Package 'AssocTests'

October 14, 2017

Type Package

Title Genetic Association Studies

Version 0.0-4

Date 2017-10-14

Author Lin Wang [aut],
Wei Zhang [aut],
Qizhai Li [aut],
Weicheng Zhu [ctb]

Maintainer Lin Wang <wanglin2009@amss.ac.cn>

Depends cluster, mvtnorm, combinat, fExtremes, R(>= 2.10.0)

Description Some procedures including EIGENSTRAT (a procedure for detecting and correcting for population stratification through searching for the eigenvectors in genetic association studies), PCoC (a procedure for correcting for population stratification through calculating the principal coordinates and the clustering of the subjects), Tracy-Widom test (a procedure for detecting the significant eigenvalues of a matrix), distance regression (a procedure for detecting the association between a distance matrix and some independent variants of interest), single-marker test (a procedure for identifying the association between the genotype at a biallelic marker and a trait using the Wald test or the Fisher's exact test), MAX3 (a procedure for testing for the association between a single nucleotide polymorphism and a binary phenotype using the maximum value of the three test statistics derived for the recessive, additive, and dominant models), nonparametric trend test (a procedure for testing for the association between a genetic variant and a non-normal distributed quantitative trait based on the nonparametric risk), and nonparametric MAX3 (a procedure for testing for the association between a biallelic single nucleotide polymorphism and a quantitative trait using the maximum value of the three nonparametric trend tests derived for the recessive, additive, and dominant models), which are commonly used in genetic association studies.

License GPL-2

Collate 'AssociationTestWithCorrectPS.R' 'CalculateGapK.R' 'CalculateWAll.R' 'CalExpect.R' 'ChangeX.R' 'CorrMatNRTest.R' 'DR_main.R' 'EigenStrat_main.R' 'FindCNumRandom.R' 'MAX3_main.R' 'MAX3Sign.R' 'MDS.R' 'PCoC_main.R' 'ModifyNormalization.R' 'NMAX3_main.R' 'NPT_main.R' 'RhombusFormula.R' 'ScoreTest.R' 'SimilarityMatrix.R'

'SMT_main.R' 'Str2Num.R' 'TrendTest.R' 'TW_main.R'
'UniformSample.R' 'UniformTest.R' 'WaldTest.R'

RoxygenNote 6.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2017-10-14 07:06:18 UTC

R topics documented:

`dr` *Distance regression*

Description

Conduct the distance regression with or without the adjustment of the covariates to detect the association between a distance matrix and some independent variants of interest.

Usage

```
dr(simi.mat, null.space, x.mat, permute = TRUE, n.MonteCarlo = 1000,
  seed = NULL)
```

Arguments

`simi.mat`	a similarity matrix among the subjects.
`null.space`	a numeric vector to show the column numbers of the null space in `x.mat`.
`x.mat`	the covariate matrix which combines the null space and the matrix of interest.
`permute`	logical. If `TRUE`, the Monte Carlo sampling is used without replacement; otherwise, with replacement. The default is `TRUE`.
`n.MonteCarlo`	the number of times for the Monte Carlo procedure. The default is `1000`.
`seed`	if it is not `NULL`, set the random number generator state for random number generation. The default is `NULL`.

Details

The pseudo *F* statistic based on the distance regression with or without the adjustment of the covariates detects the association between a distance matrix and some independent variants of interest. A distance matrix can be transformed into a similarity matrix easily.

Value

A list with class "`htest`" containing the following components:

`statistic`
: the observed value of the test statistic.

`p.value`
: the p-value for the test.

`alternative`
: a character string describing the alternative hypothesis.

`method`
: a character string indicating the type of test performed.

`data.name`
: a character string giving the names of the data.

Author(s)

Lin Wang, Wei Zhang, and Qizhai Li.

References

Q Li, S Wacholder, DJ Hunter, RN Hoover, S Chanock, G Thomas, and K Yu. Genetic Background Comparison Using Distance-Based Regression, with Applications in Population Stratification Evaluation and Adjustment. *Genetic Epidemiology*. 2009; 33(5): 432-441.

