Statistics and Its Interface Volume 7 (2014) 571–582

A new Bayesian lasso

Himel Mallick[*,†] and Nengjun Yi[‡]

Park and Casella (2008) provided the Bayesian lasso for linear models by assigning scale mixture of normal (SMN) priors on the parameters and independent exponential priors on their variances. In this paper, we propose an alternative Bayesian analysis of the lasso problem. A different hierarchical formulation of Bayesian lasso is introduced by utilizing the scale mixture of uniform (SMU) representation of the Laplace density. We consider a fully Bayesian treatment that leads to a new Gibbs sampler with tractable full conditional posterior distributions. Empirical results and real data analyses show that the new algorithm has good mixing property and performs comparably to the existing Bayesian method in terms of both prediction accuracy and variable selection. An ECM algorithm is provided to compute the MAP estimates of the parameters. Easy extension to general models is also briefly discussed.

AMS 2000 subject classifications: Primary 62F15.
Keywords and phrases: Lasso, Bayesian lasso, Scale mixture of uniform, MCMC, Variable selection.

1. INTRODUCTION

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

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

where $\boldsymbol{y}$ 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 $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (X'X)^{-1}X'\boldsymbol{y}$, which is obtained by minimizing the residual sum of squares (RSS) $= (\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta})$. It is well known that the OLS estimator is highly unstable in the presence of multicollinearity. Also, if $p \gg n$, it is known to produce a non-unique estimator as X is less than full rank. To improve upon the prediction accuracy of OLS, least squares regression methods with various penalties have been developed. Ridge regression [15] minimizes RSS subject to a constraint $\sum_{j=1}^{p} |\beta_j|^2 \leq t$. While ridge regression often achieves better prediction accuracy by shrinking OLS coefficients, it cannot do variable selection as it naturally keeps all the predictors. Frank and Friedman [8] introduced bridge regression which minimizes RSS subject to a constraint $\sum_{j=1}^{p} |\beta_j|^{\alpha} \leq t, \alpha \geq 0$. It includes ridge regression with $\alpha = 2$ and subset selection with $\alpha = 0$ as special cases. Among other developments, Fan and Li [6] proposed the Smoothly Clipped Absolute Deviation (SCAD) penalty and Zhang [33] introduced the Minimax Concave Penalty (MCP), both of which result in consistent, sparse and continuous estimators in linear models [6, 33].

Among penalized regression techniques, probably the most widely used method in statistical literature is the Least Absolute Shrinkage and Selection Operator (LASSO), which is a special case of bridge estimator with $\alpha = 1$. The lasso of Tibshirani [28] is obtained by minimizing

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

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 [5, 31, 9]. Given the tuning parameter(s), these algorithms are extremely fast. However, none of these algorithms provide a valid measure of standard error [18], which is arguably a major drawback of these approaches.

Very recently, much work has been done in the direction of Bayesian framework. Tibshirani [28] 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 [1] have been proposed [7, 2]. Park and Casella [24] introduced Gibbs sampling using a conditional Laplace prior specification of the form

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

and non-informative scale-invariant marginal prior on σ^2, i.e. $\pi(\sigma^2) \propto 1/\sigma^2$. Park and Casella [24] devoted serious efforts to address the important unimodality issue. They pointed out that conditioning on σ^2 is important for uni-

[*] Corresponding author.
[†] PhD candidate.
[‡] Professor.

modality and lack of unimodality might slow down the convergence of the Gibbs sampler and make the point estimates less meaningful [18]. Other methods based on Laplace prior include the Bayesian lasso via reversible-jump MCMC [3] and the Bayesian lasso regression [13]. Unlike their frequentist counterparts, Bayesian methods usually provide a valid measure of standard error based on a geometrically ergodic Markov chain [18]. Moreover, an MCMC-based Bayesian framework provides a flexible way of estimating the tuning parameter along with other parameters in the model.

In this paper, along the same line as Park and Casella [24], we propose a new hierarchical representation of Bayesian lasso. 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. The remainder of the paper is organized as follows. In Section 2, we briefly review the SMU distribution. The Gibbs sampler is presented in Section 3. Some empirical studies and real data analyses are presented in Sections 4 and 5 respectively. Easy extension to general models is provided in Section 6 and an ECM algorithm is described in Section 7. In Section 8, we provide conclusions and further discussions in this area. Some proofs and related derivations are included in an appendix.

2. SCALE MIXTURE OF UNIFORM DISTRIBUTION

Proposition. *A Laplace density can be written as a scale mixture of uniform distribution, the mixing density being a particular gamma distribution, i.e.*

$$\frac{\lambda}{2}e^{-\lambda|x|} = \int_{u>|x|} \frac{1}{2u}\frac{\lambda^2}{\Gamma(2)}u^{2-1}e^{-\lambda u}du, \quad \lambda > 0. \tag{4}$$

Proof of this result is straightforward and included in Appendix A.

SMU distribution for regression models has been used in a few occasions in literature. Walker et al. [29] used SMU distribution in normal regression models in non-Bayesian framework. Qin et al. [25] provided Gibbs sampler by using SMU in variance regression models and also to derive Gibbs sampler for autocorrelated heteroscedastic regression models [26]. Choy et al. [4] used it in stochastic volatility model by using a two-stage scale mixture representation of the student-t distribution. However, its use has been limited in penalized regression framework. We explore this fact by observing that the lasso penalty function corresponds to a scale mixture of uniform distribution, the mixing distribution being a particular gamma distribution. Following Park and Casella [24], we consider conditional Laplace priors of the form (3) on the coefficients and scale-invariant marginal prior on σ^2. 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 $p(\boldsymbol{\beta}, \sigma^2|y)$ is exactly the same as the original Bayesian lasso model of Park and Casella [24] 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. Conditioning on σ^2 ensures unimodal full posteriors in both Bayesian lasso models.

3. THE NEW BAYESIAN LASSO

3.1 Model hierarchy and prior distributions

Using (3) and (4), we formulate our hierarchical representation as follows:

$$\begin{aligned}
&\boldsymbol{y}|X, \boldsymbol{\beta}, \sigma^2 \sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n),\\
&\boldsymbol{\beta}|\boldsymbol{u}, \sigma^2 \sim \prod_{j=1}^{p} \text{Uniform}(-\sqrt{\sigma^2}u_j, \sqrt{\sigma^2}u_j), \qquad (5)\\
&\boldsymbol{u}|\lambda \sim \prod_{j=1}^{p} \text{Gamma}(2, \lambda),\\
&\sigma^2 \sim \pi(\sigma^2).
\end{aligned}$$

3.2 Full conditional posterior distributions

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

$$\begin{aligned}
&\boldsymbol{\beta}|\boldsymbol{y}, X, \boldsymbol{u}, \lambda, \sigma^2\\
&\sim N_p(\hat{\boldsymbol{\beta}}_{\text{OLS}}, \sigma^2(X'X)^{-1}) \prod_{j=1}^{p} I\{|\beta_j| < \sqrt{\sigma^2}u_j\}, \qquad (6)
\end{aligned}$$

$$\boldsymbol{u}|\boldsymbol{y}, X, \boldsymbol{\beta}, \lambda, \sigma^2 \sim \prod_{j=1}^{p} \text{Exponential}(\lambda) I\{u_j > \frac{|\beta_j|}{\sqrt{\sigma^2}}\}, \tag{7}$$

$$\begin{aligned}
&\sigma^2|\boldsymbol{y}, X, \boldsymbol{\beta}, \boldsymbol{u}, \lambda \sim \text{Inverse Gamma}(\frac{n-1+p}{2},\\
&\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta})) I\{\sigma^2 > \text{Max}_j(\frac{{\beta_j}^2}{{u_j}^2})\}, \qquad (8)
\end{aligned}$$

where, $I(.)$ denotes an indicator function. The derivations are included in Appendix A.

