Package 'GWASinlps'

October 20, 2024

Type Package

Title Non-Local Prior Based Iterative Variable Selection Tool for Genome-Wide Association Studies

Version 2.3

Date 2024-10-19

Description Performs variable selection with data from Genome-wide association studies (GWAS), or other high-dimensional data with continuous, binary or survival outcomes, combining in an iterative framework the computational efficiency of the structured screen-and-select variable selection strategy based on some association learning and the parsimonious uncertainty quantification provided by the use of non-local priors (see Sanyal et al., 2019 <DOI:10.1093/bioinformatics/bty472>).

License GPL (>= 2)

Depends mombf

Imports Rcpp (>= 1.0.9), RcppArmadillo, fastglm, horseshoe, survival

Suggests glmnet

LinkingTo Rcpp, RcppArmadillo

URL https://nilotpalsanyal.github.io/GWASinlps/

BugReports https://github.com/nilotpalsanyal/GWASinlps/issues

Repository CRAN

NeedsCompilation yes

Author Nilotpal Sanyal [aut, cre] (<https://orcid.org/0000-0003-4814-7602>)

Maintainer Nilotpal Sanyal <nilotpal.sanyal@gmail.com>

Date/Publication 2024-10-20 04:10:02 UTC

Contents

GWASinlps-package 2
GWASinlps 3
nlps 8

Index **11**

GWASinlps-package *Non-local prior based iterative variable selection tool for genome-wide association study data, or other high-dimensional data*

Description

The **GWASinlps** package performs variable selection for data from genome-wide association studies (GWAS), or other high-dimensional data with continuous, binary or survival outcomes, combining in an iterative framework, the computational efficiency of the structured screen-and-select variable selection strategy based on some association learning and the parsimonious uncertainty quantification provided by the use of non-local priors (see the References).

Details

Package:	GWASinlps
Type:	Package
Version:	2.3
Date:	2024-10-19
License:	GPL (>= 2)

The main function:
`GWASinlps`

The main function calls the following functions:
`nlpsLM`
`nlpsGLM`
`nlpsAFTM`

Author(s)

Nilotpal Sanyal <nilotpal.sanyal@gmail.com>

Maintainer: Nilotpal Sanyal <nilotpal.sanyal@gmail.com>

References

Sanyal et al. (2019), "GWASinlps: Non-local prior based iterative SNP selection tool for genome-wide association studies". Bioinformatics, 35(1), 1-11.

Sanyal, N. (2022). "Iterative variable selection for high-dimensional data with binary outcomes". arXiv preprint arXiv:2211.03190.

GWASinlps *Non-local prior based iterative variable selection for GWAS data, or other high-dimensional data*

Description

GWASinlps performs variable selection with data from Genome-wide association studies (GWAS), or other high-dimensional data with continuous, binary or survival outcomes, combining in an iterative framework, the computational efficiency of the structured screen-and-select variable selection strategy based on some association learning and the parsimonious uncertainty quantification provided by the use of non-local priors (see the References).

Usage

```
GWASinlps(
  y,
  event,
  x,
  family = c("normal","binomial","survival"),
  method = c("rigorous","quick"),
  cor_xy = NULL,
  mmle_xy = NULL,
  mu_xy = NULL,
  prior = c("mom", "imom", "emom", "zellner", "horseshoe"),
  tau,
  priorDelta = modelbbprior(1,1),
  k0,
  m,
  rxx,
  nskip = 3,
  niter = 2000,
  verbose = FALSE,
  seed = NULL,
  tau.hs.method = "halfCauchy",
  sigma.hs.method = "Jeffreys"
)
```

Arguments

y	The vector of continuous response (phenotype) for linear models (LM), or binary response (phenotype) for generalized linear models (GLM), or survival times for accelerated failure time models (AFTM). Binary response values must be 0 or 1.
event	Only for AFTM. The vector of status indicator for the survival data.
x	The design matrix with subjects in rows and independent variables (e.g., SNPs) in columns. Missing values are not accepted currently.