J Wessel and NJ Schork. Generalized Genomic Distance-Based Regression Methodology for Multilocus Association Analysis. *American Journal of Human Genetics*. 2006; 79(5): 792-806.

MA Zapala and NJ Schork. Multivariate Regression Analysis of Distance Matrices for Testing Associations Between Gene Expression Patterns and Related Variables. *Proceedings of the National Academy of Sciences of the United States of America*. 2006; 103(51): 19430-19435.

Examples

```
data(drS.eg)
null.space <- 1
x.mat <- matrix(c(rep(1, 600), rep(0, 200)), ncol=2)
dr(drS.eg, null.space, x.mat, permute = TRUE, n.MonteCarlo = 50, seed = NULL)
```

drS.eg *A toy similarity matrix for dr*

Description

This data set is the toy similarity matrix among the subjects calculated from a toy genotype data set which contains of the genotypes of 400 subjects at 200 markers.

Usage

```
data(drS.eg)
```

eigenstrat *EIGENSTRAT for correcting for population stratification*

Description

Find the eigenvectors of the similarity matrix among the subjects used for correcting for population stratification in the population-based genetic association studies.

Usage

```
eigenstrat(genoFile, outFile.Robj = "out.list", outFile.txt = "out.txt",
  rm.marker.index = NULL, rm.subject.index = NULL, miss.val = 9,
  num.splits = 10, topK = NULL, signt.eigen.level = 0.01,
  signal.outlier = FALSE, iter.outlier = 5, sigma.thresh = 6)
```

Arguments

genoFile	a txt file containing the genotypes (0, 1, 2, or 9). The element of the file in Row *i* and Column *j* represents the genotype at the *i*th marker of the *j*th subject. 0, 1, and 2 denote the number of risk alleles, and 9 (default) is for the missing genotype.
outFile.Robj	the name of an R object for saving the list of the results which is the same as the return value of this function. The default is "out.list".
outFile.txt	a txt file for saving the eigenvectors corresponding to the top significant eigenvalues.
rm.marker.index	a numeric vector for the indices of the removed markers. The default is NULL.
rm.subject.index	a numeric vector for the indices of the removed subjects. The default is NULL.
miss.val	the number representing the missing data in the input data. The default is 9. The element 9 for the missing data in the genoFile should be changed according to the value of miss.val.

num.splits	the number of groups into which the markers are split. The default is 10.
topK	the number of eigenvectors to return. If NULL, it is calculated by the Tracy-Widom test. The default is NULL.
signt.eigen.level	a numeric value which is the significance level of the Tracy-Widom test. It should be 0.05, 0.01, 0.005, or 0.001. The default is 0.01.
signal.outlier	logical. If TRUE, delete the outliers of the subjects; otherwise, do not search for the outliers. The default is FALSE.
iter.outlier	a numeric value that is the iteration time for finding the outliers of the subjects. The default is 5.
sigma.thresh	a numeric value that is the lower limit for eliminating the outliers. The default is 6.

Details

Suppose that a total of *n* cases and controls are randomly enrolled in the source population and a panel of *m* single-nucleotide polymorphisms are genotyped. The genotype at a marker locus is coded as 0, 1, or 2, with the value corresponding to the copy number of risk alleles. All the genotypes are given in the form of a *m*n* matrix, in which the element in the *i*th row and the *j*th column represents the genotype of the *j*th subject at the *i*th marker. This function calculates the top eigenvectors or the eigenvectors with significant eigenvalues of the similarity matrix among the subjects to infer the potential population structure. See also tw.

Value

eigenstrat returns a list, which contains the following components:

num.markers	the number of markers excluding the removed markers.
num.subjects	the number of subjects excluding the outliers.
rm.marker.index	the indices of the removed markers.
rm.subject.index	the indices of the removed subjects.
TW.level	the significance level of the Tracy-Widom test.
signal.outlier	dealing with the outliers in the subjects or not.
iter.outlier	the iteration time for finding the outliers.
sigma.thresh	the lower limit for eliminating the outliers.
num.outliers	the number of outliers.
outliers.index	the indices of the outliers.
num.used.subjects	the number of the used subjects.
used.subjects.index	the indices of the used subjects.
similarity.matrix	the similarity matrix among the subjects.
eigenvalues	the eigenvalues of the similarity matrix.
eigenvectors	the eigenvectors corresponding to the eigenvalues.
topK	the number of significant eigenvalues.
TW.stat	the observed values of the Tracy-Widom statistics.
topK.eigenvalues	the top eigenvalues.
topK.eigenvectors	the eigenvectors corresponding to the top eigenvalues.
runtime	the running time of this function.

