Package 'SAIGEgds'

April 12, 2022

Type Package

Title Scalable Implementation of Generalized mixed models using GDS files in Phenome-Wide Association Studies

Version 1.8.1

Date 2021-09-17

Depends R (>= 3.5.0), gdsfmt (>= 1.20.0), SeqArray (>= 1.31.8), Rcpp

LinkingTo Rcpp, RcppArmadillo, RcppParallel (>= 5.0.0)

Imports methods, stats, utils, RcppParallel, SPAtest (>= 3.0.0)

Suggests parallel, crayon, RUnit, knitr, markdown, rmarkdown, BiocGenerics, SNPRelate

Description Scalable implementation of generalized mixed models with highly optimized C++ implementation and integration with Genomic Data Structure (GDS) files. It is designed for single variant tests in large-scale phenome-wide association studies (PheWAS) with millions of variants and samples, controlling for sample structure and case-control imbalance.
The implementation is based on the original SAIGE R package (v0.29.4.4 for single variant tests, Zhou et al. 2018).
SAIGEgds also implements some of the SPAtest functions in C to speed up the calculation of Saddlepoint approximation. Benchmarks show that SAIGEgds is 5 to 6 times faster than the original SAIGE R package.

License GPL-3

SystemRequirements C++11, GNU make

VignetteBuilder knitr

ByteCompile TRUE

URL https://github.com/AbbVie-ComputationalGenomics/SAIGEgds

biocViews Software, Genetics, StatisticalMethod

git_url https://git.bioconductor.org/packages/SAIGEgds

git_branch RELEASE_3_14

git_last_commit 13e4275

git_last_commit_date 2022-03-29

Date/Publication 2022-04-12

Author Xiuwen Zheng [aut, cre] (<https://orcid.org/0000-0002-1390-0708>),
Wei Zhou [ctb] (the original author of the SAIGE R package),
J. Wade Davis [ctb]

Maintainer Xiuwen Zheng <xiuwen.zheng@abbvie.com>

R topics documented:

`SAIGEgds-package` *Scalable Implementation of Generalized mixed models in Phenome-Wide Association Studies using GDS files*

Description

Scalable and accurate implementation of generalized mixed mode with the support of Genomic Data Structure (GDS) files and highly optimized C++ implementation. It is designed for single variant tests in large-scale phenome-wide association studies (PheWAS) with millions of variants and hundreds of thousands of samples, e.g., UK Biobank genotype data, controlling for case-control imbalance and sample structure in single variant association studies.

The implementation of SAIGEgds is based on the original SAIGE R package (v0.29.4.4) [Zhou et al. 2018] https://github.com/weizhouUMICH/SAIGE/releases/tag/v0.29.4.4. All of the calculation with single-precision floating-point numbers in SAIGE are replaced by the double-precision calculation in SAIGEgds. SAIGEgds also implements some of the SPAtest functions in C to speed up the calculation of Saddlepoint Approximation.

Details

Package:	SAIGEgds
Type:	Package
License:	GPL version 3

Author(s)

Xiuwen Zheng <xiuwen.zheng@abbvie.com>, Wei Zhou (the original author of the SAIGE R package, https://github.com/weizhouUMICH/SAIGE)

References

Zheng X, Davis J.Wade. SAIGEgds – an efficient statistical tool for large-scale PheWAS with mixed models. *Bioinformatics* (2020). DOI: 10.1093/bioinformatics/btaa731.

Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A, Bastarache LA, Wei WQ, Denny JC, Lin M, Hveem K, Kang HM, Abecasis GR, Willer CJ, Lee S. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* (2018). Sep;50(9):1335-1341.

Zheng X, Gogarten S, Lawrence M, Stilp A, Conomos M, Weir BS, Laurie C, Levine D. SeqArray – A storage-efficient high-performance data format for WGS variant calls. *Bioinformatics* (2017). DOI: 10.1093/bioinformatics/btx145.

Examples

```
# open the GDS file
fn <- system.file("extdata", "grm1k_10k_snp.gds", package="SAIGEgds")
gdsfile <- seqOpen(fn)

# load phenotype
phenofn <- system.file("extdata", "pheno.txt.gz", package="SAIGEgds")
pheno <- read.table(phenofn, header=TRUE, as.is=TRUE)
head(pheno)

# fit the null model
glmm <- seqFitNullGLMM_SPA(y ~ x1 + x2, pheno, gdsfile, trait.type="binary")

# p-value calculation
assoc <- seqAssocGLMM_SPA(gdsfile, glmm, mac=10)

head(assoc)

# close the GDS file
seqClose(gdsfile)
```

---

glmmHeritability *Heritability estimation*

Description

Get the heritability estimate from the SAIGE model.

