A Comparative Study of Bayesian Lasso

(Project of MTH535A)

Submitted by:

Rajdeep Saha (201380)
Sagnik Dey (201397)
Saumyadip Bhowmick (201408)
Shuvam Gupta (201421)
Soumik Karmakar (201428)

Supervised by,

Dr. Arnab Hazra

Contents

1 Introduction

In a normal linear regression setup, we have the following model

$$y = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where y is the $n \times 1$ vector of centered responses; X is the $n \times p$ matrix of standardized regressors; $\boldsymbol{\beta}$ is the $p \times 1$ vector of coefficients to be estimated and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of independent and identically distributed normal errors with mean 0 and variance σ^2.
The classical estimator in linear regression is the Ordinary Least Squares (OLS) estimator

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y$$

, which is obtained by minimizing the residual sum of squares $(RSS) = (y - X\beta)'(y - X\beta)$.
For very high dimensional data there may be multicollinearity between the variables. In those cases the estimate of the coefficients becomes unstable and the variance of the estimated increases, making the estimates unreliable. Also, if $p >> n$, X is less than full rank, so the solution for the coefficients becomes non-unique. So, several penalty terms are introduced in the least squares regression methods to improve upon the prediction accuracy of OLS.
Some example of penalized regression are Ridge Regression, LASSO regression etc.
Frank and Friedman [3] introduced Bridge regression which minimizes RSS subject to a constraint $\sum_{j=1}^{p} |\beta_j^{\alpha}| \leq t, \alpha \geq 0$. For $\alpha = 2$ it is just Ridge regression i.e Ridge regression minimizes RSS subject to constraint $\sum_{j=1}^{p} |\beta_j^{2}| \leq t$. Although ridge regression often achieves better prediction accuracy by shrinking OLS coefficients, it cannot do variable selection as it naturally keeps all the predictors.
Among penalized regression techniques, probably the most widely used method in statistical literature is the Least Absolute Shrinkage and Selection Operator (LASSO) by Tibshirani [2] , which is a special case of bridge estimator with $\alpha = 1$. The lasso estimate is obtained by minimizing

$$Q(\beta) = (y - X\beta)'(y - X\beta) + \lambda\|\beta\|_1, \lambda > 0 \tag{2}$$

Compared to ridge regression a remarkable property of lasso is that it can shrink some coefficients exactly to zero, which facilitates automatic variable selection. Various computationally efficient algorithms have been proposed to obtain the lasso and related estimators. For a given value of the tuning parameter(s), these algorithms are quite fast.

2 Bayesian Lasso

As discussed in the Introduction, although penalized regression is a very efficient and accurate way of prediction, still none of these algorithms provide a valid measure of standard error, which is arguably a major drawback of these approaches.
Very recently, much work has been done in the direction of Bayesian framework. Noting the form of penalty term in (2) Tibshirani [2] suggested that lasso estimates can be interpreted as posterior mode estimates when the regression parameters are assigned independent and identical **Laplace priors**.
Motivated by this, different approaches based on **scale mixture of normal (SMN)** distributions with independent exponentially distributed variances have been proposed. Park and Casella [1] introduced Gibbs sampling using a conditional Laplace prior specification of the form

$$\pi(\beta|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp -\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}} \tag{3}$$

and non-informative scale-invariant marginal prior on σ^2, i.e. $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$.

Note:

A continuous mixture distributions of R.V. X with pdf $p(x|a)$ parameterized by a and mixing density being $\pi(a)$ is written as follows,

$$f(x) = \int_A p(x|a).\pi(a)da$$

Park and Casella [1] devoted serious efforts to address the important issue of unimodality. They pointed out that conditioning on σ^2 is important for unimodality and lack of unimodality might slow down the convergence of the Gibbs sampler and make the point estimates less meaningful. Unlike their frequentist counterparts, Bayesian methods usually provide a valid measure of standard error.
The Gibbs sampler for the Bayesian Lasso exploits the following representation of the Laplace distribution as a **scale mixture of normals** (with an exponential mixing density).

$$\frac{a}{2}e^{-a|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-\frac{z^2}{2s}}.\frac{a^2}{2}e^{-\frac{a^2 s}{2}} ds$$

Following from the **Note** above we can say that here $p(x|s) = N(0, s)$ and $\pi(s) = \frac{a^2}{2}e^{-\frac{a^2 s}{2}}$ (Mixing density is $Exponential(\frac{a^2}{2})$).
This suggests the following hierarchical representation of the full model:

$$y|\mu, X, \beta, \sigma^2 \sim N_n(\mu 1_n + X\beta, \sigma^2 I_n),$$

$$\beta|\sigma^2, \tau_1^2, ..., \tau_p^2 \sim N_p(0_p, \sigma^2 D_\tau),$$

$$D_\tau = \text{diag}(\tau_1^2, ..., \tau_p^2), \sigma^2, \tau_1^2, ..., \tau_p^2 \sim \pi(\sigma^2)d\sigma^2 \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\frac{\lambda^2 \tau_j^2}{2}}$$

,

$$\sigma^2, \tau_1^2, ..., \tau_p^2 > 0$$

We use improper prior density $\pi(\sigma^2) = \frac{1}{\sigma^2}$, but any inverse-gamma prior for σ^2 would maintain conjugacy.

3 New Bayesian Lasso

In this paper, like Park and Casella [1], a new hierarchical representation of Bayesian lasso is proposed. A new Gibbs sampler is put forward utilizing the **scale mixture of uniform (SMU)** representation of the Laplace density. Empirical studies and real data analyses show that the new algorithm inherits good mixing property and yields satisfactory performance comparable to the existing Bayesian method. All statistical analyses and illustrations were conducted in R.

3.1 Scale Mixture of Uniform Distribution

Now we want to express equation (3) as a scale mixture of Uniform distribution with mixing density being a Gamma distribution, i.e.

$$\frac{\lambda}{2\sqrt{\sigma^2}} e^{-\frac{\lambda|x|}{\sqrt{\sigma^2}}} = \int_{|x|<u\sqrt{\sigma^2}} \frac{1}{2u\sqrt{\sigma^2}} . \frac{\lambda^2}{\Gamma(2)} u^{2-1} e^{-\lambda u} du \tag{4}$$

Proof. Here we consider a Uniform distribution as $Uniform(-u\sqrt{\sigma^2}, u\sqrt{\sigma^2})$, u being scale parameter and mixing density as Gamma distribution which is given by

$$\pi(u) = \frac{\lambda^2}{\Gamma(2)} u^{2-1} e^{-\lambda u}, \quad u > 0$$

Now, we know that

$$\int_{z>\frac{|x|}{\sqrt{\sigma^2}}} \lambda e^{-\lambda z} dz = e^{-\frac{\lambda|x|}{\sqrt{\sigma^2}}}$$

