Package 'gyriq'

January 7, 2016

Type Package

Title Kinship-Adjusted Survival SNP-Set Analysis

Version 1.0.2

Date 2016-01-06

Author Martin Leclerc and Lajmi Lakhal Chaieb

Maintainer Martin Leclerc <martin.leclerc.5@ulaval.ca>

Description
SNP-set association testing for censored phenotypes in the presence of intrafamilial correlation.

Imports CompQuadForm, irlba, mvtnorm, survival

Suggests snowfall

License GPL (>= 2)

NeedsCompilation yes

Depends R (>= 2.10)

Repository CRAN

Date/Publication 2016-01-07 13:55:34

R topics documented:

gyriq-package *Kinship-Adjusted Survival SNP-Set Analysis*

Description

SNP-set association testing for censored phenotypes in the presence of intrafamilial correlation

Details

Package:	gyriq
Type:	Package
Version:	1.0.2
Date:	2016-01-06
License:	GPL (>= 2)

This variance-components test between a set of SNPs and a survival trait is valid for both common and rare variants. A proportional hazards Cox model (written as a transformation model with censored data; Cheng et al., 1995) is specified for the marginal distribution of the survival trait. The familial dependence is modelled via a Gaussian copula with a correlation matrix expressed in terms of the kinship matrix. The statistical procedure has been described in full detail by Leclerc et al. (2015).

Censored values are treated as partially missing data and a multiple imputation procedure is employed to estimate vectors of residuals. These residuals and the SNPs in the genomic region under study are used to compute measures of phenotypic and genotypic similarity between pairs of subjects. The contribution to the score statistic is maximal when these measures are both high which corresponds to departure from the null hypothesis of no association between the set of SNPs and the survival outcome. The selection of the SNPs forming the SNP set can be based on biological information such as linkage disequilibrium (LD) blocks or rely on a sliding window method.

The procedure is convenient for GWAS as the multiple imputation procedure for the estimation of a completed vector of residuals has to be performed only once using the function genComplResid. A sliding window approach can then be used to examine the evidence of association across the SNP set. In each run, the p-value is computed with the function testGyriq.

Author(s)

Martin Leclerc <martin.leclerc.5@ulaval.ca> and Lajmi Lakhal Chaieb <lakhal@mat.ulaval.ca>

References

Cheng SC, Wei LJ, Ying Z. 1995. Analysis of transformation models with censored data. Biometrika 82:835-845.

Leclerc M, The Consortium of Investigators of Modifiers of BRCA1/2, Simard J, Lakhal-Chaieb L. 2015. SNP set association testing for survival outcomes in the presence of intrafamilial correlation. Genetic Epidemiology 39:406-414.

Lin X, Zhou Q. 2015. coxKM: Cox kernel machine SNP-set association test. R package version 0.3, URL http://www.hsph.harvard.edu/xlin/software.html#coxkm.

Lin X, Cai T, Wu M, Zhou Q, Liu G, Christiani D, Lin X. 2011. Survival kernel machine SNP-set analysis for genome-wide association studies. Genetic Epidemiology 35:620-631.

Cai T, Tonini G, Lin X. 2011. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. Biometrics 67:975-986.

Examples

```
data(simGyriq)
for (i in seq_along(simGyriq)) assign(names(simGyriq)[i], simGyriq[[i]])

cr <- genComplResid(U, Delta, Phi, blkID, m=50, X)
testGyriq(cr$compResid, G, w, ker="LIN", asv=NULL, method="davies",
starResid=NULL, bsw, tsw, pos)
```

---

genComplResid *genComplResid*

Description

Generates a completed vector of residuals

Usage

```
genComplResid(U, Delta, Phi, blkID, m = 50, X = NULL)
```

Arguments

U	a nx1 vector containing the survival times. `U = min(C, T)` where `C` is the censoring time, and `T` the failure time
Delta	a nx1 vector containing the censoring indicator
Phi	a nxn kinship matrix
blkID	a nx1 vector with entries identifying correlated groups of observations. The number of censored individuals in each group cannot exceed 1000 (see **Details**)
m	default=50. Number of imputations used to generate the completed vector of residuals
X	a nxp **matrix** of p covariates. Each row represents a different individual, and each column represents a different numeric covariate. If no covariates are present, `X` can be left as `NULL`

Details

This function involves three steps. The first two are similar in spirit to the two-stage procedure of Othus and Li (2010).