Usage

```
glmmHeritability(modobj, adjust=TRUE)
```

Arguments

modobj	an R object for SAIGE model parameters
adjust	if TRUE and binary outcomes, uses adjusted tau estimate for the heritability estimation

Details

In SAIGE, penalized quasi-likelihood (PQL) is used to estimate the variance component parameter tau. It is known to produce biased estimate of the variance component tau using PQL. If adjust=TRUE for binary outcomes, tau is adjusted based prevalence and observed tau using the data in Supplementary Table 7 (Zhou et al. 2018) to reduce the bias of PQL estimate of variance component.

Value

Return a liability scale heritability.

Author(s)

Xiuwen Zheng

See Also

seqFitNullGLMM_SPA

Examples

```
# open a GDS file
fn <- system.file("extdata", "grm1k_10k_snp.gds", package="SAIGEgds")
gdsfile <- seqOpen(fn)

# load phenotype
phenofn <- system.file("extdata", "pheno.txt.gz", package="SAIGEgds")
pheno <- read.table(phenofn, header=TRUE, as.is=TRUE)
head(pheno)

# fit the null model
glmm <- seqFitNullGLMM_SPA(y ~ x1 + x2, pheno, gdsfile, trait.type="binary")

glmmHeritability(glmm)

seqClose(gdsfile)
```

pACAT *Cauchy Combination Test*

Description

P-value calculation from Cauchy combination test.

Usage

```
pACAT(p, w=NULL)
pACAT2(p, maf, wbeta=c(1,25))
```

Arguments

p	a numeric vector for p-values
w	weight for each p-value
maf	minor allele frequency for each p-value
wbeta	weights for per-variant effect, using beta distribution `dbeta()` according to variant's MAF

Value

Return a single number for the combined p-value.

References

Liu Y., Cheng S., Li Z., Morrison A.C., Boerwinkle E., Lin X.; ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. Am J Hum Genetics 104, 410-421 (2019).

See Also

seqFitNullGLMM_SPA, seqAssocGLMM_SPA

Examples

```
p1 <- 10^-4
p2 <- 10^-5
p3 <- 10^-(3:20)
sapply(p3, function(p) pACAT(c(p1, p2, p)))

pACAT2(c(10^-4, 10^-6), c(0.01, 0.005))
```

seqAssocGLMM_SPA *P-value calculation*

Description

P-value calculations using variance approximation and an adjustment of Saddlepoint approximation.

Usage

```
seqAssocGLMM_SPA(gdsfile, modobj, maf=NaN, mac=10, missing=0.1, dsnode="",
    spa.pval=0.05, var.ratio=NaN, res.savefn="", res.compress="LZMA",
    parallel=FALSE, verbose=TRUE)
```

Arguments

gdsfile	a SeqArray GDS filename, or a GDS object
modobj	an R object for SAIGE model parameters
maf	minor allele frequency threshold (checking >= maf), NaN for no filter
mac	minor allele count threshold (checking >= mac), NaN for no filter
missing	missing threshold for variants (checking <= missing), NaN for no filter
dsnode	"" for automatically searching the GDS nodes "genotype" and "annotation/format/DS", or use a user-defined GDS node in the file
spa.pval	the p-value threshold for SPA adjustment, 0.05 by default (since normal approximation performs well when the test statistic is close to the mean)
var.ratio	NaN for using the estimated variance ratio in the model fitting, or a user-defined variance ratio
res.savefn	an RData or GDS file name, "" for no saving
res.compress	the compression method for the output file, it should be one of LZMA, LZMA_RA, ZIP, ZIP_RA and none
parallel	FALSE (serial processing), TRUE (multicore processing), a numeric value for the number of cores, or other value; parallel is passed to the argument cl in seqParallel, see seqParallel for more details
verbose	if TRUE, show information

Details

The original SAIGE R package uses 0.05 as a threshold for unadjusted p-values (based on asymptotic normality) to further calculate adjusted p-values (Saddlepoint approximation, SPA). If var.ratio=NaN, the average of variance ratios (mean(modobj$var.ratio$ratio)) is used instead. For more details of SAIGE algorithm, please refer to the SAIGE paper [Zhou et al. 2018] (see the reference section).

