Package ‘rvsel’

May 31, 2016

Type Package

Title Rare Variant Selection Procedure

Version 0.2.1

Date 2016-05-31

Author Chunghui Lee and Hokeun Sun

Maintainer Hokeun Sun <hsun@pusan.ac.kr>

Depends R (>= 3.1.0)

Description
When a gene or a genetic region is significantly associated with a disease or a trait, the rare variant selection procedure is able to distinguish causal (risk or protective) rare variants from noncausal rare variants located within the same gene or the same genetic region.

License GPL-2

Repository CRAN

NeedsCompilation yes

R topics documented:

`rvsel` *Rare variant selection procedure*

Description

When a gene or a genetic region is significantly associated with a disease/trait, the rvsel procedure can distinguish causal (risk or protective) rare variants from noncausal rare variants located within the same gene or the same genetic region.

The most outcome-related rare variants are selected within a gene or a genetic region, considering all possible combinations of rare variants. First, genetic data of each individual is combined into one dimensional numeric vector based on one of the following methods; a weighted linear combination of a subset of rare variants, an adaptive weighted linear combination of a subset of rare variants and a combined multivariate and collapsing method.

Next, one of the selection procedures such as exhaustive search, forward selection, backward selection, and forward based both risk and protective search is conducted to identify causal rare variants within a gene or a genetic region.

Usage

```
rvsel(x,y,cx=NULL,weight=NULL,family=c("gaussian","binomial"),method=
   c("asum","sum","cmc"),selection=c("exhaustive","forward","backward","Fsel"),
    ad.alpha=0.1,lambda=0)
```

Arguments

`x` The number of genetic mutations with n samples and p variants, where $x = 0, 1$, or 2. It should be a n x p matrix and $p > 1$.

`y` A phenotype outcome is coded as 1 for cases and 0 for controls if the phenotype is case-control binary data. Otherwise, it is considered as a quantitative outcome.

`cx` Covariates such as gender and age. It should be a n x m matrix, where m is the number of covariates.

`weight` User defined weights for p variants. It should be the p-dimensional vector. Default is 1.

`family` A type of phenotype data. "`binomial`" is for a case-control binary outcome and "`gaussian`" for a quantitative outcome. Default is "`gaussian`".

`method` A way to combine genetic data. "`sum`" conducts a weighted linear combinations of rare variants, "`asum`" first conducts a pre-screening of potential protective variants via a marginal association test. If a potential protective variant is detected, the coding of `x` for the variant is flipped out and then a weighted linear combinations of rare variants is performed. "`cmc`" conducts a combined multivariate and collapsing method. Default is "`asum`". See details.

`selection` A type of selection procedure. "`exhaustive`" performs a complete search of the power set of the subset of all rare variants, where the combination of the most outcome-related variants is finally selected. "`forward`" conducts a forward based selection, where the most outcome-related variant is sequentially added from a null model. "`backward`" performs a backward based selection, where the most insignificant variant is sequentially removed from the full model. "`Fsel`" is also based on a forward selection procedure, but it considers both cases of "`x`" and the flipped coding of "`x`" so it can detect protective variants as well as risk variants without a pre-screening test of "`asum`". When "`Fsel` is selected, a weighted linear combination of rare variants is conducted, regardless of the choice of "`method`". See details.

`ad.alpha` A significance level of a marginal association test to detect potential protective variants when "`method`" is "`asum`". Default is 0.1.

`lambda` A tuning parameter value used for a stopping rule, when "`selection`" is "`forward`", "`backward`" or "`Fsel`". A larger value of "`lambda`" induces a smaller model selection, where false positives can be reduced down while some true positives are missed. Default is 0.

Details

The method "`sum`" employs a weighted linear combination of the subset of p rare variants to combine the rare variants. The weighted linear combination of the ith individual is defined as

$$z_i = \sum_{j=1}^{p} \xi_j w_j x_{ij},$$

where $\xi_j = 1$ if the jth variant is included in a model, otherwise $\xi_j = 0$. w_j is a user defined weight of the jth variant. The method "`asum`" replace x_{ij} by x^*_{ij}, where $x^*_{ij} = 1 - x_{ij}$ if the jth variant is potentially protective variant. Otherwise, $x^*_{ij} = x_{ij}$. If the p-value of an marginal association test between the jth variant and a phenotype outcome is less than "`ad.alpha`" and they have a negative relationship, the jth variant is considered as potentially protective. The method "`cmc`" combines the p rare variants such as

$$z_i = I\left(\sum_{j=1}^{p} \xi_j x_{ij} > 0\right),$$

where $I(\cdot)$ is an indicator function.

