jtGWAS

Efficient Jonckheere-Terpstra Test Statistics

2016-07-07

Outline

Introduction

- This document provides an example for using the `jtGWAS` package to calculate the Jonckheere-Terpstra test statistics for large data sets (multiple markers and genome-wide SNPs) commonly encountered in GWAS.
- The calculation of the standardized test statistic employs the null variance equation as defined by Hollander and Wolfe (1999, eq. 6.19) to account for ties in the data.
- The major algorithm in this package is written in C++, which is ported to `R` by `Rcpp`, to facilitate fast computation.
- Features of this package include:
  1. `OpenMP` supported parallelization
  2. Customized output of top m significant SNPs for each marker
  3. $O(N \times \log(N))$ computational complexity (where N is the number of the samples)

jtGWAS

```
res <- jtGWAS(X, G, outTopN=15, numThreads=1, standardized=TRUE)
```

Function aguments:

- `X`: Matrix of marker levels, with sample IDs as row names and marker IDs as column names.
- `G`: Matrix of genotypes, with sample IDs as row names and SNP IDs as column names.
- `outTopN`: Number of top statistics to return (i.e., the largest *n* standardized statistics). The default value is 15. If `outTopN` is set to `NA`, all results will be returned.
- `numThreads`: Number of threads to use for parallel computation. The default value is 1 (sequential computation).
- `standardized`: A boolean to specify whether to return standardized statistics or non-standardized statistics. The default value is `TRUE`, returning standardized statistics.

Users may wish to consider the `dplyr::recode()` function for converting non-numeric group indices into ordinal values for argument `G`.

Returned Values

Function returns:

- `J`: A matrix of standardized/non-standardized Jonckheere Terpstra test statistics, depending on option `standardized`, with column names from input `X`. If `outTopN` is not `NA`, results are sorted within each column.
- `gSnipID`: If `outTopN` is not `NA`, this is a matrix of column names from `G` associated with top standardized Jonckheere Terpstra test statistics from `J`. Otherwise this is an unsorted vector of column names from input `G`.

Simulate Data

1. Define the number of markers, patients, and SNPs:

```
num_patient     <- 100
num_marker      <- 4
num_SNP         <- 50
```

2. Create two matrices containing marker levels and genotype information.
   a. `X_pat_mak` contains the patients' marker levels.
   b. `G_pat_SNP` contains the patients' genotypes.

```
set.seed(12345);
X_pat_mak <- matrix(rnorm(num_patient*num_marker),
                    num_patient,
                    num_marker)
G_pat_SNP <- matrix(rbinom(num_patient*num_SNP,2,0.5),
                    num_patient,
                    num_SNP)
colnames(X_pat_mak) <- colnames(X_pat_mak, do.NULL = FALSE, prefix = "Mrk:" )
colnames(G_pat_SNP) <- colnames(G_pat_SNP, do.NULL = FALSE, prefix = "SNP:" )
```

Load Package

Load `jtGWAS` (after installing its dependent packages):

```
library(jtGWAS)
```

Example Execution

```
JTStat <- jtGWAS(X_pat_mak, G_pat_SNP, outTopN=10)
summary(JTStat, marker2Print=1:4, SNP2Print=1:5)
```

```
##
##
##                   Johckheere-Terpstra Test for Large Matrices
##                     P-values for Top Standardized Statistics
## ================================================================================
##
##               Mrk:1|              Mrk:2|              Mrk:3|              Mrk:4|
## --------------------------------------------------------------------------------
##       SNPID P-value|      SNPID P-value|      SNPID P-value|      SNPID P-value|
## --------------------------------------------------------------------------------
##      SNP:35 1.7e-02|     SNP:35 2.0e-02|     SNP:20 1.9e-02|     SNP:46 1.2e-02|
##      SNP:17 1.7e-02|      SNP:7 5.7e-02|     SNP:49 3.5e-02|     SNP:34 1.7e-02|
##      SNP:27 7.0e-02|     SNP:46 9.1e-02|     SNP:26 3.8e-02|     SNP:47 3.7e-02|
##      SNP:28 7.0e-02|     SNP:40 9.4e-02|     SNP:47 5.6e-02|     SNP:23 4.5e-02|
##      SNP:14 8.8e-02|     SNP:29 1.5e-01|     SNP:30 9.7e-02|     SNP:16 5.1e-02|
```