3.3 MCMC sampling for the new Bayesian lasso

3.3.1 Sampling coefficients and latent variables

(6), (7) and (8) lead us to an exact Gibbs sampler that starts at initial guesses for $\boldsymbol{\beta}$ and σ^2 and iterates the following steps:

1. Generate u_j from the left-truncated exponential distribution (7) using an inversion method which can be done as follows:
   a) Generate u_j^* from an exponential distribution with rate parameter λ.
   b) Set $u_j = u_j^* + \frac{|\beta_j|}{\sqrt{\sigma^2}}$.
2. Generate $\boldsymbol{\beta}$ from a truncated multivariate normal distribution proportional to (6). This step can be done by implementing an efficient sampling technique developed by Li and Ghosh [21].
3. Generate σ^2 from a left-truncated Inverse Gamma distribution proportional to (8). 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 ${\sigma^2}^*$ from the right-truncated gamma distribution proportional to

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

   and replacing $\sigma^2 = \frac{1}{{\sigma^2}^*}$ we can mimic sampling from the targeted left-truncated Inverse Gamma distribution (8).

3.3.2 Sampling hyperparameters

To update the tuning parameter λ, we work directly with the Laplace density marginalizing out the latent variables u_j's. From (5), we observe that the posterior for λ given $\boldsymbol{\beta}$ is conditionally independent of $\boldsymbol{y}$ and takes the form

$$\pi(\lambda|\boldsymbol{\beta}) \propto \lambda^{2p} \exp\{-\lambda \sum_{j=1}^{p} |\beta_j|\}\pi(\lambda).$$

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

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

Thus, we update the tuning parameter along with other parameters in the model by generating samples from Gamma$(a+2p, b+\sum_{j=1}^{p}|\beta_j|)$.

4. SIMULATION STUDIES

4.1 Prediction

In this section, we investigate the prediction accuracy of our method (NBLasso) and compare its performance with both original Bayesian lasso (OBLasso) and frequentist lasso (Lasso) across varied simulation scenarios. LARS algorithm of Efron et al. [5] is used for lasso, in which 10-fold cross-validation is used to select the tuning parameter, as implemented in the R package **lars**. 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 [18]. The Bayesian estimates are posterior means using 10,000 samples of the Gibbs sampler after burn-in. To decide on the burn-in number, we make use of the potential scale reduction factor [11]. Once $\hat{R} < 1.1$ for all parameters of interest, we continue to draw 10,000 iterations to obtain samples from the joint posterior distribution. The response is centered and the predictors are normalized to 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

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \text{N}(\boldsymbol{0}, \sigma^2 I).$$

Each simulated sample is partitioned into a training set and a testing set. Models are fitted on the training set and MSE's are calculated on the testing set. In all examples, detailed comparisons with both ordinary and Bayesian lasso methods are presented.

Example 1 (Simple Example - I): Here we consider a simple sparse situation which was also used by Tibshirani [28] in his original lasso paper. Here we set $\boldsymbol{\beta} = (\boldsymbol{0^T}, \boldsymbol{2^T}, \boldsymbol{0^T}, \boldsymbol{2^T})^T$, where $\boldsymbol{0}$ and $\boldsymbol{2}$ are vectors of length 10 with each entry equal to 0 and 2 respectively. The design matrix X is generated from the multivariate normal distribution with mean 0, variance 1 and pairwise correlations between x_i and x_j equal to 0.5. We experiment with four different scenarios by varying the sample size and σ^2. We simulate datasets wih $\{n_T, n_P\} = \{100, 400\}$ and $\{200{,}200\}$ 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 $\sigma : \sigma \in \{9, 25\}$. The simulation results are summarized in Table 1 which clearly suggest that NBLasso outperforms both Lasso and OBLasso across all scenarios of this example.

Example 2 (Difficult Example - I): In this example, we consider a complicated model which exhibits a substantial amount of data collinearity. A similar example was presented in the elastic net paper by Zou and Hastie [36]. Here we simulate Z_1, Z_2 and Z_3 independently from N(0,1). Then, we let $x_i = Z_1 + \epsilon_i,\ i = 1, \ldots, 5$; $x_i = Z_2 + \epsilon_i,\ i = 6, \ldots, 10$; $x_i = Z_3 + \epsilon_i,\ i = 11, \ldots, 15$ and $x_i \sim N(0, 1),\ i = 16, \ldots, 30$, where $\epsilon_i \sim N(0, 0.01),\ i = 1(1)15$. We set $\boldsymbol{\beta} = (\boldsymbol{3^T}, \boldsymbol{3^T}, \boldsymbol{3^T}, \boldsymbol{0^T})^T$ where $\boldsymbol{3}$ and $\boldsymbol{0}$ are vectors of length 5 and 15 with each entry equal to 3 and 0 respectively. We experiment with the same values of σ^2 and $\{n_T, n_P\}$ as in Example 1. The simulation results are presented in Table 2. It

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}	225	279.72	249.79	244.4
{200, 200}	81	101.2	93.89	92.93
{100, 400}	225	354.19	268.74	259.85
{100, 400}	81	131.62	104.17	102.32

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

$\{n_T, n_P\}$	σ^2	Lasso	OBLasso	NBLasso
{200, 200}	225	242.97	240.85	240.35
{200, 200}	81	90.01	88.98	88.92
{100, 400}	225	250.35	254.71	253.84
{100, 400}	81	93.36	95.34	94.49

can be observed that NBLasso is competitive with OBLasso in terms of prediction accuracy in all the scenarios presented in this example.

Example 3 (High Correlation Example - I): Here we consider a sparse model with a strong level of correlation. We set $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\sigma^2 = \{1, 9\}$. The design matrix X is generated from the multivariate normal distribution with mean 0, variance 1 and pairwise correlations between x_i and x_j equal to 0.95 $\forall\, i \neq j$. We simulate datasets with $n_T = \{20, 50, 100, 200\}$ for the training set and $n_P = 200$ for the testing set. Table 3 summarizes our experimental results for this example. We can see that both Bayesian lassos yield similar performance and usually outperform their frequentist counterpart. As σ decreases and n_T increases, all the three methods yield equivalent performance.

