Bayesian Analysis (2017) **12**, Number 3, pp. 753–778

# Selection of Tuning Parameters, Solution Paths and Standard Errors for Bayesian Lassos

Vivekananda Roy[*] and Sounak Chakraborty[†]

**Abstract.** Penalized regression methods such as the lasso and elastic net (EN) have become popular for simultaneous variable selection and coefficient estimation. Implementation of these methods require selection of the penalty parameters. We propose an empirical Bayes (EB) methodology for selecting these tuning parameters as well as computation of the regularization path plots. The EB method does not suffer from the "double shrinkage problem" of frequentist EN. Also it avoids the difficulty of constructing an appropriate prior on the penalty parameters. The EB methodology is implemented by efficient importance sampling method based on multiple Gibbs sampler chains. Since the Markov chains underlying the Gibbs sampler are proved to be geometrically ergodic, Markov chain central limit theorem can be used to provide asymptotically valid confidence band for profiles of EN coefficients. The practical effectiveness of our method is illustrated by several simulation examples and two real life case studies. Although this article considers lasso and EN for brevity, the proposed EB method is general and can be used to select shrinkage parameters in other regularization methods.

**MSC 2010 subject classifications:** primary 62F15, 62J07; secondary 60J05.

**Keywords:** Bayesian lasso, elastic net, empirical Bayes, geometric ergodicity, importance sampling, Markov chain Monte Carlo, shrinkage, standard errors.

## 1 Introduction

Consider the standard linear model $\boldsymbol{y} = \mu \mathbf{1}_n + X\beta + \epsilon$, where $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of responses, $\mu \in \mathbb{R}$ is the overall mean, $\mathbf{1}_n$ is the $n \times 1$ vector of 1's, $X = (X_1, X_2, \ldots, X_p)$ is the $n \times p$ (*standardized*) covariate matrix, $\beta \in \mathbb{R}^p$ is the unknown vector of regression coefficients, and $\epsilon$ is the $n \times 1$ vector of iid normal errors with mean zero and unknown variance parameter $\sigma^2$. The ordinary least square (OLS) method for estimating $\beta$ and $\sigma^2$ is not applicable when $p > n$, which is very common in the modern data sets arising in genetics, medical science, and other scientific disciplines. OLS also has problems when "$n$ is not much larger than $p$" (James et al., 2013, p. 203). The *shrinkage* (also known as *regularization*) approach, where regression coefficients are shrunken toward zero, can be used to analyze these types of data sets. For example, ridge regression (Hoerl and Kennard, 1970) penalizes large values of the coefficients using $L_2$ norm. The result of the ridge regression, however, is not sparse and all the regression coefficients will remain nonzero at the end of the analysis. The extremely popular *least absolute shrinkage and selection operator* (*lasso*) (Tibshirani, 1996), which is based on $L_1$ norm regularization,

[*] Department of Statistics, Iowa State University, Ames, Iowa, vroy@iastate.edu

[†] Department of Statistics, University of Missouri, Columbia, Missouri, chakrabortys@missouri.edu

DOI: 10.1214/16-BA1025

can simultaneously perform shrinkage and variable selection as many of the coefficients can be estimated to be exactly zero. The lasso estimate of $\beta$ is obtained by solving

$$\min_{\beta} (\boldsymbol{y} - \mu \mathbf{1}_n - X\beta)^T (\boldsymbol{y} - \mu \mathbf{1}_n - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{1}$$

for some shrinkage parameter $\lambda \in \mathbb{R}$. The lasso estimator, although has shown to be very useful in many situations, does have some shortcomings. Since lasso does convex optimization, it can not select more variables than the sample size. But, many problems, for example the micro array experiments, involve much more predictors than the available sample size. Also lasso performs unsatisfactorily in the situations where predictors are highly correlated. Finally, if there is some group structure among the variables, the lasso tends to select only one variable from a group ignoring others. Zou and Hastie (2005) proposed the *Elastic Net* (EN) to achieve better performance in the above three scenarios where lasso has limitations. The EN estimator is obtained by solving

$$\min_{\beta} (\boldsymbol{y} - \mu \mathbf{1}_n - X\beta)^T (\boldsymbol{y} - \mu \mathbf{1}_n - X\beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2, \tag{2}$$

where $\lambda_1$ and $\lambda_2$ are tuning parameters. From (2) we see that the elastic net uses both an $L_1$ penalty as in lasso and an $L_2$ penalty as in the ridge regression. A variety of other penalty terms have been proposed in the literature for incorporating the grouping structure among the variables and to overcome other limitations of lasso. For example, the grouped lasso penalty in Yuan and Lin (2006), the adaptive lasso of Zou (2006), and the octagonal shrinkage and clustering of Bondell and Reich (2008).

One of the problems that we consider in this paper is the selection of the shrinkage parameters in the penalized regression methods. The tuning parameters control the impact of the penalty terms. For example, in (2) if $\lambda_1 = 0 = \lambda_2$ the penalty terms have no effect and in this case EN produces the OLS estimates. On the other hand, as $\lambda_1$, or $\lambda_2 \to \infty$, the impact of the penalty terms increase and the EN estimates approach zero. Typically cross validation methods are used for selecting the shrinkage parameters. However, there are problems with the use of cross validation methods for selecting the tuning parameters. For example, in the context of EN, the cross validation method described in Zou and Hastie (2005) "selects $\lambda_1$ and $\lambda_2$ sequentially instead of simultaneously and causes the double shrinkage problem" (Li and Lin, 2010, p. 152). In the Bayesian framework, the shrinkage parameters can be estimated using either an empirical Bayes (EB) approach or a fully Bayesian analysis with appropriate priors on $\lambda_1$ and $\lambda_2$. There are several problems with the full Bayesian approach, for example, choosing an appropriate prior may not be easy and as seen in Section 4, the choice of prior can have influence on subsequent inference. Also, sampling from the full conditional distributions of $\lambda_1$ and $\lambda_2$, which is required in the Markov chain Monte Carlo (MCMC) sampling for the full Bayesian analysis, is computationally expensive. As shown in Section 2.1, the full conditional distributions of $\lambda_1$ and $\lambda_2$ mentioned in Kyung et al. (2010) are incorrect. The correct conditionals of $\lambda_1$ and $\lambda_2$ are nonstandard, complex distributions. Thus MCMC sampling for full Bayesian analysis of EN is computationally demanding.

It has been recently shown by Khare and Hobert (2013) that when $\lambda$ in (1) is assumed fixed, the Gibbs sampler Markov chain for Bayesian lasso of Park and Casella (2008) has a *geometric rate* of convergence, while no such convergence results is currently known about the MCMC algorithm for the full Bayesian lasso model. (See Section 2 for the definition of geometric convergence.) In section 2, we prove that, the Bayesian elastic net Gibbs sampler that do not update $\lambda_1$ and $\lambda_2$ is geometrically ergodic.

Park and Casella (2008) mention that a Monte Carlo EM algorithm can be used for calculating the maximum marginal likelihood estimate of $\lambda$ in the Bayesian lasso. Li and Lin (2010) use the Monte Carlo EM for jointly estimating the tuning parameters of the EN. In the Monte Carlo EM algorithm, a *new* fully convergent Markov chain is to be run in each iteration of the EM algorithm—which is computationally inefficient. In Section 4, we see that the EM algorithm can be extremely slow. Also, we may want to obtain the entire marginal likelihood surface instead of just the maximizing value as available from the Monte Carlo EM algorithm. Park and Casella (2008, p. 683) mention that an importance sampling method can be used to approximate the ratio of marginal likelihoods near the maximizer of $\lambda$. As we explain in Section 3 this naive importance sampling method is in general inefficient. In this paper we develop an EB approach with efficient generalized importance sampling methods based on multiple Markov chains for estimating the shrinkage parameters in penalized regression methods.

In penalized regression methods such as lasso and EN, a plot of the profiles of the estimated regression coefficients as a function of the penalty parameter is used to display the amount of shrinkage corresponding to different tuning parameter values. These plots are also useful for comparing different shrinkage methods. We show that the proposed generalized importance sampling method can efficiently compute such regularization path plots.

The other issue that we consider in this paper is the standard error estimation of the EN estimator. In the frequentist analysis, various standard error estimates have been proposed for the lasso estimator. The problem with the ridge regression approximation suggested by Tibshirani (1996) or the sandwich estimator of Fan and Li (2001) is that they produce the standard error estimate to be zero when the regression parameter estimate is zero. Kyung et al. (2010) have shown that the bootstrap method of standard error estimation does not "attach valid standard error estimates to the values of the lasso that are shrunk to zero, in the sense that these estimators are inconsistent" (see also Knight and Fu, 2000). On the other hand, as mentioned above, Khare and Hobert (2013) have recently shown that the Bayesian lasso Gibbs sampler is geometrically ergodic—which allows for calculation of asymptotically valid standard errors of lasso estimates (see Section 2 for details). Following Khare and Hobert (2013), in this paper we prove that the Bayesian EN Gibbs sampler is geometrically ergodic. This implies that there is a central limit theorem (CLT) for posterior EN estimates based on the Gibbs sampler. Moreover, it also justifies the use of batch means estimator proposed in Roy et al. (2015) for producing confidence bands for marginal likelihood surface estimates as well as regularization path estimates.

