Package ‘ebreg’

October 29, 2019

Type Package

Title Implementation of the Empirical Bayes Method

Version 0.1.0

Author Yiqi Tang, Ryan Martin

Maintainer Yiqi Tang <ytang22@ncsu.edu>

Description
   Implements a Bayesian-like approach to the high-dimensional sparse linear regression problem based on an empirical or data-dependent prior distribution, which can be used for estimation/inference on the model parameters, variable selection, and prediction of a future response. The method was first presented in Martin, Ryan and Mess, Raymond and Walker, Stephen G (2017) <doi:10.3150/15-BEJ797>. More details focused on the prediction problem are given in Martin, Ryan and Tang, Yiqi (2019) <arXiv:1903.00961>.

License GPL-3

Encoding UTF-8

LazyData true

Depends lars, stats

RoxygenNote 6.1.1

Imports Rdpack

RdMacros Rdpack

Suggests testthat, roxygen2

NeedsCompilation no

Repository CRAN

Date/Publication 2019-10-29 15:00:02 UTC

R topics documented:

ebreg ................................................................. 2

Index 5
Implements the empirical Bayes method in high-dimensional linear model setting for inference and prediction

Description

The function ebreg implements the method first presented in Martin, Mess, and Walker (2017) for Bayesian inference and variable selection in the high-dimensional sparse linear regression problem. The chief novelty is the manner in which the prior distribution for the regression coefficients depends on data; more details, with a focus on the prediction problem, are given in Martin and Tang (2019).

Usage

```
ebreg(y, X, XX, alpha, gam, sig2, prior = TRUE, igpar, log.f, M, sample.beta = FALSE, pred = FALSE, conf.level = 0.95)
```

Arguments

- `y`: vector of response variables for regression
- `X`: matrix of predictor variables
- `XX`: vector to predict outcome variable, if `pred=TRUE`
- `alpha`: numeric value between 0 and 1, likelihood fraction
- `gam`: numeric value between 0 and 1, conditional prior precision parameter
- `sig2`: numeric value for error variance. If NULL (default), variance is estimated from data
- `prior`: logical. If TRUE, a prior is used for the error variance
- `igpar`: the parameters for the inverse gamma prior on the error variance
- `log.f`: log of the prior for the model size
- `M`: integer value to indicate the Monte Carlo sample size (burn-in of size 0.2 * M automatically added)
- `sample.beta`: logical. If TRUE, samples of beta are obtained
- `pred`: logical. If TRUE, predictions are obtained
- `conf.level`: numeric value between 0 and 1, confidence level for the marginal credible interval if `sample.beta=TRUE`, and for the prediction interval if `pred=TRUE`

Details

Consider the classical regression problem

\[ y = X\beta + \sigma\epsilon, \]

where \( y \) is a \( n \)-vector of responses, \( X \) is a \( n \times p \) matrix of predictor variables, \( \beta \) is a \( p \)-vector of regression coefficients, \( \sigma > 0 \) is a scale parameter, and \( \epsilon \) is a \( n \)-vector of independent and identically
distributed standard normal random errors. Here we allow $p \geq n$ (or even $p \gg n$) and accommodate
the high dimensionality by assuming $\beta$ is sparse in the sense that most of its components are zero.
The approach described in Martin, Mess, and Walker (2017) and in Martin and Tang (2019) starts
by decomposing the full $\beta$ vector as a pair $(S, \beta_S)$ where $S$ is a subset of indices $1, 2, \ldots, p$ that
represents the location of active variables and $\beta_S$ is the $|S|$-vector of non-zero coefficients. The
approach proceeds by specifying a prior distribution for $S$ and then a conditional prior distribution
for $\beta_S$, given $S$. This latter prior distribution here is taken to depend on data, hence "empirical". A
prior distribution for $\sigma^2$ can also be introduced, and this option is included in the function.

Value
A list with components
• beta - matrix with rows containing sampled beta, if sample.beta=TRUE, otherwise NULL
• beta.mean - vector containing the posterior mean of sampled beta, if sample.beta=TRUE, otherwise NULL
• ynew - matrix containing predicted responses, if pred=TRUE, otherwise NULL
• ynew.mean - vector containing the predictions for the predictor values tested, XX, if pred=TRUE, otherwise NULL
• S - matrix with rows containing the sampled models
• incl.prob - vector containing inclusion probabilities of the predictors
• sig2 - estimated error variance, if prior=FALSE, otherwise NULL
• PI - prediction interval, confidence level specified by the user, if pred=TRUE, otherwise NULL
• CI - matrix containing marginal credible intervals, confidence level specified by the user, if sample.beta=TRUE, otherwise NULL

Author(s)
Yiqi Tang
Ryan Martin

References

Examples
n <- 70
p <- 100
beta <- rep(1, 5)
s0 <- length(beta)
sig2 <- 1
d <- 1
log.f <- function(x) -x * (log(1) + 0.05 * log(p)) + log(x <= n)
X <- matrix(rnorm(n * p), nrow=n, ncol=p)
X.new <- matrix(rnorm(p), nrow=1, ncol=p)
y <- as.numeric(X[, 1:s0] %*% beta[1:s0]) + sqrt(sig2) * rnorm(n)

o <- ebreg(y, X, X.new, .99, .005, NULL, FALSE, igpar=c(0.01, 4), log.f, M=5000, TRUE, FALSE, .95)

incl.pr <- o$incl.prob
plot(incl.pr, xlab="Variable Index", ylab="Inclusion Probability", type="h", ylim=c(0,1))
Index

ebreg, 2