family	"normal" for continuous data, "binomial" for binary data (logit link is used), "survival" for survival data.
method	Applies only when family="binomial". The rigorous method uses logistic regression based analysis which is theoretically appropriate but can be slow. The quick method uses a curious combination of linear model and logistic regression based analyses and is considerably faster. See Details.
cor_xy	Used only when family="normal". Vector of (Pearson) correlation coefficients of y with the individual columns of x.
mmle_xy	Used only when family="binomial". Vector of maximum marginal likelihood estimates of the regression parameters corresponding to the x variables (e.g., SNPs). These are obtained from GLM fits of y with the individual columns of x including an intercept.
mu_xy	Used only when family="survival". Vector of marginal utility estimates of the variables (SNPs) in x. These may be obtained by fitting AFT model to y with individual columns of x using the survreg function of the package survival.
prior	"mom" for pMOM prior, "imom" for piMOM prior, "emom" for peMOM prior, "zellner" for Zellner's g-prior, "horseshoe" for horseshoe prior. For GLM, "zellner" considers group Zellner prior and "emom" and "horseshoe" are not available. For AFTM, "horseshoe" is not available.
tau	The value of the scale parameter tau of the non-local prior.
priorDelta	Prior for model space. Defaults to modelbbprior(1,1), which is beta-binomial(1,1) prior.
k0	GWASinlps tuning parameter denoting the number of leading SNPs/variables (see Details).
m	GWASinlps tuning parameter, denoting the maximum number of SNPs/variables to be selected.
rxx	GWASinlps tuning parameter denoting the correlation threshold to determine leading sets (see References).
nskip	GWASinlps tuning parameter denoting the maximum allowed count of skipping an iteration that does not select any variable (SNP) (see References).
niter	Number of MCMC iterations for non-local prior based Bayesian variable selection. Defaults to 2000.
verbose	If TRUE, prints result from the iterations progressively. FALSE by default.
seed	For reproducibility. If provided, the random seed is set to this value at the beginning of the function.
tau.hs.method	Necessary only when prior="horseshoe". See horseshoe function reference.
sigma.hs.method	Necessary only when prior="horseshoe". See horseshoe function reference.

Details

The GWASinlps method selects variables (SNPs) iteratively.

For continuous response:

For continuous response (phenotype), the procedure starts with an initial set of independent variables (SNPs), a design matrix (SNP genotype matrix) `x` and a response (phenotype) vector `y`.

- An iteration proceeds by determining the `k0` *leading SNPs/variables* having the highest association with `y`. The measure of association is the absolute value of the Pearson's correlation coefficient `cor_xy`. These `k0` leading SNPs/variables, in turn, determine `k0` *leading sets*, where each leading set consists of all SNPs/variables with absolute correlation coefficient more than or equal to `rxx` with the correspondng leading SNP/variable.

- Within each leading set, non-local prior based Bayesian variable selection for linear models is performed (using package **mombf**). The variables (SNPs) appearing in the highest posterior probability model (HPPM) are considered selected in the current iteration. Note that a single variable (SNP) can be selected from multiple leading sets. The selected variables (SNPs) are regressed out from `y` using `lm` fit. The variables (SNPs) that are included in one or more of the *leading sets* but do not appear in any HPPM are dropped from further analysis.

- With updated `y` and variable (SNP) set, next iteration proceeds similarly. And so on like this. The procedure continues until the stopping point, which is determined by the GWASinlps tuning parameters `m`, `rxx`, and `nskip`, is reached. For more details, see References.

For binary response:

For binary response (phenotype), the procedure starts with an initial set of variables (SNPs), a design matrix (SNP genotype matrix) `x` and a binary response (phenotype) vector `y`. If `method="rigorous"`,

- The first iteration proceeds by determining the `k0` *leading SNPs/variables* having the highest association with `y`. The measure of association is the absolute value of the maximum marginal likelihood estimate `mmle_xy`. These `k0` leading SNPs/variables, in turn, determine `k0` *leading sets*, where each leading set consists of all SNPs/variables with absolute correlation coefficient more than or equal to `rxx` with the correspondng leading SNP.