Example Execution: Statistics in the Summary

```
summary(JTStat, marker2Print=1:4, SNP2Print=1:5, printP=FALSE)
```

```
##
##
##                   Johckheere-Terpstra Test for Large Matrices
##                           Top Standardized Statistics
## ================================================================================
##
##               Mrk:1|              Mrk:2|              Mrk:3|              Mrk:4|
## --------------------------------------------------------------------------------
##       SNPID      J*|      SNPID      J*|      SNPID      J*|      SNPID      J*|
## --------------------------------------------------------------------------------
##      SNP:35  -2.390|     SNP:35  -2.331|     SNP:20   2.350|     SNP:46   2.505|
##      SNP:17  -2.388|      SNP:7   1.905|     SNP:49  -2.106|     SNP:34   2.396|
##      SNP:27  -1.813|     SNP:46   1.693|     SNP:26  -2.072|     SNP:47  -2.089|
##      SNP:28   1.813|     SNP:40   1.676|     SNP:47   1.914|     SNP:23   2.000|
##      SNP:14   1.706|     SNP:29   1.432|     SNP:30  -1.662|     SNP:16   1.955|
```

Example Execution: Sorting in the Summary

```
JTAll <- jtGWAS(X_pat_mak, G_pat_SNP, outTopN=NA)
summary(JTAll, marker2Print=1:4, SNP2Print=1:3)
summary(JTAll, marker2Print=1:4, outTopN=3)
```

```
##
##       Johckheere-Terpstra Test
##     P-values Based on Standardized Statistics
##
##           Mrk:1     Mrk:2     Mrk:3     Mrk:4
## SNP:1 0.2931953 0.7711424 0.5877522 0.1610595
## SNP:2 0.9120503 0.6085816 0.8017917 0.3169385
## SNP:3 0.5081104 0.3280014 0.5546797 0.2823776
##
##
##                   Johckheere-Terpstra Test for Large Matrices
##                     P-values for Top Standardized Statistics
## ================================================================================
##
##               Mrk:1|              Mrk:2|              Mrk:3|              Mrk:4|
## --------------------------------------------------------------------------------
##       SNPID P-value|      SNPID P-value|      SNPID P-value|      SNPID P-value|
## --------------------------------------------------------------------------------
##      SNP:35 1.7e-02|     SNP:35 2.0e-02|     SNP:20 1.9e-02|     SNP:46 1.2e-02|
##      SNP:17 1.7e-02|      SNP:7 5.7e-02|     SNP:49 3.5e-02|     SNP:34 1.7e-02|
##      SNP:27 7.0e-02|     SNP:46 9.1e-02|     SNP:26 3.8e-02|     SNP:47 3.7e-02|
```

References

Hollander, M. and Wolfe, D. A., *Nonparametric Statistical Methods*. New York, Wiley, 2nd edition, 1999.

Session Information

- R version 3.3.0 (2016-05-03), `x86_64-pc-linux-gnu`
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils
- Other packages: jtGWAS 1.4, knitr 1.13
- Loaded via a namespace (and not attached): Rcpp 0.12.4, evaluate 0.8, formatR 1.2.1, highr 0.5.1, magrittr 1.5, stringi 1.0-1, stringr 1.0.0, tools 3.3.0

```
## [1] "Start Time Thu Jul  7 14:40:08 2016"
## [1] "End Time   Thu Jul  7 14:40:09 2016"
```