Hence, the pdf of Laplace distribution with mean 0 and Variance $\frac{2\sqrt{\sigma^2}}{\lambda}$ can be written as,

$$\begin{aligned}\frac{\lambda}{2\sqrt{\sigma^2}}e^{-\frac{\lambda|x|}{\sqrt{\sigma^2}}} &= \int_{u>\frac{|x|}{\sqrt{\sigma^2}}} \lambda e^{-\lambda u}du \\ &= \int_{-u\sqrt{\sigma^2}<x<u\sqrt{\sigma^2}} \frac{1}{2u\sqrt{\sigma^2}}\frac{\lambda^2}{\Gamma(2)}u^{2-1}e^{-\lambda u}du \\ &= \int_{-u\sqrt{\sigma^2}<x<u\sqrt{\sigma^2}} U(-u\sqrt{\sigma^2}, u\sqrt{\sigma^2}) \times \pi(u)du\end{aligned}$$

Rewriting the Laplace priors as scale mixtures of uniform distributions and introducing the gamma mixing densities result in a new hierarchy. Under this new hierarchical representation, the posterior distribution of interest $p(\beta, \sigma^2|y)$ is exactly the same as the original Bayesian lasso model of Park and Casella [1] and therefore, the resulting estimates should exactly be the same 'theoretically' for both Bayesian lasso models. We establish this fact by simulation studies and real data analyses.

3.2 Model Hierarchy and Prior Distribution

The hierarchical representation is presented as follows:

$$\begin{aligned}&y|X, \beta, \sigma^2 \sim N_n(X\beta, \sigma^2 I_n) \\ &\beta|u, \sigma^2 \sim \prod_{j=1}^{p} Uniform(-\sqrt{\sigma^2}u_j, \sqrt{\sigma^2}u_j) \\ &u|\lambda \sim \prod_{j=1}^{p} Gamma(2, \lambda) \\ &\sigma^2 \sim \pi(\sigma^2)\end{aligned}$$

3.3 Full Conditional Posterior Distributions

Introduction of $u = (u_1, u_2, \ldots, u_p)$ enables us to derive the tractable full conditional posterior distributions, which are given as:

$$\beta|y, X, u, \lambda, \sigma^2 \sim N_p(\hat{\beta}_{OLS}, \sigma^2(X'X)^{-1}) \prod_{j=1}^{p} \mathbb{I}\left\{|\beta_j| < \sqrt{\sigma^2}u_j\right\} \tag{5}$$

$$u|y, X, \beta, \lambda, \sigma^2 \sim \prod_{j=1}^{p} Exponential(\lambda)\mathbb{I}\left\{u_j > \frac{|\beta_j|}{\sqrt{\sigma^2}u_j}\right\} \tag{6}$$

$$\sigma^2|y, X, \beta, u, \lambda \sim InverseGamma(\frac{n-1+p}{2}, \frac{1}{2}(y-X\beta)'(y-X\beta))\mathbb{I}\left\{\sigma^2 > Max_j\left(\frac{\beta_j^2}{u_j^2}\right)\right\} \tag{7}$$

where, $\mathbb{I}(.)$ denotes an indicator function. The proofs of the posteriors are in the appendix at the end.

3.4 Simulation Studies for New Bayesian Lasso

3.4.1 Sampling Coefficients and Latent Variables

(5), (6) and (7) lead us to an exact Gibbs sampler that starts at initial guesses for β and σ^2 and iterates the following steps:

1. Generate u_j from the left-truncated exponential distribution (6) using an inversion method which can be done as follows:
   (a) Generate u^* from an exponential distribution with rate parameter λ.
   (b) Set $u_j = u_j^* + \frac{|\beta_j|}{\sqrt{\sigma^2}}$
2. Generate β from a truncated multivariate normal distribution proportional to (5) using the following intuition that the full condition densities of $\beta_j|\beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p$ with constraints on β_j as $|\beta_j| < \sqrt{\sigma^2}u_j, j = 1, 2, \ldots, p$ follows truncated univariate normal distribution.Then apply gibbs sampling using this full conditional densities to generate β. The mentioned full conditional distribution can be found in the paper of Li and Ghosh. [5]
3. Generate σ^2 from a left-truncated Inverse Gamma distribution proportional to (7). This step can be done by utilizing the fact that the inverse of a left truncated Inverse Gamma distribution is a right truncated Gamma distribution. By generating σ^{2^*} from the right-truncated gamma distribution proportional to

$$Gamma\left(\frac{n-1+p}{2}, \frac{1}{2}(y - X\beta)'(y - X\beta)\right)\mathbb{I}\{\sigma^{2^*} < \frac{1}{\text{Max}_j\left(\frac{\beta_j^2}{\sigma_j^2}\right)}\}$$

and replacing $\sigma^2 = \frac{1}{\sigma^{2*}}$ we can mimic sampling from the targeted left-truncated Inverse Gamma distribution.

3.4.2 Sampling Hyperparameters

If λ has a $Gamma(a, b)$ prior, its conditional posterior will also be a gamma distribution, i.e.

$$\lambda|y, X, \beta, \sigma^2 \propto \lambda^{a+2p-1}\exp\{-\lambda(b + \sum_{j=1}^{p}|\beta_j|)\}$$

. Thus, we update the tuning parameter along with other parameters in the model by generating samples from

$$Gamma\left(a + 2p, \lambda(b + \sum_{j=1}^{p}|\beta_j|)\right)$$

4 Simulation Studies

In this section, we investigate the prediction accuracy of our method (NBLasso) and compare its performance with that of both original Bayesian lasso (OBLasso) and frequentist lasso (Lasso) across varied simulation scenarios.
LARS algorithm of Efron et al. is used for lasso. For Bayesian lassos, we estimate the tuning parameter λ by using a gamma prior distribution with shape parameter $a = 1$ and scale parameter $b = 0.1$, which is relatively flat and results in high posterior probability near the MLE. The Bayesian estimates are posterior means using 10,000 samples of the Gibbs sampler after burn-in of 1000 samples.
The response is centered and also the predictors are normalized so that they have zero means and unit variances before applying any model selection method. For the prediction errors, we calculate the **median of mean squared errors (MMSE)** for the simulated examples based on 100 replications.
We simulate data from the true model

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)$$

. For 3 examples we have taken different β vectors and for each case X is generated from slightly different populations. Also, we are taking different values of σ^2 and different training and test set sizes. Each simulated sample is randomly partitioned into a training set and a test set. Models are fitted on the training set and MSE's are calculated on the test set.
In all examples, detailed comparisons with both ordinary and Bayesian lasso methods are presented.

