Package 'IGG'

April 9, 2018

Type Package

Title Inverse Gamma-Gamma

Version 1.0

Date 2018-04-04

Author Ray Bai, Malay Ghosh

Maintainer Ray Bai <raybai07@ufl.edu>

Description Implements Bayesian linear regression, normal means estimation, and variable selection using the inverse gamma-gamma prior, as introduced by Bai and Ghosh (2018) <arXiv:1710.04369>.

License GPL-3

LazyData true

Depends R (>= 3.1.0), Matrix, MASS, glmnet, pscl, GIGrvg

NeedsCompilation yes

Repository CRAN

Date/Publication 2018-04-09 12:11:15 UTC

R topics documented:

IGG-package *Inverse Gamma-Gamma*

Description

This package contains the functions, `igg` and `igg.normalmeans`, for implementing Bayesian linear regression, sparse normal means estimation, and variable selection with the inverse gamma-gamma (IGG) prior, as introduced by Bai and Ghosh (2018) <arXiv:1710.04369>

Details

The DESCRIPTION file: This package was not yet installed at build time.

Index of help topics:

```
IGG-package             Inverse Gamma-Gamma
diabetes                Blood and other measurements in diabetics
igg                     Inverse Gamma-Gamma Regression
igg.normalmeans         Normal Means Estimation and Classification with
                        the IGG Prior
singh2002               Prostate Cancer Study of Singh et al. (2002)
```

This package implements the IGG model for sparse Bayesian linear regression and the normal means problem. Our package performs both estimation and model selection. The `igg` and `igg.normalmeans` functions also returns the endpoints of the credible intervals (i.e. the 2.5th and 97.5th percentiles) for every single parameter of interest, so that uncertainty quantification can be assessed.

Author(s)

Ray Bai and Malay Ghosh

Maintainer: Ray Bai <raybai07@ufl.edu>

References

Bai, R. and Ghosh, M. (2018). "The Inverse Gamma-Gamma Prior for Optimal Posterior Contraction and Multiple Hypothesis Testing." Submitted, arXiv:1711.07635.

diabetes *Blood and other measurements in diabetics*

Description

The diabetes data frame has 442 rows and 3 columns. These are the data used in the Efron et al. "Least Angle Regression" paper and is also available in the lars package.

Usage

```
data(diabetes)
```

Format

This data frame consists of the following columns:

x: is a design matrix with 10 columns (no interactions).

y: is a numeric vector.

x2: is a design matrix with 64 columns (includes interactions.

Details

The x matrix has been standardized to have unit L2 norm in each column and zero mean. The matrix x2 consists of x plus certain interactions.

Source

https://cran.r-project.org/web/packages/lars/

References

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2003). "Least Angle Regression" (with discussion). *Annals of Statistics*, **32**(2): 407-499.

igg *Inverse Gamma-Gamma Regression*

Description

This function provides a fully Bayesian approach for obtaining a sparse estimate of the $p \times 1$ vector, β in the univariate linear regression model,

$$y = X\beta + \epsilon,$$

where $\epsilon \sim N_n(0, \sigma^2 I_n)$. This is achieved by placing the inverse gamma-gamma (IGG) prior on the coefficients of β. In the case where $p > n$, we utilize an efficient sampler from Bhattacharya et al. (2016) to reduce the computational cost of sampling from the full conditional density of β to $O(n^2 p)$.

Usage

```
igg(X, y, a=NA, b=NA, sigma2=NA, max.steps=10000, burnin=5000)
```

Arguments

X	$n \times p$ design matrix.
y	$n \times 1$ response vector.
a	The parameter for $IG(a, 1)$. If not specified (a=NA), defaults to $1/2 + 1/p$. User may specify a value for a between 0 and 1.
b	The parameter for $G(b, 1)$. If not specified (b=NA), defaults to $1/p$. User may specify a value for b between 0 and 1.
sigma2	The variance parameter. If the user does not specify this (sigma2=NA), the Gibbs sampler will estimate this using Jeffreys prior. If σ^2 is known or estimated separately (e.g. through empirical Bayes), the user may also specify it.
max.steps	The total number of iterations to run in the Gibbs sampler. Defaults to 10,000.
burnin	The number of burn-in iterations for the Gibbs sampler. Defaults to 5,000.

Details

The function performs sparse estimation of β in the standard linear regression model and variable selection from the p covariates. Variable selection is performed by assessing the 95 percent marginal posterior credible intervals. The lower and upper endpoints of the 95 percent credible intervals for each of the p covariates are also returned so that the user may assess uncertainty quantification. The full model is:

$$Y|(X, \beta) \sim N_n(X\beta, \sigma^2 I_n),$$
$$\beta_i|(\lambda_i, \xi_i, \sigma^2) \sim N(0, \sigma^2\lambda_i\xi_i), i = 1, ..., p,$$
$$\lambda_i \sim IG(a, 1), i = 1, ..., p,$$
$$\xi_i \sim G(b, 1), i = 1, ..., p,$$
$$\sigma^2 \propto 1/\sigma^2.$$

If σ^2 is known or estimated separately, the Gibbs sampler does not sample from the full conditional for σ^2.