Value

Return a `data.frame` with the following components if not saving to a file:

`id`	variant ID in the GDS file;
`chr`	chromosome;
`pos`	position;
`rs.id`	the RS IDs if it is available in the GDS file;
`ref`	the reference allele;
`alt`	the alternative allele;
`AF.alt`	allele frequency for the alternative allele; the minor allele frequency is `pmin(AF.alt,1-AF.alt)`;
`mac`	minor allele count; the allele count for the alternative allele is `ifelse(AF.alt<=0.5,mac,2*num-mac)`;
`num`	the number of samples with non-missing genotypes;
`beta`	beta coefficient, odds ratio if binary outcomes (alternative allele vs. reference allele);
`SE`	standard error for beta coefficient;
`pval`	adjusted p-value with the Saddlepoint approximation method;
`p.norm`	p-values based on asymptotic normality (could be 0 if it is too small, e.g., `pnorm(-50) = 0` in R; used for checking only
`converged`	whether the SPA algorithm converges or not for adjusted p-values.

Author(s)

Xiuwen Zheng

References

Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A, Bastarache LA, Wei WQ, Denny JC, Lin M, Hveem K, Kang HM, Abecasis GR, Willer CJ, Lee S. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet (2018). Sep;50(9):1335-1341.

See Also

`seqAssocGLMM_SPA`, `seqSAIGE_LoadPval`

Examples

```
# open a GDS file
fn <- system.file("extdata", "grm1k_10k_snp.gds", package="SAIGEgds")
gdsfile <- seqOpen(fn)

# load phenotype
phenofn <- system.file("extdata", "pheno.txt.gz", package="SAIGEgds")
pheno <- read.table(phenofn, header=TRUE, as.is=TRUE)
head(pheno)
```

```
# fit the null model
glmm <- seqFitNullGLMM_SPA(y ~ x1 + x2, pheno, gdsfile, trait.type="binary")

# p-value calculation
assoc <- seqAssocGLMM_SPA(gdsfile, glmm, mac=10)
head(assoc)

# close the GDS file
seqClose(gdsfile)
```

`seqAssocGLMM_spaACAT_O` *ACAT-V tests*

Description

ACAT-O combined p-value calculations using mixed models and the Saddlepoint approximation method for case-control imbalance.

Usage

```
seqAssocGLMM_spaACAT_O(gdsfile, modobj, units, wbeta=AggrParamBeta,
    burden.mac=10, burden.summac=3, dsnode="", spa.pval=0.05, var.ratio=NaN,
    res.savefn="", res.compress="LZMA", parallel=FALSE,
    verbose=TRUE, verbose.maf=TRUE)
```

Arguments

`gdsfile`	a SeqArray GDS filename, or a GDS object
`modobj`	an R object for SAIGE model parameters
`units`	a list of units of selected variants, with S3 class `"SeqUnitListClass"` defined in the SeqArray package
`wbeta`	weights for per-variant effect, using beta distribution `dbeta()` according to variant's MAF; a length-two vector, or a matrix with two rows for multiple beta parameters; by default, using beta(1,1) and beta(1,25) both
`burden.mac`	a threshold of minor allele count for using burden test instead of single variant test if `mac < burden.mac`
`burden.summac`	a threshold for the weighted sum of minor allele counts in burden test (checking `>= burden.summac`)
`dsnode`	"" for automatically searching the GDS nodes "genotype" and "annotation/format/DS", or use a user-defined GDS node in the file
`spa.pval`	the p-value threshold for SPA adjustment, 0.05 by default
`var.ratio`	`NaN` for using the estimated variance ratio in the model fitting, or a user-defined variance ratio
`res.savefn`	an RData or GDS file name, "" for no saving

res.compress	the compression method for the output file, it should be one of LZMA, LZMA_RA, ZIP, ZIP_RA and none
parallel	FALSE (serial processing), TRUE (multicore processing), a numeric value for the number of cores, or other value; parallel is passed to the argument cl in seqParallel, see seqParallel for more details
verbose	if TRUE, show information
verbose.maf	if TRUE, show summary of MAFs in units

Details

The original SAIGE R package uses 0.05 as a threshold for unadjusted p-values to further calculate SPA-adjusted p-values. If var.ratio=NaN, the average of variance ratios (mean(modobj$var.ratio$ratio)) is used instead. For more details of SAIGE algorithm, please refer to the SAIGE paper [Zhou et al. 2018] (see the reference section). No SKAT implementation.