Example 4 (Small n Large p Example): Here we consider a case where $p \geq n$. We let $\boldsymbol{\beta}_{1:q} = (5, \ldots, 5)^T$, $\boldsymbol{\beta}_{q+1:p} = \mathbf{0}$, $p = 20$, $q = 10$, $\sigma = \{1, 3\}$. The design matrix X is generated from the multivariate normal distribution with mean 0, variance 1 and pairwise correlations between x_i and x_j equal to 0.95 $\forall\, i \neq j$. We simulate datasets with $n_T = \{10, 20\}$ for the training set and $n_P = 200$ for the testing set. It is evident from the results presented in Table 4 that the proposed method performs better than both OBLasso and Lasso in most of the situations. In one situation, OBLasso performs slightly better. Overall, Bayesian lassos significantly outperform frequentist lasso in terms of prediction accuracy.

4.2 Variable selection

In this section, we investigate the model selection performance of our method (NBLasso) and compare its performance with both original Bayesian lasso (OBLasso) and frequentist lasso (Lasso). Note that, the lasso was originally developed as a variable selection tool. However, in Bayesian framework, this attractive property vanishes as Bayesian lassos usually do not set any coefficient to zero. One way to tackle this problem is to use the credible interval criterion as suggested by Park and Casella [24] in their seminal paper. However, it brings the problem of threshold selection. Moreover, credible intervals are not uniquely defined. Therefore, we will seek out an alternative strategy here. In Bayesian paradigm, it is a standard procedure to fully explore the posterior distribution and estimate λ by posterior median or posterior mean. Therefore, we can plug in the posterior estimate of λ in (2) and solve (2) to carry out variable selection. This strategy was recently used by Leng et al. [19]. For the optimization problem (2), we make use of the LARS algorithm of [5].

For each simulated dataset, we apply three different lasso methods viz. NBLasso, OBLasso and Lasso and record the frequency of correctly-fitted models over 100 replications. For the Bayesian lassos, we assign a Gamma $(1, 0.1)$ prior on λ to estimate the tuning parameter. Based on the posterior samples (10,000 MCMC samples after burn-in), we calculate two posterior quantities of interest, viz. posterior mean and posterior median. Then we plug-in either posterior mean or posterior median estimate of λ in (2) and solve (2) to get the estimates of the coefficients. We refer to these different strategies as NBLasso-Mean, OBLasso-Mean, NBLasso-Median and OBLasso-Median, where NBLasso-Mean refers to NBLasso coupled with λ estimated by posterior mean, OBLasso-Median refers to OBLasso coupled with λ estimated by posterior median and so on. LARS algorithm is used for frequentist lasso, in which 10-fold cross-validation is used to select the tuning parameter. The response is centered and the predictors are normalized to have zero

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

n_T	σ	Lasso	OBLasso	NBLasso
20	3	11.61	10.4	10.4
50	3	10.03	9.8	9.8
100	3	9.6	9.55	9.54
200	3	9.4	9.29	9.29
20	1	1.79	1.6	1.6
50	1	1.28	1.27	1.27
100	1	1.19	1.18	1.18
200	1	1.1	1.1	1.1

Table 4. Median mean squared error (MMSE) based on 100 replications for Example 4

n_T	σ	Lasso	OBLasso	NBLasso
10	3	91.39	77.0	77.4
10	1	81.4	69.47	68.81
20	3	86.04	41.66	41.59
20	1	46.94	30.98	30.71

Table 5. Frequency of correctly-fitted models over 100 replications for Example 5

n_T	Lasso	OBLasso - Mean	NBLasso - Mean	OBLasso - Median	NBLasso - Median
20	15	13	22	11	17
50	11	14	20	12	15
100	16	21	25	20	25
200	11	16	17	16	16

Table 6. Frequency of correctly-fitted models over 100 replications for Example 6

n_T	Lasso	OBLasso - Mean	NBLasso - Mean	OBLasso - Median	NBLasso - Median
20	32	10	28	7	17
50	34	13	26	12	22
100	27	20	31	19	27
200	20	14	21	12	18

Table 7. Frequency of correctly-fitted models over 100 replications for Example 7

n_T	σ	Lasso	OBLasso - Mean	NBLasso - Mean	OBLasso - Median	NBLasso - Median
120	5	0	6	6	5	6
300	3	0	10	8	12	10
300	1	0	9	12	12	12

means and unit variances before applying any model selection method.

Example 5 (Simple Example - II): This example was used in the original lasso paper to systematically compare the predictive performance of lasso and ridge regression. Here we set $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\sigma^2 = 9$. The design matrix X is generated from the multivariate normal distribution with mean 0, variance 1 and pairwise correlations between x_i and x_j equal to $0.5^{|i-j|}$ $\forall\, i \neq j$. We simulate datasets with $n_T = \{20, 50, 100, 200\}$ for the training set and $n_P = 200$ for the testing set. The simulation results are summarized in Table 5. From Table 5 it can be seen that NBLasso performs reasonably well outperforming both frequentist and original Bayesian lasso.

Example 6 (Simple Example - III): We consider another simple example from Tibshirani's original lasso paper. We set $\boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)^T$ and $\sigma^2 = 9$. The design matrix X is generated from the multivariate normal distribution with mean 0, variance 1 and pairwise correlations between x_i and x_j equal to $0.5^{|i-j|}$ $\forall\, i \neq j$. We simulate datasets with $n_T = \{20, 50, 100, 200\}$ for the training set and $n_P = 200$ for the testing set. The simulation results are presented in Table 6. We see that NBLasso always performs better than OBLasso for this example although outperformed by frequentist lasso in many 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.

Example 7 (Difficult Example - II): Here we consider a situation where lasso does not give consistent model selection. This example is taken from the adaptive lasso paper by Zou [35]. Here we set $\boldsymbol{\beta} = (5.6, 5.6, 5.6, 0)^T$ and the correlation matrix of X is such that $\text{Cor}(x_i, x_j) = -0.39$, $i < j < 4$ and $\text{Cor}(x_i, x_4) = 0.23$, $i < 4$. Zou [35] showed that for this example lasso is inconsistent regardless of the sample size. The experimental results are summarized in Table 7. None of the methods perform well for this example. Both Bayesian lassos yield similar performance and behave better than frequentist lasso in selecting the correct model.

Example 8 (High Correlation Example - II): Here we consider a simple model with a strong level of correlation. We set $\boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)^T$ and $\sigma^2 = 9$. The design matrix X is generated from the multivariate normal distribution with mean 0, variance 1 and pairwise correlations between x_i and x_j equal to 0.95 $\forall\, i \neq j$. We simulate datasets with $n_T = \{20, 50, 100, 200\}$ for the training set and $n_P = 200$ for the testing set. Table 8 summarizes our experimental results for this example. It can be seen from the table that both Bayesian lassos yield similar performance and outperform frequentist lasso.