The rest of the paper is organized as follows. Section 2 presents the Gibbs samplers as well as the statement regarding their convergence rates. Section 3 describes an efficient importance sampling method for selecting the shrinkage parameters. It also provides regularization path estimators. Section 4 contains numerical results involving simulation studies and real data application. Some concluding remarks appear in Section 5. Proofs of results are relegated to the Web based supplementary materials (Roy and Chakraborty, 2016).

# 2 Hierarchical models and Gibbs samplers for lasso and elastic net

## 2.1 Gibbs samplers for lasso and elastic net

Tibshirani (1996) noted that lasso estimates can be obtained as posterior mode when the regression parameters have independent and identical Laplace priors. Using a Gaussian mixture representations (with exponential mixing density) of the Laplace distribution, Park and Casella (2008) consider a hierarchical Bayesian lasso model as follows

$$\begin{aligned} \boldsymbol{y}|\mu,\beta,\sigma^2 &\sim N_n(\mu\mathbf{1}_n + X\beta, \sigma^2 I_n), \\ \pi(\mu)\propto 1; \beta|\sigma^2,\tau_1^2,\ldots,\tau_p^2 &\sim N_p(\mathbf{0}_p, \sigma^2 D_\tau),\ \text{where } D_\tau \equiv \text{Diag}(\tau_1^2,\ldots,\tau_p^2) \\ \sigma^2 \sim \text{Inverse-Gamma}(\alpha,\xi); \tau_j^2|\lambda &\overset{iid}{\sim} \text{Exponential}(\lambda^2/2), \text{for } j=1,2,\ldots,p. \end{aligned} \tag{3}$$

The improper prior $\pi(\sigma^2) = 1/\sigma^2$ (used in Park and Casella (2008)) is obtained by replacing $\alpha = 0, \xi = 0$ in the above model. The connection of the above hierarchical representation with the lasso penalty in (1) can be seen from the following result of Andrews and Mallows (1974)

$$\frac{a}{2}\exp(-a|z|) = \int_0^\infty \frac{1}{\sqrt{2\pi s}}\exp\Big(-\frac{z^2}{2s}\Big)\frac{a^2}{2}\exp\Big(-\frac{a^2}{2}s\Big)ds. \tag{4}$$

From (4) it follows that by integrating out $\tau_1^2,\ldots,\tau_p^2$, the conditional density of $\beta|\sigma^2$ is $\prod_{i=1}^p(\lambda/2\sigma)\exp(-\lambda|\beta_i|/\sigma)$. Let $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \bar{\boldsymbol{y}}\mathbf{1}_n$. Since the columns of $X$ are centered, easy calculation shows that

$$f(\tilde{\boldsymbol{y}}|\beta,\sigma^2) \equiv \int_{\mathbb{R}} f(\boldsymbol{y}|\mu,\beta,\sigma^2)\pi(\mu)d\mu = \frac{1}{(2\pi)^{(n-1)/2}\sigma^{n-1}}\exp\Big[-\frac{(\tilde{\boldsymbol{y}}-X\beta)^T(\tilde{\boldsymbol{y}}-X\beta)}{2\sigma^2}\Big], \tag{5}$$

that is, marginalization over $\mu$ does not break the conjugacy offered by the use of conjugate priors in the above hierarchical model. In fact, the full conditional distributions of $\beta,\tau^2,\sigma^2$ are given by

$$\begin{aligned} &\beta|\sigma^2,\tau^2,\boldsymbol{y} \sim N_p((X^TX + D_\tau^{-1})^{-1}X^T\tilde{\boldsymbol{y}}, \sigma^2(X^TX + D_\tau^{-1})^{-1}) \\ &\frac{1}{\tau_j^2}|\beta,\sigma^2,\boldsymbol{y} \overset{ind}{\sim} \text{Inverse-Gaussian}\Big(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}},\lambda^2\Big) \text{ for } j=1,2,\ldots,p \\ &\sigma^2|\beta,\tau^2,\boldsymbol{y} \sim \text{Inverse-Gamma}\Big(\frac{n-1+p+2\alpha}{2}, \frac{(\tilde{\boldsymbol{y}}-X\beta)^T(\tilde{\boldsymbol{y}}-X\beta)+\beta^TD_\tau^{-1}\beta+2\xi}{2}\Big). \end{aligned}$$

The Bayesian lasso Gibbs sampler is a fixed scan Gibbs sampling algorithm which updates the parameters $(\beta, \tau^2, \sigma^2)$ in each iteration using draws from the above three conditional distributions sequentially.

Following Park and Casella (2008), Kyung et al. (2010) present hierarchical models for other generalized lasso methods including the elastic net. We consider the hierarchical model for the Bayesian elastic net:

$$\begin{aligned}
&\boldsymbol{y}|\mu, \beta, \sigma^2 \sim N_n(\mu \mathbf{1}_n + X\beta, \sigma^2 I_n),\\
&\pi(\mu) \propto 1; \beta|\sigma^2, \tau_1^2, \ldots, \tau_p^2, \lambda_2 \sim N_p(\mathbf{0}_p, \sigma^2 D_\tau^*),\\
&\qquad\qquad \text{where } D_\tau^* \equiv \mathrm{Diag}((\tau_1^{-2} + \lambda_2)^{-1}, \ldots, (\tau_p^{-2} + \lambda_2)^{-1})\\
&\sigma^2 \sim \text{Inverse-Gamma}(\alpha, \xi)\\
&\quad \tau_j^2 | \lambda_1, \lambda_2 \overset{iid}{\sim} \frac{1}{C(\lambda_1, \lambda_2)} \exp\Big(-\frac{\lambda_1^2 \tau_j^2}{2}\Big) \frac{1}{\sqrt{1 + \lambda_2 \tau_j^2}}, \text{for } j = 1, 2, \ldots, p,
\end{aligned} \tag{6}$$

with

$$C(\lambda_1, \lambda_2) = (1/\lambda_1)\sqrt{(2/\lambda_2)} \exp[\lambda_1^2/(2\lambda_2)]\Gamma(1/2, \lambda_1^2/\{2\lambda_2\}),$$

where $\Gamma(a, x) = \int_x^\infty t^{a-1} \exp(-t) dt$ is the incomplete Gamma function. Here we also allow the improper prior, $1/\sigma^2$ (corresponds to $\alpha = 0, \xi = 0$ in (6)) for $\sigma^2$. The connection of the above hierarchical representation with (2) can be seen from the conditional prior density of $\beta|\sigma^2$ given by

$$\pi(\beta|\sigma^2) \propto \prod_{i=1}^p \exp\Big\{-\frac{\lambda_1 |\beta_i|}{\sigma} - \frac{\lambda_2 \beta_i^2}{2\sigma^2}\Big\}, \tag{7}$$

which is obtained using the mixture representation (4). Note that the independent exponential prior of $\tau_j^2$ used in the Bayesian EN model by Kyung et al. (2010) does not lead to the prior density $\beta|\sigma^2$ in (7). Kyung et al. (2010) assumed independent exponential $(\lambda_1^2/2)$ priors on $\tau_j^2, j = 1, \ldots, p$. Consequently, with independent Gamma priors on $\lambda_1^2$ and $\lambda_2$, the full conditionals of $\lambda_1^2$ and $\lambda_2$ become Gamma distributions. As shown above, the prior of $\tau_j^2$ depends on both $\lambda_1$ and $\lambda_2$, and in this case, the conditional distributions of $\lambda_1^2$ and $\lambda_2$ are complicated. Thus MCMC sampling for the full Bayesian analysis of EN is computationally expensive (Lee et al., 2015).

As in the Bayesian lasso model, the parameter $\mu$ can be analytically integrated out from the joint posterior distribution corresponding to (6). The full conditional distributions of $\beta, \tau^2, \sigma^2$ are similar to the Bayesian lasso and are given by

$$\beta|\sigma^2, \tau^2, \lambda_2, \boldsymbol{y} \sim N_p((X^T X + D_\tau^{*-1})^{-1} X^T \tilde{\boldsymbol{y}}, \sigma^2 (X^T X + D_\tau^{*-1})^{-1}) \tag{8}$$

$$\frac{1}{\tau_j^2} | \beta, \sigma^2, \lambda_1, \boldsymbol{y} \overset{ind}{\sim} \text{Inverse-Gaussian}\Big(\sqrt{\frac{\lambda_1^2 \sigma^2}{\beta_j^2}}, \lambda_1^2\Big) \text{ for } j = 1, 2, \ldots, p \tag{9}$$

$$\begin{aligned}
&\sigma^2 | \beta, \tau^2, \lambda_2, \boldsymbol{y}\\
&\quad \sim \text{Inverse-Gamma}\Big(\frac{n - 1 + p + 2\alpha}{2}, \frac{(\tilde{\boldsymbol{y}} - X\beta)^T (\tilde{\boldsymbol{y}} - X\beta) + \beta^T D_\tau^{*-1} \beta + 2\xi}{2}\Big).
\end{aligned} \tag{10}$$

In Section 2.2 we analyze the Gibbs sampler which is run by updating $(\beta, \tau^2, \sigma^2)$ in each iteration using draws from the above three conditional distributions.