Value

Return a data.frame with the following components if not saving to a file: chr, chromosome; start, a starting position; end, an ending position; numvar, the number of variants in a window; summac, the weighted sum of minor allele counts; beta, beta coefficient, odds ratio if binary outcomes); SE, standard error for beta coefficient; pval, adjusted p-value with Saddlepoint approximation;

p.norm	p-values based on asymptotic normality (could be 0 if it is too small, e.g., pnorm(-50) = 0 in R; used for checking only

cvg, whether the SPA algorithm converges or not for adjusted p-value.

Author(s)

Xiuwen Zheng

References

Liu Y., Chen S., Li Z., Morrison A.C., Boerwinkle E., Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. Am J Hum Genetics 104, 410-421 (2019).

See Also

seqAssocGLMM_spaBurden, seqAssocGLMM_spaACAT_V

Examples

```
# open a GDS file
fn <- system.file("extdata", "grm1k_10k_snp.gds", package="SAIGEgds")
gdsfile <- seqOpen(fn)

# load phenotype
phenofn <- system.file("extdata", "pheno.txt.gz", package="SAIGEgds")
```

```
pheno <- read.table(phenofn, header=TRUE, as.is=TRUE)
head(pheno)

# fit the null model
glmm <- seqFitNullGLMM_SPA(y ~ x1 + x2, pheno, gdsfile, trait.type="binary")

# get a list of variant units for burden tests
units <- seqUnitSlidingWindows(gdsfile, win.size=500, win.shift=250)

assoc <- seqAssocGLMM_spaACAT_O(gdsfile, glmm, units)
head(assoc)

# close the GDS file
seqClose(gdsfile)
```

`seqAssocGLMM_spaACAT_V` *ACAT-V tests*

Description

ACAT-V p-value calculations using mixed models and the Saddlepoint approximation method for case-control imbalance.

Usage

```
seqAssocGLMM_spaACAT_V(gdsfile, modobj, units, wbeta=AggrParamBeta,
    burden.mac=10, burden.summac=3, dsnode="", spa.pval=0.05, var.ratio=NaN,
    res.savefn="", res.compress="LZMA", parallel=FALSE,
    verbose=TRUE, verbose.maf=TRUE)
```

Arguments

`gdsfile`	a SeqArray GDS filename, or a GDS object
`modobj`	an R object for SAIGE model parameters
`units`	a list of units of selected variants, with S3 class `"SeqUnitListClass"` defined in the SeqArray package
`wbeta`	weights for per-variant effect, using beta distribution `dbeta()` according to variant's MAF; a length-two vector, or a matrix with two rows for multiple beta parameters; by default, using beta(1,1) and beta(1,25) both
`burden.mac`	a threshold of minor allele count for using burden test instead of single variant test if `mac` < `burden.mac`
`burden.summac`	a threshold for the weighted sum of minor allele counts in burden test (checking >= `burden.summac`)
`dsnode`	"" for automatically searching the GDS nodes "genotype" and "annotation/format/DS", or use a user-defined GDS node in the file

spa.pval	the p-value threshold for SPA adjustment, 0.05 by default
var.ratio	NaN for using the estimated variance ratio in the model fitting, or a user-defined variance ratio
res.savefn	an RData or GDS file name, "" for no saving
res.compress	the compression method for the output file, it should be one of LZMA, LZMA_RA, ZIP, ZIP_RA and none
parallel	FALSE (serial processing), TRUE (multicore processing), a numeric value for the number of cores, or other value; parallel is passed to the argument cl in seqParallel, see seqParallel for more details
verbose	if TRUE, show information
verbose.maf	if TRUE, show summary of MAFs in units

Details

Liu Y., Chen S., Li Z., Morrison A.C., Boerwinkle E., Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. Am J Hum Genetics 104, 410-421 (2019).