4.3 Some comments

We have considered a variety of experimental situations to investigate the predictive and model selection performance of NBLasso. Most of the simulation examples considered here have previously appeared in other lasso and related papers. From our extensive simulation experiments it is evident that NBLasso performs as well as, or better than OBLasso 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. Note

Table 8. Frequency of correctly-fitted models over 100 replications for Example 8

n_T	Lasso	OBLasso - Mean	NBLasso - Mean	OBLasso - Median	NBLasso - Median
20	8	19	22	17	19
50	11	18	20	18	19
100	8	17	17	17	17
200	5	16	16	16	16

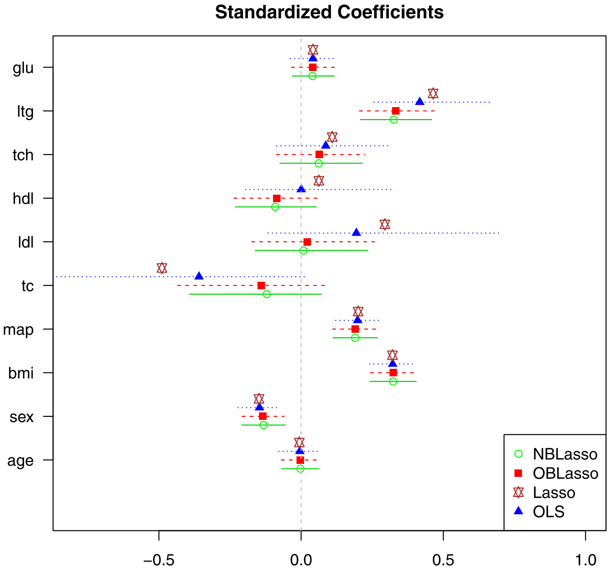

Figure 1. Posterior mean Bayesian lasso estimates (computed over a grid of λ values, using 10,000 samples after burn-in) and corresponding 95% credible intervals (equal-tailed) of Diabetes data ($n = 442$) covariates. The hyperprior parameters were chosen as $a = 1,\ b = 0.1$. OLS estimates with corresponding 95% confidence intervals are also reported. For the lasso estimates, the tuning parameter was chosen by 10-fold CV of the LARS algorithm.

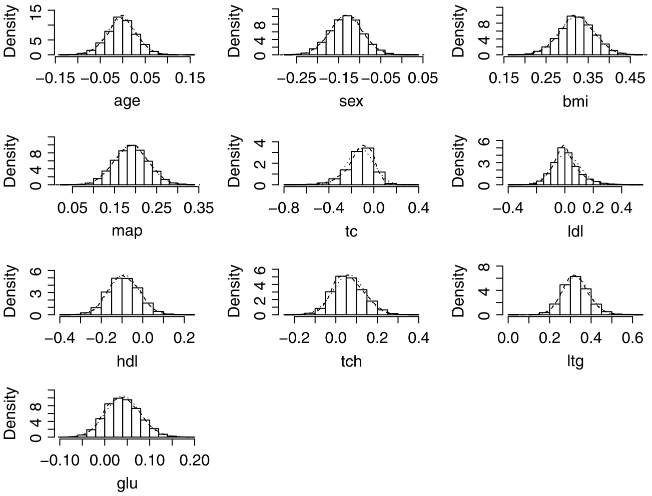

Figure 2. Histograms based on posterior samples of Diabetes data covariates.

that, superiority of Bayesian lasso and related methods is already well-established in literature [18, 20, 19]. We have found a similar conclusion in this paper. For all the simulated examples, convergence of the corresponding MCMC chain was assessed by trace plots of the generated samples and calculating the Gelman-Rubin scale reduction factor [11] using the **coda** package in R. For $n \leq p$ situation, none of the methods performed well in model selection. Therefore, those results are omitted. In summary, based on our experimental results, it can be concluded that NBLasso is as effective as OBLasso with respect to both model selection and prediction performance.

5. REAL DATA ANALYSES

In this section, two real data analyses are conducted using the proposed and the existing lasso methods. 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 burn-in. To decide on the burn-in number, we make use of the potential scale reduction factor [11]. Once $\hat{R} < 1.1$ for all parameters of interest, we continue to draw 10,000 iterations to obtain samples from the joint posterior distribution. 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. The convergence of our MCMC is checked by trace plots of the generated samples and calculating the Gelman-Rubin scale reduction factor [11] using the **coda** package in R. For the frequenstist lasso, 10-fold cross-validation (CV) is used to select the shrinkage parameter. The response is centered and the predictors are normalized to have zero means and unit variances before applying any model selection method.

5.1 The diabetes example

We analyze the benchmark diabetes dataset [5] which contains $n = 442$ measurements from diabetes patients. Each measurement has ten baseline predictors: 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. Figure 1 gives