- Within each leading set, non-local prior based Bayesian variable selection for logistic regression model is performed (using package **mombf**). The variables (SNPs) appearing in the HPPM are considered selected in the first iteration. Note that a single variable (SNP) can be selected from multiple leading sets. The variables (SNPs) which are included in one or more *leading sets* but do not appear in any HPPM are dropped from further analysis. After this, the selected variables (SNPs) are accounted for by including them in `glm` fits of `y` with each of the remaining variables (SNPs). The `glm` coefficients of the remaining variables, thus obtained, reflect their contribution in presence of the selected variables (SNPs) of the first iteration.

- Considering the absolute values of these `glm` coefficients as the measure of association, we proceed with the second iteration with updated variable (SNP) set. And so on in this manner. The procedure continues until the stopping point, which is determined by the GWASinlps tuning parameters `m`, `rxx`, and `nskip`, is reached.

If `method="quick"`, the procedure is similar to above except at the following points. In this method, non-local prior based Bayesian variable selection using logistic regression model is performed until one or more variables (SNP) are selected in an iteration. Until a variable is selected, there is no need to account for anything, so the initial maximum marginal likelihood estimates continue to be used. After the first selections (if any) are made, a `glm` fit of `y` on the selected variables is performed and the deviance residuals are computed. In the subsequent iterations, considering these (continuous) deviance residuals as response, non-local prior based Bayesian variable selection for linear models is performed till the stopping point is reached.

For survival data:

For survival data, the procedure starts with an initial set of variables (SNPs), a design matrix (SNP genotype matrix) `x` and a binary response (phenotype) vector `y`.

- The first iteration proceeds by determining the `k0` *leading SNPs/variables* having the highest association with `y`. The measure of association is the absolute value of the marginal utility `mu_xy`. These `k0` leading SNPs/variables, in turn, determine `k0` *leading sets*, where each leading set consists of all SNPs with absolute correlation coefficient more than or equal to `rxx` with the correspondng leading SNP.

- Within each leading set, non-local prior based Bayesian variable selection for accelerated failure time model is performed (using package **mombf**). The variables (SNPs) appearing in the HPPM are considered selected in the first iteration. Note that a single variable (SNP) can be selected from multiple leading sets. The variables (SNPs) which are included in one or more *leading sets* but do not appear in any HPPM are dropped from further analysis. After this, to account for the selected variables (SNPs), conditional utilities of each of the remaining variables (SNPs) are computed in the presence of the selected variables (SNPs) in the model. These conditional utilities reflect the contribution of the remaining variables (SNPs) in presence of the selected variables (SNPs) of the first iteration.

- Considering the absolute values of these conditional utilities as the measure of association, we proceed with the second iteration with updated variable (SNP) set. And so on in this manner. The procedure continues until the stopping point, which is determined by the GWASinlps tuning parameters `m`, `rxx`, and `nskip`, is reached.

For horseshoe prior, package **horseshoe** is used.

Value

A list containing

`selected` Vector with names of the GWASinlps selected variables (SNPs) in the order they were selected.

`selected_iterwise` List with selected variables (SNPs) from each iteration.

Author(s)

Nilotpal Sanyal <nilotpal.sanyal@gmail.com>

References

Sanyal et al. (2019), "GWASinlps: Non-local prior based iterative SNP selection tool for genome-wide association studies". Bioinformatics, 35(1), 1-11.

Sanyal, N. (2022). "Iterative variable selection for high-dimensional data with binary outcomes". arXiv preprint arXiv:2211.03190.