Value

Return a data.frame with the following components if not saving to a file: chr, chromosome; start, a starting position; end, an ending position; numvar, the number of variants in a window; summac, the weighted sum of minor allele counts; beta, beta coefficient, odds ratio if binary outcomes); SE, standard error for beta coefficient; pval, adjusted p-value with Saddlepoint approximation;

p.norm	p-values based on asymptotic normality (could be 0 if it is too small, e.g., pnorm(-50) = 0 in R; used for checking only

cvg, whether the SPA algorithm converges or not for adjusted p-value.

Author(s)

Xiuwen Zheng

References

XX

See Also

seqAssocGLMM_spaBurden, seqAssocGLMM_spaACAT_O

Examples

```
# open a GDS file
fn <- system.file("extdata", "grm1k_10k_snp.gds", package="SAIGEgds")
gdsfile <- seqOpen(fn)

# load phenotype
phenofn <- system.file("extdata", "pheno.txt.gz", package="SAIGEgds")
pheno <- read.table(phenofn, header=TRUE, as.is=TRUE)
head(pheno)

# fit the null model
glmm <- seqFitNullGLMM_SPA(y ~ x1 + x2, pheno, gdsfile, trait.type="binary")

# get a list of variant units for burden tests
units <- seqUnitSlidingWindows(gdsfile, win.size=500, win.shift=250)

assoc <- seqAssocGLMM_spaACAT_V(gdsfile, glmm, units)
head(assoc)

# close the GDS file
seqClose(gdsfile)
```

---

seqAssocGLMM_spaBurden

Burden tests

Description

Burden p-value calculations using mixed models and the Saddlepoint approximation method for case-control imbalance.

Usage

```
seqAssocGLMM_spaBurden(gdsfile, modobj, units, wbeta=AggrParamBeta,
    summac=3, dsnode="", spa.pval=0.05, var.ratio=NaN, res.savefn="",
    res.compress="LZMA", parallel=FALSE, verbose=TRUE, verbose.maf=TRUE)
```

Arguments

gdsfile	a SeqArray GDS filename, or a GDS object
modobj	an R object for SAIGE model parameters
units	a list of units of selected variants, with S3 class "SeqUnitListClass" defined in the SeqArray package
wbeta	weights for per-variant effect, using beta distribution dbeta() according to variant's MAF; a length-two vector, or a matrix with two rows for multiple beta parameters; by default, using beta(1,1) and beta(1,25) both
summac	a threshold for the weighted sum of minor allele counts (checking >= summac)

dsnode	"" for automatically searching the GDS nodes "genotype" and "annotation/format/DS", or use a user-defined GDS node in the file
spa.pval	the p-value threshold for SPA adjustment, 0.05 by default
var.ratio	NaN for using the estimated variance ratio in the model fitting, or a user-defined variance ratio
res.savefn	an RData or GDS file name, "" for no saving
res.compress	the compression method for the output file, it should be one of LZMA, LZMA_RA, ZIP, ZIP_RA and none
parallel	FALSE (serial processing), TRUE (multicore processing), a numeric value for the number of cores, or other value; parallel is passed to the argument cl in seqParallel, see seqParallel for more details
verbose	if TRUE, show information
verbose.maf	if TRUE, show summary of MAFs in units

Details

The original SAIGE R package uses 0.05 as a threshold for unadjusted p-values to further calculate SPA-adjusted p-values. If var.ratio=NaN, the average of variance ratios (mean(modobj$var.ratio$ratio)) is used instead. For more details of SAIGE algorithm, please refer to the SAIGE paper [Zhou et al. 2018] (see the reference section).

Value

Return a data.frame with the following components if not saving to a file: chr, chromosome; start, a starting position; end, an ending position; numvar, the number of variants in a window; summac, the weighted sum of minor allele counts; beta, beta coefficient, odds ratio if binary outcomes); SE, standard error for beta coefficient; pval, adjusted p-value with Saddlepoint approximation;

p.norm	p-values based on asymptotic normality (could be 0 if it is too small, e.g., pnorm(-50) = 0 in R; used for checking only

cvg, whether the SPA algorithm converges or not for adjusted p-value.

Author(s)

Xiuwen Zheng

References

Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A, Bastarache LA, Wei WQ, Denny JC, Lin M, Hveem K, Kang HM, Abecasis GR, Willer CJ, Lee S. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet (2018). Sep;50(9):1335-1341.