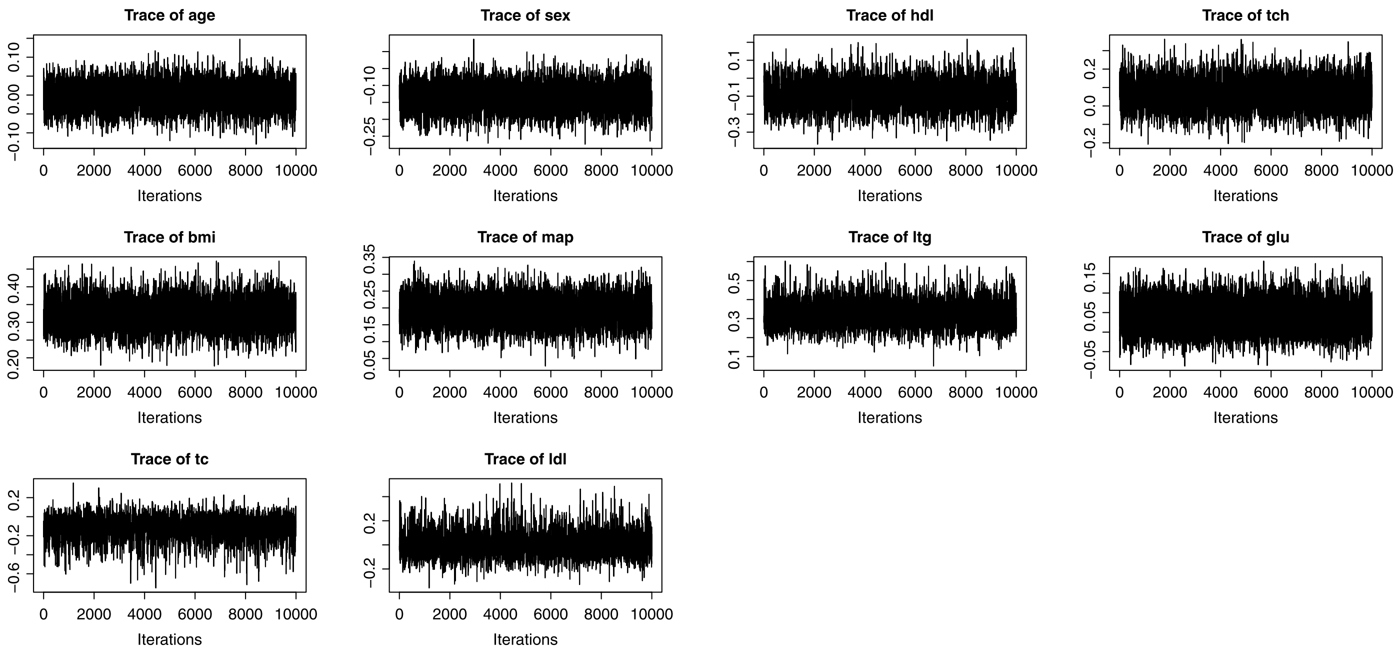

Figure 3. Trace plots of Diabetes data covariates.

the 95% equal-tailed credible intervals along with posterior mean Bayesian lasso estimates of Diabetes data covariates along with frequentist lasso estimates, overlaid with OLS estimates with corresponding 95% confidence intervals. The estimated λ's are 5.1 (2.5, 9.1) and 4.0 (2.2, 6.4) for NBLasso and OBLasso respectively. Figure 1 shows that two Bayesian lassos behave similarly for all the coefficients of this dataset. The 95% credible intervals are also similar. Any observed differences in parameter estimates can be contributed (up to Monte Carlo error) to the properties of the different Gibbs samplers used to obtain samples from the corresponding posterior distributions. The histograms of the Diabates data covariates based on posterior samples of 10,000 iterations are illustrated in Figure 2. 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 [10]. Trace plot is a good visual indicator of the mixing property. This plot is shown in Figure 3 for the Diabetes data covariates. It is highly satisfactory to observe that for this benchmark dataset the samples traverse the posterior space very fast. We also conduct the Geweke's convergence diagnosis test and all the individual chains pass the tests. All these illustrate that the new Gibbs sampler has good mixing property.

5.2 The prostate example

The data in this example is taken from a prostate cancer study [27]. Following Zou and Hastie [36], we analyze the data by dividing it into a training set with 67 observations and a test set with 30 observations. Model fitting is carried out on the training data and performance is evaluated with the prediction error (MSE) on the test data. The response of interest 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). Figure 4 shows the 95% equal-tailed credible intervals for regression parameters of Prostate data based on the posterior mean Bayesian lasso estimates with point estimates of frequentist lasso and OLS estimates with corresponding 95% confidence intervals. The estimated λ's are 3.5 (1.6, 7.3) and 3.1 (1.5, 5.3) for NBLasso and OBLasso respectively. The predictors in this dataset are known to be more correlated than those in the Diabetes data. Even for this dataset, the proposed method performs impressively. Figure 4 reveals that two Bayesian lasso estimates are strikingly similar and the corresponding 95% credible intervals are almost identical for this dataset. Also, it is interesting to note that all the estimates are inside the credible intervals which indicates that the resulting conclusion will be similar regardless of which method is used. Moreover, the new method outperforms both OBLasso and Lasso in terms of prediction accuracy (Table 9).

The trace plot shown in Figure 5 demonstrates that the sampler jumps from one remote region of the posterior space to another in relatively few steps. Here also, we conduct

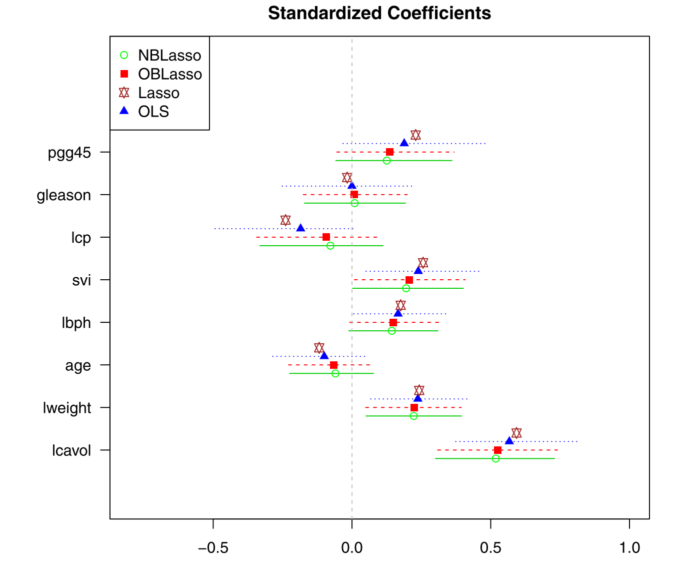

Figure 4. Posterior mean Bayesian lasso estimates (computed over a grid of λ values, using 10,000 samples after burn-in) and corresponding 95% credible intervals (equal-tailed) of Prostate data ($n = 67$) covariates. The hyperprior parameters were chosen as $a = 1,\ b = 0.1$. OLS estimates with corresponding 95% confidence intervals are also reported. For the lasso estimates, the tuning parameter was chosen by 10-fold CV of the LARS algorithm.

Table 9. Prostate Cancer Data Analysis - Mean squared prediction errors based on 30 observations of the test set for four methods: New Bayesian Lasso (NBLasso), Original Bayesian Lasso (OBLasso), Lasso and OLS

Method	NBLasso	OBLasso	Lasso	OLS
MSE	0.4696	0.4729	0.4856	0.5212

Geweke's convergence diagnosis test and all the individual chains pass the test which establishes good mixing property of the proposed Gibbs sampler. The histograms of the Prostate data covariates based on 10,000 posterior samples (Figure 6) reveal that the conditional posterior distributions are the desired stationary distributions viz. truncated univariate normals, which further validate our findings.