## 2.2 Geometric convergence of the elastic net Gibbs sampler

Let $\theta \equiv (\beta, \tau^2, \sigma^2)$, where $\tau^2 \equiv (\tau_1^2, \dots, \tau_p^2)$. Let $\{\theta_m\}_{m\geq 0}$ be the Markov chain underlying the elastic net Gibbs sampling algorithm discussed in the previous section. That is, at iteration $m$ given $\theta_m \equiv (\beta_m, \tau_m^2, \sigma_m^2)$, this Markov chain uses three steps to move to the new value $\theta_{m+1} \equiv (\beta_{m+1}, \tau_{m+1}^2, \sigma_{m+1}^2)$: draw $\sigma_{m+1}^2|\beta_m, \tau_m^2$ from (10) given $(\beta_m, \tau_m^2)$, then draw $\tau_{m+1}^2|\beta_m, \sigma_{m+1}^2$ from (9) given $(\beta_m, \sigma_{m+1}^2)$, and finally draw $\beta_{m+1}|\tau_{m+1}^2, \sigma_{m+1}^2$ from (8) given $(\tau_{m+1}^2, \sigma_{m+1}^2)$. (Note that $\tau_m^2$ is used to denote both $m$th component of $\tau^2$ as well as the value of $\tau^2$ at iteration $m$.) The Markov transition density (Mtd) of this Gibbs sampler is

$$k((\beta, \tau^2, \sigma^2)|(\beta_0, \tau_0^2, \sigma_0^2)) = \pi(\beta|\tau^2, \sigma^2, \boldsymbol{y})\pi(\tau^2|\beta_0, \sigma^2, \boldsymbol{y})\pi(\sigma^2|\beta_0, \tau_0^2, \boldsymbol{y}), \qquad (11)$$

where $\pi(\beta|\tau^2, \sigma^2, \boldsymbol{y})$, $\pi(\tau^2|\beta, \sigma^2, \boldsymbol{y})$ and $\pi(\sigma^2|\beta, \tau^2, \boldsymbol{y})$ are the full conditional densities corresponding to the joint posterior density

$$\pi(\beta, \tau^2, \sigma^2|\boldsymbol{y}, \lambda_1, \lambda_2) = \frac{f(\tilde{\boldsymbol{y}}|\beta, \sigma^2)\pi(\beta, \sigma^2, \tau^2|\lambda_1, \lambda_2)}{m_{\lambda_1,\lambda_2}(\boldsymbol{y})}, \qquad (12)$$

where $f(\tilde{\boldsymbol{y}}|\beta, \sigma^2)$ is given in (5), $\pi(\beta, \sigma^2, \tau^2|\lambda_1, \lambda_2)$ is the prior density of $(\beta, \sigma^2, \tau^2)$ given in (6), and

$$m_{\lambda_1,\lambda_2}(\boldsymbol{y}) = \int_{\mathbb{R}_+}\int_{\mathbb{R}_+^p}\int_{\mathbb{R}^p} f(\tilde{\boldsymbol{y}}|\beta, \sigma^2)\pi(\beta, \sigma^2, \tau^2|\lambda_1, \lambda_2)d\beta d\tau^2 d\sigma^2 \qquad (13)$$

is the normalizing constant of the posterior density (12). Standard calculations show that the posterior density (12) is the invariant density for the EN Gibbs Markov chain $\{\beta_m, \tau_m^2, \sigma_m^2\}_{m\geq 0}$. Since the Mtd $k$ is strictly positive, the EN Gibbs chain is irreducible with respect to the Lebesgue measure on $\mathbb{R}^p \times \mathbb{R}_+^p \times \mathbb{R}_+$ (Meyn and Tweedie, 1993, Chapter 4). Then from Asmussen and Glynn (2011) it follows that $\{\beta_m, \tau_m^2, \sigma_m^2\}_{m\geq 0}$ is a *positive Harris recurrent* Markov chain. Thus, the EN Gibbs chain can be used to produce strongly consistent estimators (Meyn and Tweedie, 1993, Chapter 17). In particular, if $E_\pi|g| < \infty$, that is, if $g$ is an integrable function with respect to the posterior density (12), then $\bar{g}_m := \sum_{i=0}^{m-1} g(\beta_i, \tau_i^2, \sigma_i^2)/m$ is strongly consistent for $E_\pi g$ no matter what is the initial distribution of $(\beta_0, \tau_0^2, \sigma_0^2)$. However, this estimator is useful in practice only if an associated standard error is provided. In fact if there is a CLT for $g$, that is,

$$\sqrt{m}(\bar{g}_m - E_\pi g) \xrightarrow{\text{d}} N(0, \psi_g^2), \ \text{ as } m \to \infty, \qquad (14)$$

and if $\hat{\psi}_g^2$ is a consistent estimator of the asymptotic variance $\psi_g^2$, then an asymptotic standard error for $\bar{g}_m$ is $\hat{\psi}_g/\sqrt{m}$. Unfortunately, Harris recurrence of a Markov chain does not guarantee the existence of such a CLT. The only standard method for establishing a Markov chain CLT and obtaining consistent estimates of the asymptotic variance is to prove that the underlying Markov chain is geometrically ergodic (see Jones and Hobert, 2001; Roy and Hobert, 2007; Flegal and Jones, 2010). The EN Gibbs chain is called *geometrically ergodic* if there exists a positive real valued function $M$ and a constant $r \in (0,1)$ such that, for all $m$,

$$\left\| K^m\left((\beta,\tau^2,\sigma^2),\cdot\right) - \Pi(\cdot|\boldsymbol{y})\right\|_{\text{TV}} \leq W(\beta,\tau^2,\sigma^2) r^m \ , \tag{15}$$

where $K^m((\beta,\tau^2,\sigma^2),\cdot)$ denotes the probability distribution of the Markov chain started at $(\beta,\tau^2,\sigma^2)$ after $m$ steps, $\Pi(\cdot|\boldsymbol{y})$ denotes the probability measure corresponding to the posterior density (12), and $\|\cdot\|_{\text{TV}}$ denotes the total variation norm. Note that the function $W$ and the constant $r$ may depend on $\lambda_1, \lambda_2$. It is known that if the EN Gibbs chain is geometrically ergodic, then there is a CLT (14) for every function $g$ such that $E_\pi |g|^{2+\delta} < \infty$, for some $\delta > 0$ (Roberts and Rosenthal, 1997). We have the following result.

**Proposition 1.** *The elastic net Gibbs Markov chain is geometrically ergodic for every* $n \geq 4, p, X, \lambda_1$ *and* $\lambda_2$.

Hence the CLT (14) holds for the EN Gibbs chain and it can be used to construct asymptotic standard errors for posterior estimates. We establish geometric ergodicity of the EN Gibbs chain by establishing the so-called *drift condition*, and an associated *minorization condition*, which we now describe. (See Jones and Hobert (2001) for an introduction to these ideas.)

We begin with a drift condition. Consider the following function

$$V(\beta,\tau^2,\sigma^2) = (\tilde{\boldsymbol{y}} - X\beta)^T(\tilde{\boldsymbol{y}} - X\beta) + \beta^T D_\tau^{*-1}\beta + \sum_{j=1}^{p} \tau_j^2 .$$

Let $(KV)(\beta_0,\tau_0^2,\sigma_0^2)$ denote the expectation of $V(\cdot)$ with respect to the Mtd $k$ in (11), that is,

$$(KV)(\beta_0,\tau_0^2,\sigma_0^2) = \int_{\mathbb{R}_+}\int_{\mathbb{R}_+^p}\int_{\mathbb{R}^p} V(\beta,\tau^2,\sigma^2) k((\beta,\tau^2,\sigma^2)|(\beta_0,\tau_0^2,\sigma_0^2)) d\beta d\tau^2 d\sigma^2 \ .$$

We use $V(\beta,\tau^2,\sigma^2)$ to establish the following drift condition.

**Lemma 1.** *If* $n \geq 4$, *then there exists constants* $0 \leq \gamma < 1$ *and* $d > 0$ *such that*

$$(KV)(\beta_0,\tau_0^2,\sigma_0^2) \leq \gamma V(\beta_0,\tau_0^2,\sigma_0^2) + d \tag{16}$$

*for every* $(\beta_0,\tau_0^2,\sigma_0^2) \in \mathbb{R}^p \times \mathbb{R}_+^p \times \mathbb{R}_+$.

We also establish an associated minorization condition to the geometric drift condition (16). For every $L > 0$, let $B_{V,L} = \{(\beta,\tau^2,\sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+^p \times \mathbb{R}_+ : V(\beta,\tau^2,\sigma^2) \leq L\}$. Below we present the required minorization condition.

**Lemma 2.** *For every* $(\beta_0,\tau_0^2,\sigma_0^2) \in B_{V,L}$, *we have*

$$k((\beta,\tau^2,\sigma^2)|(\beta_0,\tau_0^2,\sigma_0^2)) \geq \epsilon u(\beta,\tau^2,\sigma^2), \tag{17}$$

*where* $u(\cdot)$ *is a probability density function on* $\mathbb{R}^p \times \mathbb{R}_+^p \times \mathbb{R}_+$, *and* $\epsilon \equiv \epsilon(V,L) \in (0,1)$ *is a constant.*

The drift and minorization conditions in Lemma 1 and Lemma 2 imply that the EN Gibbs sampler is geometrically ergodic, that is, Proposition 1 holds (Rosenthal, 1995). Moreover, Rosenthal's (1995) Theorem 12 provides a computable upper bound on $W(\beta, \tau^2, \sigma^2)r^m$ in (15) that involves the functions and constants from the drift and minorization conditions. Since the drift function $V(\cdot)$ depends on $\lambda_2$ and the constants $\gamma, d, \epsilon$, and the pdf $u(\cdot)$ may depend on $\lambda_1, \lambda_2$, the upper bound may also depend on $\lambda_1, \lambda_2$. The proofs of Lemma 1 and Lemma 2 are given in the Web based supplementary materials.