4.1 Example 1:

Here we consider a simple sparse situation. Here we set $\beta = (\mathbf{0}^T, \mathbf{2}^T, \mathbf{0}^T, \mathbf{2}^T)^T$, where $\mathbf{0}$ and $\mathbf{2}$ are vectors of length 5 with each entry equal to 0 and 2 respectively.
The design matrix X is generated from the multivariate normal distribution with mean $\mathbf{0}$, variance 1 and pairwise correlations between x_i and x_j equal to 0.5. i.e.

$$\Sigma = ((\sigma_{ij})) = \begin{cases} 1, i = j \\ 0.5, i \neq j \end{cases}$$

We experiment with four different scenarios by varying the sample size and σ^2. We simulate datasets wih $(n_T, n_P) = (200, 200)$ and $(300, 100)$ respectively, where n_T denotes the size of the training set and n_P denotes the size of the testing set. We consider two values of $\sigma : \sigma \in \{9, 15\}$. The simulation results are summarized in Table 1.

The trace plot for the 20 coefficients is given below:

Table 1: Median mean squared error (MMSE) based on 100 replications for Example 1

$\{n_T, n_p\}$	σ^2	LASSO	OBLASSO	NBLASSO
$\{200, 200\}$	81	94.793012	92.268992	90.710142
$\{200, 200\}$	225	260.566008	250.336158	239.397948
$\{300, 100\}$	81	94.674226	85.612109	84.559162
$\{300, 100\}$	225	281.149024	278.208848	267.319121

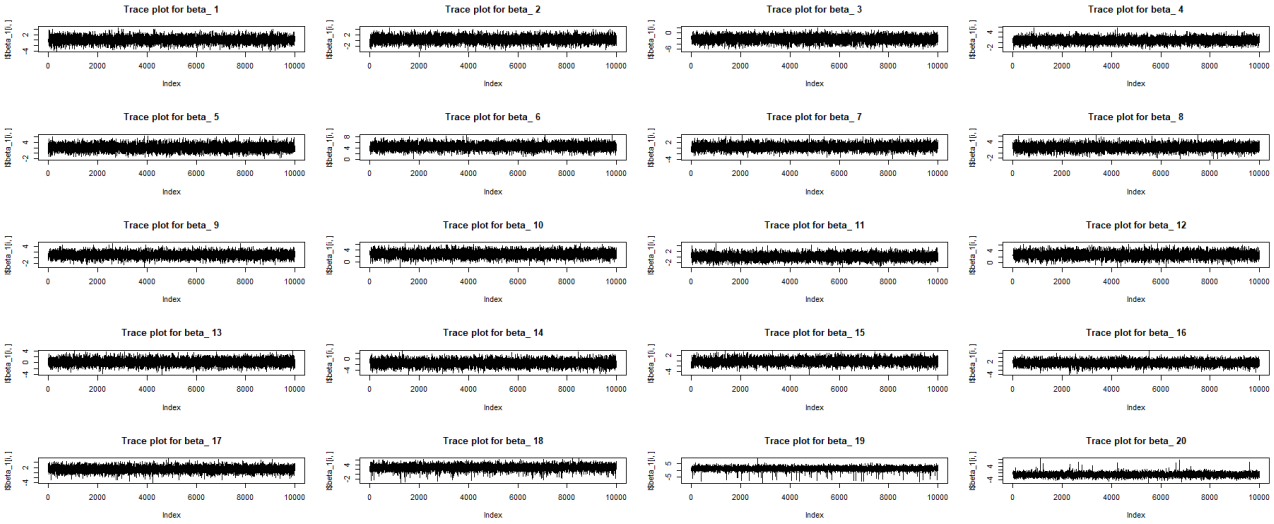

Figure 1: Trace plot for the covariates of the Example 1

From this example we can see that NBLasso outperforms both Lasso and OBLasso across all scenarios. As in this case $p = 20$, so in many cases $(X'X)^{-1}$ had a very low condition number and hence, the matrix $(X'X)^{-1}$ was non invertible. So, we rejected those cases and sampled again.

4.2 Example 2:

In this case we consider a sparse model with a strong level of correlation. We set $\beta = (3, 1.5, 2, 0, 1, 0, 0, 0)^T$. We consider the value of σ to be $\sigma = \{1, 3, 9\}$
The design matrix X is generated from the multivariate normal distribution with mean $\mathbf{0}$, variance 1 and pairwise correlations between x_i and x_j equal to $0.95, \forall i \neq j$. i.e.

$$\Sigma = ((\sigma_{ij})) = \begin{cases} 1, & i = j \\ 0.95, & i \neq j \end{cases}$$

We simulate datasets with $n_T = \{200, 300\}$ for the training set and $n_P = \{200, 100\}$ for the test set. Table 2 summarizes our experimental results for this example.

The trace plot for the 8 coefficients is given below:

Table 2: Median mean squared error (MMSE) based on 100 replications for Example 1

$\{n_T, n_p\}$	σ^2	LASSO	OBLASSO	NBLASSO
{200, 200}	1	1.093662	5.399895	1.2334945
{200, 200}	9	9.382442	13.170800	9.594651
{200, 200}	81	84.935448	88.2708881	84.942775
{300, 100}	1	1.039977	3.107135	1.085084
{300, 100}	9	9.189321	12.876020	9.149883
{300, 100}	81	86.656262	86.230068	84.415223

