Package 'iSKAT'

October 14, 2015

Version 1.2

Date 2015-10-14

Title interaction Sequence/SNP-set Kernel Association Test (iSKAT) / Gene-Environment Set Association Test (GESAT)

Author
Xinyi (Cindy) Lin <xinyilin@mail.harvard.edu>, Shawn Lee <phila78@gmail.com>

Maintainer Xinyi (Cindy) Lin <xinyilin@mail.harvard.edu>, Shawn Lee <phila78@gmail.com>

Depends SKAT, penalized

Description Test for Interactions Between a Genetic Marker Set and Environment. Please also install SKAT (last compatibility version check V0.78) if you want to use SSD related functions. Prior versions have major bugs, please do not use versions <0.3. Versions 1.0 onwards implements both GESAT and iSKAT. Note functions name change from versions 1.0 onwards. In versions <1.0, iSKAT() function implements GESAT (Lin et al., Biostatistics 2013). In versions 1.0 onwards, iSKAT() function implements the optimal test (Lin et al., Biometrics, in press) while GESAT() function implements GESAT (Lin et al., Biostatistics 2013). For versions <1.2, if is_check_genotype==FALSE and is_dosage== FALSE, please ensure that Z has no missingness (versions 1.2 onwards will give an error message, but earlier version would not).

License GPL (version 2 or later).

R topics documented:

GESAT *Gene-Environment Set Association Test*

Description

Test for interactions between a set of SNPS/genes and Environment. GESAT/iSKAT tests for ZxE (gene-environment interactions), after accounting for main effects of Z (gene), main effects of E (environment) and X (covariates). If the appropriate arguments are set (i.e. same scale.Z, weights.Z, weights.V, MAF_cutoff), iSKAT with r.corr=0 corresponds to GESAT. Note that the default function values for GESAT() and iSKAT() can be different. Warning: Current implementation of GESAT/iSKAT assumes large sample asymptotics. We do not recommend using GESAT/iSKAT for SNP-sets where p/n > 1/3 where p and n are defined as below.

Usage

```
GESAT(Z, Y, E, X=NULL, type="davies",
        lower=1e-20, upper=sqrt(nrow(Y))/log(nrow(Y)), nintervals=5,
        plotGCV=FALSE, plotfile=NA, scale.Z=TRUE, weights.Z=NULL,
        weights.V=NULL, out_type="C", impute.method = "fixed",
        is_check_genotype=TRUE, is_dosage=FALSE, missing_cutoff=0.15,
        SetID=NULL)

GESAT.SSD.OneSet(SSD.INFO, SetID, ...)

GESAT.SSD.OneSet_SetIndex(SSD.INFO, SetIndex, ... )
```

Arguments

Z	a n x p numeric genotype matrix with each row as a different individual and each column as a separate gene/snp. Each genotype should be coded as 0, 1, 2, and 9 (or NA) for AA, Aa, aa, and missing, where A is a major allele and a is a minor allele. Missing genotypes will be imputed by the simple Hardy-Weinberg equilibrium (HWE) based imputation.
Y	a n x 1 matrix of the phenotype. Cannot contain missing values.
E	a n x r numeric environmental matrix with each row as a different individual and each column as an environmental variable. Cannot contain missing values. If the environmental variable is categorical or non-numerical, it has to be recoded into a numeric variable.
X	a n x q numeric covariates matrix with each row as a different individual and each column as a covariate (default=NULL). Cannot contain missing values. X should not include an intercept. X should not include the variables in E.
type	a method to compute the p-value (default= "davies"). "davies" represents an exact method that computes the p-value by inverting the characteristic function of the mixture chisq, "liu" represents an approximation method that matches the first 3 moments.
lower	a scalar for the lower bound of the tuning parameter (default=1e-20). lower has to be positive and <= upper.

`upper` a scalar for the upper bound of the tuning parameter (default=sqrt(nrow(Y))/log(nrow(Y))). upper has to be positive and >= lower.

`nintervals` a scalar for the number of tuning parameters to search over (default=5). resulting possible tuning parameters are c(exp(seq(log(lower),log(upper), length=nintervals))). Computation time depends on nintervals.

`plotGCV` TRUE/FALSE (default=FALSE). whether or not to plot the GCV tuning parameter. default=FALSE.

`plotfile` filename to save the GCV plot (default=NA). must have a .pdf extension. plotfile is ignored if plotGCV=FALSE.

`scale.Z` TRUE/FALSE (default=TRUE). whether or not to scale Z matrix to mean zero and unit variance before applying ridge penalty when adjusting for main effects of Z. If scale.Z=TRUE, weights.Z are ignored.

