Package ‘basad’

November 20, 2017

Type Package

Title Bayesian Variable Selection with Shrinking and Diffusing Priors

Version 0.2.0

Date 2017-11-15

Author Qingyan Xiang <qxiang@illinois.edu>, Naveen Narisetty <naveen@illinois.edu>

Maintainer Qingyan Xiang <qxiang@illinois.edu>

Description Provides a Bayesian variable selection approach using continuous spike and slab prior distributions. The prior choices here are motivated by the shrinking and diffusing priors studied in Narisetty & He (2014) <DOI:10.1214/14-AOS1207>.

License GPL (>= 3)

Imports Rcpp, rmutil

LinkingTo Rcpp, RcppEigen

NeedsCompilation yes

Repository CRAN

Date/Publication 2017-11-20 21:41:07 UTC

R topics documented:

basad ......................................................... 2
predict.basad ............................................. 5
print.basad ................................................. 6

Index 8
Bayesian variable selection with shrinking and diffusing priors

Description

This function performs the Bayesian variable selection procedure with shrinking and diffusing priors via Gibbs sampling. Three different prior options placed on the coefficients are provided: Gaussian, Student's t, Laplace. The posterior estimates of coefficients are returned and the final model is selected either by using the "BIC" criterion or the median probability model.

Usage

basad( x = NULL, y = NULL, K = -1, df = 5, nburn = 1000, niter = 1000,
alternative = FALSE, verbose = FALSE, nsplit = 20, tau0 = NULL, tau1 = NULL,
prior.dist = "Gauss", select.cri = "median", BIC.maxsize = 20)

Arguments

x
The matrix or data frame of covariates.
y
The response variables.
K
An initial value for the numbers of active covariates in the model. This value is related to the prior probability that a covariate is nonzero. If K is not specified greater than 3, this prior probability will be estimated by a Beta prior using Gibbs sampling (see details below).
df
The degrees of freedom of t prior when prior.dist == "t".
nburn
The number of iterations for burn-in.
niter
The number of iterations for estimation.
alternative
If TRUE, an alternative sampling scheme from Bhattacharya will be used which can accelerate the speed of the algorithm for very large p. However, when using block updating (by setting nsplit to be greater than 1) this alternative sampling will not be invoked.
verbose
If TRUE, verbose output is sent to the terminal.
nsplit
Numbers of splits for the block updating scheme.
tau0
The scale of the prior distribution for inactive coefficients (see details below).
tau1
The scale of the prior distribution for active coefficients (see details below).
prior.dist
Choice of the base distribution for spike and slab priors. If prior.dist="t", the algorithm will place Student's t prior for regression coefficients. If prior.dist="Laplace", the algorithm will place Laplace prior. Otherwise, it will place the default Gaussian priors.
select.cri
Model selection criteria. If select.cri="median", the algorithm will use the median probability model to select the active variables. If select.cri="BIC", the algorithm will use the BIC criteria to select the active variables.
BIC.maxsize
The amount of the variables that are chosen to apply BIC criteria based on the ranking of their marginal posterior probabilities. If the input sample size is less than the default value 20, all variables will be considered when applying BIC.
Details

In the package, the regression coefficients have following hierarchical structure:

\[
\beta | (Z = 0, \sigma^2) = N(0, \tau_0^2 \sigma^2), \beta | (Z = 1, \sigma^2) = N(0, \tau_1^2 \sigma^2)
\]

where the latent variable \( Z \) of value 0 or 1 indicates whether \( i \)th variable is in the slab and spike part of the prior. The package provides different prior choices for the coefficients: Gaussian, Student’s t, Laplace. Through setting the parameter \texttt{prior.dist}, the coefficients will have the corresponding prior densities as follows:

1. The Gaussian priors case:

\[
\beta | (Z = k, \sigma^2) = \frac{1}{\sqrt{2\pi \tau_k^2 \sigma^2}} e^{-\frac{\beta^2}{2\tau_k^2 \sigma^2}}
\]

2. The Student’s t prior case:

\[
\beta | (Z = k, \sigma^2) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi \nu \tau_k \sigma^2}} \left(1 + \frac{1}{\nu} \left(\frac{\beta^2}{\tau_k^2 \sigma^2}\right)\right)^{-\frac{\nu+1}{2}}
\]

Where \( \nu \) is the degrees of freedom

3. The Laplace prior case:

\[
\beta | (Z = k, \sigma^2) = \frac{1}{2\tau_k^2 \sigma^2} \exp\left(\frac{-|\beta|}{\tau_k^2 \sigma^2}\right)
\]

The \( \tau_k \) is the scale for the prior distribution. If user did not set a specific value, the prior scales are specified as follows:

\[
\tau_0^2 = \frac{1}{n} a, \quad \tau_1^2 = \max\left(100\tau_0^2, \frac{\tau_0 p_n}{(1 - p_n) \rho}\right)
\]

where \( \rho \) is the prior density evaluated at \( f_p(b_T \times \log(p_n + 1)) \), \( f_p \) is the density function for the corresponding prior distribution. The parameter \( a \) and \( b \) are \( a_T = 1 \) and \( b_T = 2.4 \) by default.