# 3 Selection of tuning parameters and computation of regularization paths using generalized importance sampling methods

In this section, we propose a method for selecting the penalty parameters of the penalized regression methods and computing the regularization paths. For brevity we describe the proposed method in the context of Bayesian elastic net model.

## 3.1 Selection of the tuning parameters

Here we consider an empirical Bayes approach for making inference on the hyperparameters $\lambda_1$ and $\lambda_2$ in the Bayesian elastic net model (6). For notational convenience we denote $(\lambda_1, \lambda_2)$ by $\lambda$. In particular, we estimate $\lambda \equiv (\lambda_1, \lambda_2)$ by

$$\hat{\lambda} = \operatorname*{argmax}_{\lambda \in \Lambda} m_\lambda(\boldsymbol{y}) \ ,$$

where $m_\lambda(\boldsymbol{y}) \equiv m_{\lambda_1,\lambda_2}(\boldsymbol{y})$ is the marginal density defined in (13), and $\Lambda = \mathbb{R}_+ \times \mathbb{R}_+$. Note that, $m_\lambda(\boldsymbol{y})$ is not available in closed form and in order to select the value of $\lambda$ which maximizes $m_\lambda(\boldsymbol{y})$, one can estimate $m_\lambda(\boldsymbol{y})$ for several (large number of) values of $\lambda$ and compute $\hat{\lambda}$ using these estimated values. Monte Carlo estimation of the marginal likelihood is extremely difficult, for example, Newton and Raftery's (1994) harmonic mean estimator is known to perform poorly (Wolpert and Schmidler, 2012). It is often much easier to estimate $\{am_\lambda(\boldsymbol{y}), \lambda \in \Lambda\}$ than $\{m_\lambda(\boldsymbol{y}), \lambda \in \Lambda\}$ for an appropriately chosen constant $a$. We calculate and subsequently compare the values of $B_{\lambda,\lambda^0} := m_\lambda(\boldsymbol{y})/m_{\lambda^0}(\boldsymbol{y})$, where $\lambda^0$ is a suitably chosen fixed value of $\lambda$. (Here $a$ is simply $1/m_{\lambda^0}(\boldsymbol{y})$.) Note that $B_{\lambda,\lambda^0}$ is the Bayes factor (BF) of the model indexed by $\lambda$ versus the model indexed by $\lambda^0$. Since $\hat{\lambda} = \operatorname*{argmax}_{\lambda \in \Lambda} B_{\lambda,\lambda^0}$, we would like to calculate and compare $B_{\lambda,\lambda^0}$ for a large number of values of $\lambda$.

Note that,

$$\begin{aligned} m_\lambda(\boldsymbol{y}) &= \int_{\mathbb{R}_+} \int_{\mathbb{R}_+^p} \int_{\mathbb{R}^p} f(\tilde{\boldsymbol{y}}|\beta, \sigma^2)\pi(\beta, \tau^2, \sigma^2|\lambda) d\beta d\tau^2 d\sigma^2 \\ &= m_{\lambda^0}(\boldsymbol{y}) \int_{\mathbb{R}_+} \int_{\mathbb{R}_+^p} \int_{\mathbb{R}^p} \frac{\pi(\beta, \tau^2, \sigma^2|\lambda)}{\pi(\beta, \tau^2, \sigma^2|\lambda^0)} \pi(\beta, \tau^2, \sigma^2|\boldsymbol{y}, \lambda^0) d\beta d\tau^2 d\sigma^2. \end{aligned}$$

Let $\{(\beta_m, \tau_m^2, \sigma_m^2)\}_{m\geq 1}$ be the Gibbs Markov chain mentioned in Section 2.2 with invariant density $\pi(\beta, \tau^2, \sigma^2|\boldsymbol{y}, \lambda^0)$, then by ergodic theorem we have a simple consistent estimator of $B_{\lambda,\lambda^0}$,

$$\frac{1}{M}\sum_{i=1}^{M}\frac{\pi(\beta_m, \tau_m^2, \sigma_m^2|\lambda)}{\pi(\beta_m, \tau_m^2, \sigma_m^2|\lambda^0)} \overset{a.s.}{\to} \frac{m_\lambda(\boldsymbol{y})}{m_{\lambda^0}(\boldsymbol{y})}, \tag{18}$$

as $M \to \infty$. Note that in (18) a single Markov chain $\{\beta_m, \tau_m^2, \sigma_m^2\}_{m=1}^M$ with stationary density $\pi(\beta, \tau^2, \sigma^2|\boldsymbol{y}, \lambda^0)$ is used to estimate $B_{\lambda,\lambda^0}$ for different values of $\lambda$. In general the naive importance sampling estimator (18) is very unstable; indeed when $\lambda$ is not close to $\lambda^0$ only a few values of the ratios $\pi(\beta_m, \tau_m^2, \sigma_m^2|\lambda)/\pi(\beta_m, \tau_m^2, \sigma_m^2|\lambda^0)$ may dominate the estimator.

Recently, Doss (2010) describes a method for efficiently computing large families of BFs (see also Geyer, 1996). The basic idea is to appropriately chose $k$ reference points $\lambda^0, \lambda^1, \ldots, \lambda^{k-1} \in \Lambda$ and replace $\pi(\beta_m, \tau_m^2, \sigma_m^2|\lambda^0)$ in (18) with a linear combinations of the prior densities evaluated at these skeleton points. We now describe this generalized importance sampling (GIS) method based on multiple Markov chains.

Let $\{\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}\}_{l=1}^{M_j}$ be a Markov chain with stationary density $\pi(\beta, \tau^2, \sigma^2|\boldsymbol{y}, \lambda^j)$ for $j = 0\ldots, k-1$. Let $M = \sum_{j=0}^{k-1} M_j$, and $M_j/M \to \alpha_j \in (0, 1)$ for $j = 0, \ldots, k-1$ with $\sum_{j=0}^{k-1}\alpha_j = 1$. Define $r^i = m_{\lambda^i}(\boldsymbol{y})/m_{\lambda^0}(\boldsymbol{y})$ for $i = 0, 1, \ldots, k-1$, with $r^0 = 1$ and $r = (r^0, r^1, \ldots, r^{k-1})$. Note that, by the ergodic theorem,

$$\begin{aligned}
\hat{B}_{\lambda,\lambda^0}(r) &\equiv \sum_{j=0}^{k-1}\sum_{l=1}^{M_j}\frac{\pi(\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}|\lambda)}{\sum_{i=0}^{k-1} M_i\pi(\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}|\lambda^i)/r^i}\\
&= \frac{1}{m_{\lambda^0}(\boldsymbol{y})}\sum_{j=0}^{k-1}\frac{1}{M_j}\sum_{l=1}^{M_j}\frac{\frac{M_j}{M}\pi(\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}|\lambda)}{\sum_{i=0}^{k-1}\frac{M_i}{M}\pi(\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}|\lambda^i)/m_{\lambda^i}(\boldsymbol{y})}\\
&\xrightarrow{\text{a.s.}} \frac{m_\lambda(\boldsymbol{y})}{m_{\lambda^0}(\boldsymbol{y})}\sum_{j=0}^{k-1}\int_{\mathbb{R}_+}\int_{\mathbb{R}_+^p}\int_{\mathbb{R}^p}\frac{\alpha_j\pi(\beta, \tau^2, \sigma^2|\boldsymbol{y}, \lambda)}{\sum_{i=0}^{k-1}\alpha_i\pi(\beta, \tau^2, \sigma^2|\boldsymbol{y}, \lambda^i)}\pi(\beta, \tau^2, \sigma^2|\boldsymbol{y}, \lambda^j)d\beta d\tau^2 d\sigma^2 \qquad (19)\\
&= \frac{m_\lambda(\boldsymbol{y})}{m_{\lambda^0}(\boldsymbol{y})} = B_{\lambda,\lambda^0}.
\end{aligned}$$

Although $B_{\lambda,\lambda^0}$ is consistently estimated by $\hat{B}_{\lambda,\lambda^0}(r)$, in practice we can not compute $\hat{B}_{\lambda,\lambda^0}(r)$ as $r$ is not available in closed form. Following Doss (2010) we consider the estimator

$$\hat{B}_{\lambda,\lambda^0}(\hat{r}) \equiv \sum_{j=0}^{k-1}\sum_{l=1}^{M_j}\frac{\pi(\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}|\lambda)}{\sum_{i=0}^{k-1} M_i\pi(\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}|\lambda^i)/\hat{r}^i}, \tag{20}$$

where $\hat{r}^0 = 1$ $\hat{r}^i$, $i = 1, 2, \ldots, k-1$ are consistent estimator of $r^i$'s obtained by the "reverse logistic regression" method proposed by Geyer (1994). In (20), naive weights (proportional to sample sizes) are used for the $k$ reference densities, although one could use more general weights (Roy et al., 2015).