Author(s)

Lin Wang, Wei Zhang, and Qizhai Li.

References

AL Price, NJ Patterson, RM Plenge, ME Weinblatt, NA Shadick, and D Reich. Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies. *Nature Genetics*. 2006; 38(8): 904-909.

N Patterson, AL Price, and D Reich. Population Structure and Eigenanalysis. *PloS Genetics*. 2006; 2(12): 2074-2093.

CA Tracy and H Widom. Level-Spacing Distributions and the Airy Kernel. *Communications in Mathematical Physics*. 1994; 159(1): 151-174.

Examples

```
eigenstratG.eg <- matrix(rbinom(3000, 2, 0.5), ncol = 30)
write.table(eigenstratG.eg, file = "eigenstratG.eg.txt", quote = FALSE,
            sep = "", row.names = FALSE, col.names = FALSE)
eigenstrat(genoFile = "eigenstratG.eg.txt", outFile.Robj = "eigenstrat.result.list",
             outFile.txt = "eigenstrat.result.txt", rm.marker.index = NULL,
             rm.subject.index = NULL, miss.val = 9, num.splits = 10,
             topK = NULL, signt.eigen.level = 0.01, signal.outlier = FALSE,
             iter.outlier = 5, sigma.thresh = 6)
```

`max3` *Maximum Test: maximum value of the three Cochran-Armitage trend tests under the recessive, additive, and dominant models*

Description

Conduct MAX3 (the maximal value of the three Cochran-Armitage trend tests derived for the recessive, additive, and dominant models) based on the trend tests without the adjustment of the covariates or based on the Wald tests with the adjustment of the covariates to test for the association between a single-nucleotide polymorphism and the binary phenotype.

Usage

```
max3(y, g, covariates = NULL, Score.test = TRUE, Wald.test = FALSE,
  rhombus.formula = FALSE)
```

Arguments

`y`	a numeric vector of the observed trait values in which the *i*th element is for the *i*th subject. The elements should be `0` or `1`.
`g`	a numeric vector of the observed genotype values (`0`, `1`, or `2` denotes the number of risk alleles) in which the *i*th element is for the *i*th subject. The missing value is represented by `NA`. `g` has the same length as `y`.

`covariates` a numeric matrix for the covariates used in the model. Each column is for one covariate. The default is `NULL`, that is, there are no covariates to be adjusted for.

`Score.test` logical. If `TRUE`, the score tests are used. One of `Score.test` and `Wald.test` should be `FALSE`, and the other should be `TRUE`. The default is `TRUE`.

`Wald.test` logical. If `TRUE`, the Wald tests are used. One of `Score.test` and `Wald.test` should be `FALSE`, and the other should be `TRUE`. The default is `FALSE`.

`rhombus.formula` logical. If `TRUE`, the p-value for the MAX3 is approximated by the rhombus formula. IF `FALSE`, the 2-fold integration is used to calculate the p-value. The default is `FALSE`.

Details

In an association study, the genetic inheritance models (recessive, additive, or dominant) are unknown beforehand. This function can account for the uncertainty of the underlying genetic models and test for the association between a single-nucleotide polymorphism and a binary phenotype with or without correcting for the covariates.

Value

A list with class "`htest`" containing the following components:

`statistic` the observed value of the test statistic.

`p.value` the p-value for the test.

`alternative` a character string describing the alternative hypothesis.

`method` a character string indicating the type of test performed.

`data.name` a character string giving the names of the data.

Author(s)

Lin Wang, Wei Zhang, and Qizhai Li.

References

Q Li, G Zheng, Z Li, and K Yu. Efficient Approximation of P Value of the Maximum of Correlated Tests, with Applications to Genome-Wide Association Studies. *Annals of Human Genetics*. 2008; 72(3): 397-406.