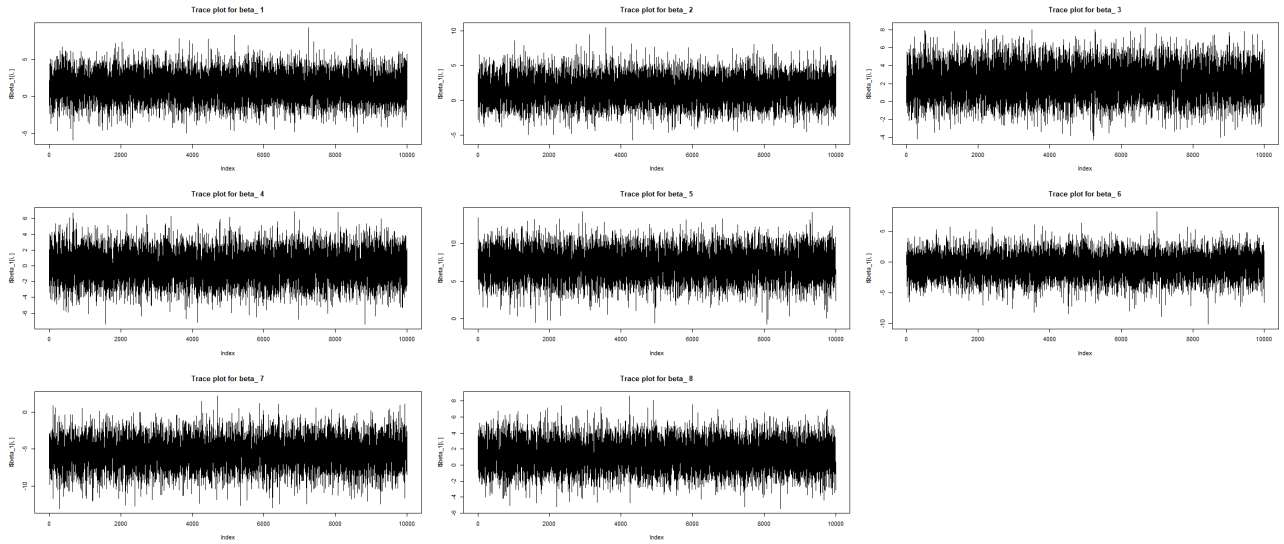

Figure 2: Trace plot for the covariates of the Example 2

We can see that NBLasso and Lasso both outperforms OBLasso in all the cases. As σ decreases and n_T increases, OBLasso is performing competitively with NBLasso and Lasso.

4.3 Example 3:

We consider another simple example from Tibshirani's original lasso paper. We set $\beta = (5, 0, 0, 0, 0, 0, 0, 0)^T$. We consider the value of σ to be $\sigma = \{1, 3\}$
The design matrix X is generated from the multivariate normal distribution with mean $\mathbf{0}$, variance 1 and pairwise correlations between x_i and x_j equal to $0.5^{|i-j|}, \forall i \neq j$. i.e.

$$\Sigma = ((\sigma_{ij})) = \begin{cases} 1, & i = j \\ 0.5^{|i-j|}, & i \neq j \end{cases}$$

We simulate datasets with $n_T = \{100, 200\}$ for the training set and $n_P = \{300, 200\}$ for the test set. Table 3 summarizes our experimental results for this example.

Table 3: Median mean squared error (MMSE) based on 100 replications for Example 1

$\{n_T, n_p\}$	σ^2	LASSO	OBLASSO	NBLASSO
{200, 200}	1	1.056328	3.330527	1.394388
{200, 200}	9	9.711258	11.470397	9.560851
{100, 300}	1	1.093054	2.696485	1.598812
{100, 300}	9	9.879831	11.321493	9.995464

The trace plot for the 8 coefficients is given below:

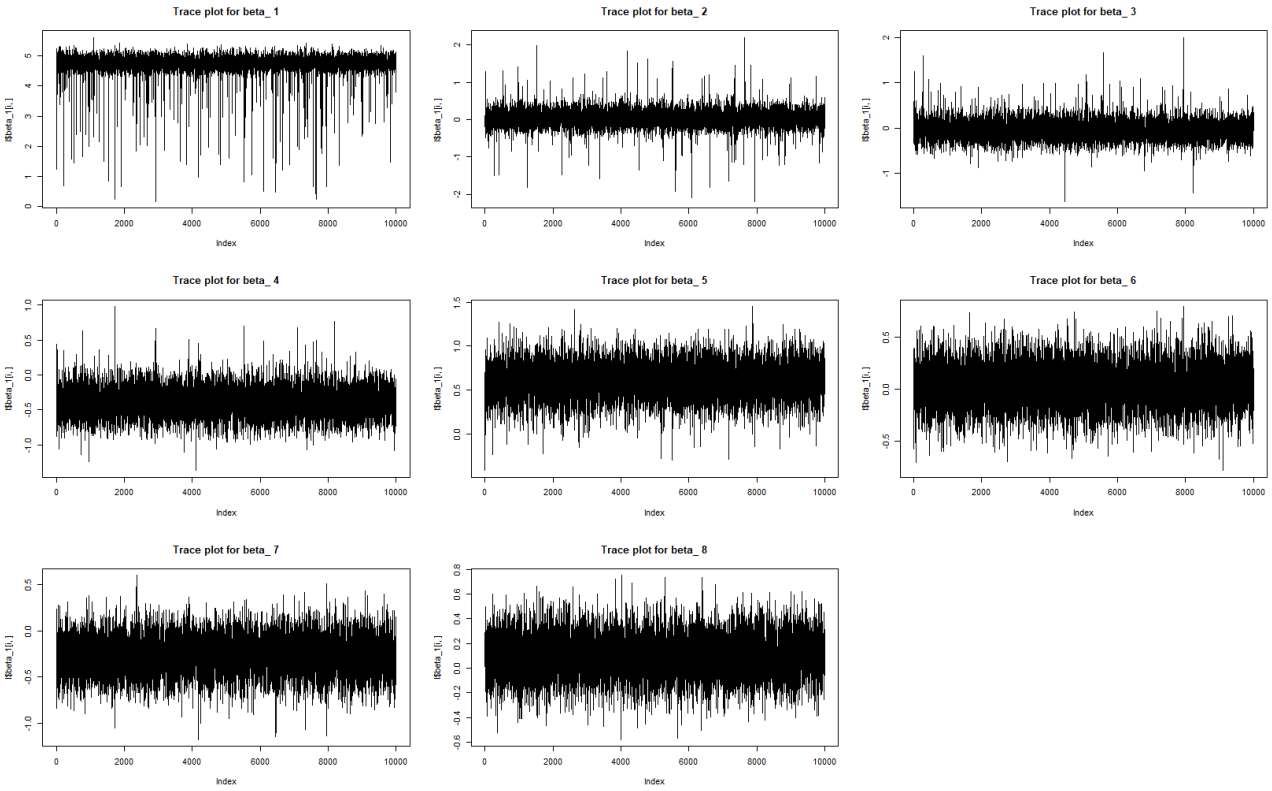

Figure 3: Trace plot for the covariates of the Example 3

We see that NBLasso always performs better than OBLasso for this example although outperformed by frequentist lasso in some situations. The reason might be contributed to the fact that not much variance is explained by introducing the priors which resulted in poor model selection performance for the Bayesian methods.

4.4 Conclusion

We have considered a variety of experimental situations to investigate the predictive and model selection performance of NBLasso. it is evident that NBLasso performs as well as, or better than OBLasso and Lasso for most of the examples. For the simple examples, NBLasso performs the best whereas for other examples, NBLasso provides comparable and slightly better performance in terms of prediction and model selection. In summary, based on our experimental results, it can be concluded that NBLasso is as effective as OBLasso and Lasso with respect to both model selection and prediction performance.