See Also

seqAssocGLMM_spaACAT_V, seqAssocGLMM_spaACAT_O

Examples

```
# open a GDS file
fn <- system.file("extdata", "grm1k_10k_snp.gds", package="SAIGEgds")
gdsfile <- seqOpen(fn)

# load phenotype
phenofn <- system.file("extdata", "pheno.txt.gz", package="SAIGEgds")
pheno <- read.table(phenofn, header=TRUE, as.is=TRUE)
head(pheno)

# fit the null model
glmm <- seqFitNullGLMM_SPA(y ~ x1 + x2, pheno, gdsfile, trait.type="binary")

# get a list of variant units for burden tests
units <- seqUnitSlidingWindows(gdsfile, win.size=500, win.shift=250)

assoc <- seqAssocGLMM_spaBurden(gdsfile, glmm, units)
head(assoc)

# close the GDS file
seqClose(gdsfile)
```

---

seqFitNullGLMM_SPA *Fit the null model with GRM*

Description

Fit the null model in the mixed model framework with genetic relationship matrix (GRM).

Usage

```
seqFitNullGLMM_SPA(formula, data, gdsfile, trait.type=c("binary", "quantitative"),
    sample.col="sample.id", maf=0.005, missing.rate=0.01, max.num.snp=1000000L,
    variant.id=NULL, inv.norm=TRUE, X.transform=TRUE, tol=0.02, maxiter=20L,
    nrun=30L, tolPCG=1e-5, maxiterPCG=500L, num.marker=30L, tau.init=c(0,0),
    traceCVcutoff=0.0025, ratioCVcutoff=0.001, geno.sparse=TRUE, num.thread=1L,
    model.savefn="", seed=200L, fork.loading=FALSE, verbose=TRUE)
```

Arguments

formula	an object of class formula (or one that can be coerced to that class), e.g., y ~ x1 + x2, see lm
data	a data frame for the formulas
gdsfile	a SeqArray GDS filename, or a GDS object
trait.type	"binary" for binary outcomes, "quantitative" for continuous outcomes
sample.col	the column name of sample IDs corresponding to the GDS file

maf	minor allele frequency for imported genotypes (checking >= maf), if variant.id=NULL; NaN for no filter
missing.rate	threshold of missing rate (checking <= missing.rate), if variant.id=NULL; NaN for no filter
max.num.snp	the maximum number of SNPs used, or -1 for no limit
variant.id	a list of variant IDs, used to construct GRM
inv.norm	if TRUE, perform inverse normal transformation on residuals for quantitative outcomes, see the reference [Sofer, 2019]
X.transform	if TRUE, perform QR decomposition on the design matrix
tol	overall tolerance for model fitting
maxiter	the maximum number of iterations for model fitting
nrun	the number of random vectors in the trace estimation
tolPCG	tolerance of PCG iterations
maxiterPCG	the maximum number of PCG iterations
num.marker	the number of SNPs used to calculate the variance ratio
tau.init	a 2-length numeric vector, the initial values for variance components, tau; for binary traits, the first element is always be set to 1. if tau.init is not specified, the second element will be 0.5 for binary traits
traceCVcutoff	the threshold for coefficient of variation (CV) for the trace estimator, and the number of runs for trace estimation will be increased until the CV is below the threshold
ratioCVcutoff	the threshold for coefficient of variation (CV) for estimating the variance ratio, and the number of randomly selected markers will be increased until the CV is below the threshold
geno.sparse	if TRUE, store the sparse structure for genotypes; otherwise, save genotypes in a 2-bit dense matrix; see details
num.thread	the number of threads
model.savefn	the filename of model output, R data file '.rda', '.RData', or '.rds'
seed	an integer as a seed for random numbers
fork.loading	load genotypes via forking or not; forking processes in Unix can reduce loading time of genotypes, but may double the memory usage; not applicable on Windows
verbose	if TRUE, show information

Details

Utilizing the sparse structure of genotypes could significantly improve the computational efficiency of model fitting, but it also increases the memory usage. For more details of SAIGE algorithm, please refer to the SAIGE paper [Zhou et al. 2018] (see the reference section).