Examples

```
y <- rep(c(0, 1), 5)
g <- sample(c(0, 1, 2), 10, replace = TRUE)
max3(y, g, covariates = NULL, Score.test = TRUE, Wald.test = FALSE,
       rhombus.formula = FALSE)
```

```
max3(y, g, covariates = matrix(sample(c(0,1), 20, replace = TRUE), ncol=2),
       Score.test = TRUE, Wald.test = FALSE, rhombus.formula = FALSE)
```

`nmax3` *NMAX3 based on the maximum value of the three nonparametric trend tests under the recessive, additive, and dominant models*

Description

Test for the association between a biallelic SNP and a quantitative trait using the maximum value of the three nonparametric trend tests derived for the recessive, additive, and dominant models. It is a robust procedure against the genetic models.

Usage

```
nmax3(y, g)
```

Arguments

`y`	a numeric vector of the observed quantitative trait values in which the *i*th element is the trait value of the *i*th subject.
`g`	a numeric vector of the observed genotype values (`0`, `1`, or `2` denotes the number of risk alleles) in which the *i*th element is the genotype value of the *i*th subject for a biallelic SNP. `g` has the same length as `y`.

Details

Under the null hypothesis of no association, the vector of the three nonparametric tests under the recessive, additive, and dominant models asymptotically follows a three-dimensional normal distribution. The p-value can be calculated using the function pmvnorm in the R package "**mvtnorm**".

This test is different from the MAX3 test using in the function max3. On one hand, the NMAX3 applies to the quantitative traits association studies. However, the MAX3 is used in the case-control association studies. On the other hand, the NMAX3 is based on the nonparametric trend test. However, the MAX3 is based on the Cochran-Armitage trend test.

Value

A list with class "`htest`" containing the following components:

`statistic`	the observed value of the test statistic.
`p.value`	the p-value for the test.
`alternative`	a character string describing the alternative hypothesis.
`method`	a character string indicating the type of test performed.
`data.name`	a character string giving the names of the data.

Author(s)

Lin Wang, Wei Zhang, and Qizhai Li.

References

W Zhang and Q Li. Nonparametric Risk and Nonparametric Odds in Quantitative Genetic Association Studies. *Science Reports (2nd revision)*. 2015.

B Freidlin, G Zheng, Z Li, and JL Gastwirth. Trend Tests for Case-Control Studies of Genetic Markers: Power, Sample Size and Robustness. *Human Heredity*. 2002; 53:146-152.

WG Cochran. Some Methods for Strengthening the Common Chi-Square Tests. *Biometrics*. 1954; 10:417-451.

P Armitage. Tests for Linear Trends in Proportions and Frequencies. *Biometrics*. 1955; 11:375-386.

Examples

```
g <- rbinom(1500, 2, 0.3)
y <- 0.5 + 0.25 * g + rgev(1500, 0, 0, 5)
nmax3(y, g)
```

`npt` *Nonparametric trend test based on the nonparametric risk under a given genetic model*

Description

Test for the association between a genetic variant and a non-normal distributed quantitative trait based on the nonparametric risk under a specific genetic model.

Usage

```
npt(y, g, varphi)
```

Arguments

`y`	a numeric vector of the observed quantitative trait values in which the *i*th element corresponds to the trait value of the *i*th subject.
`g`	a numeric vector of the observed genotype values (`0`, `1`, or `2` denotes the number of risk alleles) in which the *i*th element is the genotype value of the *i*th subject for a biallelic SNP. `g` has the same length as `y`.
`varphi`	a numeric value which represents the genetic model. It should be `0`, `0.5`, or `1`, which indicates that the calculation is performed under the recessive, additive, or dominant model, respectively.

Details

For a non-normal distributed quantitative trait, three genetic models (recessive, additive and dominant) used commonly are defined in terms of the nonparametric risk (NR). The recessive, additive, and dominant models can be classified based on the nonparametric risks. More specifically, the recessive, additive, and dominant models refer to NR20 > NR10 = 1/2, NR12 = NR10>1/2, and NR10 = NR20 > 1/2, respectively, where NR10 and NR20 are the nonparametric risks of the groups with the genotypes 1 and 2 relative to the group with the genotype 0, respectively, and NR12 is the nonparametric risk of the group with the genotype 2 relative to the group with the genotype 1.