The prior probability \( q_n = P(Z_i = 1) \) that a covariate is nonzero can be specified by value K. The K represents a prior belief of the upper bound of the true covariates in the model. When user specifies a value of K greater than 3, setting \( q_n = c/p_n \), through the calculation(see details in Naveen (2014)):

\[
\Phi \left(\frac{(K - c)/\sqrt{c}}{\sqrt{c}}\right) = 1 - \alpha
\]

The prior probability on the models with sizes greater than K will be \( \alpha \), and this \( \alpha \) is set to 0.1 in the package.

Value

An object of class \texttt{basad} with the following components:

- \texttt{all.var} Summary object for all the variables.
- \texttt{select.var} Summary object for the selected variables.
beta.names  Variable names for the coefficients.
verbose     Verbose details (used for printing).
posteriorZ A vector of the marginal posterior probabilities for the latent vector Z.
model.index A vector containing the indices of selected variables.
modelZ      A binary vector Z indicating whether the coefficient is true in the selected model.
est.B       Estimated coefficient values from the posterior distribution through Gibbs sampling.
allB        A matrix of all sampled coefficient values along the entire chain. Each row represents the sampled values under each iteration.
allZ        A matrix of all sampled posterior probabilities for the latent variable Z along the entire chain. Each row represents the sampled values under each iteration.
x           Standardized x-matrix.
y           Standardized y vector.

Author(s)

Qingyan Xiang (<qxiang@illinois.edu>)
Naveen Narisetty (<naveen@illinois.edu>)

References


Examples

#-------------------------------------------------------------
#Generate Data: The simulated high dimensional data           
#-------------------------------------------------------------

n = 100; p = 499; nz = 5

rho1=0.25; rho2=0.25; rho3=0.25  ### correlations
Bc = c( 0,seq(0.6,3,length.out=nz), array(0, p-nz))

covr1=(1- rho1)*diag(nz) + array(rho1,c(nz,nz))
covr3=(1- rho3)*diag(p-nz) + array(rho3,c(p-nz,p-nz))
covr2=array(rho2,c(nz,p-nz))
cov=rbind( cbind(covr1,covr2), cbind(t(covr2),covr3) )
predict.basad

```r
covE = eigen(covr)
covsq = covE$vectors %*% diag( sqrt(covE$values) ) %*% t(covE$vectors)

Xs = matrix( rnorm(n*p), nrow = n ); Xn = covsq %*% t(Xs)
X = cbind(array(1, n), t(Xn))
Y = X %*% Bc + rnorm(n); X <- X[,2:ncol(X)]

#-----------------------------------------------------------
#Example 1: Run the default setting of the Gussian priors
#-----------------------------------------------------------
obj <- basad(x = X, y = Y)
print(obj)

#-----------------------------------------------------------
#Example 2: Use different priors and selection criteria
#-----------------------------------------------------------
obj <- basad(x = X, y = Y, prior.dist = "t", select.cri = "BIC")
print(obj)
```

---

predict.basad  Basad prediction

**Description**

Predict the response values of test data using basad.

**Usage**

```r
## S3 method for class 'basad'
predict(object, testx = NULL, ...)
```

**Arguments**

- `object`  An object of class basad.
- `testx`  Data frame or x-matrix containing test data.
- `...`  Further arguments passed to or from other methods.

**Value**

A vector of fitted values for estimated response values.
Author(s)
Qingyan Xiang (<qxiang@illinois.edu>)
Naveen Narisetty (<naveen@illinois.edu>)

References

Examples

```r
#-----------------------------
#Generate Data: The simulated high dimensional data
#-----------------------------

n = 100; p = 499; nz = 5

rho1=0.25;rho2=0.25;rho3=0.25  ### correlations
Bc = c( 0,seq(0.6,3,length.out=nz), array(0, p-nz))

covr1=(1- rho1)*diag(nz) + array(rho1,c(nz,nz))
covr3=(1- rho3)*diag(p-nz) + array(rho3,c(p-nz,p-nz))
covr2= array(rho2,c(nz,p-nz))
covr=rbind( cbind(covr1,covr2), cbind(t(covr2),covr3) )

cove = eigen(covr)
covsq = cove$vectors %*% diag(sqrt(cove$values)) %*% t(cove$vectors)

Xs = matrix(rnorm(n*p), nrow = n); Xn = covsq %*% t(Xs)
X = cbind(array(1, n), t(Xn))
Y = X %*% Bc + rnorm(n); X <- X[,2:ncol(X)]

#-----------------------------
#Run the algorithm and then predict
#-----------------------------

obj <- basad( x = X, y = Y)
predict( obj, testx = X )
```

Description
Print summary output from basad analysis. Note that this is the default print method for the package.
print.basad

Usage

## S3 method for class 'basad'
print(x, ...)

Arguments

x               An object of class basad.
...             Further arguments passed to or from other methods.

Author(s)

Qingyan Xiang (<qxiang@illinois.edu>)
Naveen Narisetty (<naveen@illinois.edu>)

References

Index

*Topic print
  print.basad, 6
*Topic regression
  basad, 2
  predict.basad, 5

basad, 2
predict.basad, 5
print.basad, 6