processor.

For details, please refer Ray et al. (2016). We request that the reference for Ray et al. (2016) be cited if this software is used in any publication.

Value

`T.usat`	The value of the USAT test statistic (scalar).
`omg.opt`	The optimal weight ω based on a grid search over $[0,1]$.
`p.usat`	The p-value of association based on the USAT statistic.
`n.obs`	Number of individuals (with complete observations) used for testing association.
`T.manova`	The value of the MANOVA test statistic (scalar). Provided if `manova.out=TRUE`.
`p.manova`	The p-value of association based on MANOVA statistic. Provided if `manova.out=TRUE`.

Reference

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.

[1] Pan, W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology*, 33:497-507, 2009.

Example

```
source("usat_v1.21.R")
# simulate 2 phenotypes on 1000 individuals
library(MASS) # needed for multivariate normal simulation
Y<-mvrnorm(n=1000, mu=c(0,0), Sigma=matrix(c(1,0.2,0.2,1),2,2))
# simulate a single marker for 1000 individuals
X<-matrix(rbinom(n=1000, size=2, prob=0.2), ncol=1) # additive model
## apply USAT to test association
u.out<-usat(Y=Y, X=X, COV=NULL, na.check=FALSE)
# USAT test statistic and p-value
t<-u.out$T.usat
p<-u.out$p.usat
```