`varphi` can be `0`, `0.5`, or `1` for the recessive, additive, or dominant model, respectively. When `varphi` is `0`, the test is constructed under the recessive model by pooling together the subjects with the genotypes 0 and 1. Similarly, when `varphi` is `1`, the test is constructed under the dominant model by pooling together the subjects with the genotypes 1 and 2. When `varphi` is `0.5`, the test is based on the weighted sum of NR10 and NR12.

Value

A list with class "`htest`" containing the following components:

`statistic`	the observed value of the test statistic.
`p.value`	the p-value for the test.
`alternative`	a character string describing the alternative hypothesis.
`method`	a character string indicating the type of test performed.
`data.name`	a character string giving the names of the data.

Author(s)

Lin Wang, Wei Zhang, and Qizhai Li.

Examples

```
g <- rbinom(1500, 2, 0.3)
y <- 0.5 + 0.25 * g + rgev(1500, 0, 0, 5)
npt(y, g, 0.5)
```

---

pcoc *PCoC for correcting for population stratification*

Description

Identify the clustered and continuous patterns of the genetic variation using the PCoC, which calculates the principal coordinates and the clustering of the subjects for correcting for PS.

Usage

```
pcoc(genoFile, outFile.txt = "pcoc.result.txt", n.MonteCarlo = 1000,
  num.splits = 10, miss.val = 9)
```

Arguments

genoFile	a txt file containing the genotypes (0, 1, 2, or 9). The element of the file in Row *i* and Column *j* represents the genotype at the *i*th marker of the *j*th subject. 0, 1, and 2 denote the number of risk alleles, and 9 (default) is for the missing genotype.
outFile.txt	a txt file for saving the result of this function. The default is "pcoc.result.txt".
n.MonteCarlo	the number of times for the Monte Carlo procedure. The default is 1000.
num.splits	the number of groups into which the markers are split. The default is 10.
miss.val	the number representing the missing data in the input data. The default is 9. The element 9 for the missing data in the genoFile should be changed according to the value of miss.val.

Details

The hidden population structure is a possible confounding effect in the large-scale genome-wide association studies. Cases and controls might have systematic differences because of the unrecognized population structure. The PCoC procedure uses the techniques from the multidimensional scaling and the clustering to correct for the population stratification. The PCoC could be seen as an extension of the EIGENSTRAT.

Value

A list of principal.coordinates and cluster. principal.coordinates is the principal coordinates and cluster is the clustering of the subjects. If the number of clusters is only one, cluster is omitted.

Author(s)

Lin Wang, Wei Zhang, and Qizhai Li.

References

Q Li and K Yu. Improved Correction for Population Stratification in Genome-Wide Association Studies by Identifying Hidden Population Structures. *Genetic Epidemiology*. 2008; 32(3): 215-226.

KV Mardia, JT Kent, and JM Bibby. Multivariate Analysis. *New York: Academic Press*. 1976.

Examples

```
pcocG.eg <- matrix(rbinom(4000, 2, 0.5), ncol = 40)
write.table(pcocG.eg, file = "pcocG.eg.txt", quote = FALSE,
       sep = "", row.names = FALSE, col.names = FALSE)
pcoc(genoFile = "pcocG.eg.txt", outFile.txt = "pcoc.result.txt",
       n.MonteCarlo = 50, num.splits = 10, miss.val = 9)
```

smt *Single-marker test*

Description

Conduct the single-marker test in an association study to test for the association between the genotype at a biallelic marker and a trait.

Usage

```
smt(y, g, covariates = NULL, min.count = 5, missing.rate = 0.2,
  y.continuous = FALSE)
```

Arguments

y	a numeric vector of the observed trait values in which the *i*th element is for the *i*th subject. The elements could be discrete (0 or 1) or continuous. The missing value is represented by NA.
g	a numeric vector of the observed genotype values (0, 1, or 2 denotes the number of risk alleles) in which the *i*th element is for the *i*th subject. The missing value is represented by NA. g has the same length as y.
covariates	an optional data frame, list or environment containing the covariates used in the model. The default is NULL, that is, there are no covariates.
min.count	a critical value to decide which method is used to calculate the p-value when the trait is discrete and covariates = NULL. If the minimum number of the elements given a specific trait value and a specific genotype value is less than min.count, the Fisher's exact test is adopted; otherwise, the Wald test is adopted. The default is 5.
missing.rate	the highest missing value rate of the genotype values that this function can tolerate. The default is 0.2.
y.continuous	logical. If TRUE, y is continuous; otherwise, y is discrete. The default is FALSE.