5 Real Data Analyses

Now, two real data analyses are conducted using the proposed and the existing lasso methods. One is with the benchmark Diabetes dataset, and the other is taken from a prostate cancer study. Four different methods are applied to the datasets: original Bayesian lasso (OBLasso), new Bayesian lasso (NBLasso), frequentist lasso (Lasso) and ordinary least squares (OLS).
For the Bayesian methods, posterior means are calculated as estimates based on 10,000 samples after . The tuning parameter λ is estimated as posterior mean with a gamma prior with shape parameter $a = 1$ and scale parameter $b = 0.1$ in the MCMC algorithm.
Trace plot is a good visual indicator of the mixing property. This plot is shown in Figure 5 for the Diabetes data covariates. It is highly satisfactory to observe in Figure 5 the sampler converges in 10,000 iterations to stable estimates. All these illustrate that the new Gibbs sampler is performing well.

5.1 Diabetes Data

We analyze the diabetes dataset [6] which has n = 442 observations from diabetes patients. The predictors are: age, sex, body mass index (bmi), average blood pressure (map) and six blood serum measurements (tc, ldl, hdl, tch, lth, glu). The response variable is a quantity that measures progression of diabetes one year after baseline.
The histograms of the Diabates data coefficients of covariates based on posterior samples of 10,000 iterations are illustrated in Figure 4. These histograms reveal that the conditional posterior distributions are in fact the desired stationary truncated univariate normals. The mixing of an MCMC chain shows how rapidly the MCMC chain converges to the stationary distribution.