1. The vector of covariate parameters and the monotone increasing function of the transformation model with censored data (Cheng et al., 1995) are estimated under the working independence assumption following the algorithm of Chen et al. (2002) and used to compute raw residuals;
2. The polygenic heritability parameter is estimated which is a measure of the dependence between the survival traits of correlated groups that cannot be attributed to the SNP set under investigation. This estimate is used to deduce the approximate covariance matrix of the raw residuals.
3. An imputation procedure is employed to replace the censored raw residuals by the mean of multiple imputed values generated from the posterior distribution of the uncensored version with the restriction to be larger than the original censored values, componentwise. The completed vector of residuals is then deduced and standardized. A scale parameter is used to reflect the fact that we are using multiple imputed values rather than real observations.

Warning: Correlated groups identified by the vector `blkID` most often corresponds to families or blocks of the block-diagonal kinship matrix **Phi**. Larger groups such as regions of residence can be considered, for example to take into account population stratification or cryptic relatedness. However, the number of censored individuals in each group cannot exceed 1000 as the test makes use of the distribution function of the multivariate normal distribution for which the maximum dimension is 1000 in the function **pmvnorm** of the package **mvtnorm**.

Simulation studies reported in Leclerc et al. (2015) suggest that the use of `m = 50` imputations guarantees a reasonable power in practice.

Warning: No missing data is allowed for `U`, `Delta`, `Phi`, `blkID`, and `X`.

Value

The function produces a list consisting of:

`compResid`	the completed vector of residuals
`herit`	the estimate of the polygenic heritability parameter
`covPar`	the estimate of the vector of covariate parameters (if applicable)

Author(s)

Martin Leclerc <martin.leclerc.5@ulaval.ca> and Lajmi Lakhal Chaieb <lakhal@mat.ulaval.ca>

References

Chen K, Jin Z, Ying Z. 2002. Semiparametric analysis of transformation models with censored data. Biometrika 89:659-668.

Cheng SC, Wei LJ, Ying Z. 1995. Analysis of transformation models with censored data. Biometrika 82:835-845.

Leclerc M, The Consortium of Investigators of Modifiers of BRCA1/2, Simard J, Lakhal-Chaieb L. 2015. SNP set association testing for survival outcomes in the presence of intrafamilial correlation. Genetic Epidemiology 39:406-414.

Othus M, Li Y. 2010. A gaussian copula model for multivariate survival data. Stat Biosci 2:154-179.

Examples

```
data(simGyriq)
for (i in seq_along(simGyriq)) assign(names(simGyriq)[i], simGyriq[[i]])

cr <- genComplResid(U, Delta, Phi, blkID, m=50, X)
```

`simGyriq` *Simulated SNP-set*

Description

Simulated dataset of phenotypic, genotypic and kinship data.

Format

A list containing the following elements:

U 600x1 vector containing the survival times. `U = min(C, T)` where `C` is the censoring time, and `T` the failure time

Delta 600x1 vector containing the censoring indicator

Phi 600x600 kinship matrix

blkID 600x1 vector with entries identifying correlated groups of observations

X 600x2 matrix of 2 covariates

G 600x50 matrix containing the set of 50 SNPs

w 50x1 vector of weights for the 50 SNPs

bsw 4x1 vector containing the lower bounds of the 4 sliding windows considered for the SNP-set

tsw 4x1 vector containing the upper bounds of the 4 sliding windows considered for the SNP-set

pos 50x1 vector of SNP positions (used for the output only)

indResid 10,000*600x1 vector of permuted row indices

Details

This dataset was generated under conditions described in Leclerc et al. (2015).