Details

Single-marker analysis is a core in many gene-based or pathway-based procedures, such as the truncated p-value combination and the minimal p-value.

Value

smt returns a list with class "htest".

If y is continuous, the list contains the following components:

statistic	the observed value of the test statistic.
p.value	the p-value for the test.

`alternative`
: a character string describing the alternative hypothesis.

`method`
: a character string indicating the type of test performed.

`data.name`
: a character string giving the names of the data.

`sample.size`
: a vector giving the numbers of the subjects with the genotypes `0`, `1`, and `2` (`n0`, `n1`, and `n2`, respectively).

If y is discrete, the list contains the following components:

`statistic`
: the observed value of the test statistic.

`p.value`
: the p-value for the test.

`alternative`
: a character string describing the alternative hypothesis.

`method`
: a character string indicating the type of test performed.

`data.name`
: a character string giving the names of the data.

`sample.size`
: a vector giving
the number of subjects with the trait value `1` and the genotype `0` (`r0`),
the number of subjects with the trait value `1` and the genotype `1` (`r1`),
the number of subjects with the trait value `1` and the genotype `2` (`r2`),
the number of subjects with the trait value `0` and the genotype `0` (`s0`),
the number of subjects with the trait value `0` and the genotype `1` (`s1`),
and the number of subjects with the trait value `0` and the genotype `2` (`s2`).

`bad.obs`
: a vector giving the number of missing genotype values with the trait value `1` (`r.miss`), the number of missing genotype values with the trait value `0` (`s.miss`), and the total number of the missing genotype values (`n.miss`).

Author(s)

Lin Wang, Wei Zhang, and Qizhai Li.

Examples

```
y <- rep(c(0, 1), 25)
g <- sample(c(0, 1, 2), 50, replace = TRUE)
smt(y, g, covariates = NULL, min.count=5,
        missing.rate=0.20, y.continuous = FALSE)
```

tw *Tracy-Widom test*

Description

Find the significant eigenvalues of a matrix.

Usage

```
tw(eigenvalues, eigenL, criticalpoint = 2.0234)
```

Arguments

`eigenvalues`	a numeric vector whose elements are the eigenvalues of a matrix. The values should be sorted in the descending order.
`eigenL`	the number of eigenvalues.
`criticalpoint`	a numeric value corresponding to the significance level. If the significance level is 0.05, 0.01, 0.005, or 0.001, the `criticalpoint` should be set to be `0.9793`, `2.0234`, `2.4224`, or `3.2724`, accordingly. The default is `2.0234`.

Value

A list with class "`htest`" containing the following components:

`statistic`	a vector of the Tracy-Widom statistics.
`alternative`	a character string describing the alternative hypothesis.
`method`	a character string indicating the type of test performed.
`data.name`	a character string giving the name of the data.
`SigntEigenL`	the number of significant eigenvalues.

Author(s)

Lin Wang, Wei Zhang, and Qizhai Li.

References

N Patterson, AL Price, and D Reich. Population Structure and Eigenanalysis. *PloS Genetics*. 2006; 2(12): 2074-2093.

CA Tracy and H Widom. Level-Spacing Distributions and the Airy Kernel. *Communications in Mathematical Physics*. 1994; 159(1): 151-174.

A Bejan. Tracy-Widom and Painleve II: Computational Aspects and Realisation in S-Plus. In *First*

Workshop of the ERCIM Working Group on Computing and Statistics. 2008, Neuchatel, Switzerland.

www.vitrum.md/andrew/MScWrwck/codes.txt

Examples

```
tw(eigenvalues = c(5, 3, 1, 0), eigenL = 4, criticalpoint = 2.0234)
```

Index