6. EXTENSIONS

6.1 MCMC for general models

In this section, we briefly discuss how NBLasso can be extended to several other models (e.g. GLM, Cox's model, etc.) beyond the linear regression. Our extension is based on the least squares approximation (LSA) by Wang and Leng [30]. Recently, Leng et al. [19] used this approximation for original Bayesian lasso. Therefore, here we only describe the algorithm for NBLasso. The algorithm for OBLasso can be found in Leng et al. [19]. Let us denote by $L(\boldsymbol{\beta})$ the negative log-likelihood. Following Wang and Leng [30], $L(\boldsymbol{\beta})$ can be approximated by LSA as follows

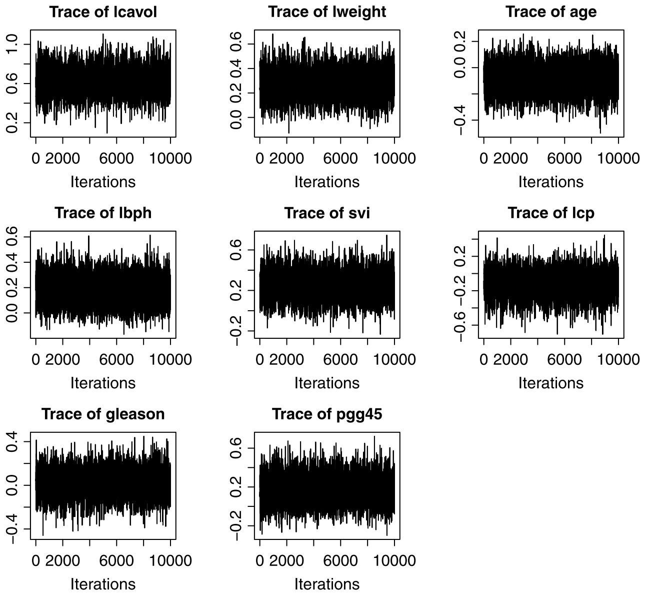

Figure 5. Trace plots of Prostate data covariates.

$$L(\boldsymbol{\beta}) \approx \frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\hat{\Sigma}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}),$$

where $\tilde{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ and $\hat{\Sigma}^{-1} = \delta^2 L(\boldsymbol{\beta})/\delta\boldsymbol{\beta}^2$. Therefore, for a general model, the conditional distribution of $\boldsymbol{y}$ is given by

$$\boldsymbol{y}|\boldsymbol{\beta} \sim \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\hat{\Sigma}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\right\}.$$

Thus, we can easily extend our method to several other models by approximating the corresponding likelihood by normal likelihood. Combining the SMU representation of the Laplace density and the LSA approximation of the general likelihood, the hierarchical presentation of NBLasso for general models can be written as

$$\begin{aligned}
&\boldsymbol{y}|\boldsymbol{\beta} \sim \exp\left\{-\tfrac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\hat{\Sigma}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\right\},\\
&\boldsymbol{\beta}|\boldsymbol{u} \sim \prod_{j=1}^{p} \text{Uniform}(-u_j, u_j),\\
&\boldsymbol{u}|\lambda \sim \prod_{j=1}^{p} \text{Gamma}(2, \lambda),\\
&\lambda \sim \text{Gamma}(a, b).
\end{aligned} \tag{9}$$

The full conditional distributions are given as

$$\boldsymbol{\beta}|\boldsymbol{y}, X, \boldsymbol{u}, \lambda \sim N_p(\tilde{\boldsymbol{\beta}}, \hat{\Sigma}) \prod_{j=1}^{p} I\{|\beta_j| < u_j\}, \tag{10}$$

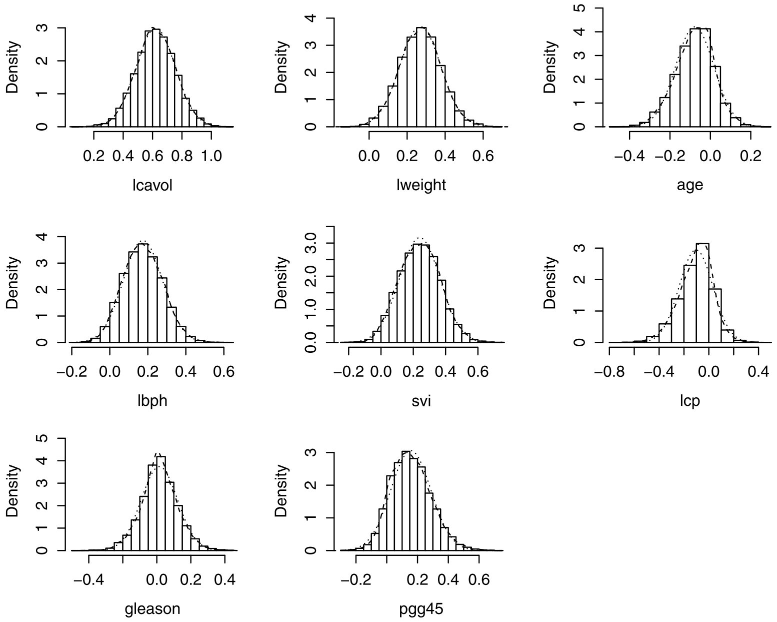

Figure 6. Histograms based on posterior samples of Prostate data covariates.

$$(11) \qquad \boldsymbol{u}|\boldsymbol{y}, X, \boldsymbol{\beta}, \lambda \sim \prod_{j=1}^{p} \text{Exponential}(\lambda) I\{u_j > |\beta_j|\},$$

$$(12) \qquad \lambda|\boldsymbol{y}, X, \boldsymbol{\beta} \sim \lambda^{a+2p-1} \exp\{-\lambda(b + \sum_{j=1}^{p} |\beta_j|)\}.$$

As before, an efficient Gibbs sampler can be easily carried out based on these full conditionals.

6.2 Simulation examples for general models

We now assess the performance of NBLasso in general models by means of two examples. For brevity, we only report the performance of various methods in terms of prediction accuracy. Three different lasso methods are applied to the simulated datasets. For the frequentist lasso, we use the R package **glmnet** which implements the coordinate descent algorithm of Friedman et al. [9], in which 10-fold cross-validation is used to select the tuning parameter. We normalize the predictors to 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. The design matrix X is generated from the multivariate normal distribution with mean 0, variance 1 and pairwise correlations between x_i and x_j equal to $0.5^{|i-j|}$ $\forall\, i \neq j$. We simulate datasets with $n_T = \{200, 400\}$ for the training set and $n_P = 500$ for the testing set.

Table 10. Simulation Results for Logistic Regression

n_T	Lasso	OBLasso	NBLasso
200	0.004	0.006	0.006
400	0.003	0.004	0.003

Example 9 (Logistic Regression Example): In this example, observations with binary response are independently generated according to the following model [30]

$$P(y_i|x_i) = \frac{\exp\{x_i^T \boldsymbol{\beta}\}}{1 + \exp\{x_i^T \boldsymbol{\beta}\}},$$

where $\boldsymbol{\beta} = (3, 0, 0, 1.5, 0, 0, 2, 0, 0)^T$. The experimental results are summarized in Table 10 which shows that NBLasso performs comparably with OBLasso. As the size of the training data increases, all the three methods yield equivalent performance.

Example 10 (Cox's Model Example): In this simulation study, independent survival data are generated according to the following hazard function [30]

$$h(t_i|x_i) = \exp\{x_i^T \boldsymbol{\beta}\},$$

where t_i is the survival time from the ith subject and $\boldsymbol{\beta} = (0.8, 0, 0, 1, 0, 0, 0.6, 0)^T$. Also, independent censoring time is generated from an exponential distribution with mean u $\exp\{x_i^T \boldsymbol{\beta}\}$, where $u \sim$ Uniform$(1, 3)$. The experimental results are summarized in Table 11 which shows that both Bayesian lassos perform comparably, outperforming frequentist lasso. Thus, it is evident from the experimental results that NBLasso is as effective as OBLasso even for the general models.