The selection procedure "`exhaustive`" generates $2^p - 1$ subsets of the power set of p rare variants, where an empty set is excluded since the selection procedure assumes that at least one variant is causal. The best combination of rare variants among the $2^p - 1$ subsets that can maximize the association with a phenotype outcome is selected as a final model. When p is relatively large, computational time of "`exhaustive`" is exponentially increased. Either "`forward`" or "`backward`" selection is desirable for a relatively large p. `Fsel` is a different selection procedure from others based on a weighted linear combination. It defines a weighted linear combination of the subset of p rare variants as

$$z_i = \sum_{j=1}^{p} \xi_j w_j x^*_{ij},$$

where $\xi_j = 1$, -1 or 0 if the jth variant is risk, protective or noncausal variant, respectively. Also, $x^*_{ij} = -(1 - x_{ij})$ if the jth variant is protective, otherwise, $x^*_{ij} = x_{ij}$. `Fsel` can be performed based only on forward selection procedure.

Value

`model` Types of "`family`", "`method`" and "`selection`" used in analysis

`selection` The selection result of p x 2 matrix. In the first column indicator values of variants are displayed where 1 for selected variants and 0 for unselected variants. In the second column the weights of individual variants used are displayed.

`score` The largest sample correlation between the combined genotypes and a phenotypic outcome, which can be replaced by a regression residual if a covariate exists).

`sequence` When "`selection`" is "`forward`", "`backward`" or "`Fsel`", the sequence of selected variants are listed.

Author(s)

Hokeun Sun <hsun@pusan.ac.kr>

References

S. Kim, K. Lee, and H. Sun (2015) *Statistical Selection Strategy for Risk and Protective Rare Variants Associated with Complex Traits*, Journal of Computational Biology 22(11), 1034–1043

H. Sun and S. Wang (2014) *A Power Set Based Statistical Selection Procedure to Locate Susceptible Rare Variants Associated with Complex Traits with Sequencing Data*, Bioinformatics 30(16), 2317–2323

Examples

```
# Generate simulation data
 n <- 2000
 p <- 10
 MAF <- runif(p,0.001,0.01)
 geno.prob <- rbind((1-MAF)^2,2*(1-MAF)*MAF,MAF^2)
 x <- apply(geno.prob,2,function(x) sample(0:2,n,prob=x,replace=TRUE))
 cx <- cbind(rnorm(n),sample(0:1,n,replace=TRUE))
 beta <- c(rep(1,4),rep(0,6))
 y <- cx %*% c(0.5,0.5)+ x %*% beta+rnorm(n)

 # method = 'asum' and selection = 'exhaustive'
 g <- rvsel(x,y,cx=cx)

 # selection = 'Fsel'
 g <- rvsel(x,y,cx=cx,selection="Fsel")

# Both risk and protective variants are present
 n <- 2000
 p <- 10
 MAF <- runif(p,0.001,0.01)
 geno.prob <- rbind((1-MAF)^2,2*(1-MAF)*MAF,MAF^2)
 x <- apply(geno.prob,2,function(x) sample(0:2,n,prob=x,replace=TRUE))
 cx <- cbind(rnorm(n),sample(0:1,n,replace=TRUE))
 beta <- c(rep(1,2),rep(-1,2), rep(0,6))
 y <- cx %*% c(0.5,0.5)+ x %*% beta+rnorm(n)

 # method = 'cmc' and selection = 'exhaustive'
 g <- rvsel(x,y,cx=cx,method="cmc")

 # selection = 'Fsel'
 g <- rvsel(x,y,cx=cx,selection="Fsel")

 # A big gene simulation
 n <- 2000
 p <- 50
 MAF <- runif(p,0.001,0.01)
 geno.prob <- rbind((1-MAF)^2,2*(1-MAF)*MAF,MAF^2)
 x <- apply(geno.prob,2,function(x) sample(0:2,n,prob=x,replace=TRUE))
 cx <- cbind(rnorm(n),sample(0:1,n,replace=TRUE))
 beta <- c(rep(1,8),rep(0,42))
 y <- cx %*% c(0.5,0.5)+ x %*% beta+rnorm(n)

 # method = 'asum' and selection = 'forward'
 ## Not run: g <- rvsel(x,y,cx=cx,selection="forward")
```

```
# selection = 'Fsel'
## Not run: g <- rvsel(x,y,cx=cx,selection="Fsel", lambda=0.01)
```

Index