Note that, (20) can be used to estimate the entire family of BFs $\{B_{\lambda,\lambda^0}, \lambda \in \Lambda\}$. The function $\hat{B}_{\lambda,\lambda^0}(\hat{r})$ can also be optimized to find $\hat{\lambda}$ instead of estimating $B_{\lambda,\lambda^0}$ for large number of values of $\lambda$. We use a quasi-Newton optimization procedure to maximize $\hat{B}_{\lambda,\lambda^0}(\hat{r})$ and estimate $\lambda$. Since the prior density of $\sigma^2$ does not depend on $\lambda$, the estimator (20) becomes

$$\begin{aligned}
&\hat{B}_{\lambda,\lambda^0}(\hat{r}) \\
&\quad\equiv \sum_{j=0}^{k-1}\sum_{l=1}^{M_j}\Bigg\{\Bigg[\exp\left(-\lambda_2\sum_{s=1}^{p}(\beta_s^{(j;l)})^2/[2\sigma^{2(j;l)}]\right)\exp\left(-\lambda_1^2\sum_{s=1}^{p}\tau_s^{2(j;l)}/2\right)\Big/[C(\lambda_1,\lambda_2)]^p\Bigg] \\
&\qquad\times\Bigg[\sum_{i=0}^{k-1}\Bigg\{M_i\exp\left(-\lambda_2^i\sum_{s=1}^{p}(\beta_s^{(j;l)})^2/[2\sigma^{2(j;l)}]\right) \\
&\qquad\times\exp\left(-[\lambda_1^i]^2\sum_{s=1}^{p}\tau_s^{2(j;l)}/2\right)\Big/\left(\hat{r}^i[C(\lambda_1^i,\lambda_2^i)]^p\right)\Bigg\}\Bigg]^{-1}\Bigg\}.
\end{aligned}\tag{21}$$

We use the two stage procedure mentioned in Doss (2010) for computing $\{\hat{B}_{\lambda,\lambda^0}(\hat{r}), \lambda \in \Lambda\}$. In stage I, we draw large MCMC samples $\{\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}\}_{l=1}^{M_j}$ from $\pi(\beta, \tau^2, \sigma^2|\boldsymbol{y}, \lambda^j)$ for each $j = 0\ldots, k-1$. Since the EN Gibbs sampler does not involve any computationally demanding calculations, these MCMC samples can be obtained quickly. We estimate $\hat{r}$ by Geyer's (1994) reverse logistic regression method using these samples. Independently of stage I, in stage II, we get *new* MCMC samples from $\pi(\beta, \tau^2, \sigma^2|\boldsymbol{y}, \lambda^j)$ for each $j = 0\ldots, k-1$ and use these stage II samples for estimating $B_{\lambda,\lambda^0}$ using $\hat{B}_{\lambda,\lambda^0}(\hat{r})$ given in (20). The reason for using this two stage procedure is that the amount of computation required to calculate $\hat{B}_{\lambda,\lambda^0}(\hat{r})$ is linear in $M$ and this rules out large $M$ in stage II (Doss, 2010). On the other hand, it is desirable to use large $M_j$ in stage I to estimate $r$ accurately. Roy et al. (2016) use this two-stage method for selecting correlation parameters and link function parameters in spatial generalized linear mixed models (See also Roy, 2014, for another application.). They use Roy et al.'s (2015) standard error estimates of $\hat{B}_{\lambda,\lambda^0}(\hat{r})$ for choosing good values of $k$, and the skeleton points $\lambda^0, \ldots, \lambda^{k-1}$ (see also Buta and Doss, 2011, p. 2671). In order to use Roy et al.'s (2016) method for choosing the skeleton points, it is required to establish a CLT result for the estimator $\hat{B}_{\lambda,\lambda^0}(\hat{r})$ defined in (20). A CLT for $\hat{B}_{\lambda,\lambda^0}(\hat{r})$ is also required for constructing (point wise) confidence interval for the BFs $B_{\lambda,\lambda^0}$. Buta and Doss's (2011) Theorem 1 presents the conditions under which $\hat{B}_{\lambda,\lambda^0}(\hat{r})$ has a CLT. The Markov chains $\{\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}\}_{l=1}^{M_j}$ need to be geometrically ergodic—which we have shown in Section 2.2. Define

$$Z(\theta) \equiv \frac{\exp(-\lambda_2\sum_{s=1}^{p}\beta_s^2/[2\sigma^2])\exp(-\lambda_1^2\sum_{s=1}^{p}\tau_s^2/2)/[C(\lambda_1,\lambda_2)]^p}{\sum_{i=0}^{k-1}\{\alpha_i\exp(-\lambda_2^i\sum_{s=1}^{p}\beta_s^2/[2\sigma^2])\exp(-[\lambda_1^i]^2\sum_{s=1}^{p}\tau_s^2/2)/(r^i[C(\lambda_1^i,\lambda_2^i)]^p)\}}. \tag{22}$$

Let $E^{\lambda}Z$ denote the posterior mean of the function $Z$ with respect to the density (12). Another condition of Buta and Doss's (2011) Theorem 1 is that there exists $\epsilon > 0$ such that $E^{\lambda^l}(Z^{2+\epsilon}) < \infty$ for $l = 0, 1, \ldots, k-1$. The following lemma provides a condition under which $Z(\cdot)$ is a bounded function, in particular, it has moments of all orders.

**Lemma 3.** *If there exists at least one* $i' \in \{0, 1, \ldots, k-1\}$ *such that* $\lambda_1^{i'} < \lambda_1$ *and* $\lambda_2^{i'} < \lambda_2$*, then* $Z(\theta)$ *is a bounded function of* $(\beta, \tau^2, \sigma^2)$*.*

Proof of Lemma 3 is given in the Web based supplementary materials.

Note that for the original lasso model (3), since the prior distributions of $\beta$ and $\sigma^2$ do not depend on the penalty parameter $\lambda$, the estimator (21) for lasso becomes

$$\sum_{j=0}^{k-1} \sum_{l=1}^{M_j} \frac{\lambda^{2p} \exp(-\lambda^2 \sum_{s=1}^p \tau_s^{2(j;l)}/2)}{\sum_{i=0}^{k-1} M_i (\lambda^i)^{2p} \exp(-(\lambda^i)^2 \sum_{s=1}^p \tau_s^{2(j;l)}/2)/\hat{\hat{r}}^i},$$

where $\hat{\hat{r}}^i$'s are the reverse logistic regression estimates for lasso model. For lasso a similar result as in Lemma 3 holds if there exists at least one $i' \in \{0, 1, \ldots, k-1\}$ such that $\lambda^{i'} < \lambda$.

## 3.2 Computing regularization paths

In penalized regression methods, a plot of the estimated regression coefficients ($\hat{\beta}$ or rather $\hat{\beta}_\lambda$) as a function of the penalty parameter $\lambda$ is useful for displaying the amount of shrinkage. In this section, we show that the proposed GIS method can be used to produce regularization path plots of Bayesian EN. Doing similar calculations as in (19) we have

$$\begin{aligned} &\sum_{j=0}^{k-1} \sum_{l=1}^{M_j} \frac{g(\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}) \pi(\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}|\lambda)}{\sum_{i=0}^{k-1} M_i \pi(\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}|\lambda^i)/r^i} \\ &\xrightarrow{\text{a.s.}} B_{\lambda,\lambda^0} \int_{\mathbb{R}_+} \int_{\mathbb{R}_+^p} \int_{\mathbb{R}^p} g(\beta, \tau^2, \sigma^2) \pi(\beta, \tau^2, \sigma^2 | \boldsymbol{y}, \lambda) d\beta d\tau^2 d\sigma^2 = B_{\lambda,\lambda^0} E^\lambda g. \end{aligned} \tag{23}$$

Hence from (19) and (23) we have

$$\hat{\eta}^{[g]} = \frac{\sum_{j=0}^{k-1} \sum_{l=1}^{M_j} \frac{g(\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}) \pi(\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}|\lambda)}{\sum_{i=0}^{k-1} M_i \pi(\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}|\lambda^i)/\hat{r}^i}}{\sum_{j=0}^{k-1} \sum_{l=1}^{M_j} \frac{\pi(\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}|\lambda)}{\sum_{i=0}^{k-1} M_i \pi(\beta^{(j;l)}, \tau^{2(j;l)}, \sigma^{2(j;l)}|\lambda^i)/\hat{r}^i}} \xrightarrow{\text{a.s.}} E^\lambda g. \tag{24}$$

Since the EN Gibbs sampler is geometrically ergodic, Roy et al.'s (2015) Theorem 3 can be used to compute standard errors of $\hat{\eta}^{[g]}$. To use this theorem, we need to show that there exists $\epsilon > 0$ such that $E^{\lambda^l}(|gZ|^{2+\epsilon}) < \infty$ for $l = 0, 1, \ldots, k-1$, where $Z$ is defined in (22). The following corollary follows from Lemma 3.

**Corollary 1.** *If there exists at least one* $i' \in \{0, 1, \ldots, k-1\}$ *such that* $\lambda_1^{i'} < \lambda_1$ *and* $\lambda_2^{i'} < \lambda_2$*, and* $E^{\lambda^l}|g|^{2+\epsilon} < \infty$ *then* $E^{\lambda^l}|gZ|^{2+\epsilon} < \infty$*.*

We use the GIS estimator $\hat{\eta}^{[g]}$ to efficiently compute the profile plot of $E^\lambda \beta$, the posterior mean of the regression coefficients for Bayesian lasso and EN. Also Roy et al.'s (2015) Theorem 3 is used to produce asymptotically valid confidence band (point wise) for these regularization path estimates. A step-by-step summary of the proposed inferential procedure is given in the Web based supplementary materials.

