Package ‘Blend’

July 21, 2025

Type Package

Title Robust Bayesian Longitudinal Regularized Semiparametric Mixed Models

Version 0.1.1.1

Date 2025-01-29

Description Our recently developed fully robust Bayesian semiparametric mixed-effect model for high-dimensional longitudinal studies with heterogeneous observations can be implemented through this package. This model can distinguish between time-varying interactions and constant-effect-only cases to avoid model misspecifications. Facilitated by spike-and-slab priors, this model leads to superior performance in estimation, identification and statistical inference. In particular, robust Bayesian inferences in terms of valid Bayesian credible intervals on both parametric and nonparametric effects can be validated on finite samples. The Markov chain Monte Carlo algorithms of the proposed and alternative models are efficiently implemented in 'C++'.

Depends R (>= 4.2.0)

License GPL-2

Encoding UTF-8

URL `https://github.com/kunfa/Blend`

LinkingTo Rcpp, RcppArmadillo

Imports Rcpp, splines, stats, ggplot2

RoxygenNote 7.3.2

NeedsCompilation yes

Author Kun Fan [aut, cre],
Cen Wu [aut]

Maintainer Kun Fan <`kfan@ksu.edu`>

Repository CRAN

Date/Publication 2025-01-29 21:40:06 UTC

Contents

`Blend-package` *Robust Bayesian Longitudinal Regularized Semiparametric Mixed Model*

Description

In this package, we further extend the sparse robust Bayesian mixed models to nonlinear longitudinal interactions. Specifically, the proposed Bayesian semiparametric model is robust not only to outliers and heavy-tailed distributions of the response variable, but also to the misspecification of interaction effect in the forms other than non-linear interactions. We have developed the Gibbs sampler with the spike-and-slab priors to promote sparse identification of appropriate forms of main and interaction effects. In addition to the default method, users can also choose different selection structures for separation of constant and varying effects or not, methods without spike–and–slab priors and non-robust methods. In total, *Blend* provides 8 different methods (4 robust and 4 non-robust) under the random intercept and slope model. All the methods in this package are developed for the first time. Please read the Details below for how to configure the method used.

Details

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

robust:	whether to use robust methods for modelling.
structural:	whether to incorporate structural identification(separation of constant and varying effects) .
sparse:	whether to use the spike-and-slab priors to impose sparsity.

The function Blend() returns a Blend object that contains the posterior estimates of each coefficients and other useful information for selection(). S3 generic functions selection() and print() are implemented for Blend objects. selection() takes a Blend object and returns the variable selection results.

References

Fan, K., Ren, J., Ma, Shuangge and Wu, C. (2025+). Robust Bayesian Regularized Semiparametric Mixed Models in Longitudinal Studies. (submitted)

Fan, K., Subedi, S., Yang, G., Lu, X., Ren, J., and Wu, C. (2024). Is Seeing Believing? A Practitioner's Perspective on High-Dimensional Statistical Inference in Cancer Genomics Studies. Entropy, 26(9), 794.

Ren, J., Zhou, F., Li, X., Ma, S., Jiang, Y. and Wu, C. (2023). Robust Bayesian variable selection for gene-environment interactions. *Biometrics*, 79(2), 684-694 doi:10.1111/biom.13670

Zhou, F., Ren, J., Ma, S. and Wu, C. (2023). The Bayesian regularized quantile varying coefficient model. Computational Statistics & Data Analysis, 187, 107808.

Zhou, F., Lu, X., Ren, J., Fan, K., Ma, S., & Wu, C. (2022). Sparse group variable selection for gene–environment interactions in the longitudinal study. Genetic epidemiology, 46(5-6), 317-340.