Table 11. Simulation Results for Cox's Model

n_T	Lasso	OBLasso	NBLasso
200	0.41	0.3	0.3
400	0.15	0.1	0.1

7. COMPUTING MAP ESTIMATES

7.1 ECM algorithm for linear models

It is well known that the conditional distributions of a truncated multivariate normal distribution are truncated univariate normals. This fact motivates us to develop an ECM algorithm to estimate the conditional posterior mode of $\boldsymbol{\beta}$. At each step, we treat the latent variables u_j's and the tuning parameter λ as missing parameters and average over them to estimate β_j's and σ^2 by maximizing the expected log conditional posterior distributions.

The complete data log-likelihood is

$$\begin{aligned}
&\log\, p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{u}, \lambda|\boldsymbol{y}, X)\\
&\qquad \propto \sum_{i=1}^{n} \log\, p(y_i|X_i\boldsymbol{\beta}, \sigma^2) + \sum_{j=1}^{p} \log\, p(\beta_j|u_j, \sigma^2)\\
&\qquad\quad + \sum_{j=1}^{p} \log\, p(u_j|\lambda) + \ \log\, p(\sigma^2) + \ \log\, p(\lambda)\\
&\qquad \propto \{\sum_{i=1}^{n} -\frac{1}{2\sigma^2}(y_i - X_i\boldsymbol{\beta})^2 + (a + 2p - 1)\log\lambda \qquad (13)\\
&- \lambda(b + \sum_{j=1}^{p} u_j) - \frac{(n-1+p)}{2}\ \log\,\sigma^2\} I\left\{u_j > \frac{|\beta_j|}{\sqrt{\sigma^2}}, \forall j\right\}.
\end{aligned}$$

We initialize the algorithm by starting with a guess of $\boldsymbol{\beta}$ and σ^2. Then, at each step of the algorithm we replace $\boldsymbol{u}$ and λ in the log joint posterior (13) by their expected values conditional on the current estimates of $\boldsymbol{\beta}$ and σ^2. Finally, we update $\boldsymbol{\beta}$ and σ^2 by maximizing the expected log conditional posterior distributions. The algorithm proceeds as follows:

E-Step:

$$u_j^{(t)} = \left(\frac{1}{\lambda}\right)^{(t)} + \frac{|\beta_j^{(t)}|}{\sqrt{\sigma^{2(t)}}},\ j = 1, .., p,$$

$$\left(\frac{1}{\lambda}\right)^{(t)} = \frac{1}{(b + \sum_{j=1}^{p} |\beta_j^{(t)}|)(a + 2p - 1)};$$

CM-Steps:

$$\beta_j^{(t+1)} = \begin{cases} -u_j^{(t)}\sqrt{\sigma^{2(t)}} & \text{if } \beta_j^{(t)} < -u_j^{(t)}\sqrt{\sigma^{2(t)}}\\ \hat{\beta}_{j\,OLS} & \text{if } -u_j^{(t)}\sqrt{\sigma^{2(t)}} < \beta_j^{(t)} < u_j^{(t)}\sqrt{\sigma^{2(t)}}\\ u_j^{(t)}\sqrt{\sigma^{2(t)}} & \text{if } \beta_j^{(t)} > u_j^{(t)}\sqrt{\sigma^{2(t)}}, \quad j = 1(1)p, \end{cases}$$

$$\sigma^{2(t+1)} = \text{Max}\left\{\frac{(\boldsymbol{y} - X\boldsymbol{\beta}^{(t+1)})'(\boldsymbol{y} - X\boldsymbol{\beta}^{(t+1)})}{n + p + 1}, \text{Max}_j \left(\frac{\beta_j^{(t+1)}}{u_j{}^{(t)}}\right)^2\right\}.$$

At convergence of the algorithm, we summarize the inferences using the latest estimates of $\boldsymbol{\beta}$ and their variances [17].

7.2 ECM algorithm for general models

Similarly, an approximate ECM algorithm for general models can be given as follows:

1. E-Step:

$$u_j^{(t)} = \frac{1}{(b + \sum_{j=1}^{p} |\beta_j^{(t)}|)(a + 2p - 1)} + |\beta_j^{(t)}|,\ j = 1, .., p;$$

2. CM-Steps: $\beta_j^{(t+1)} = \begin{cases} -u_j^{(t)} & \text{if } \beta_j^{(t)} < -u_j^{(t)}\\ \hat{\beta}_{j\,\text{MLE}} & \text{if } -u_j^{(t)} < \beta_j^{(t)} < u_j^{(t)}\\ u_j^{(t)} & \text{if } \beta_j^{(t)} > u_j^{(t)}, \quad j = 1(1)p. \end{cases}$

3. Repeat 1 & 2 until convergence.

8. CONCLUDING REMARKS

In this paper, we have introduced a new hierarchical representation of Bayesian lasso using SMU distribution. It is to be noted that the posterior distribution of interest $p(\boldsymbol{\beta}, \sigma^2|y)$ is exactly the same for both original Bayesian lasso (OBLasso) and new Bayesian lasso (NBLasso) models. As such, all inference and prediction that is based on the posterior distribution should be exactly the same 'theoretically'. Any observed differences must be attributed (up to Monte Carlo error) to the properties of the different Gibbs samplers used to obtain samples from the corresponding posterior distributions. We establish this fact by real data analyses and empirical studies. Our results indicate that both Bayesian lassos perform comparably in different empirical and real scenarios. In many situations, the new method is competitive in terms of either prediction accuracy or variable selection. Moreover, NBLasso performs quite satisfactorily for general models beyond the linear regression. Furthermore, the proposed Gibbs sampler inherits good mixing properties as evident from both empirical studies (data not shown due to too many predictors) and real data analyses. Note that, we do not have a theoretical result on the posterior convergence of our MCMC. Therefore, despite our encouraging findings, theoretical research is needed to investigate the posterior convergence of the proposed MCMC algorithm.

One should be aware that both non-Bayesian and Bayesian lasso are essentially optimization methods with the common goal of determining the model parameters that maximize some objective function. Therefore, a Bayesian approach can often lead to very different results than a traditional penalized likelihood approach [14]. Apart from the advantages discussed above, the proposed Bayesian lasso also has some limitations as it carries forward all the drawbacks of frequentist lasso [18]. To overcome those limitations, one can easily adopt the adaptive lasso [35] by choosing variable-specific tuning parameter in the MCMC step. However, adopting the new hierarchical representation based on SMU distribution to other regularization methods viz. bridge estimator [8], group lasso [32, 22], elastic net [36], group bridge [16], adaptive group bridge [23] and adaptive elastic net [34, 12] remains an active area for future research.