# 4 Applications

In this section we perform simulations to compare our empirical Bayes Lasso and empirical Bayes Elastic-Net models fitted using GIS with other competing models. We also apply these models to two real data sets.

## 4.1 Simulation study

We compare the performance of our two models against several frequentist and Bayesian versions of lasso and elastic-net. They are as follows i) lasso (Tibshirani, 1996), ii) elastic-net (EN) (Zou and Hastie, 2005), iii) full hierarchical Bayes lasso (FB-lasso) (Park and Casella, 2008), iv) full hierarchical Bayes elastic-net (FB-EN) (Kyung et al., 2010), v) empirical Bayes lasso fitted using Monte Carlo EM (EM-lasso) (Park and Casella, 2008), and vi) empirical Bayes elastic-net fitted using Monte Carlo EM (EM-EN) (Li and Lin, 2010). For each simulation scenario, we generate a training set and a test set separately. We fit GIS-lasso, GIS-EN and all other competing models (i–vi) using the training data and obtain the parameter estimates. Then we use the fitted models from the training set to calculate the prediction accuracy in the test set. The five simulation setups we choose to use are very similar to those in the original EN paper (Zou and Hastie, 2005) and in the OSCAR paper (Bondell and Reich, 2008).

The five simulation scenarios are as follows:

1. **Scenario 1:** $\boldsymbol{y} = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$, where we set $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$, $\sigma = 3$, and $X_k \overset{iid}{\sim} MVN(0, \Sigma)$ for $k = 1, 2, \ldots, p$. The covariance matrix is set as $\Sigma = (\sigma_{ij}) = 0.5^{|i-j|}$. We take training set sample size $= 20$, and test set sample size $= 200$.

2. **Scenario 2:** $\boldsymbol{y} = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$, where we set $\beta = (3, 0, 0, 1.5, 0, 0, 0, 2)$, $\sigma = 3$, and $X_k \overset{iid}{\sim} MVN(0, \Sigma)$ for $k = 1, 2, \ldots, p$. The covariance matrix is set as $\Sigma = (\sigma_{ij}) = 0.7^{|i-j|}$. Training set sample size $= 20$; Test set sample size $= 200$.

3. **Scenario 3:** Same as Simulation 1, except $\beta = (\underbrace{0.85, \ldots, 0.85}_{8})$.

4. **Scenario 4:** $\boldsymbol{y} = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$, where we set $\beta = (\underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10})$, $\sigma = 15$, and $X_k \overset{iid}{\sim} MVN(0, \Sigma)$ for $k = 1, 2, \ldots, p$. The covariance matrix is set such that all variances are one and pairwise correlation is 0.5. Training set sample size $= 100$; Test set sample size $= 400$.

5. **Scenario 5:** $\boldsymbol{y} = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2 I)$, where we set $\beta = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25})$ and $\sigma = 15$. Letting $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, 0.16)$ for $i = 1, \ldots, 15$, the predictors are generated

from

$$\begin{cases} X_i = Z_1 + \epsilon_i, & Z_1 \sim \mathrm{N}(0,1), \quad i = 1, \ldots, 5 \\ X_i = Z_2 + \epsilon_i, & Z_2 \sim \mathrm{N}(0,1), \quad i = 6, \ldots, 10 \\ X_i = Z_3 + \epsilon_i, & Z_3 \sim \mathrm{N}(0,1), \quad i = 11, \ldots, 15 \\ X_i \overset{iid}{\sim} \mathrm{N}(0,1), & i = 16, \ldots, 40. \end{cases}$$

Training set sample size = 100; Test set sample size = 400.

For each of the five above simulation scenarios, we generate 100 training data sets and 100 test data sets from the respective models.

The tuning parameters for frequentist lasso and elastic-net are selected using five fold cross-validation on the training set. We use the *glmnet*() package in R to fit the frequentist lasso and elastic-net models. In FB-lasso and FB-EN the prior distributions for the tuning parameters for $\lambda$, $\lambda_1$, $\lambda_2$ are selected following the recommendations provided in Park and Casella (2008) and Kyung et al. (2010). To study if there is any effect of different choices of the prior distribution for the tuning parameters we have fitted the Bayesian lasso and Bayesian elastic-net under several choices of the priors for $\lambda$, $\lambda_1$ and $\lambda_2$. In FB-lasso we assign the priors (a) $\lambda^2 \sim Gamma(1, .01)$, (b) $\lambda^2 \sim Gamma(1, .1)$, and (c) $\lambda^2 \sim Gamma(1, 1)$. In FB-EN we assign the priors (a) $\lambda_1^2 \sim Gamma(1, .01)$ and $\lambda_2 \sim Gamma(1, .01)$, (b) $\lambda_1^2 \sim Gamma(1, .1)$ and $\lambda_2 \sim Gamma(1, .1)$, and (c) $\lambda_1^2 \sim Gamma(1, 1)$ and $\lambda_2 \sim Gamma(1, 1)$. All these choices provide us near diffuse priors for the tuning parameters (Kyung et al., 2010) and thus to some extent can ensure the objectivity of the analysis. For EM-lasso and EM-EN the tuning parameters are estimated through maximizing the marginal likelihood, which is implemented with the EM-Gibbs algorithm proposed by Kyung et al. (2010). The FB-lasso, FB-EN, EM-lasso, and EM-EN are all fitted using the computer codes obtained from Prof. Minjung Kyung (of Kyung et al., 2010). The convergence of the MCMC chains are checked via trace plots.

Our GIS-lasso and GIS-EN do not require any prior distribution assignment for the tuning parameters. The improper prior $1/\sigma^2$ is chosen for $\sigma^2$. Due to the use of this improper prior, $m_\lambda(\boldsymbol{y})$ is not uniquely defined. Nevertheless, the Bayes factor among any two models, say $m_\lambda(\boldsymbol{y})/m_{\lambda^0}(\boldsymbol{y})$, is well-defined because the same improper prior is assigned to the shared parameters of the two models (see e.g. Kass and Raftery (1995, Section 5) and Liang et al. (2008, Section 2)). We need to specify the skeleton points needed to calculate the marginal likelihood. In all five simulation studies we have around 10 to 20 skeleton points and they produce highly satisfactory results. We choose the skeleton points using the method described in Roy et al. (2016). The two-stage procedure described in Section 3.1 is used to estimate the entire profile of the marginal likelihood of the tuning parameters. Then the marginal likelihood is maximized to find the estimates of the tuning parameters. In the first stage of generating parameter samples for selected skeleton points we ran MCMC chain of length 10000 with first 5000 as burn-in. In the second stage we ran MCMC chains of length 1000 iteration with the first 500 as burn in. In all our simulation settings and the two real data analysis these choices of MCMC chain lengths produced fast and accurate estimates. Finally, MCMC

| Simulation | | lasso | FB-lasso | | | EM-lasso | GIS-lasso |
|---|---|---|---|---|---|---|---|
| | | | (1,.01) | (1,.1) | (1,1) | | |
| 1 | Ave root-MSE | 3.76 | 3.63 | 3.60 | 3.67 | 3.58 | 3.49 |
| | SD | 0.52 | 0.37 | 0.34 | 0.36 | 0.37 | 0.36 |
| | Ave $\hat{\lambda}$ | 0.48 | 3.99 | 2.75 | 1.43 | 2.69 | 2.65 |
| | Range of $\hat{\lambda}$ | [.002,2.89] | [.96,8.03] | [.80,3.74] | [.69,1.63] | [.54,5.74] | [.55,5.58] |
| 2 | Ave root-MSE | 3.75 | 3.59 | 3.63 | 3.59 | 3.57 | 3.56 |
| | SD | .53 | .36 | .33 | .37 | .35 | .36 |
| | Ave $\hat{\lambda}$ | .36 | 3.73 | 2.69 | 1.36 | 2.57 | 2.46 |
| | Range of $\hat{\lambda}$ | [.002,2.18] | [1.06,8.80] | [1.07,3.90] | [.84,1.62] | [.78,9.12] | [.71,7.98] |
| 3 | Ave root-MSE | 3.90 | 3.47 | 3.47 | 3.61 | 3.45 | 3.44 |
| | SD | .73 | .29 | .30 | .45 | .32 | .30 |
| | Ave $\hat{\lambda}$ | .29 | 4.16 | 2.89 | 1.45 | 2.90 | 2.88 |
| | Range of $\hat{\lambda}$ | [.005,1.81] | [1.89, 8.99] | [1.70,4.04] | [1.11,1.69] | [1.31,8.23] | [1.32,6.47] |
| 4 | Ave root-MSE | 16.66 | 16.33 | 17.41 | 18.77 | 15.91 | 15.83 |
| | SD | .75 | .68 | .70 | .81 | .68 | .67 |
| | Ave $\hat{\lambda}$ | 1.41 | 8.89 | 6.77 | 3.12 | 8.73 | 8.85 |
| | Range of $\hat{\lambda}$ | [.43,2.90] | [6.67,11.08] | [5.27,7.12] | [2.81,3.31] | [6.35,11,64] | [6.30,14.23] |
| 5 | Ave root-MSE | 17.43 | 16.33 | 17.41 | 18.77 | 15.85 | 15.87 |
| | SD | 4.74 | 1.34 | 2.66 | 4.01 | .79 | .83 |
| | Ave $\hat{\lambda}$ | 3.23 | 15.41 | 7.93 | 3.44 | 29.46 | 25.70 |
| | Range of $\hat{\lambda}$ | [1.05,5.03] | [11.29,19.46] | [7.11,8.80] | [3.23,3.63] | [12.16,139.46] | [7.88,44.46] |

Table 1: lasso: Test MSEs of the simulation studies. The results are based on 100 replicated data sets.

samples from $\pi(\beta, \tau^2, \sigma^2|\boldsymbol{y}, \hat{\lambda})$, the posterior density corresponding to estimated tuning parameters are used to estimate other parameters $(\beta, \tau^2, \sigma^2)$ and predict $\boldsymbol{y}$ at new covariate vector $x$.