See Also

`nlpsLM`, `nlpsGLM`, `nlpsAFTM`, `modelSelection`, `horseshoe`

Examples

```
n = 200
p = 1000
m = 10

# Generate design matrix (genotype matrix)
set.seed(1)
f = runif( p, .1, .2 ) # simulate minor allele frequency
x = matrix( nrow = n, ncol = p )
colnames(x) = 1:p
for(j in 1:p)
  x[,j] = rbinom( n, 2, f[j] )

# Generate true effect sizes
causal_snps = sample( 1:p, m )
beta = rep( 0, p )
set.seed(1)
beta[causal_snps] = rnorm(m, mean = 0, sd = 2 )

# Generate continuous (phenotype) data
y = x %*% beta + rnorm(n, 0, 1)

# Fix scale parameter tau
tau = 0.2

# GWASinlps analysis
inlps = GWASinlps(y=y, x=x, family="normal", prior="mom", tau=tau, k0=1,
        m=50, rxx=0.2)
cat( "GWASinlps selected", length(inlps$selected), "SNPs with",
    length(intersect(inlps$selected, causal_snps)), "true positive(s) and",
    length(setdiff(causal_snps, inlps$selected)), "false negative(s) out
    of a pool of", p, "SNPs with data from", n, "persons." )

# Compare with LASSO
library(glmnet)
fit.cvlasso = cv.glmnet( x, y, alpha = 1 )
l.min = fit.cvlasso $lambda.min # lambda that gives minimum cvm
l.1se = fit.cvlasso $lambda.1se  # largest lambda such that error is
                                 # within 1 se of the minimum

lasso_min = which( as.vector( coef( fit.cvlasso, s = l.min ) )[-1] != 0 )
cat( "LASSO with lambda.min selected", length(lasso_min), "SNPs with",
     length(intersect(lasso_min, causal_snps)), "true positives and",
    length(setdiff(causal_snps, inlps$selected)), "false negative(s)." )

lasso_1se = which( as.vector( coef( fit.cvlasso, s = l.1se ) )[-1] != 0 )
cat( "LASSO with lambda.1se selected", length(lasso_1se), "SNPs with",
     length(intersect(lasso_1se, causal_snps)), "true positives and",
    length(setdiff(causal_snps, inlps$selected)), "false negative(s)." )
```

nlps *Non-local prior based single-step variable selection for high-dimensional data*

Description

nlpsLM, nlpsGLM, nlpsAFTM perform variable selection in a single iteration respectively for continuous, binary and survival outcomes, combining the computational efficiency of the 'structured screen-and-select' variable selection strategy based on some association learning and the parsimonious uncertainty quantification provided by the use of non-local priors (see the References).

Usage

```
nlpsLM(y, x, cor_xy, prior = c("mom", "imom", "emom", "zellner",
  "horseshoe"), tau, priorDelta = modelbbprior(1,1),
  k0, rxx, niter = 2000, verbose = F,
  tau.hs.method = "halfCauchy", sigma.hs.method = "Jeffreys" )

nlpsGLM(y, x, mmle_xy, prior = c("mom", "imom", "zellner"),
  tau, priorDelta = modelbbprior(1,1),
  k0, rxx, niter = 2000, verbose = F )

nlpsAFTM(y, event, x, mu_xy, prior = c("mom", "imom", "emom",
  "zellner"), tau, priorDelta = modelbbprior(1,1),
  k0, rxx, niter = 2000, verbose = F )
```

Arguments

y	The vector of continuous response (phenotype) for linear models (LM), or binary response (phenotype) for generalized linear models (GLM), or survival times for accelerated failure time models (AFTM). Binary response values must be 0 or 1.
event	Only for AFTM. The vector of status indicator for the survival data.
x	The design matrix with subjects in rows and independent variables (SNPs) in columns. Missing values are not accepted currently.
cor_xy	Only for LM. Vector of (Pearson) correlations of y with the columns of x.
mmle_xy	Only for GLM. Vector of maximum marginal likelihood estimates of the regression parameters for the variables (SNPs) in x. These may be obtained from individual GLM fits of y with the columns of x.
mu_xy	Only for AFTM. Vector of marginal utility estimates of the variables (SNPs) in x. These may be obtained by fitting AFT model to y with individual columns of x using the survreg function of the package survival.
prior	"mom" for pMOM prior, "imom" for piMOM prior, "emom" for peMOM prior, "zellner" for Zellner's g-prior, "horseshoe" for horseshoe prior. For GLM, "zellner" considers group Zellner prior and "emom" and "horseshoe" are not available. For AFTM, "horseshoe" is not available.