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. (2020). roben: Robust Bayesian Variable Selection for Gene-Environment Interactions. R package version 0.1.1. `https://CRAN.R-project.org/package=roben`

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 doi:10.1007/9781-071609477_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

`Blend`

`Blend` *fit a robust Bayesian longitudinal regularized semi-parametric mixed model*

Description

fit a robust Bayesian longitudinal regularized semi-parametric mixed model

Usage

```
Blend(
  y,
  x,
  t,
  J,
  kn,
  degree,
  iterations = 10000,
  burn.in = NULL,
  robust = TRUE,
  sparse = "TRUE",
  structural = TRUE
)
```

Arguments

`y`	the vector of repeated - measured response variable. The current version of mixed only supports continuous response.
`x`	the matrix of repeated - measured predictors (genetic factors) with intercept. Each row should be an observation vector for each measurement.
`t`	the vector of scheduled time points.
`J`	the vector of number of repeated measurement for each subject.
`kn`	the number of interior knots for B-spline.
`degree`	the degree of B spline basis.
`iterations`	the number of MCMC iterations.
`burn.in`	the number of iterations for burn-in.
`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.
`structural`	logical flag. If TRUE, the coefficient functions with varying effects and constant effects will be penalized separately.

Details

Consider the data model described in "data":

$$Y_{ij} = \alpha_0(t_{ij}) + \sum_{k=1}^{m} \beta_k(t_{ij}) X_{ijk} + \boldsymbol{Z}_{\boldsymbol{ij}}^{\top} \boldsymbol{\zeta_i} + \epsilon_{ij}.$$

The basis expansion and changing of basis with B splines will be done automatically:

$$\beta_k(\cdot) \approx \gamma_{k1} + \sum_{u=2}^{q} B_{ku}(\cdot)\gamma_{ku}$$

where $B_{ku}(\cdot)$ represents B spline basis. γ_{k1} and $(\gamma_{k2}, \ldots, \gamma_{kq})^{\top}$ correspond to the constant and varying parts of the coefficient functional, respectively. q=kn+degree+1 is the number of basis functions. By default, kn=degree=2. User can change the values of kn and degree to any other positive integers. When ‘structural=TRUE‘(default), the coefficient functions with varying effects and constant effects will be penalized separately. Otherwise, the coefficient functions with varying effects and constant effects will be penalized together.

When ‘sparse="TRUE"‘ (default), spike-and-slab priors are imposed on individual and/or group levels to identify important constant and varying effects. Otherwise, Laplacian shrinkage will be used.

When ‘robust=TRUE‘ (default), the distribution of ϵ_{ij} is defined as a Laplace distribution with density.

$f(\epsilon_{ij}|\theta, \tau) = \theta(1-\theta)\exp\{-\tau\rho_\theta(\epsilon_{ij})\}$, $(i = 1, \ldots, n, j = 1, \ldots, J_i)$, where $\theta = 0.5$. If ‘robust=FALSE‘, ϵ_{ij} follows a normal distribution.

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

Value

an object of class ‘Blend’ is returned, which is a list with component:

`posterior`	the posteriors of coefficients.
`coefficient`	the estimated coefficients.
`burn.in`	the total number of burn-ins.
`iterations`	the total number of iterations.

See Also

data

Examples

```
data(dat)

## default method
fit = Blend(y,x,t,J,kn,degree)
fit$coefficient
```

```
## alternative: robust non-structural
fit = Blend(y,x,t,J,kn,degree, structural=FALSE)
fit$coefficient

## alternative: non-robust structural
fit = Blend(y,x,t,J,kn,degree, robust=FALSE)
fit$coefficient

## alternative: non-robust non-structural
fit = Blend(y,x,t,J,kn,degree, robust=FALSE, structural=FALSE)
fit$coefficient
```

`Coverage` *95% coverage for a Blend object with structural identification*

Description

calculate 95% coverage for varying effects and constant effects under example data

Usage

```
Coverage(x)
```

Arguments

`x`	Blend object.

Value

coverage

See Also

`Blend`

Examples

```
data(dat)
fit = Blend(y,x,t,J,kn,degree)
Coverage(fit)
```

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

Description

Simulated gene expression data for demonstrating the features of Blend.

Format

The data object consists of 8 components: y, x, t, J, kn and degree.