Each simulation scenario is repeated for 100 times and each time a separate training set and a test set are generated. In Tables 1 and 2 we report the average test set mean squared error of prediction (MSEP) along with their corresponding standard deviations for our two models and all other competing lasso and elastic-net models. The ranges of estimates of the tuning parameters are also included in Tables 1 and 2. We can clearly

| Simulation | | EN | FB-EN | | | EM-EN | GIS-EN |
|---|---|---|---|---|---|---|---|
| $\lambda_1$ priors | | | (1,.01) | (1,.1) | (1,1) | | |
| $\lambda_2$ priors | | | (1,.01) | (1,.1) | (1,1) | | |
| 1 | Ave root-MSE | 3.52 | 3.43 | 3.45 | 3.51 | 3.38 | 3.37 |
| | SD | 0.42 | 0.31 | 0.34 | 0.35 | 0.33 | 0.34 |
| | Ave $\hat{\lambda}_1$ | 0.33 | 2.79 | 1.57 | 1.13 | 2.16 | 2.21 |
| | Range of $\hat{\lambda}_1$ | [.003,2.67] | [.91,6.53] | [.86,3.45] | [.56,1.99] | [.68,5.07] | [.72,5.29] |
| | Ave $\hat{\lambda}_2$ | 0.88 | 3.77 | 2.79 | 2.43 | 3.16 | 3.30 |
| | Range of $\hat{\lambda}_2$ | [.008,3.45] | [.87,10.42] | [.97,6.55] | [.78,3.22] | [.57,6.07] | [.57,6.42] |
| 2 | Ave root-MSE | 3.47 | 3.49 | 3.56 | 3.59 | 3.42 | 3.36 |
| | SD | .51 | .38 | .37 | .34 | .38 | .32 |
| | Ave $\hat{\lambda}_1$ | .39 | 2.73 | 1.99 | 1.06 | 2.34 | 2.30 |
| | Range of $\hat{\lambda}_1$ | [.002,2.23] | [1.05,6.95] | [1.07,3.97] | [.74,1.82] | [.70,4.62] | [.75,4.87] |
| | Ave $\hat{\lambda}_2$ | .72 | 3.93 | 2.98 | 1.86 | 2.69 | 2.81 |
| | Range of $\hat{\lambda}_2$ | [.004,2.83] | [1.05,7.60] | [1.09,3.40] | [.79,2.92] | [.79,7.03] | [.86,6.80] |
| 3 | Ave root-MSE | 3.10 | 2.97 | 3.09 | 3.33 | 2.88 | 2.76 |
| | SD | .75 | .23 | .32 | .41 | .29 | .29 |
| | Ave $\hat{\lambda}_1$ | .32 | 2.26 | 1.77 | 1.04 | 2.35 | 2.53 |
| | Range of $\hat{\lambda}_1$ | [.004,1.65] | [1.55, 8.23] | [1.60,4.99] | [.95,1.48] | [1.76,8.44] | [1.80,7.39] |
| | Ave $\hat{\lambda}_2$ | 0.92 | 4.51 | 2.98 | 2.54 | 3.19 | 3.23 |
| | Range of $\hat{\lambda}_2$ | [.005,2.01] | [1.60, 9.41] | [1.98,5.14] | [1.87,3.09] | [1.95,10.88] | [1.96,9.94] |
| 4 | Ave root-MSE | 14.89 | 12.87 | 12.33 | 12.92 | 12.83 | 12.78 |
| | SD | .71 | .63 | .72 | .87 | .65 | .65 |
| | Ave $\hat{\lambda}_1$ | 1.90 | 6.42 | 4.37 | 3.02 | 6.75 | 6.76 |
| | Range of $\hat{\lambda}_1$ | [.39,3.90] | [4.60,14.18] | [4.70,6.33] | [2.11,3.89] | [3.59,15.99] | [3.64,15.91] |
| | Ave $\hat{\lambda}_2$ | 3.62 | 10.92 | 8.45 | 6.01 | 11.45 | 11.67 |
| | Range of $\hat{\lambda}_2$ | [2.45,5.92] | [7.52,16.81] | [4.79,10.55] | [4.53,8.30] | [8.34,14.99] | [8.68,14.44] |
| 5 | Ave root-MSE | 14.90 | 15.13 | 15.67 | 15.40 | 14.53 | 14.46 |
| | SD | 4.17 | 1.49 | 2.87 | 4.23 | .89 | .85 |
| | Ave $\hat{\lambda}_1$ | 2.97 | 10.22 | 8.48 | 4.16 | 14.32 | 14.18 |
| | Range of $\hat{\lambda}_1$ | [1.11,5.52] | [7.76,18.32] | [6.44,13.15] | [3.49,8.65] | [9.09,30.22] | [10.47,34.19] |
| | Ave $\hat{\lambda}_2$ | 5.60 | 15.23 | 10.27 | 6.08 | 19.10 | 18.79 |
| | Range of $\hat{\lambda}_2$ | [3.14,9.31] | [10.29,23.60] | [8.67,14.51] | [4.75,9.10] | [12.98,39.20] | [12.42,37.10] |

Table 2: Elastic-net: Test MSEs of the simulation studies. The results are based on 100 replicated data sets.

see from the table that our method GIS-lasso and GIS-EN are highly competitive under all five simulation scenarios. Under all five simulation scenarios both GIS-lasso and GIS-EN have significantly lower MSEP than the competing methods. Moreover, we can see that the tuning parameter estimates of full hierarchical Bayes lasso and elastic-net are quite sensitive to the choice of prior parameters. (Here we report posterior mean estimates of the tuning parameters for FB-lasso and FB-EN.) This points out the fact that although the recommended priors on the tuning parameters are supposed to provide objective analysis but in reality there is a noticeable difference in the estimates of the tuning parameters depending on the chosen prior. This makes the problem of choosing appropriate prior distribution for the tuning parameters extremely difficult. On the other hand EM-lasso and EM-EN tries to circumvent the problem of choosing the appropriate priors for the tuning parameters by maximizing marginal maximum likelihood through Monte Carlo EM algorithm. EM-lasso and EM-EN are painstakingly slow due to the nature of the Monte Carlo EM algorithm and the solution obtained by it is highly unstable and suboptimal. Our GIS-lasso and GIS-EN are considerably faster than their counterparts EM-lasso and EM-EN. Thus our GIS-lasso and GIS-EN give us three distinct advantages. Firstly it does not require any prior distribution specifications on the tuning parameters, secondly it estimates the full profile of the marginal likelihood of the tuning parameters and thirdly, it provides the whole solution path plots along with (point wise) confidence bands. Estimation of the tuning parameter and solution path are important because the choice of tuning parameter leads to the appropriate level of sparsity. In Figures 1–4 we have provided the full profile of the marginal likelihood and solution paths for the two real data examples.