Value

The function returns a list containing the following components:

beta.hat	The posterior mean estimate of β.
beta.med	The posterior median estimate of β.
beta.intervals	The lower and upper endpoints of the 95 percent credible intervals for all p estimates in β.
igg.classifications	A p-dimensional binary vector with "1" if the covariate is selected and "0" if it is deemed irrelevant.

Author(s)

Ray Bai and Malay Ghosh

References

Bai, R. and Ghosh, M. (2018). "The Inverse Gamma-Gamma Prior for Optimal Posterior Contraction and Multiple Hypothesis Testing." Submitted, arXiv:1711.07635.

Bhattacharya, A., Chakraborty, A., and Mallick, B.K. (2016). "Fast Sampling with Gaussian Scale Mixture Priors in High-Dimensional Regression." *Biometrika*, **69**(2): 447-457.

Examples

```
######################
# Load diabetes data #
######################
data(diabetes)
attach(diabetes)
X <- scale(diabetes$x)
y <- scale(diabetes$y)

###############################
# Fit the IGG regression model #
###############################
igg.model <- igg(X=X, y=y, max.steps=5000, burnin=2500)

############################
# Posterior median estimates #
############################
igg.model$beta.med

########################################
# 95 percent posterior credible intervals #
########################################
igg.model$beta.intervals

######################
# Variable selection #
######################
igg.model$igg.classifications
```

`igg.normalmeans` *Normal Means Estimation and Classification with the IGG Prior*

Description

This function provides a fully Bayesian approach for obtaining a sparse estimate of $\theta = (\theta_1, ..., \theta_n)$ in the normal means problem,

$$X_i = \theta_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$. This is achieved by placing the inverse gamma-gamma (IGG) prior on the individual θ_i's. Variable selection can also be performed by using either thresholding the shrinkage factor in the posterior mean, or by examining the marginal 95 percent credible intervals.

Usage

```
igg.normalmeans(x, a=NA, b=NA, sigma2=NA, var.select = c("threshold", "intervals"),
max.steps=10000, burnin=5000)
```

Arguments

`x`	an $n \times 1$ multivariate normal vector.
`a`	The parameter for $IG(a, 1)$. If not specified (`a=NA`), defaults to $1/2 + 1/n$. User may specify a value for `a` between 0 and 1.
`b`	The parameter for $G(b, 1)$. If not specified (`b=NA`), defaults to $1/n$. User may specify a value for `b` between 0 and 1.
`sigma2`	The variance parameter. If the user does not specify this (`sigma2=NA`), the Gibbs sampler will estimate this using Jeffreys prior. If $sigma^2$ is known or estimated separately (e.g. through empirical Bayes), the user may also specify it.
`var.select`	The method of variable selection. `threshold` selects variables by thresholding the shrinkage factor in the posterior mean. `intervals` will classify entries of x as either signals ($x_i \neq 0$) or as noise ($x_i = 0$) by examining the 95 percent marginal credible intervals.
`max.steps`	The total number of iterations to run in the Gibbs sampler. Defaults to 10,000.
`burnin`	The number of burn-in iterations for the Gibbs sampler. Defaults to 5,000.

Details

The function performs sparse estimation of $\theta = (\theta_1, ..., \theta_n)$ in normal means problem. The full model is:

$$X|\theta \sim N_n(\theta, \sigma^2 I_n),$$

$$\theta_i|(\lambda_i, \xi_i, \sigma^2) \sim N(0, \sigma^2\lambda_i\xi_i), i = 1, ..., n,$$

$$\lambda_i \sim IG(a, 1), i = 1, ..., n,$$

$$\xi_i \sim G(b, 1), i = 1, ..., n,$$

$$\sigma^2 \propto 1/\sigma^2.$$

If σ^2 is known or estimated separately, the Gibbs sampler does not sample from the full conditional for σ^2.

As described in Bai and Ghosh (2018), the posterior mean is of the form $[E(1-\kappa_i|X_i)]X_i, i = 1, ..., n$. The "threshold" method for variable selection is to classify θ_i as signal ($\theta_i \neq 0$) if

$$E(1-\kappa_i|X_i) > 1/2,$$

and to classify θ_i as noise ($\theta_i = 0$) if

$$E(1-\kappa_i|X_i) \leq 1/2.$$

Value

The function returns a list containing the following components:

`theta.hat` The posterior mean estimate of θ.

