Package 'rolong'

April 7, 2023

Type Package

Title Bayesian Quantile Variable Selection with Mixed Effects Model in Longitudinal Study

Version 1.0

Date 2023-04-03

Description In longitudinal studies, the same subjects are measured repeatedly over time, leading to correlations among the repeated measurements. Properly accounting for the inner cluster correlations in the presence of data heterogeneity and long tailed distributions of the disease phenotype is challenging, especially in the context of high dimensional regressions. Here, we aim at developing novel Bayesian regularized quantile mixed effect models to tackle these challenges.In this package, we have implemented the Gibbs samplers of Bayesian quantile variable selection method in the mixed effects model with the spike-and-slab priors. The Markov Chain Monte Carlo algorithms of the proposed and alternative models can be efficiently performed by using the package.

Depends R (>= 4.2.0)

License GPL-2

Encoding UTF-8

URL https://github.com/kunfa/rolong

Imports Rcpp

LinkingTo Rcpp, RcppArmadillo

RoxygenNote 7.2.3

R topics documented:

`rolong-package` *Bayesian Quantile Variable Selection with Mixed Effects Model in Longitudinal Study*

Description

In longitudinal studies, the same subjects are measured repeatedly over time, leading to correlations among the repeated measurements. Properly accounting for the inner cluster correlations in the presence of data heterogeneity and long tailed distributions of the disease phenotype is challenging, especially in the context of high dimensional regressions. Here, we aim at developing novel Bayesian regularized quantile mixed effect models to tackle these challenges. In this package, we have implemented the Gibbs samplers of Bayesian quantile variable selection method in the mixed effects model with the spike-and-slab priors. The Markov Chain Monte Carlo algorithms of the proposed and alternative models can be efficiently performed by using the package.

Details

The user friendly, integrated interface **rolong()** allows users to flexibly choose the fitting methods by specifying the following parameter:

slope:	whether to use random intercept and slope model.
robust:	whether to use robust methods for modelling.
quant:	to specify different quantiles when using robust methods.
sparse:	whether to use the spike-and-slab priors to impose sparsity.

The function rolong() returns a rolong object that contains the posterior estimates of each coefficients.

References

Fan, K., Wang, W. and Wu, C. (2022). Bayesian Quantile Variable Selection with Mixed Effects Model in the Longitudinal Study. (to be submitted)

Zhou, F., Ren, J., Li, G., Jiang, Y., Li, X., Wang, W. and Wu, C. (2019). Penalized Variable Selection for Lipid-Environment Interactions in a Longitudinal Lipidomics Study. *Genes*, 10(12), 1002 doi:10.3390/genes10121002

Ren, J., Zhou, F., Li, X., Ma, S., Jiang, Y. and Wu, C. (2022). Robust Bayesian variable selection for gene-environment interactions. *Biometrics*, (in press) doi:10.1111/biom.13670

Wu, C., and Ma, S. (2015). A selective review of robust variable selection with applications in bioinformatics. *Briefings in Bioinformatics*, 16(5), 873–883 doi:10.1093/bib/bbu046

Zhou, F., Ren, J., Lu, X., Ma, S. and Wu, C. (2021). Gene–Environment Interaction: a Variable Selection Perspective. *Epistasis. Methods in Molecular Biology.* 2212:191–223 https://link.springer.com/protocol/10.1007/978-1-0716-0947-7_13

Ren, J., Zhou, F., Li, X., Chen, Q., Zhang, H., Ma, S., Jiang, Y. and Wu, C. (2020) Semi-parametric Bayesian variable selection for gene-environment interactions. *Statistics in Medicine*, 39: 617– 638 doi:10.1002/sim.8434

Ren, J., Zhou, F., Li, X., Wu, C. and Jiang, Y. (2019) spinBayes: Semi-Parametric Gene-Environment Interaction via Bayesian Variable Selection. R package version 0.1.0. https://CRAN.R-project.org/package=spinBayes

Wu, C., Jiang, Y., Ren, J., Cui, Y. and Ma, S. (2018). Dissecting gene-environment interactions: A penalized robust approach accounting for hierarchical structures. *Statistics in Medicine*, 37:437–456 doi:10.1002/sim.7518

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

Wu, C., Zhong, P.S. and Cui, Y. (2013). High dimensional variable selection for gene-environment interactions. *Technical Report. Michigan State University.*

See Also

rolong

`data` *simulated data for demonstrating the features of rolong*

Description

Simulated gene expression data for demonstrating the features of rolong.

Format

The data object consists of six components: y, e, g, w ,k and coeff. coeff contains the true values of parameters used for generating Y.

Details

The data and model setting

Consider a longitudinal study on n subjects with k repeated measurement for each subject. Let Y_{ij} be the measurement for the ith subject at each time point $j(1 \leq i \leq n, 1 \leq j \leq k)$ .We use a m-dimensional vector G_{ij} to denote the genetics factors, where$L_{ij} = (L_{ij1}, ..., L_{ijm})^T$. Also, we use p-dimensional vector E_{ij} to denote the treatment factors, where $T_{ij} = (T_{ij1}, ..., T_{ijp})^T$. $X_{ij} = (1, j, j^2)^T$. Z_{ij} is a $h \times 1$ covariate associated with random effects and α_i is a $h \times 1$ vector of random effects. At the beginning, the interaction effects is modeled as the product of genomics features and treatment factor with 4 different levels. After representing the treatment factors as three dummy variables, the identification of the lipid by treatment interaction needs to be performed as group level. Combing the genetics factors, treatment factors and their interactions that associated with the longitudinal phenotype, we have the following mixed-effects model:

$$Y_{ij} = X_{ij}^T\gamma_0 + E_{ij}^T\gamma_1 + G_{ij}^T\gamma_2 + (G_{ij} \bigotimes E_{ij})^T\gamma_3 + Z_{ij}^T\alpha_i + \epsilon_{ij}.$$

where γ_1,γ_2,γ_3 are p,m and mp dimensional vectors that represent the coefficients of the treatment effects, the genetics effects and interactions effects, respectively. Accommodating the Kronecker product of the m - dimensional vector L_{ij} and the p-dimensional vector E_{ij}, the interactions between genetics and treatment factors can be expressed as a mp-dimensional vector, denoted as the following form:

$$G_{ij} \bigotimes E_{ij} = [E_{ij1}E_{ij1}, E_{ij2}E_{ij2}, ..., E_{ij1}E_{ijp}, E_{ij2}E_{ij1}, ..., E_{ijm}E_{ijp}]^T.$$

When $h = 1$, the model becomes a mixed-effects model with random intercept only.

See Also

rolong

Examples

```
data(data)
dim(y)
dim(g)
dim(e)
dim(w)
print(k)
print(coeff)
```

---

rolong *fit a Bayesian quantile variable selection with mixed effects model in longitudinal study*

Description

fit a Bayesian quantile variable selection with mixed effects model in longitudinal study

Usage

```
rolong(
  y,
  e,
  g,
  w,
  k,
  iterations = 10000,
  burn.in = NULL,
  slope = TRUE,
  robust = TRUE,
  quant = 0.5,
  sparse = TRUE,
  structure = c("group", "individual")
)
```

Arguments

y	the matrix of response variable. The current version of rolong only supports continuous response.
e	the matrix of a group of dummy environmental factors variables.
g	the matrix of predictors (genetic factors) without intercept. Each row should be an observation vector.
w	the matrix of interactions between genetic factors and environmental factors.
k	the total number of time points.
iterations	the number of MCMC iterations.
burn.in	the number of iterations for burn-in.

slope	logical flag. If TRUE, random intercept and slope model will be used.
robust	logical flag. If TRUE, robust methods will be used.
quant	specify different quantiles when applying robust methods.
sparse	logical flag. If TRUE, spike-and-slab priors will be used to shrink coefficients of irrelevant covariates to zero exactly.
structure	structure for interaction effects, two choices are available. "group" for selection on group-level only. "individual" for selection on individual-level only.

Details

Consider the data model described in "data":

$$Y_{ij} = X_{ij}^T\gamma_0 + E_{ij}^T\gamma_1 + \sum_{l=1}^{p} G_{ijl}\gamma_{2l} + \sum_{l=1}^{p} W_{ijl}^T\gamma_{3l} + Z_{ij}^T\alpha_i + \epsilon_{ij}.$$

where γ_{2l} is the main effect of the lth genetic variant. The interaction effects is corresponding to the coefficient vector $\gamma_{3l} = (\gamma_{3l1}, \gamma_{3l2}, \ldots, \gamma_{3lm})^\top$.

When structure="group", group-level selection will be conducted on $||\gamma_{3l}||_2$. If structure="individual", individual-level selection will be conducted on each γ_{3lq}, $(q = 1, \ldots, m)$.

When slope=TRUE (default), random intercept and slope model will be used as the mixed effects model.

When sparse=TRUE (default), spike–and–slab priors are imposed on individual and/or group levels to identify important main and interaction effects. Otherwise, Laplacian shrinkage will be used.

When robust=TRUE (default), the distribution of ϵ_i is defined as a Laplace distribution with density. $f(\epsilon_{ij}|\theta,\tau) = \theta(1-\theta)\exp\{-\tau\rho_\theta(\epsilon_{ij})\}$, $(i = 1, \ldots, n, j = 1, \ldots, k)$, which leads to a Bayesian formulation of quantile regression. If robust=FALSE, ϵ_{ij} follows a normal distribution.

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

See Also

data

Examples

```
data(data)

## default method
fit=rolong(y,e,g,w,k,structure=c("group"))
fit$coefficient

## Compute TP and FP
b = selection(fit,sparse=TRUE)
index = which(coeff!=0)
pos = which(b != 0)
tp = length(intersect(index, pos))
fp = length(pos) - tp
list(tp=tp, fp=fp)

## alternative: robust individual selection
fit=rolong(y,e,g,w,k,structure=c("individual"))
fit$coefficient
```

```
## alternative: non-robust group selection
fit=rolong(y,e,g,w,k,robust=FALSE, structure=c("group"))
fit$coefficient

## alternative: robust group selection under random intercept model
fit=rolong(y,e,g,w,k,slope=FALSE, structure=c("group"))
fit$coefficient
```

`selection` *Variable selection for a rolong object*

Description

Variable selection for a rolong object

Usage

```
selection(obj, sparse)
```

Arguments

`obj`	rolong 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 median probability model (MPM) (Barbieri and Berger, 2004) is used to identify predictors that are significantly associated with the response variable. 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 'selection' is returned, which is a list with component:

`inde`	a vector of indicators of selected effects.

References

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

Barbieri, M.M. and Berger, J.O. (2004). Optimal predictive model selection. *Ann. Statist*, 32(3):870–897

See Also

`rolong`

Examples

```
data(data)
## sparse
fit=rolong(y,e,g,w,k,structure=c("group"))
selected=selection(fit,sparse=TRUE)
selected


## non-sparse
fit=rolong(y,e,g,w,k,sparse=FALSE,structure=c("group"))
selected=selection(fit,sparse=FALSE)
selected
```

Index