Package 'marble'

May 10, 2023

Type Package

Title Robust Marginal Bayesian Variable Selection for Gene-Environment Interactions

Version 0.0.2

Date 2023-05-09

Description Recently, multiple marginal variable selection methods have been developed and shown to be effective in Gene-Environment interactions studies. We propose a novel marginal Bayesian variable selection method for Gene-Environment interactions studies. In particular, our marginal Bayesian method is robust to data contamination and outliers in the outcome variables. With the incorporation of spike-and-slab priors, we have implemented the Gibbs sampler based on Markov Chain Monte Carlo. The core algorithms of the package have been developed in 'C++'.

Depends R (>= 3.5.0)

License GPL-2

Encoding UTF-8

URL https://github.com/xilustat/marble

LazyData true

LinkingTo Rcpp, RcppArmadillo

Imports Rcpp, stats

RoxygenNote 7.2.3

NeedsCompilation yes

Repository CRAN

Author Xi Lu [aut, cre],
Cen Wu [aut]

Maintainer Xi Lu <xilu@ksu.edu>

Date/Publication 2023-05-10 19:30:02 UTC

R topics documented:

`marble-package` *Robust Marginal Bayesian Variable Selection for Gene-Environment Interactions*

Description

In this package, we provide a set of robust marginal Bayesian variable selection methods for gene-environment interaction analysis. A Bayesian formulation of the quantile regression has been adopted to accommodate data contamination and heavy-tailed distributions in the response. The proposed method conducts a robust marginal variable selection by accounting for structural sparsity. In particular, the spike-and-slab priors are imposed to identify important main and interaction effects. In addition to the default method, users can also choose different structures (robust or non-robust), methods without spike-and-slab priors.

Details

The user friendly, integrated interface **marble()** allows users to flexibly choose the fitting methods they prefer. There are two arguments in marble() that control the fitting method: robust: whether to use robust methods; sparse: whether to use the spike-and-slab priors to create sparsity. The function marble() returns a marble object that contains the posterior estimates of each coefficients. Moreover, it also provides a rank list of the genetic factors and gene-environment interactions. Functions GxESelection() and print.marble() are implemented for marble objects. GxESelection() takes a marble object and returns the variable selection results.

References

Lu, X., Fan, K., Ren, J., and Wu, C. (2021). Identifying Gene–Environment Interactions With Robust Marginal Bayesian Variable Selection. *Frontiers in Genetics*, 12:667074 doi:10.3389/fgene.2021.667074

Ren, J., Zhou, F., Li, X., Ma, S., Jiang, Y. and Wu, C. (2020). Robust Bayesian variable selection for gene-environment interactions. doi:10.1111/biom.13670

Zhou, F., Ren, J., Lu, X., Ma, S. and Wu, C. (2020). Gene–Environment Interaction: a Variable Selection Perspective. Epistasis. Methods in Molecular Biology. *Humana Press* (Accepted) https://arxiv.org/abs/2003.02930

Wu, C., Cui, Y., and Ma, S. (2014). Integrative analysis of gene–environment interactions under a multi–response partially linear varying coefficient model. *Statistics in Medicine*, 33(28), 4988–4998 doi:10.1002/sim.6287

Shi, X., Liu, J., Huang, J., Zhou, Y., Xie, Y. and Ma, S. (2014). A penalized robust method for identifying gene–environment interactions. *Genetic epidemiology*, 38(3), 220-230 doi:10.1002/gepi.21795

Chai, H., Zhang, Q., Jiang, Y., Wang, G., Zhang, S., Ahmed, S. E. and Ma, S. (2017). Identifying gene-environment interactions for prognosis using a robust approach. *Econometrics and statistics*, 4, 105-120 doi:10.1016/j.ecosta.2016.10.004

See Also

`marble`

`dat` *simulated data for demonstrating the features of marble.*

Description

Simulated gene expression data for demonstrating the features of marble.

Usage

```
data("dat")
```

Format

dat consists of four components: X, Y, E, clin.