`theta.med` The posterior median estimate of θ.

`theta.intervals`
The lower and upper endpoints of the 95 percent credible intervals for all n components of θ.

`igg.classifications`
An n-dimensional binary vector with "1" if the covariate is selected and "0" if it is deemed irrelevant.

Author(s)

Ray Bai and Malay Ghosh

References

Bai, R. and Ghosh, M. (2018). "The Inverse Gamma-Gamma Prior for Optimal Posterior Contraction and Multiple Hypothesis Testing." Submitted, arXiv:1711.07635.

Examples

```
#####################################################
#####################################################
## Example on synthetic data.                      ##
## 5 percent of entries in a sparse vector theta ##
## are set equal to signal value A =7.             ##
#####################################################
#####################################################

n <- 100
sparsity.level <- 5
A <- 7

# Initialize theta vector of all zeros
theta.true <- rep(0,n)

# Set (sparsity.level)% of them to be A
q <- floor(n*(sparsity.level/100))
```

```
# Pick random indices of theta.true to equal A
signal.indices <- sample(1:n, size=q, replace=FALSE)

######################
# Generate true theta #
######################
theta.true[signal.indices] <- A

#########################################################
# Generate data X by corrupting theta.true with noise #
#########################################################
X <- theta.true + rnorm(n,0,1)

#########################
# Run the IGG model on X #
#########################
# For optimal performance, should set max.steps=10,000 and burnin=5000.

igg.model <- igg.normalmeans(X, var.select="threshold", max.steps=2000, burnin=1000)

################################
# Calculate mean squared error #
################################
igg.mse <- sum((igg.model$theta.med-theta.true)^2)/n
igg.mse

# To compute misclassification probability and false discovery rate
true.classifications <- rep(0,n)
signal.indicies <- which(theta.true != 0)
true.classifications[signal.indices] <- 1
igg.classifications <- igg.model$igg.classifications

false.pos <- length(which(igg.classifications != 0 & true.classifications == 0))
num.pos <- length(which(igg.classifications != 0))
false.neg <- length(which(igg.classifications == 0 & true.classifications != 0))

#########################################################
# Calculate FDR and misclassification probability (MP) #
#########################################################
igg.fdr <- false.pos/num.pos
igg.fdr

igg.mp <- (false.pos+false.neg)/n
igg.mp

## Not run:
#
#
#####################################################
#####################################################
## Prostate cancer data analysis.                  ##
```

```
## Running this code will allow you to reproduce    ##
## the results in Section 6 of Bai and Ghosh (2018) ##
######################################################
######################################################

# Load the data
data(singh2002)
attach(singh2002)

# Only look at the gene expression data.
# First 50 rows are the cancer patients,
# and the last 52 rows are the control subjects.d

prostate.data <- singh2002$x

# Perform 2-sample t-test and obtain z-scores
n <- ncol(prostate.data)
test.stat <- rep(NA,n)
z.scores <- rep(NA, n)

# Fill in the vectors
for(i in 1:n){
  test.stat[i] <- t.test(prostate.data[51:102,i],
                         prostate.data[1:50,i])$statistic
  z.scores[i] <- qnorm(pt(test.stat[i],100))
}


#######################################
# Apply IGG model on the z-scores.    #
# Here sigma2 is known with sigma2= 1 #
#######################################

igg.model <- igg.normalmeans(z.scores, sigma2=1, var.select="threshold")

#########################################
# How many genes flagged as significant? #
#########################################
num.sig <- sum(igg.model$igg.classifications != 0)
num.sig

#########################################################
# Estimated effect size for 10 most significant genes #
#########################################################
most.sig <- c(610,1720,332,364,914,3940,4546,1068,579,4331)
igg.model$theta.hat[most.sig]

## End(Not run)
```

---

singh2002 *Prostate Cancer Study of Singh et al. (2002)*

Description

Gene expression dat (6033 genes for 102 samples) from the microarray study of Singh et al. (2002). Also available in the sda package.

Usage

```
data(singh2002)
```

Format

A list of two components.

x: is a 102 $\times$ 6033 matrix containing the expression levels. The rows contain the samples and the columns the genes.

y: is a factor containing the diagnosis for each sample ("cancer" or "healthy").

Details

This data set contains measurements of the gene expression of 6033 genes for 102 observations. The first 52 rows are for the cancer patients and the last 50 rows are for the normal control subjects.

Source

The data is described in Singh et al. (2001) and are provided in exactly the form as used by Efron (2010).

References

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Maonla, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R., and Sellers, W.R. (2002). "Gene expression correlates of clinical prostate cancer behavior." *Cancer Cell*, **1**(2): 203-209.

Efron, B. (2010). "The Future of Indirect Evidence." *Statistical Science*, **25**(2): 145-157.

Index