Samples of n = 600 individuals from 120 families were generated: 40 families of two parents and one child, 40 families of two parents and two children, and 40 families of three generations (two grand-parents, four parents, and two grandchildren). The coefficients of the block diagonal kinship matrix were fixed at their expected theoretical values. The number of biallelic SNPs was set to s = 50. The minor allele frequencies were randomly sampled from Unif(0.001, 0.1). The genotypes of the 50 SNPs were simulated assuming a linkage disequilibrium corresponding to a squared correlation coefficient of r^2 = 0.5 between consecutive SNPs.

The two covariates follow Bernoulli(0.5) and Uniform(-0.2, 0.2) distributions respectively. The polygenic heritability parameter was fixed at 0.5. Each covariate parameter was set equal to 1 and

the monotone increasing function of the transformation model with censored data (Cheng et al., 1995) was fixed at H(t) = log(t) in order to generate the survival traits. The censoring rate was equal to 50%. The weight of each SNP was defined as the density function of the Beta (1, 25) evaluated at the corresponding minor allele frequency.

The dataset includes simulated positions for the 50 SNPs, and the lower and upper bounds of 4 sliding windows. Each window includes 10 SNPs, overlapping with the previous and subsequent windows. A vector of size B*n of permuted row indices is also included, where B=10,000. This is to be used to compute the p-value of the test following the standard or matching moments permutation approach.

References

Cheng SC, Wei LJ, Ying Z. 1995. Analysis of transformation models with censored data. Biometrika 82:835-845.

Leclerc M, The Consortium of Investigators of Modifiers of BRCA1/2, Simard J, Lakhal-Chaieb L. 2015. SNP set association testing for survival outcomes in the presence of intrafamilial correlation. Genetic Epidemiology 39:406-414.

Examples

```
data(simGyriq)
for (i in seq_along(simGyriq)) assign(names(simGyriq)[i], simGyriq[[i]])

cr <- genComplResid(U, Delta, Phi, blkID, m=50, X)
testGyriq(cr$compResid, G, w, ker="LIN", asv=NULL, method="davies",
starResid=NULL, bsw, tsw, pos)
```

`testGyriq` *testGyriq*

Description

Calculates the p-value of the kinship-adjusted SNP-set association test for censored traits

Usage

```
testGyriq(compResid, G, w, ker = "LIN", asv = NULL, method = "davies",
  starResid = NULL, bsw = NULL, tsw = NULL, pos = NULL, sf = FALSE,
  fileOut = "outGyriq.out")
```

Arguments

`compResid`	a nx1 vector containing the completed residuals
`G`	a nxs matrix containing the set of SNPs. Each row represents a different individual and each column represents a separate SNP. The SNP genotypes should be equal to the number of copies of the minor allele (0, 1 or 2).
`w`	a sx1 vector of weights for the s SNPs

`ker` (default="LIN") Type of kernel matrix: weighted linear ("LIN") or weighted identical-by-state ("IBS")

`asv` (default=NULL) Number of approximate eigenvalues to be estimated for the kernel matrix using the implicitly-restarted Lanczos bidiagonalization implemented in the package **irlba** (Baglama and Reichel, 2005). If the spectral decomposition of the matrix is to be conducted using the R base function **eigen**, `asv` can be left as `NULL`. This argument has no effect if `method` is not equal to "davies".

`method` (default="davies") Procedure used to obtain the p-value of the test. "davies" represents the approximation of Davies (1980), "rspMom" represents the permutation approach based on matching moments described in Lee et al. (2012), and "rspOrd" represents the standard permutation procedure.

`starResid` (default=NULL) a Bxn **matrix** of permuted residuals used to obtain the p-value of the test following a permutation procedure (method based on matching moments or standard permutation method). Each row represents a different permutation sample, and each column represents a different individual. This argument has no effect if `method` is not equal to "rspOrd" or "rspMom".