`weights.Z` a p x 1 vector of weights for Z in main effects (default=NULL). If scale.Z=TRUE, weights.Z are ignored.

`weights.V` a p x 1 vector of weights for Z in interaction effects (default=NULL).

`out_type` "C" for continuous phenotype Y and "D" for binary phenotype Y (default="C").

`impute.method` a method to impute missing genotypes (default= "fixed"). "random" imputes missing genotypes by generating binomial(2,p) random variables (p is the MAF), and "fixed" imputes missing genotypes by assigning the mean genotype value (2p). If you use "random", you will have different p-values for different runs because imputed values are randomly assigned. Can use set.seed() to replicate results.

`is_check_genotype` a logical value indicating whether to check the validity of the genotype matrix Z (default= TRUE). If you use non-SNP type data and want to run iSKAT, please set it to FALSE. If you use SNP data or imputed data, please set it to TRUE. Note that if is_check_genotype=FALSE, missing values in Z have to be coded as NA as 9 will not be treated as missing.

`is_dosage` a logical value indicating whether the matrix Z is a dosage matrix (default= FALSE). If it is TRUE, GESAT will ignore "is_check_genotype" and "impute.method" and GESAT will check the genotype matrix and set impute.method="fixed". Note that GESAT will also treat 9 as missing in Z.

`missing_cutoff` a cutoff of the missing rates of SNPs (default=0.15). Any SNP with missing rates higher than cutoff will be excluded from the analysis.

`SetID` a character value of Set ID. You can find a set ID of each set from SetInfo object of SSD.INFO

`SSD.INFO` an SSD_INFO object returned from Open_SSD.

`SetIndex` a numeric value of Set index. You can find a set index of each set from SetInfo object of SSD.INFO

`...` furthuer arguments to be passed to "GESAT"

Details

Data Format: Y, E, Z, X(if not NULL) should all be matrices with the same no. of rows. Y, E, X cannot have any missing values. Please remove all individuals with missing Y, E, X prior to analysis (If plink files are used, these individuals have to be removed from all the plink files, e.g. prior to

generating the SSD files.). Missingness in Z is allowed and imputation will be used as described above.

SSD Files: If you want to use the SSD file, open it first, and then use either GESAT.SSD.OneSet or GESAT.SSD.OneSet_SetIndex. Set index is a numeric value and it is automatically assigned to each set (from 1).

Tuning Parameter: Upper should not be set to too large a value as if the chosen tuning parameter is too large, the main effects of Z are effectively shrunk to zero. This results in testing for ZxE without accounting for main effects of Z.

Value

`pvalue`	the p-value of GESAT.
`Is_Converge`	an indicator of the convergence. 1 indicates the method converged, and 0 indicates the method did not converge. When Is_Converge=0 (no convergence), "liu" method is used to compute p-value. Note that if method="liu", Is_converge=1 always.
`lambda`	chosen tuning parameter.
`n.G.test`	number of SNPs in the genotype matrix used in the test. It can be different from ncol(Z) when some markers are monomorphic or have higher missing rates than the missing_cutoff.
`n.GE.test`	number of columns in the ZxE interaction matrix used in the test. It can be different from ncol(Z)*ncol(E) when some markers are monomorphic or have higher missing rates than the missing_cutoff. It can also be different from n.G.test*ncol(E) as the resulting ZxE variable might have no variability or might be perfectly collinear with columns of Z. Note that singletons are adjusted for in the main effects but are not tested for interactions. Likewise, columns of ZxE perfectly collinear with columns of Z are not tested for interactions.
`Error`	1= Error, 0=No Error.

Author(s)

Xinyi (Cindy) Lin

References

Lin, X., Lee, S., Christiani, D. C., and Lin, X. (2013). Test for the Interaction between a Genetic Marker Set and Environment in Generalized Linear Models. Biostatistics, 14: 667-681. doi:10.1093/biostatistics/kxt006.

Examples

```
############################################################
# Generate data
############################################################
set.seed(1)
n <- 1000
p <- 10
Y <- matrix(rnorm(n), ncol=1)
Z <- matrix(rbinom(n*p, 2, 0.3), nrow=n)
E <- matrix(rnorm(n))
X <- matrix(rnorm(n*2), nrow=n)
set.seed(2)
```

```
Ybinary <- matrix(rbinom(n, 1,0.5), ncol=1)


############################################################
# Compute the P-value of GESAT  - without covariates
############################################################
GESAT(Z, Y, E)
GESAT(Z, Ybinary, E, out_type="D")


############################################################
# Compute the P-value of GESAT  - with covariates
############################################################
GESAT(Z, Y, E, X)
GESAT(Z, Ybinary, E, X, out_type="D")
```

`GESAT.SSD.All` *Gene-Environment Set Association Test*

Description

Iteratively test for interactions between a set of SNPS/genes and Environment for SNP sets in SSD file.

