Package ‘lmmprobe’

June 1, 2024

Type Package

Title Sparse high-dimensional linear mixed modeling with a partitioned empirical Bayes ECM algorithm (LMM-PROBE).

Version 1.0

Date 2023-09-07

Author Anja Zgodic, Alex McLain, Peter Olejua

Maintainer Anja Zgodic <azgodic@email.sc.edu>

Description Linear Mixed Modeling using the PROBE algorithm. This package is in development.

License GPL (>= 2)

Encoding UTF-8

RoxygenNote 7.2.0

Depends R (>= 3.5.0), Rcpp (>= 1.0.8.3), snow (>= 0.4.4), snowfall (>= 1.84.6.1), tidyr (>= 1.2.0), lme4 (>= 1.1.29)

LinkingTo Rcpp, RcppArmadillo

NeedsCompilation yes

R topics documented:

`lmmprobe-package` *Sparse high-dimensional linear mixed modeling with a partitioned empirical Bayes ECM algorithm (LMM-PROBE).*

Description

Linear Mixed Modeling using the PROBE algorithm. This package is in development.

Package Content

Index of help topics:

```
SLE                     High-dimensional dataset for sparse linear
                        mixed modeling.
lmmprobe                Sparse high-dimensional linear mixed modeling
                        with PaRtitiOned empirical Bayes ECM
                        (LMM-PROBE) algorithm.
lmmprobe-package        Sparse high-dimensional linear mixed modeling
                        with a partitioned empirical Bayes ECM
                        algorithm (LMM-PROBE).
```

Maintainer

Anja Zgodic <azgodic@email.sc.edu>

Author(s)

Anja Zgodic, Alex McLain, Peter Olejua

`lmmprobe` *Sparse high-dimensional linear mixed modeling with PaRtitiOned empirical Bayes ECM (LMM-PROBE) algorithm.*

Description

Sparse high-dimensional linear mixed modeling with PaRtitiOned empirical Bayes ECM (LMM-PROBE) algorithm. Currently, the package offers functionality for two scenarios. Scenario 1: only a random intercept, each unit has the same number of observations; Scenario 2: a random intercept and a random slope, each unit has the same number of observations. We are actively expanding the package for more flexibility and scenarios.

Arguments

`Y`	A training-data matrix containing the outcome `Y`.
`Z`	A training-data matrix containing the sparse fixed-effect predictors on which to apply the lmmprobe algorithm. The first columns should be the "id" column.
`V`	A training-data matrix containing non-sparse predictors for the random effects. This matrix is currently only programmed for two scenarios. Scenario 1: only a random intercept, where V is a matrix with one column containing ID's and each unit has the same number of observations. Scenario 2: a random intercept and a random slope, where V is a matrix with two columns. The first column is ID and the second column is a continuous variable (e.g. time) for which a random slope is to be estimated. Each unit has the same number of observations.
`ID_data`	A factor vector of IDs for subjects in the training set.
`Y_test`	A testing-data matrix containing the outcome `Y`. Default is NULL.
`Z_test`	A training-data matrix containing the sparse fixed-effect predictors on which to apply the lmmprobe algorithm. The first columns should be the "id" column. Default is NULL.

`V_test` A training-data matrix containing non-sparse predictors for the random effects. This matrix is currently only programmed for two scenarios. Scenario 1: only a random intercept, where V is a matrix with one column containing ID's and each unit has the same number of observations. Scenario 2: a random intercept and a random slope, where V is a matrix with two columns. The first column is ID and the second column is a continuous variable (e.g. time) for which a random slope is to be estimated. Each unit has the same number of observations. Default is NULL.

`ID_test` A factor vector of IDs for subjects in the testing set.

`alpha` Type I error; significance level.

`ep` Value against which to compare convergence criterion, we recommend 0.05.

`B` The number of groups to categorize estimated coefficients in to calculate correlation ρ. We recommend five.

`adj` A factor multiplying Silverman's 'rule of thumb' in determining the bandwidth for density estimation, same as the 'adjust' argument of R's density function. Default is three.

`maxit` Maximum number of iterations the algorithm will run for. Default is 10000.

`cpus` The number of CPUS user would like to use for parallel computations. Default is four.

`LR` A learning rate parameter `r`. Using zero corresponds to the implementation described in Zgodic et al.

`C` A learning rate parameter `c`. Using one corresponds to the implementation described in Zgodic et al.

`sigma_init` An initial value for the residual variance parameter. Default is NULL which corresponds to the sample variance of Y.

Value

A list of the output of the lmmprobe function, including

`beta_hat,beta_hat_var` MAP estimates of the posterior expectation (beta_hat) and variance (beta_hat_var) of the prior mean (β) of the regression coefficients assuming $\gamma = 1$,

`gamma` the posterior expectation of the latent γ variables,

`preds` predictions of Y,

`PI_lower,PI_upper` lower and upper prediction intervals for the predictions,

`sigma2_est` MAP estimate of the residual variance,

`random_var` MAP estimate of the random effect(s) variance,

`random_intercept` estimated random intercept terms,

`random_slope` estimated random slope terms, if applicable.

References

Zgodic A, Bay R, Zhang J, Olejua P, and McLain AC (2023). Sparse high-dimensional linear mixed modeling with a partitioned empirical Bayes ECM algorithm. arXiv preprint arXiv:2310.12285.

Examples

```
library(lmmprobe)
data(SLE)
Y <- matrix(real_data[,"y"], ncol=1)
Z <- real_data[,4:ncol(real_data)]
V <- matrix(real_data[,"id"], ncol=1)
ID_data <- as.numeric(as.character(real_data$id))
full_res <- lmmprobe(Y = Y, Z = Z, V = V, ID_data = ID_data)
```

SLE *High-dimensional dataset for sparse linear mixed modeling.*

Description

This is the Systemic Lupus Erythematosus (SLE) data used in the data analysis section of the LMM-PROBE reference. The dataset has 353 observations, 125 subjects, and 15397 predictors.

Usage

```
data(SLE)
```

Format

A data frame 'real_data' with 353 observations.

References

Banchereau, R., Hong, S., Cantarel, B., Baldwin, N., Baisch, J., Edens, M., Cepika, A.-M., Acs, P., Turner, J., Anguiano, E., Vinod, P., Khan, S., Obermoser, G., Blankenship, D., Wakeland, E., Nassi, L., Gotte, A., Punaro, M., Liu, Y.-J., Banchereau, J., Rossello-Urgell, J., Wright, T., and Pascual, V. (2016), "Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients," Cell, 165, 551–565.

Zgodic A, Bay R, Zhang J, Olejua P, and McLain AC (2023). Sparse high-dimensional linear mixed modeling with a partitioned empirical Bayes ECM algorithm. arXiv preprint arXiv:2310.12285.

Examples

```
data(SLE)
Y <- matrix(real_data[,"y"], ncol=1)
Z <- real_data[,4:ncol(real_data)]
V <- matrix(real_data[,"id"], ncol=1)
ID_data <- as.numeric(as.character(real_data$id))
```

Index