Details

The data and model setting

Consider a longitudinal study on n subjects with J_i repeated measurements for each subject. Let Y_{ij} be the measurement for the i-th subject at each time point t_{ij}, $(1 \leq i \leq n, 1 \leq j \leq J_i)$. We use an m-dimensional vector X_{ij} to denote the genetic factors, where $X_{ij} = (X_{ij1}, ..., X_{ijm})^\top$. Z_{ij} is a 2×1 covariate associated with random effects and ζ_i is a 2×1 vector of random effects corresponding to the random intercept and slope model. We have the following semi-parametric quantile mixed-effects model:

$Y_{ij} = \alpha_0(t_{ij}) + \sum_{k=1}^{m} \beta_k(t_{ij}) X_{ijk} + Z_{ij}^\top \zeta_i + \epsilon_{ij}, \zeta_i \sim N(0, \Lambda)$

where the fixed effects include: (a) the varying intercept $\alpha_0(t_{ij})$, and (b) the varying coefficients $\beta(t_{ij})$.

The varying intercept and the varying coefficients for the genetic factors can be further expressed as $\alpha_0(t_{ij})$ and $\beta(t_{ij}) = (\beta_1(t_{ij}), ..., \beta_m(t_{ij}))^\top$.

For the random intercept and slope model, $Z_{ij}^\top = (1, j)$ and $\zeta_i = (\zeta_{i1}, \zeta_{i2})^\top$.

Furthermore, $Z_{ij}^\top \zeta_i$ can be expressed as $(b_i^\top \otimes Z_{ij}^\top) J_2 \delta$, where $\zeta_i = \Delta b_i$, $\Lambda = \Delta \Delta^\top$, and

$b_i^\top \otimes Z_{ij}^\top = (b_{i1} Z_{ij1}, b_{i1} Z_{ij2}, b_{i2} Z_{ij1}, b_{i2} Z_{ij2})^\top$.

In the simulated data,

$$Y = \alpha_0(t) + \beta_1(t) X_1 + \beta_2(t) X_2 + \beta_3(t) X_3 + \beta_4(t) X_4 + 0.8 X_5 - 1.2 X_6 + 0.7 X_7 - 1.1 X_8 + \epsilon$$

where $\epsilon \sim N(0,1)$, $\alpha_0(t) = 2 + \sin(2\pi t)$, $\beta_1(t) = 2.5 \exp(2.5t - 1)$,$\beta_2(t) = 3t^2 - 2t + 2$,$\beta_3(t) = -4t^3 + 3$ and $\beta_4(t) = 3 - 2t$

See Also

Blend

Examples

```
data(dat)
length(y)
dim(x)
length(t)
length(J)
print(t)
print(J)
print(kn)
print(degree)
```

`plot_Blend` *plot a Blend object*

Description

plot the identified varying effects

Usage

```
plot_Blend(x, sparse, prob=0.95)
```

Arguments

`x`	Blend object.
`sparse`	sparsity.
`prob`	probability for credible interval, between 0 and 1. e.g. prob=0.95 leads to 95% credible interval

Value

plot

See Also

`Blend`

Examples

```
data(dat)
fit = Blend(y,x,t,J,kn,degree)
plot_Blend(fit,sparse=TRUE)
```

`selection` *Variable selection for a Blend object*

Description

Variable selection for a Blend object

Usage

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

Arguments

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

Details

If sparse, the median probability model (MPM) (Barbieri and Berger, 2004) is used to identify predictors that are significantly associated with the response variable. Otherwise, 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:

`method`	posterior samples from the MCMC
`indices`	a list of indices and names of selected variables
`summary`	a summary of selected variables

References

Ren, J., Zhou, F., Li, X., Ma, S., Jiang, Y. and Wu, C. (2023). Robust Bayesian variable selection for gene-environment interactions. *Biometrics*, 79(2), 684-694 doi:10.1111/biom.13670

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

See Also

`Blend`

Examples

```
data(dat)
## sparse
fit = Blend(y,x,t,J,kn,degree)
selected=selection(fit,sparse=TRUE)
selected


## non-sparse
fit = Blend(y,x,t,J,kn,degree,sparse="FALSE")
selected=selection(fit,sparse=FALSE)
selected
```

Index