Value

Returns a list with the following components:

`coefficients`	the beta coefficients for fixed effects;
`tau`	a numeric vector of variance components 'Sigma_E' and 'Sigma_G';
`linear.predictors`	the linear fit on link scale;
`fitted.values`	fitted values from objects returned by modeling functions using `glm.fit`;
`residuals`	residuals;
`cov`	covariance matrix of beta coefficients;
`converged`	whether the model is fitted or not;
`obj.noK`	internal use, returned object from the SPAtest package;
`var.ratio`	a data.frame with columns 'id' (variant.id), 'maf' (minor allele frequency), 'mac' (minor allele count), 'var1' (the variance of score statistic), 'var2' (a variance estimate without accounting for estimated random effects) and 'ratio' (var1/var2, estimated variance ratio for variance approximation);
`trait.type`	either "binary" or "quantitative";
`sample.id`	the sample IDs used in the model fitting;
`variant.id`	the variant IDs used in the model fitting.

Author(s)

Xiuwen Zheng

References

Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A, Bastarache LA, Wei WQ, Denny JC, Lin M, Hveem K, Kang HM, Abecasis GR, Willer CJ, Lee S. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet (2018). Sep;50(9):1335-1341.

T Sofer, X Zheng, SM Gogarten, CA Laurie, etc. A fully adjusted two-stage procedure for rank-normalization in genetic association studies. 2019. Genetic Epidemiology 43(3), 263-275

See Also

seqAssocGLMM_SPA

Examples

```
# open a GDS file
fn <- system.file("extdata", "grm1k_10k_snp.gds", package="SAIGEgds")
gdsfile <- seqOpen(fn)

# load phenotype
phenofn <- system.file("extdata", "pheno.txt.gz", package="SAIGEgds")
pheno <- read.table(phenofn, header=TRUE, as.is=TRUE)
```

```
head(pheno)

# fit the null model
glmm <- seqFitNullGLMM_SPA(y ~ x1 + x2, pheno, gdsfile, trait.type="binary")
glmm

# close the GDS file
seqClose(gdsfile)
```

`seqGLMM_GxG_spa` *SNP Interaction Testing*

Description

SNP interaction testing with Saddlepoint approximation method in the mixed framework.

Usage

```
seqGLMM_GxG_spa(formula, data, gds_grm, gds_assoc, snp_pair,
    trait.type=c("binary", "quantitative"), sample.col="sample.id", maf=0.005,
    missing.rate=0.01, max.num.snp=1000000L, variant.id=NULL, inv.norm=TRUE,
    X.transform=TRUE, tol=0.02, maxiter=20L, nrun=30L, tolPCG=1e-5,
   maxiterPCG=500L, tau.init=c(0,0), use_approx_tau=FALSE, glm_threshold=FALSE,
    traceCVcutoff=0.0025, ratioCVcutoff=0.001, geno.sparse=TRUE, num.thread=1L,
    model.savefn="", seed=200L, fork.loading=FALSE, verbose=TRUE,
    verbose.detail=TRUE)
```

Arguments

`formula`	an object of class `formula` (or one that can be coerced to that class), e.g., `y ~ x1 + x2`, see `lm`
`data`	a data frame for the formulas
`gds_grm`	a SeqArray GDS filename, or a GDS object
`gds_assoc`	a SeqArray GDS filename, a GDS object, or a 0/1/2/NA matrix with row names for sample IDs
`snp_pair`	a `data.frame` with the first two columns for the variant IDs in `gds_assoc`
`trait.type`	"binary" for binary outcomes, "quantitative" for continuous outcomes
`sample.col`	the column name of sample IDs corresponding to the GDS file
`maf`	minor allele frequency for imported genotypes (checking >= maf), if `variant.id=NULL`; `NaN` for no filter
`missing.rate`	threshold of missing rate (checking <= missing.rate), if `variant.id=NULL`; `NaN` for no filter
`max.num.snp`	the maximum number of SNPs used, or -1 for no limit
`variant.id`	a list of variant IDs, used to construct GRM