Usage

```
GESAT.SSD.All(SSD.INFO, ...)
```

Arguments

`SSD.INFO`	an SSD_INFO object returned from Open_SSD.
`...`	further arguments to be passed to "GESAT".

Details

Returns a data frame, where each row gives pvalue, Is_converge, lambda, n.G.test, n.GE.test, Error for that particular SETID. Please see GESAT for details.

Value

`SetID`	SetID.
`pvalue`	the p-value of GESAT.
`Is_Converge`	an indicator of the convergence. 1 indicates the method converged, and 0 indicates the method did not converge. When Is_Converge=0 (no convergence), "liu" method is used to compute p-value. Note that if method="liu", Is_converge=1 always.
`lambda`	the chosen tuning parameter.
`n.G.test`	the number of SNPs in the genotype matrix used in the test. It can be different from ncol(Z) when some markers are monomorphic or have higher missing rates than the missing_cutoff.

n.GE.test	the number of columns in the ZxE interaction matrix used in the test. It can be different from ncol(Z)*ncol(E) when some markers are monomorphic or have higher missing rates than the missing_cutoff. It can also be different from n.G.test*ncol(E) as the resulting ZxE variable might have no variability or might be perfectly collinear with columns of Z. Note that singletons are adjusted for in the main effects but are not tested for interactions. Likewise, columns of ZxE perfectly collinear with columns of Z are not tested for interactions.
Error	1= Error, 0=No Error.

Author(s)

Xinyi (Cindy) Lin

---

iSKAT *interaction Sequence/SNP-set Kernel Association Test*

Description

Test for interactions between a set of SNPS/genes and Environment. GESAT/iSKAT tests for ZxE (gene-environment interactions), after accounting for main effects of Z (gene), main effects of E (environment) and X (covariates). If the appropriate arguments are set (i.e. same scale.Z, weights.Z, weights.V, MAF_cutoff), iSKAT with r.corr=0 corresponds to GESAT. Note that the default function values for GESAT() and iSKAT() can be different. Warning: Current implementation of GESAT/iSKAT assumes large sample asymptotics. We do not recommend using GESAT/iSKAT for SNP-sets where p/n > 1/3 where p and n are defined as below.

Usage

```
iSKAT(Z, Y, E, X=NULL, type="davies",
        lower=1e-20, upper=sqrt(nrow(Y))/log(nrow(Y)), nintervals=5,
        plotGCV=FALSE, plotfile=NA, scale.Z=FALSE, weights.Z=NULL,
        weights.V=NULL, out_type="C", impute.method = "fixed",
        is_check_genotype=TRUE, is_dosage=FALSE, missing_cutoff=0.15,
r.corr=(0:10)/10, weights.beta=c(1,25), MAF_cutoff=0.05,
        SetID=NULL)

iSKAT.SSD.OneSet(SSD.INFO, SetID, ...)

iSKAT.SSD.OneSet_SetIndex(SSD.INFO, SetIndex, ... )
```

Arguments

Z	a n x p numeric genotype matrix with each row as a different individual and each column as a separate gene/snp. Each genotype should be coded as 0, 1, 2, and 9 (or NA) for AA, Aa, aa, and missing, where A is a major allele and a is a minor allele. Missing genotypes will be imputed by the simple Hardy-Weinberg equilibrium (HWE) based imputation.
Y	a n x 1 matrix of the phenotype. Cannot contain missing values.

`E` a n x r numeric environmental matrix with each row as a different individual and each column as an environmental variable. Cannot contain missing values. If the environmental variable is categorical or non-numerical, it has to be recoded into a numeric variable.

`X` a n x q numeric covariates matrix with each row as a different individual and each column as a covariate (default=NULL). Cannot contain missing values. X should not include an intercept. X should not include the variables in E.

`type` a method to compute the p-value (default= "davies"). "davies" represents an exact method that computes the p-value by inverting the characteristic function of the mixture chisq, "liu" represents an approximation method that matches the first 3 moments.

`lower` a scalar for the lower bound of the tuning parameter (default=1e-20). lower has to be positive and <= upper.

`upper` a scalar for the upper bound of the tuning parameter (default=sqrt(nrow(Y))/log(nrow(Y))). upper has to be positive and >= lower.