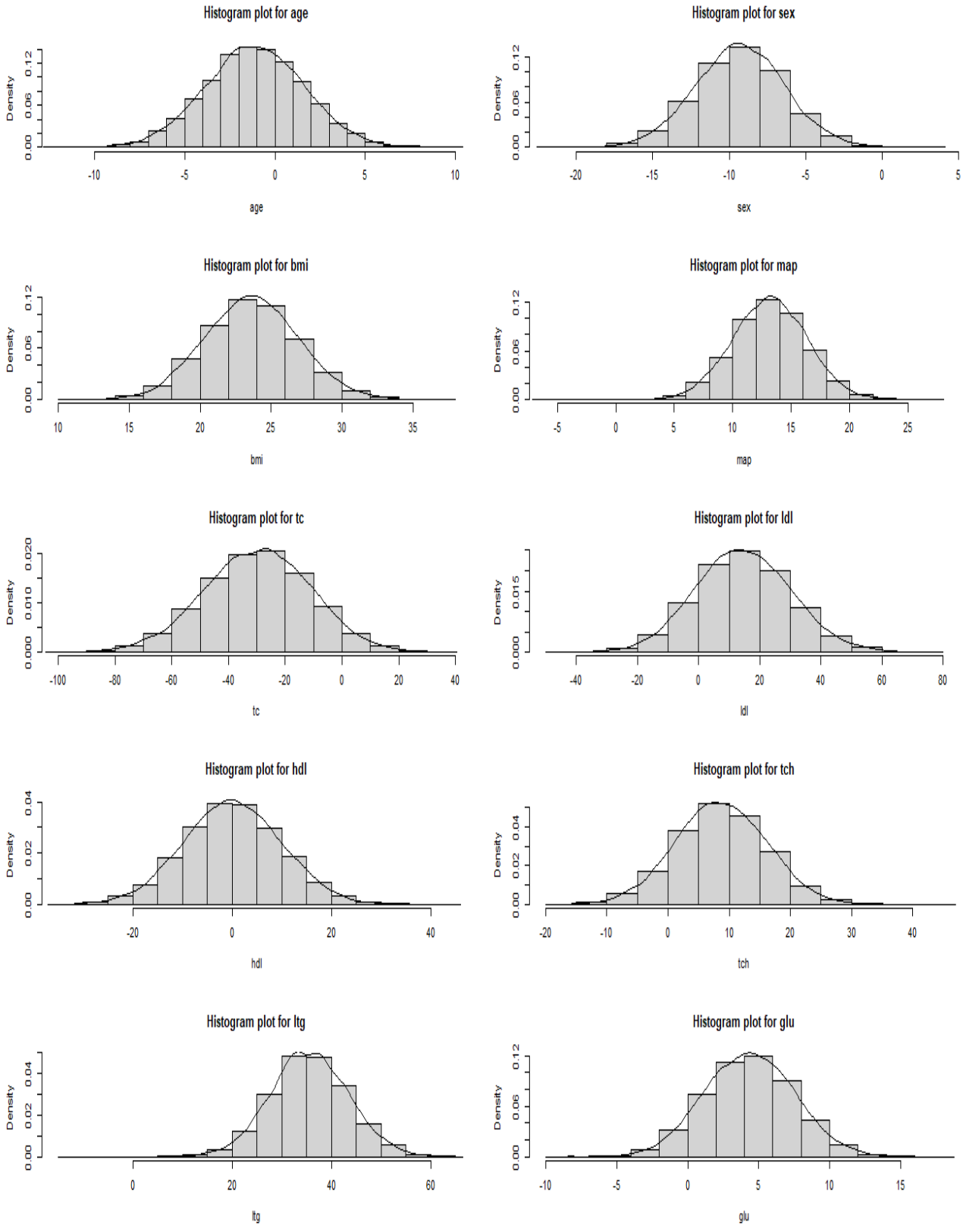

Figure 4: Histogram plot for the covariates of the diabetes data

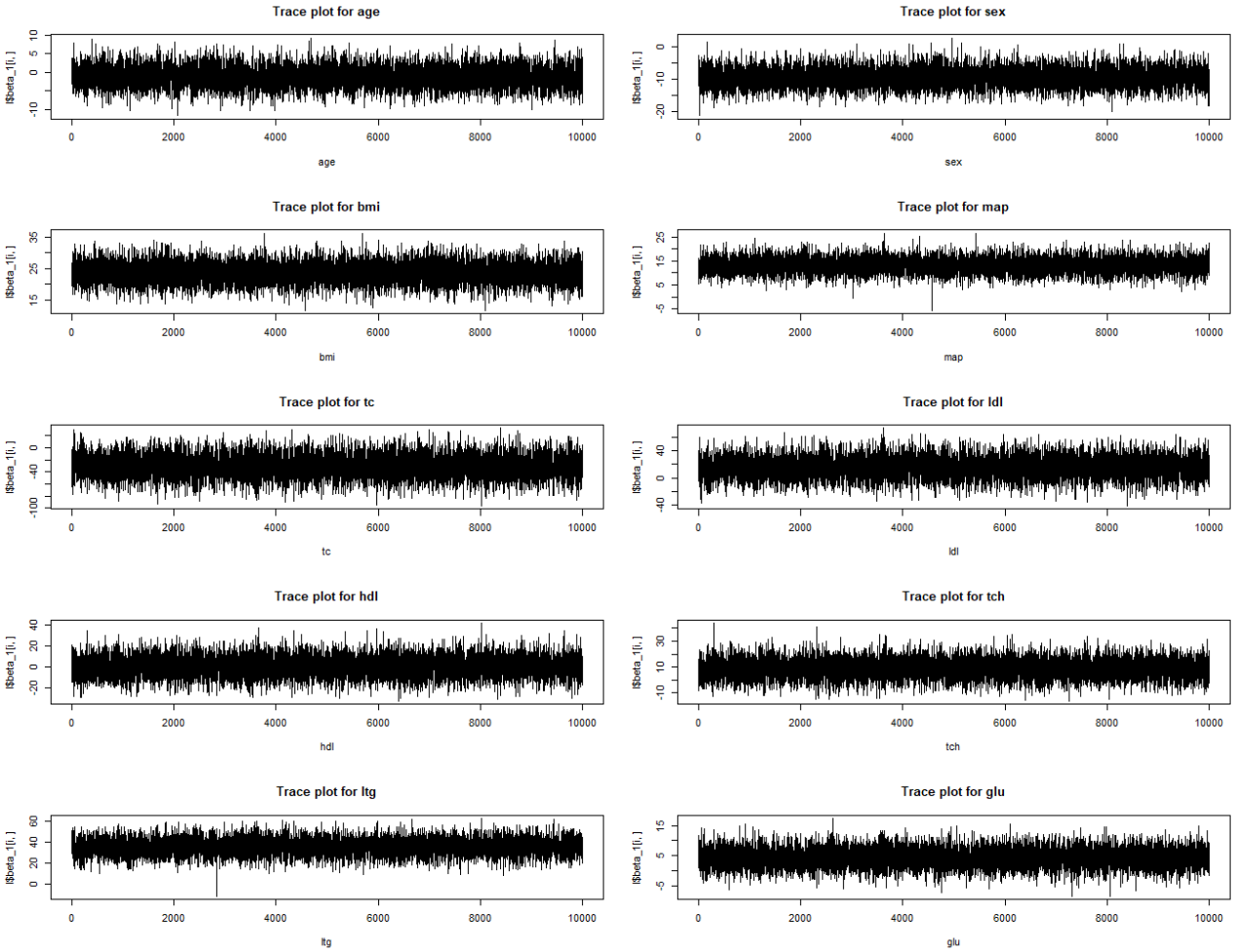

Figure 5: Trace plot for the covariates of the diabetes data

5.2 The Prostate Example

The data in this example is taken from a prostate cancer study [7]. We analyze the data by splitting it into a training set with 73 observations and a test set with 24 observations. Model fitting is done on the training data and performance is evaluated with the prediction error (MSE) on the test data.

Here,the response variable is the logarithm of prostate-specific antigen. The predictors are eight clinical measures: the logarithm of cancer volume (lcavol), the logarithm of prostate weight (lweight), age, the logarithm of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), the logarithm of capsular penetration (lcp), the Gleason score (gleason) and the percentage Gleason score 4 or 5 (pgg45).

For this dataset, the proposed method new bayesian Lasso performs impressively.

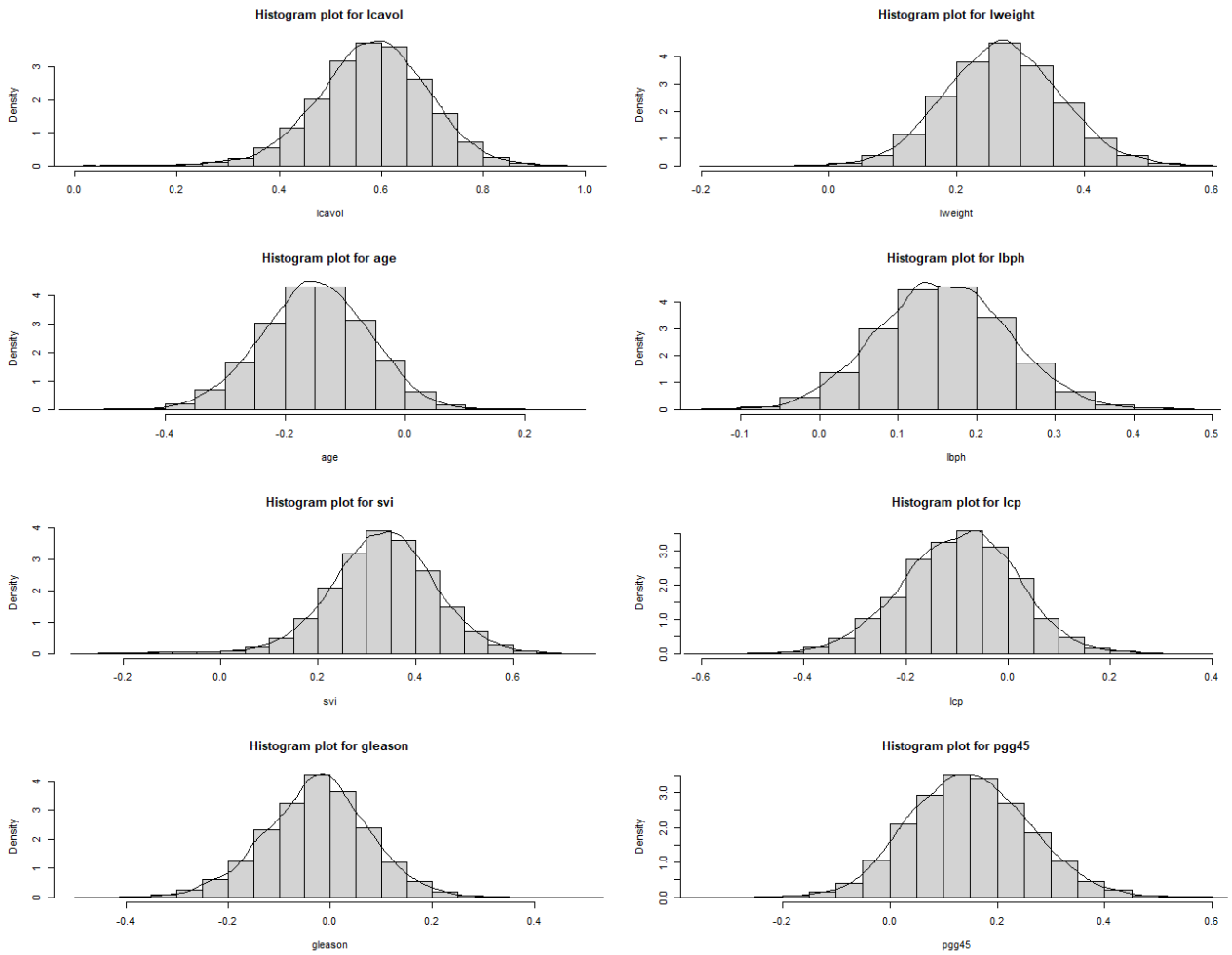

Figure 6: Histogram for the covariates of the prostate data

Table 4: Mean squared error (MSE) based on 20 replications for Prostate Data

Method	LASSO	OBLASSO	NBLASSO	OLS
MSE	0.454061	0.577838	0.41895	0.4228937

From the Table(4) we can see that NBLasso is performing the best even though all the algorithms are performing somewhat similar. The MSE is low for all the algorithms, but NBLasso in this case is performing better than OLS, OBLasso and the Lasso.
The trace plot shown in Figure 7 demonstrates that the sampler converges in 10,000 iterations to stable estimates. The histograms (Figure 6) of the Prostate data covariates based on 10,000 posterior samples reveal that the conditional posterior distributions are the desired stationary distributions viz. truncated univariate normals, which further validate our findings.