inv.norm	if TRUE, perform inverse normal transformation on residuals for quantitative outcomes, see the reference [Sofer, 2019]
X.transform	if TRUE, perform QR decomposition on the design matrix
tol	overall tolerance for model fitting
maxiter	the maximum number of iterations for model fitting
nrun	the number of random vectors in the trace estimation
tolPCG	tolerance of PCG iterations
maxiterPCG	the maximum number of PCG iterations
tau.init	a 2-length numeric vector, the initial values for variance components, tau; for binary traits, the first element is always be set to 1. if tau.init is not specified, the second element will be 0.5 for binary traits
use_approx_tau	if TRUE, fit the model defined in formula without any SNP markers for the interactions to provide the estimated tau value (variance component estimates)
glm_threshold	FALSE, TRUE or a numeric value for p-value threshold; if TRUE use 0.01 as a threshold
traceCVcutoff	the threshold for coefficient of variation (CV) for the trace estimator, and the number of runs for trace estimation will be increased until the CV is below the threshold
ratioCVcutoff	the threshold for coefficient of variation (CV) for estimating the variance ratio, and the number of randomly selected markers will be increased until the CV is below the threshold
geno.sparse	if TRUE, store the sparse structure for genotypes; otherwise, save genotypes in a 2-bit dense matrix; see details
num.thread	the number of threads
model.savefn	the filename of model output, R data file '.rda', '.RData', '.rds', '.txt' or '.csv'
seed	an integer as a seed for random numbers
fork.loading	load genotypes via forking or not; forking processes in Unix can reduce loading time of genotypes, but may double the memory usage; not applicable on Windows
verbose	if TRUE, show information
verbose.detail	if TRUE, show the details for model fitting

Details

For more details of SAIGE algorithm, please refer to the SAIGE paper [Zhou et al. 2018] (see the reference section).

Value

Return a data.frame with the following components:

id1	variant ID for the first SNP in the GDS file;
snp1	includes chromosome, position, reference & alterative alleles for SNP1;

maf1	minor allele frequency for the first SNP;
id2	variant ID for the second SNP in the GDS file;
snp2	includes chromosome, position, reference & alterative alleles for SNP2;
maf2	minor allele frequency for the second SNP;
beta	beta coefficient, odds ratio if binary outcomes;
SE	standard error for beta coefficient;
n_nonzero	the number of non-zero values in the interaction term;
pval	adjusted p-value with the Saddlepoint approximation method;
p.norm	p-values based on asymptotic normality (could be 0 if it is too small, e.g., `pnorm(-50) = 0` in R; used for checking only
converged	whether the SPA algorithm converges or not for adjusted p-values.
p.glm	glm p-value with SPA calculation
p.glm.norm	glm p-value without SPA calculation

Author(s)

Xiuwen Zheng

References

Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A, Bastarache LA, Wei WQ, Denny JC, Lin M, Hveem K, Kang HM, Abecasis GR, Willer CJ, Lee S. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet (2018). Sep;50(9):1335-1341.

See Also

seqFitNullGLMM_SPA, seqAssocGLMM_SPA

Examples

```
# open the GDS file for genetic relationship matrix (GRM)
grm_fn <- system.file("extdata", "grm1k_10k_snp.gds", package="SAIGEgds")
(grm_gds <- seqOpen(grm_fn))

# load phenotype
phenofn <- system.file("extdata", "pheno.txt.gz", package="SAIGEgds")
pheno <- read.table(phenofn, header=TRUE, as.is=TRUE)
head(pheno)

# define the SNP pairs
snp_pair <- data.frame(s1=2:3, s2=6:7, note=c("F1", "F2"))

seqGLMM_GxG_spa(y ~ x1 + x2, pheno, grm_gds, grm_fn, snp_pair,
    trait.type="binary", verbose.detail=FALSE)
```

`seqSAIGE_LoadPval` *Load the association results*

Description

Load the association results from an RData, RDS or GDS file.

Usage

```
seqSAIGE_LoadPval(fn, varnm=NULL, index=NULL, verbose=TRUE)
```

Arguments

`fn`	RData, RDS or GDS file names, merging datasets if multiple files
`varnm`	NULL, or a character vector to include the column names; e.g., `c("chr","position","rs.id","ref","`
`index`	NULL, or a logical/numeric vector for a set of rows
`verbose`	if `TRUE`, show information

Value

Return a `data.frame` including p-values.

Author(s)

Xiuwen Zheng

See Also

`seqFitNullGLMM_SPA`, `seqAssocGLMM_SPA`

Examples

```
(fn <- system.file("unitTests", "saige_pval.rds", package="SAIGEgds"))
pval <- seqSAIGE_LoadPval(fn)

names(pval)
#  [1] "id"         "chr"        "pos"        "rs.id"      "ref"
#  [6] "alt"        "AF.alt"     "AC.alt"     "num"        "beta"
# [11] "SE"         "pval"       "pval.noadj" "converged"

head(pval)
```

Index