`bsw` (default=NULL) a vx1 vector containing the lower bounds of the v sliding windows considered for the SNP-set, taking values between 1 and s

`tsw` (default=NULL) a vx1 vector containing the upper bounds of the v sliding windows considered for the SNP-set, taking values between 1 and s

`pos` (default=NULL) a sx1 vector of SNP positions

`sf` (default=FALSE) logical: indicates whether or not cluster computing is used via the package **snowfall** in order to reduce wall-clock time. Initialisation and loading of the package **gyriq** on all nodes including master must be called beforehand using the functions `sfInit` and `sfLibrary` respectively. See the reference manual of `snowfall` for details. When cluster computing is used, the p-value for each sliding window is computed on a separate node.

`fileOut` (default="outGyriq.out") a string containing the name and path of the output file where the results are printed (used only if lower and upper bounds of sliding windows are also given as input; the file is appended for each sliding window in order to reduce resource wastage)

Details

If the lower and upper bounds of sliding windows are not provided, the test is performed once on the whole SNP-set `G`. Otherwise, the score statistic and the p-value are computed for each window sequentially.

In each run, the score statistic, which has a quadratic form following a mixture of chi-squared variables, is calculated from the completed vector of residuals and a kernel matrix. The p-value is obtained using a permutation approach based on matching moments described in Lee et al. (2012), a standard permutation procedure or the Davies approximation (Davies, 1980) implemented in the package **CompQuadForm** (Duchesne and Lafaye De Micheaux, 2010).

Warning: No missing data is allowed for `compResid`, `G`, `w` and `starResid`.

Value

If the lower and upper bounds of sliding windows are not provided, the function produces a list consisting of:

`score`	the score statistic of the test
`pVal`	the p-value

Otherwise, the function produces a data frame where each row represents a sliding window tested. For each window, the following information is provided:

- `FirstSNP`: Rank of the SNP corresponding to the lower bound of the sliding window in the SNP-set
- `LastSNP`: Rank of the SNP corresponding to the upper bound of the sliding window in the SNP-set
- `winSize`: Number of SNPs in the sliding window
- `Start`: Position of the SNP corresponding to the lower bound of the sliding window
- `Stop`: Position of the SNP corresponding to the upper bound of the sliding window
- `Score`: Score statistic of the association test
- `P-value`: P-value of the association test
- `Message`: If the calculation of the p-value failed, the corresponding error message is given. Otherwise, "OK" is displayed.

Author(s)

Martin Leclerc <martin.leclerc.5@ulaval.ca> and Lajmi Lakhal Chaieb <lakhal@mat.ulaval.ca>

References

Baglama J, Reichel L. 2005. Augmented implicitly restarted Lanczos bidiagonalization methods. SIAM J Sci Comput 27:19-42.

Davies RB. 1980. The distribution of a linear combination of χ^2 random variables. J R Stat Soc Ser C 29:323-333.

Lee S, Emond MJ, Bamshad MJ et al. 2012. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet 91:224-237.

Duchesne P, Lafaye De Micheaux P. 2010. Computing the distribution of quadratic forms: further comparisons between the Liu-Tang-Zhang approximation and exact methods. Comput Stat Data Anal 54:858-862.

Lin X, Zhou Q. 2015. coxKM: Cox kernel machine SNP-set association test. R package version 0.3, URL http://www.hsph.harvard.edu/xlin/software.html#coxkm.

Lin X, Cai T, Wu M, Zhou Q, Liu G, Christiani D, Lin X. 2011. Survival kernel machine SNP-set analysis for genome-wide association studies. Genetic Epidemiology 35:620-631.

Cai T, Tonini G, Lin X. 2011. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. Biometrics 67:975-986.

Examples

```
data(simGyriq)
for (i in seq_along(simGyriq)) assign(names(simGyriq)[i], simGyriq[[i]])

cr <- genComplResid(U, Delta, Phi, blkID, m=50, X)
testGyriq(cr$compResid, G, w, ker="LIN", asv=NULL, method="davies",
starResid=NULL, bsw, tsw, pos)
```

Index