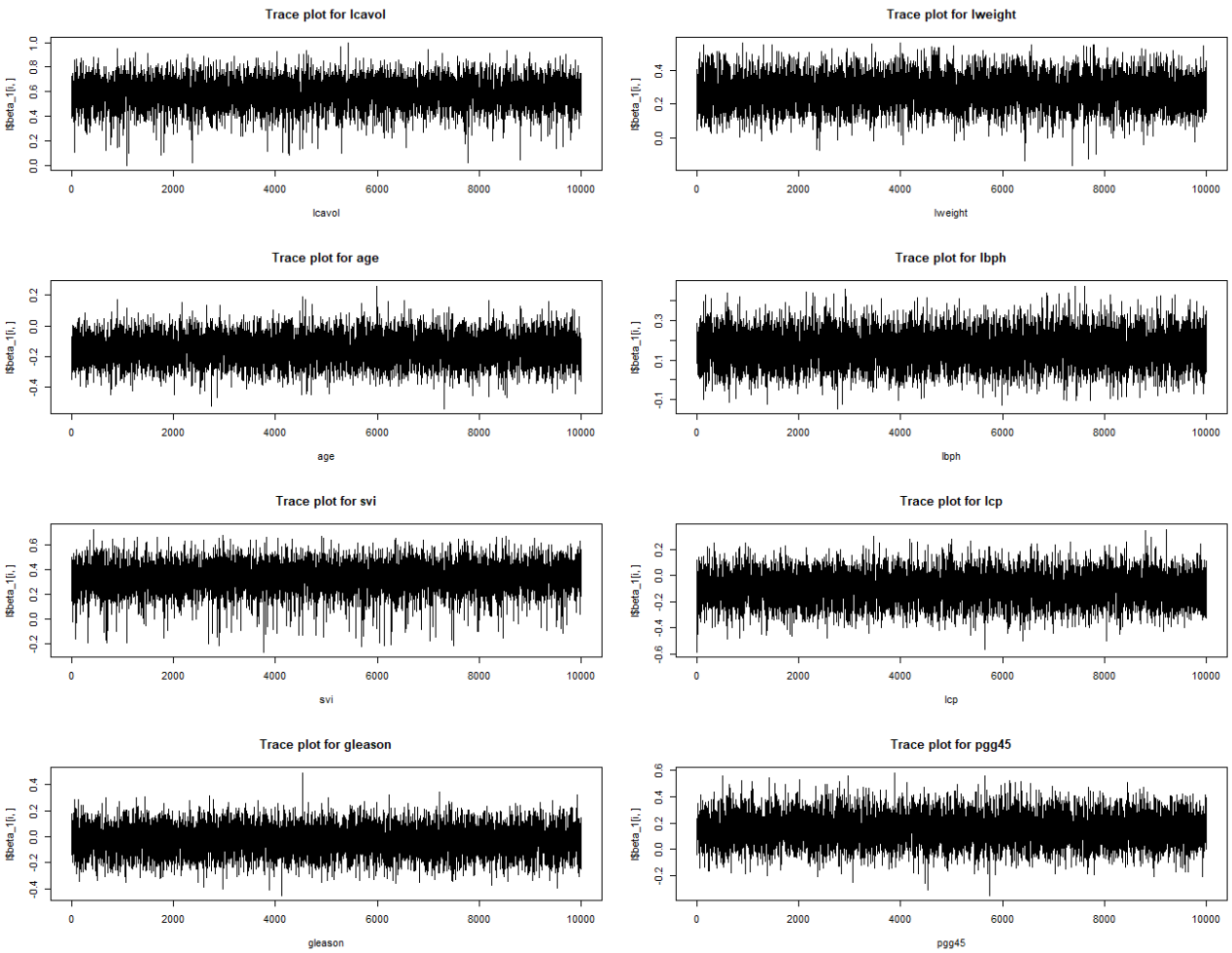

Figure 7: Trace plot for the covariates of the prostate data

6 Concluding Remarks

- NBLasso and OBLasso is performing comparably in simulation studies and also in real data analysis.
- In most of the cases NBLasso performing better than OBLasso. Theoritically both of them should perform exactly same. Practically the difference may have occured due to the use of different Gibbs Sampler.
- Full conditional distribution of β has variance $\sigma^2(X'X)^{-1}$. In many cases $(X'X)^{-1}$ had a very low condition number and hence, the matrix $(X'X)^{-1}$ was non invertible.
- Implementation of Gibbs sampler involves sampling repeatedly from Multivariate Truncated Normal distribution, which makes the MCMC process quite slow.

7 Appendix

Proof of the posterior distributions. Assuming the priors of different parameters are independent, we can express the joint Distribution of all parameters as,

$$\pi(\beta, u, \lambda, \sigma^2|y, X) \propto \pi(y|X, \beta, \sigma^2).\pi(\beta|u, \sigma^2).\pi(u|\lambda).\pi(\lambda).\pi(\sigma^2)\ d\sigma^2$$

7.1 Posterior of β

Posterior of β can be obtained by conditioning on $y, X, u, \lambda, \sigma^2$ of the joint distribution.
i.e. given the value of $y, X, u, \lambda, \sigma^2$, $\pi(u|\lambda$, $\pi(\lambda)$, $\pi(\sigma^2)$ are fixed.
Hence,

$$\begin{aligned}\pi(\beta|u, \lambda, \sigma^2, y, X) &\propto \pi(y|X, \beta, \sigma^2).\pi(\beta|\sigma^2) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right)\prod_{j=1}^{p}\mathbb{I}\left\{|\beta_j| < \sqrt{\sigma^2}u_j\right\}.\frac{1}{2\sqrt{\sigma^2}u_j}\end{aligned}$$

Now, $(y - X\beta)'(y - X\beta)$ can expanded into the following form,

$$(y - X\beta)'(y - X\beta) = (y - X\hat{\beta} + X\hat{\beta} - X\beta)'(y - X\hat{\beta} + X\hat{\beta} - X\beta)$$