`nintervals` a scalar for the number of tuning parameters to search over (default=5). resulting possible tuning parameters are c(exp(seq(log(lower),log(upper), length=nintervals))). Computation time depends on nintervals.

`plotGCV` TRUE/FALSE (default=FALSE). whether or not to plot the GCV tuning parameter. default=FALSE.

`plotfile` filename to save the GCV plot (default=NA). must have a .pdf extension. plotfile is ignored if plotGCV=FALSE.

`scale.Z` TRUE/FALSE (default=FALSE). whether or not to scale Z matrix to mean zero and unit variance before applying ridge penalty when adjusting for main effects of Z. If scale.Z=TRUE, weights.Z are ignored.

`weights.Z` a p x 1 vector of weights for Z in main effects (default=NULL). If scale.Z=TRUE, weights.Z are ignored.

`weights.V` a p x 1 vector of weights for Z in interaction effects (default=NULL).

`out_type` "C" for continuous phenotype Y and "D" for binary phenotype Y (default="C").

`impute.method` a method to impute missing genotypes (default= "fixed"). "random" imputes missing genotypes by generating binomial(2,p) random variables (p is the MAF), and "fixed" imputes missing genotypes by assigning the mean genotype value (2p). If you use "random", you will have different p-values for different runs because imputed values are randomly assigned. Can use set.seed() to replicate results.

`is_check_genotype` a logical value indicating whether to check the validity of the genotype matrix Z (default= TRUE). If you use non-SNP type data and want to run iSKAT, please set it to FALSE. If you use SNP data or imputed data, please set it to TRUE. Note that if is_check_genotype=FALSE, missing values in Z have to be coded as NA as 9 will not be treated as missing.

`is_dosage` a logical value indicating whether the matrix Z is a dosage matrix (default= FALSE). If it is TRUE, iSKAT will ignore “is_check_genotype” and “impute.method” and iSKAT will check the genotype matrix and set impute.method="fixed". Note that iSKAT will also treat 9 as missing in Z.

`missing_cutoff` a cutoff of the missing rates of SNPs (default=0.15). Any SNP with missing rates higher than cutoff will be excluded from the analysis.

r.corr	the ρ parameter of new class of kernels with compound symmetric correlation structure for interaction effects (default= (0:10)/10). If you give a vector value, iSKAT will conduct the optimal test.
weights.beta	a numeric vector of parameters of beta weights. It is only used for main effects of Z when scale.Z=FALSE and weights.Z=NULL. It is only used for interaction effects when weights.V=NULL. If you want to use your own weights, please specify weights.Z and weights.V accordingly.
MAF_cutoff	a cutoff of the MAFs of the SNPs (default=0.05). Any SNP with MAFs higher than cutoff will be excluded from the analysis.
SetID	a character value of Set ID. You can find a set ID of each set from SetInfo object of SSD.INFO
SSD.INFO	an SSD_INFO object returned from Open_SSD.
SetIndex	a numeric value of Set index. You can find a set index of each set from SetInfo object of SSD.INFO
...	furthuer arguments to be passed to "iSKAT"

Details

Data Format: Y, E, Z, X(if not NULL) should all be matrices with the same no. of rows. Y, E, X cannot have any missing values. Please remove all individuals with missing Y, E, X prior to analysis (If plink files are used, these individuals have to be removed from all the plink files, e.g. prior to generating the SSD files.). Missingness in Z is allowed and imputation will be used as described above.

SSD Files: If you want to use the SSD file, open it first, and then use either iSKAT.SSD.OneSet or iSKAT.SSD.OneSet_SetIndex. Set index is a numeric value and it is automatically assigned to each set (from 1).

Tuning Parameter: Upper should not be set to too large because if the chosen tuning parameter is too large, the main effects of Z are effectively shrunk to zero. This results in testing for ZxE without accounting for main effects of Z.

GESAT vs. iSKAT: iSKAT()$param$minp with appropriate arguments and r.corr=0 is identical to GESAT() p-value. iSKAT()$pvalue with appropriate arguments and r.corr=0 is similar (but may not be identical) to GESAT() p-value. For more details on how the two are related, see examples below.

Value

pvalue	the p-value of iSKAT.
param	estimated parameters.
lambda	chosen tuning parameter.
n.G.test	number of SNPs in the genotype matrix used in the test. It can be different from ncol(Z) when some markers are monomorphic or have higher missing rates than the missing_cutoff.
n.GE.test	number of columns in the ZxE interaction matrix used in the test. It can be different from ncol(Z)*ncol(E) when some markers are monomorphic or have higher missing rates than the missing_cutoff. It can also be different from n.G.test*ncol(E) as the resulting ZxE variable might have no variability or might be perfectly collinear with columns of Z. Note that singletons are adjusted for in the main effects but are not tested for interactions. Likewise, columns of ZxE perfectly collinear with columns of Z are not tested for interactions.
Error	1= Error, 0=No Error.