Details

The data model for generating Y

Use subscript i to denote the ith subject. Let $(Y_i, X_i, E_i, clin_i)$ $(i = 1, \ldots, n)$ be independent and identically distributed random vectors. Y_i is a continuous response variable representing the phenotype. X_i is the p–dimensional vector of genetic factors. The environmental factors and clinical factors are denoted as the q-dimensional vector E_i and the m-dimensional vector $clin_i$, respectively. The ϵ follows some heavy-tailed distribution. For X_{ij} $(j = 1, \ldots, p)$, the measurement of the jth genetic factor on the jth subject, considering the following model:

$$Y_i = \alpha_0 + \sum_{k=1}^{q} \alpha_k E_{ik} + \sum_{t=1}^{m} \gamma_t clin_{it} + \beta_j X_{ij} + \sum_{k=1}^{q} \eta_{jk} X_{ij} E_{ik} + \epsilon_i,$$

where α_0 is the intercept, α_k's and γ_t's are the regression coefficients corresponding to effects of environmental and clinical factors, respectively. The β_j's and η_{jk}'s are the regression coefficients of the genetic variants and G×E interactions effects, correspondingly. The G×E interactions effects are defined with $W_j = (X_j E_1, \ldots, X_j E_q)$. With a slight abuse of notation, denote $\tilde{W} = W_j$. Denote $\alpha = (\alpha_1, \ldots, \alpha_q)^T$, $\gamma = (\gamma_1, \ldots, \gamma_m)^T$, $\beta = (\beta_1, \ldots, \beta_p)^T$, $\eta = (\eta_1^T, \ldots, \eta_p^T)^T$, $\tilde{W} = (\tilde{W}_1, \ldots, \tilde{W}_p)$. Then model can be written as

$$Y_i = E_i \alpha + clin_i \gamma + X_{ij} \beta_j + \tilde{W}_i \eta_j + \epsilon_i.$$

See Also

marble

Examples

```
data(dat)
dim(X)
```

GxESelection *Variable selection for a marble object*

Description

Variable selection for a marble object

Usage

```
GxESelection(obj, sparse)
```

Arguments

obj	marble object.
sparse	logical flag. If TRUE, spike-and-slab priors will be used to shrink coefficients of irrelevant covariates to zero exactly.

Details

For class 'Sparse', the inclusion probability is used to indicate the importance of predictors. Here we use a binary indicator ϕ to denote the membership of the non-spike distribution. Take the main effect of the jth genetic factor, X_j, as an example. Suppose we have collected H posterior samples from MCMC after burn-ins. The jth G factor is included in the marginal G×E model at the jth MCMC iteration if the corresponding indicator is 1, i.e., $\phi_j^{(h)} = 1$. Subsequently, the posterior probability of retaining the jth genetic main effect in the final marginal model is defined as the average of all the indicators for the jth G factor among the H posterior samples. That is, $p_j = \hat{\pi}(\phi_j = 1|y) = \frac{1}{H}\sum_{h=1}^{H}\phi_j^{(h)},\ j = 1, \ldots, p$. A larger posterior inclusion probability of jth indicates a stronger empirical evidence that the jth genetic main effect has a non-zero coefficient, i.e., a stronger association with the phenotypic trait. Here, we use 0.5 as a cutting-off point. If $p_j > 0.5$, then the jth genetic main effect is included in the final model. Otherwise, the jth genetic main effect is excluded in the final model. For class 'NonSparse', variable selection is based on 95% credible interval. Please check the references for more details about the variable selection.

Value

an object of class 'GxESelection' is returned, which is a list with components:

method	method used for identifying important effects.
effects	a list of indicators of selected effects.

References

Lu, X., Fan, K., Ren, J., and Wu, C. (2021). Identifying Gene–Environment Interactions With Robust Marginal Bayesian Variable Selection. *Frontiers in Genetics*, 12:667074 doi:10.3389/fgene.2021.667074

See Also

marble

Examples

```
data(dat)
max.steps=5000
## sparse
fit=marble(X, Y, E, clin, max.steps=max.steps)
selected=GxESelection(fit,sparse=TRUE)
selected

## non-sparse
fit=marble(X, Y, E, clin, max.steps=max.steps, sparse=FALSE)
selected=GxESelection(fit,sparse=FALSE)
selected
```

`marble` *fit a robust Bayesian variable selection model for G×E interactions.*

Description

fit a robust Bayesian variable selection model for G×E interactions.

Usage

```
marble(
  X,
  Y,
  E,
  clin,
  max.steps = 10000,
  robust = TRUE,
  sparse = TRUE,
  debugging = FALSE
)
```

Arguments

X	the matrix of predictors (genetic factors). Each row should be an observation vector.
Y	the continuous response variable.
E	a matrix of environmental factors. E will be centered. The interaction terms between X (genetic factors) and E will be automatically created and included in the model.
clin	a matrix of clinical variables. Clinical variables are not subject to penalize. Clinical variables will be centered and a column of 1 will be added to the Clinical matrix as the intercept.
max.steps	the number of MCMC iterations.
robust	logical flag. If TRUE, robust methods will be used.
sparse	logical flag. If TRUE, spike-and-slab priors will be used to shrink coefficients of irrelevant covariates to zero exactly.
debugging	logical flag. If TRUE, progress will be output to the console and extra information will be returned.