tau	the value of the scale parameter tau of the non-local prior.
priorDelta	Prior for model space. Defaults to modelbbprior(1,1), which is beta-binomial(1,1) prior.
k0	GWASinlps tuning parameter denoting the number of leading SNPs (see Details).
rxx	GWASinlps tuning parameter denoting the correlation threshold to determine leading sets (see References).
niter	Number of MCMC iterations for non-local prior based Bayesian variable selection. Defaults to 2000.
verbose	If TRUE, prints result from the iterations progressively. FALSE by default.
tau.hs.method	Necessary only when prior="horseshoe". See horseshoe function reference.
sigma.hs.method	Necessary only when prior="horseshoe". See horseshoe function reference.

Details

The nlpsLM, nlpsGLM and nlpsAFTM functions perform SNP selection in one iteration for continuous data, binary data, and survival data, respectively. The GWASinlps function repeatedly calls these functions. For details of the procedure, see the reference for the GWASinlps function.

Value

A list with elements

hppm	The names of variables (SNPs) appearing in the highest posterior probability model (HPPM) of at least one leading set.
not.selected	The names of variables (SNPs) appearing in at least one leading set but in none of the HPPMs.

Author(s)

Nilotpal Sanyal <nilotpal.sanyal@gmail.com>

References

Sanyal et al. (2019), "GWASinlps: Non-local prior based iterative SNP selection tool for genome-wide association studies". Bioinformatics, 35(1), 1-11.

Sanyal, N. (2022). "Iterative variable selection for high-dimensional data with binary outcomes". arXiv preprint arXiv:2211.03190.

See Also

GWASinlps, modelSelection, horseshoe

Examples

```
n = 100
p = 1000
m = 10

# Generate design matrix (genotype matrix)
set.seed(1)
f = runif( p, .1, .2 ) # simulate minor allele frequency
x = matrix( nrow = n, ncol = p )
colnames(x) = 1:p
for(j in 1:p)
  x[,j] = rbinom( n, 2, f[j] )

# Generate true effect sizes
causal_snps = sample( 1:p, m )
beta = rep( 0, p )
set.seed(1)
beta[causal_snps] = rnorm(m, mean = 0, sd = 2 )

# Generate continuous (phenotype) data
y.cont = x %*% beta + rnorm(n, 0, 1)

# Generate binary (phenotype) data
prob = exp(x %*% beta)/(1 + exp(x %*% beta))
y.bin = sapply(1:n, function(i)rbinom(1,1,prob[i]) )

# Fix scale parameter tau
tau = 0.022

# GWASinlps analysis
cor_xy = c(cor(x,y.cont))
names(cor_xy) = colnames(x)
nlps_cont = nlpsLM(y.cont, x, cor_xy=cor_xy, prior="mom",
  tau=tau, k0=2, rxx=0.3, niter=10000, verbose=TRUE)
nlps_cont

library(fastglm)
mode(x) = "double"  #needed for fastglm() function below
mmle_xy = apply( x, 2, function(z) coef( fastglm(y=y.bin,
x=cbind(1,matrix(z)), family = binomial(link = "logit")) )[2] )
nlps_bin = nlpsGLM(y.bin, x, mmle_xy=mmle_xy, prior="mom",
  tau=tau, k0=2, rxx=0.3, niter=10000, verbose=TRUE)
nlps_bin
```

Index

GWASinlps, *2*, 3, *9*
GWASinlps-package, 2

horseshoe, *6*, *9*

modelSelection, *6*, *9*

nlps, 8
nlpsAFTM, *2*, *6*
nlpsAFTM *(*nlps*)*, 8
nlpsGLM, *2*, *6*
nlpsGLM *(*nlps*)*, 8
nlpsLM, *2*, *6*
nlpsLM *(*nlps*)*, 8