Author(s)

Xinyi (Cindy) Lin

References

Lin, X., Lee, S., Wu, M., Wang, C., Chen H., Li, Z. and Lin, X. Test for rare variants by environment interactions in sequencing association studies. Biometrics, in press.

Examples

```
############################################################
# Generate data
############################################################
set.seed(1)
n <- 1000
p <- 10
Y <- matrix(rnorm(n), ncol=1)
Z <- matrix(rbinom(n*p, 2, 0.3), nrow=n)
E <- matrix(rnorm(n))
X <- matrix(rnorm(n*2), nrow=n)
Zrare <- matrix(rbinom(n*p, 2, 0.03), nrow=n)
set.seed(2)
Ybinary <- matrix(rbinom(n, 1,0.5), ncol=1)


############################################################
# iSKAT()$param$minp with appropriate arguments and r.corr=0 gives GESAT() p-value
# Compare $param$minp here with the examples in GESAT()
############################################################
iSKAT(Z, Y, E, scale.Z=TRUE, r.corr=0, MAF_cutoff=1, weights.beta=NULL)
iSKAT(Z, Ybinary, E, out_type="D", scale.Z=TRUE, r.corr=0, MAF_cutoff=1,
weights.beta=NULL)
iSKAT(Z, Y, E, X, scale.Z=TRUE, r.corr=0, MAF_cutoff=1, weights.beta=NULL)
iSKAT(Z, Ybinary, E, X, out_type="D", scale.Z=TRUE, r.corr=0, MAF_cutoff=1,
 weights.beta=NULL)

# More comparisons
iSKAT(Zrare, Y, E, scale.Z=TRUE, r.corr=0, MAF_cutoff=1, weights.beta=NULL)
GESAT(Zrare, Y, E)
iSKAT(Zrare, Ybinary, E, out_type="D", scale.Z=TRUE, r.corr=0, MAF_cutoff=1,
 weights.beta=NULL)
GESAT(Zrare, Ybinary, E, out_type="D")
iSKAT(Zrare, Y, E, X, scale.Z=TRUE, r.corr=0, MAF_cutoff=1, weights.beta=NULL)
GESAT(Zrare, Y, E, X)
iSKAT(Zrare, Ybinary, E, X, out_type="D", scale.Z=TRUE, r.corr=0, MAF_cutoff=1,
 weights.beta=NULL)
GESAT(Zrare, Ybinary, E, X, out_type="D")


############################################################
# iSKAT() for testing rare variants by environment interactions
############################################################
iSKAT(Zrare, Y, E)
iSKAT(Zrare, Ybinary, E, out_type="D")
iSKAT(Zrare, Y, E, X)
iSKAT(Zrare, Ybinary, E, X, out_type="D")
```

`iSKAT.SSD.All` *interaction Sequence/SNP-set Kernel Association Test*

Description

Iteratively test for interactions between a set of SNPS/genes and Environment for SNP sets in SSD file.

Usage

```
iSKAT.SSD.All(SSD.INFO, ...)
```

Arguments

`SSD.INFO`	an SSD_INFO object returned from Open_SSD.
`...`	further arguments to be passed to "iSKAT".

Details

Returns a data frame, where each row gives pvalue, rho_est, lambda, n.G.test, n.GE.test, Error for that particular SETID. Please see iSKAT for details.

Value

`SetID`	SetID.
`pvalue`	the p-value of iSKAT.
`rho_est`	the estimated ρ parameter.
`lambda`	the chosen tuning parameter.
`n.G.test`	the number of SNPs in the genotype matrix used in the test. It can be different from ncol(Z) when some markers are monomorphic or have higher missing rates than the missing_cutoff.
`n.GE.test`	the number of columns in the ZxE interaction matrix used in the test. It can be different from ncol(Z)*ncol(E) when some markers are monomorphic or have higher missing rates than the missing_cutoff. It can also be different from n.G.test*ncol(E) as the resulting ZxE variable might have no variability or might be perfectly collinear with columns of Z. Note that singletons are adjusted for in the main effects but are not tested for interactions. Likewise, columns of ZxE perfectly collinear with columns of Z are not tested for interactions.
`Error`	1= Error, 0=No Error.

Author(s)

Xinyi (Cindy) Lin

Index