, where $\hat{\beta} = \hat{\beta_{OLS}}$, $\hat{\beta_{OLS}}$ is the solution of $X'X\beta = X'y$

$$= (y - X\hat{\beta})^T(y - X\hat{\beta}) + (y - X\hat{\beta})^T X(\hat{\beta} - \beta) + (\hat{\beta} - \beta)'X^T(y - X\hat{\beta}) + (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)$$

$$= (y - X\hat{\beta})^T(y - X\hat{\beta}) + (y'X - \hat{\beta}'X'X)(\hat{\beta} - \beta) + (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)$$

Now, $X'y - X'X\hat{\beta} = 0$

$$= (y - X\hat{\beta})^T(y - X\hat{\beta}) + (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)$$

Putting this above,

$$\begin{aligned}\pi(\beta|u, \lambda, \sigma^2, y, X) \propto \exp\Bigg(-\frac{1}{2\sigma^2}(y - X\hat{\beta})^T(y - X\hat{\beta}) + (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta).& \\ \prod_{j=1}^{p}\mathbb{I}\left\{|\beta_j| < \sqrt{\sigma^2}u_j\right\}.\frac{1}{2\sqrt{\sigma^2}u_j}&\end{aligned}$$

As, y, X are given $(y - X\hat{\beta})^T(y - X\hat{\beta})$ is fixed and σ^2 and u are also fixed.
Hence,

$$\pi(\beta|u, \lambda, \sigma^2, y, X) \propto \exp\left(-\frac{1}{2\sigma^2}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)\right)\prod_{j=1}^{p}\mathbb{I}\left\{|\beta_j| < \sqrt{\sigma^2}u_j\right\}$$

Thus,

$$\beta|u,\lambda,\sigma^2,y,X \sim N_p\left(\hat{\beta}_{OLS},\sigma^2(X'X)^{-1}\right)\prod_{j=1}^{p}\mathbb{I}\left\{|\beta_j|<\sqrt{\sigma^2}u_j\right\}$$

7.2 Posterior of u

Conditioning on β,λ,σ^2 of the joint distribution i.e. $\pi(y|X,\beta,\sigma^2),\pi(\lambda),pi(\sigma^2)$ is fixed. So,

$$\begin{aligned}
\pi(u|y,X,\beta,\lambda,\sigma^2) &\propto \pi(\beta|u,\sigma^2).\pi(u|\lambda)\\
&\propto \prod_{j=1}^{p}\mathbb{I}\left\{|\beta_j|<\sqrt{\sigma^2}u_j\right\}.\frac{1}{2\sqrt{\sigma^2}u_j}\frac{\lambda^2}{2}u_je^{-\lambda u_j}\\
&\propto \prod_{j=1}^{p}\mathbb{I}\left\{u_j>\frac{|\beta_j|}{\sqrt{\sigma^2}}\right\}\frac{1}{u_j}u_je^{-\lambda u_j}\\
&\propto \prod_{j=1}^{p}\mathbb{I}\left\{u_j>\frac{|\beta_j|}{\sqrt{\sigma^2}}\right\}e^{-\lambda u_j}\\
&\propto \prod_{j=1}^{p}Exponential(\lambda)\prod_{j=1}^{p}\mathbb{I}\left\{u_j>\frac{|\beta_j|}{\sqrt{\sigma^2}}\right\}
\end{aligned}$$

Hence, $u|y,X,\beta,\lambda,\sigma^2 \sim \prod_{j=1}^{p}Exponential(\lambda)\prod_{j=1}^{p}\mathbb{I}\left\{u_j>\frac{|\beta_j|}{\sqrt{\sigma^2}}\right\}$

7.3 Posterior of σ^2

Conditioning on (y,X,β,u,λ) on joint distribution. i.e $\pi(u|\lambda)$, $\pi(\lambda_)$ are fixed. Hence,

$$\begin{aligned}
\pi(\sigma^2|y,X,\beta,u,\lambda) &\propto \pi(y|X,\beta,\sigma^2).\pi(\beta|u,\sigma^2).\pi(\sigma^2)d\sigma^2\\
&\propto \frac{1}{|\sigma^2I_n|^{\frac{1}{2}}}\exp\left(-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right).\prod_{j=1}^{p}\frac{1}{2\sqrt{\sigma^2}u_j}.\mathbb{I}\left\{|\beta_j|<\sqrt{\sigma^2}u_j\right\}.\frac{1}{\sigma^2}d\sigma^2\\
&\propto (\sigma^2)^{-\frac{n}{2}}(\sigma^2)^{-\frac{p}{2}}.\exp\left(-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right).\prod_{j=1}^{p}\mathbb{I}\left\{\sigma^2>\frac{\beta_j^2}{u_j^2}\right\}.\frac{1}{\sigma^2}2\sigma\\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n+p}{2}+1}.\exp\left(-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right).\mathbb{I}\left\{\sigma^2>Max_j\frac{\beta_j^2}{u_j^2}\right\}.\left(\frac{1}{\sigma^2}\right)^{-\frac{1}{2}}\\
&\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n-1+p}{2}+1}.\exp\left(-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right).\mathbb{I}\left\{\sigma^2>Max_j\frac{\beta_j^2}{u_j^2}\right\}
\end{aligned}$$

Therefore, $\sigma^2|y,X,\beta,u,\lambda \sim$ Inverse-Gamma $\left(\frac{n-1+p}{2},\frac{1}{2}(y-X\beta)'(y-X\beta)\right)\mathbb{I}\left\{\sigma^2>Max_j\frac{\beta_j^2}{u_j^2}\right\}$

8 Bibliography

Contribution of the group members, Implementation of theory and algorithm by Shuvam Gupta, simulation study and real data analysis by Soumik Karmakar and Saumyadip Bhowmick, preparing project report and presentation by Rajdeep Saha and Sagnik Dey. We are thankful that we have received exemplary guidance from Dr. Arnab Hazra. His supervision, guidance and blessings has helped us in fulfillment of this project.

9 References

[1] Park, T. and Casella, G. The bayesian lasso. Journal of the American Statistical Association, 103:681–686, 2008.
[2] Tibshirani, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58:267–288, 1996.
[3] Frank, I. and Friedman, J. H. A statistical view of some chemometrics regression tools (with discussion). Technometrics, 35:109–135, 1993.
[4] Mallick, H., and N. Yi. 2014. A new bayesian lasso. Statistics and Its Interface 7 (4):571, 2014
[5]Li, Y. and Ghosh, S. K. Efficient sampling methods for truncated multivariate normal and student-t distributions subject to linear inequality constraints. Technical report, North Carolina State University Department of Statistics, 2013.
[6] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. The Annals of Statistics, 32(2):407–99, 2004.
[7] Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Frieha, F., Redwine, E., and Yang, N. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate ii: Radical prostatectomy treated patients. Journal of Urology, 16:1076–1083, 1989.