Details

Consider the data model described in "dat":

$$Y_i = \alpha_0 + \sum_{k=1}^{q} \alpha_k E_{ik} + \sum_{t=1}^{m} \gamma_t clin_{it} + \beta_j X_{ij} + \sum_{k=1}^{q} \eta_{jk} X_{ij} E_{ik} + \epsilon_i,$$

Where α_0 is the intercept, α_k's and γ_t's are the regression coefficients corresponding to effects of environmental and clinical factors. And β_j's and η_{jk}'s are the regression coefficients of the genetic variants and G×E interactions effects, correspondingly.

When sparse=TRUE (default), spike–and–slab priors are imposed to identify important main and interaction effects. If sparse=FALSE, Laplacian shrinkage will be used.

When robust=TRUE (default), the distribution of ϵ_i is defined as a Laplace distribution with density $f(\epsilon_i|\nu) = \frac{\nu}{2}\exp\{-\nu|\epsilon_i|\}$, $(i = 1, \ldots, n)$, which leads to a Bayesian formulation of LAD regression. If robust=FALSE, ϵ_i follows a normal distribution.

Here, a rank list of the main and interaction effects is provided. For method incorporating spike-and-slab priors, the inclusion probability is used to indicate the importance of predictors. We use a binary indicator ϕ to denote the membership of the non-spike distribution. Take the main effect of the jth genetic factor, X_j, as an example. Suppose we have collected H posterior samples from MCMC after burn-ins. The jth G factor is included in the marginal G×E model at the jth MCMC iteration if the corresponding indicator is 1, i.e., $\phi_j^{(h)} = 1$. Subsequently, the posterior probability of retaining the jth genetic main effect in the final marginal model is defined as the average of all the indicators for the jth G factor among the H posterior samples. That is, $p_j = \hat{\pi}(\phi_j = 1|y) = \frac{1}{H}\sum_{h=1}^{H}\phi_j^{(h)}$, $j = 1, \ldots, p$. A larger posterior inclusion probability jth indicates a stronger empirical evidence that the jth genetic main effect has a non-zero coefficient, i.e., a stronger association with the phenotypic trait. For method without spike-and-slab priors, variable selection is based on different level of credible intervals.

Both X, $clin$ and E will be standardized before the generation of interaction terms to avoid the multicollinearity between main effects and interaction terms.

Please check the references for more details about the prior distributions.

Value

an object of class 'marble' is returned, which is a list with component:

`posterior`	the posterior samples of coefficients from the MCMC.
`coefficient`	the estimated value of coefficients.
`ranklist`	the rank list of main and interaction effects.
`burn.in`	the total number of burn-ins.
`iterations`	the total number of iterations.
`design`	the design matrix of all effects.

References

Lu, X., Fan, K., Ren, J., and Wu, C. (2021). Identifying Gene–Environment Interactions With Robust Marginal Bayesian Variable Selection. *Frontiers in Genetics*, 12:667074 doi:10.3389/fgene.2021.667074

See Also

GxESelection

Examples

```
data(dat)

## default method
max.steps=5000
fit=marble(X, Y, E, clin, max.steps=max.steps)

## coefficients of parameters
fit$coefficient

## Estimated values of main G effects
fit$coefficient$G

## Estimated values of interactions effects
fit$coefficient$GE

## Rank list of main G effects and interactions
fit$ranklist


## alternative: robust selection
fit=marble(X, Y, E, clin, max.steps=max.steps, robust=TRUE, sparse=FALSE)
fit$coefficient
fit$ranklist
```

```
## alternative: non-robust sparse selection
fit=marble(X, Y, E, clin, max.steps=max.steps, robust=FALSE, sparse=FALSE)
fit$coefficient
fit$ranklist
```

`print.GxESelection` *print a GxESelection object*

Description

Print a summary of a GxESelection object

Usage

```
## S3 method for class 'GxESelection'
print(x, digits = max(3, getOption("digits") - 3), ...)
```

Arguments

`x`	GxESelection object.
`digits`	significant digits in printout.
`...`	other print arguments.

Value

No return value, called for side effects.

See Also

`GxESelection`

`print.marble` *print a marble object*

Description

Print a summary of a marble object

Usage

```
## S3 method for class 'marble'
print(x, digits = max(3, getOption("digits") - 3), ...)
```

Arguments

x	marble object.
digits	significant digits in printout.
...	other print arguments.

Value

No return value, called for side effects.

See Also

marble

Index