APPENDIX A. APPENDIX SECTION

Proof of Proposition: It is well known that

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

Therefore, the pdf of a Laplace distribution with mean 0 and variance $\sqrt{\sigma^2}/\lambda$ can be written as

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

Hence this proves (4).

Posterior Distributions: Assuming that priors for different parameters are independent, we can express the joint posterior distribution of all parameters as:

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

Conditional on $\boldsymbol{y}, X, \boldsymbol{u}, \lambda, \sigma^2$, the posterior distribution of $\boldsymbol{\beta}$ is

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

Hence,

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

Similarly,

$$\pi(\boldsymbol{u}|\boldsymbol{y}, X, \boldsymbol{\beta}, \lambda, \sigma^2)$$
$$\propto \pi(\boldsymbol{\beta}|\boldsymbol{u}, \sigma^2)\pi(\boldsymbol{u}|\lambda) \propto \prod_{j=1}^{p} e^{-\lambda u_j} I\{u_j > \frac{|\beta_j|}{\sqrt{\sigma^2}}\}.$$

Therefore,

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

Similarly,

$$\pi(\sigma^2|\boldsymbol{y}, X, \boldsymbol{\beta}, \boldsymbol{u}, \lambda) \propto \pi(\boldsymbol{y}|X, \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}|\boldsymbol{u}, \sigma^2)\pi(\sigma^2)d\sigma^2$$
$$\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n-1+p}{2}+1} e^{-\frac{1}{2\sigma^2}(\boldsymbol{y}-X\boldsymbol{\beta})'(\boldsymbol{y}-X\boldsymbol{\beta})} I\{\sigma^2 > \text{Max}_j(\frac{{\beta_j}^2}{{u_j}^2})\}.$$

Therefore,

$$\sigma^2|\boldsymbol{y}, X, \boldsymbol{\beta}, \boldsymbol{u}, \lambda$$
$$\sim \text{Inverse Gamma}(\frac{n-1+p}{2},$$
$$\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta}))I\{\sigma^2 > \text{Max}_j(\frac{{\beta_j}^2}{{u_j}^2})\}.$$

ACKNOWLEDGEMENTS

This work was supported in part by the research grant NIH 5R01GM069430-08. Himel Mallick was supported in part by a cooperative agreement grant U01 NS041588 from the National Institute of Neurological Disorders and Stroke, National Institutes of Health and Department of Health and Human Services. We are grateful to the reviewers for their kind comments and insightful criticisms that significantly improved the quality of the paper.

Received 15 August 2013

REFERENCES

[1] Andrews, D. F. and Mallows, C. L. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36:99–102, 1974. MR0359122

[2] Bae, K. and Mallick, B. Gene selection using a two-level hierarchical bayesian model. *Bioinformatics*, 20(18):3423–3430, 2004.

[3] Chen, X., Wang, J. Z., and McKeown, J. M. A bayesian lasso via reversible-jump mcmc. *Signal Processing*, 91(8):1920–1932, 2011.

[4] Choy, S. T. B., Wan, W., and Chan, C. Bayesian student-t stochastic volatility models via scale mixtures. *Advances in Econometrics*, 23:595–618, 2008.

[5] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of Statistics*, 32(2):407–99, 2004. MR2060166

[6] Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001. MR1946581

[7] Figueiredo, M. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1150–59, 2003.

[8] Frank, I. and Friedman, J. H. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35:109–135, 1993.

[9] Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[10] Gelman, A., Carlin, J., Stern, H., and Rubin, D. *Bayesian Data Analysis*. Chapman & Hall, London, 2003. MR2027492

[11] Gelman, A. and Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[12] Ghosh, S. On the grouped selection and model complexity of the adaptive elastic net. *Statistics and Computing*, 21(3):452–461, 2011. MR2806621

[13] Hans, C. M. Bayesian lasso regression. *Biometrika*, 96(4):835–845, 2009. MR2564494

[14] Hans, C. M. Model uncertainty and variable selection in bayesian lasso regression. *Statistics and Computing*, 20:221–9, 2010. MR2610774

[15] Hoerl, A. E. and Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[16] Huang, J., Ma, S., Xie, H., and Zhang, C. A group bridge approach for variable selection. *Biometrika*, 96(2):339–355, 2009. MR2507147

[17] Johnson, N. L., Kotz, S., and Balakrishnan, N. *Continuous Univariate Distributions*. John Wiley and Sons, New York, 1994.

[18] Kyung, M., Gill, J., Ghosh, M., and Casella, G. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5:369–412, 2010. MR2719657

[19] Leng, C., Tran, M., and Nott, D. Bayesian adaptive lasso. *Annals of the Institute of Mathematical Statistics*, 66(2):221–244, 2014. MR3171404

[20] Li, Q. and Lin, N. The bayesian elastic net. *Bayesian Analysis*, 5(1):151–70, 2010. MR2596439

[21] 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.

[22] Meier, L., van de Geer, S., and Buhlmann, P. The group lasso for logistic regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 70:53–71, 2008. MR2412631

[23] Park, C. and Yoon, Y. J. Bridge regression: Adaptivity and group selection. *Journal of Statistical Planning and Inference*, 141:3506–3519, 2011. MR2817359

[24] Park, T. and Casella, G. The bayesian lasso. *Journal of the American Statistical Association*, 103:681–686, 2008. MR2524001

[25] Qin, Z., Walker, S., and Damien, P. Uniform scale mixture models with applications to variance regression. Working papers series, University of Michigan Ross School of Business, 1998.

[26] Qin, Z., Walker, S., and Damien, P. Uniform scale mixture models with applications to bayesian inference. Working papers series, University of Michigan Ross School of Business, 1998a.

[27] 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.

[28] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996. MR1379242

[29] Walker, S., Damien, P., and Meyer, M. On scale mixtures of uniform distributions and the latent weighted least squares method. Working papers series, University of Michigan Ross School of Business, 1997.

[30] Wang, H. and Leng, C. Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048, 2007. MR2411663

[31] Wu, T. and Lange, K. Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008. MR2415601

[32] Yuan, M. and Lin, N. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 68:49–67, 2006. MR2212574

[33] Zhang, C. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010. MR2604701

[34] Zhou, H., Alexander, D. H., Sehi, M. E., Sinsheimer, J. S., Sobel, E. M., and Lange, K. Penalized regression for genome-wide association screening of sequence data. *Bioinformatics*, 26: 2375–82, 2010.

[35] Zou, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006. MR2279469

[36] Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Methodological)*, 67:301–320, 2005. MR2137327

Himel Mallick
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, AL 35294
USA
Phone: (205) 201-0252
Fax: (205) 975-2540
E-mail address: himel@uab.edu

Nengjun Yi
Department of Biostatistics
University of Alabama at Birmingham
Birmingham, AL 35294
USA
Phone: (205) 934-4924
Fax: (205) 975-2540
E-mail address: nyi@ms.soph.uab.edu