## 4.2 Real data sets

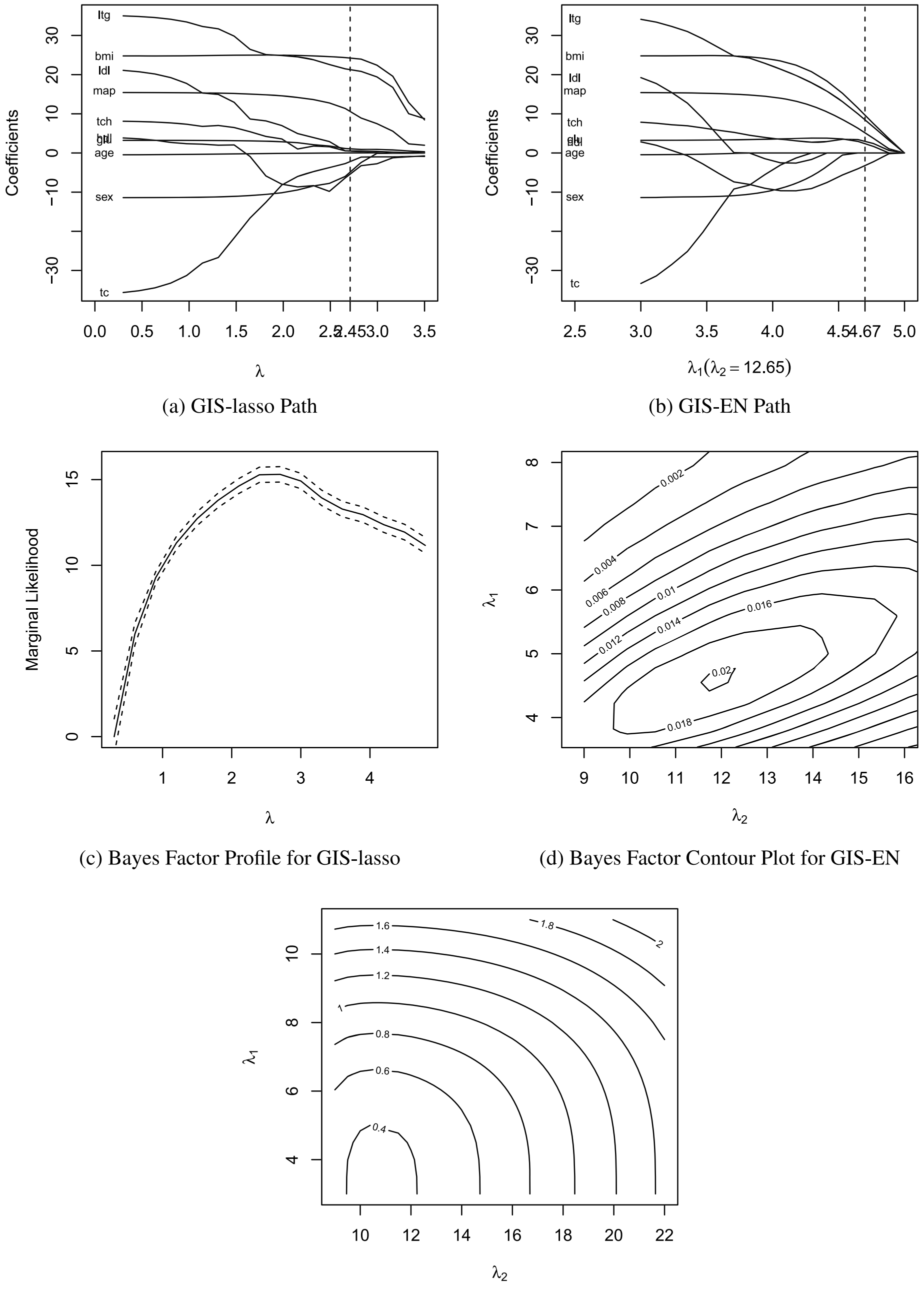
In this section, we apply our GIS-lasso and GIS-EN models on two real data sets, namely diabetes data (Efron et al., 2004) and soil data (Bondell and Reich, 2008). We also fit all other competing models as discussed in the simulation study section. The tuning parameters for frequentist models are all selected by a 5 fold cross validation like before. The FB-lasso is fitted by adopting the prior choice $\lambda^2 \sim Gamma(1, .01)$. The FB-EN is fitted with the prior choice $\lambda_1^2 \sim Gamma(1, .01)$ and $\lambda_2 \sim Gamma(1, .01)$. As in the simulation examples, we use posterior mean estimates of $\lambda$ $(\lambda_1, \lambda_2)$ for FB-lasso (FB-EN).

### Diabetes data

This data set arises from the study of 442 diabetes patients (Efron et al., 2004). The predictor or baseline variables are age, sex, body mass index (bmi), average blood pressure, and six blood serum measurements. The response variable is a quantitative measure of disease progression in one year after measuring the baseline variables. The main goal here is to give a good prediction of the disease progression along with detecting important baseline variables. We randomly split the data with 300 observations in the training set and the remaining 142 observations in the test set. From the results reported in Table 3, we see that due to the presence of moderate correlation among the

| Method | Test Set MSE | Tuning Parameter Estimates |
|---|---|---|
| lasso | 0.586 | $\lambda = 0.77$ |
| FB-lasso | 0.563 | $\lambda = 2.17$ |
| EM-lasso | 0.517 | $\lambda = 2.42$ |
| GIS-lasso | 0.512 | $\lambda = 2.46$ |
| EN | 0.529 | $\lambda_1 = 0.31, \lambda_1 = 1.09$ |
| FB-EN | 0.543 | $\lambda_1 = 3.01, \lambda_2 = 16.23$ |
| EM-EN | 0.488 | $\lambda_1 = 4.61, \lambda_2 = 12.32$ |
| GIS-EN | 0.423 | $\lambda_1 = 4.67, \lambda_2 = 12.65$ |

Table 3: Diabetes Data.



(a) GIS-lasso Path

(b) GIS-EN Path

(c) Bayes Factor Profile for GIS-lasso

(d) Bayes Factor Contour Plot for GIS-EN

(e) SE Relative to BF Estimates for GIS-EN

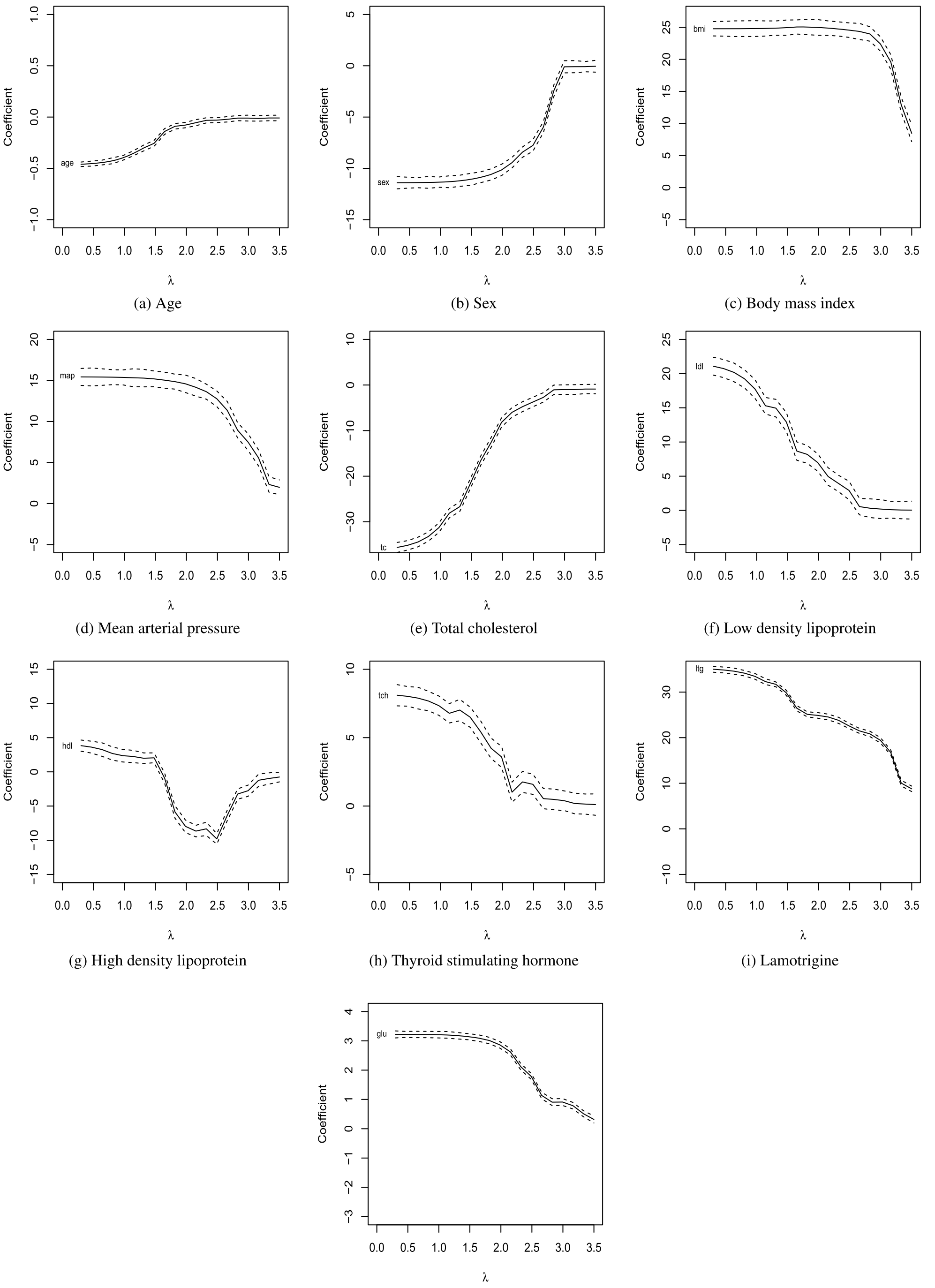
Figure 1: Diabetes Data.

covariate all of the lasso type models produced much higher MSE than the elastic-net based models. Our GIS-EN can effectively reduce the out of sample MSE by at least 7% than all other competing models. The GIS-lasso solution path in Figure 1(a) selects bmi, lamotrigine (ltg), mean arterial pressure (map), sex, total cholesterol (tc), and high density lipoprotein (hdl) to be important predictors. Whereas the solution path of GIS-EN Figure 1(b) identifies only ltg, bmi, map, hdl, and tch (thyroid stimulating hormone) to be important. We also include the full profile of the BFs in GIS-Lasso along with a 95% confidence band in Figure 1 subplot (c). In the case of GIS-EN since we have two tuning parameters ($\lambda_1$ and $\lambda_2$) we have included the contour plot of the BF and the corresponding estimate of the standard error (SE) relative to the Bayes Factor in subplots Figure 1(d) and Figure 1(e) respectively. For better illustration, in Figure 2 and Figure 3 we include the 95% confidence bands for GIS-lasso and GIS-EN solution paths for each individual coefficients. These plots illustrate the ability of our method to quantify the uncertainty over the full solution path of the coefficients.

In our GIS-Lasso and GIS-EN we can obtain the full profile of the BF and its corresponding SEs (Figure 1(c)–(e)) which give us a clear idea about the role of optimal choice of the tuning parameters. Moreover the plots (Figure 1(c)–(e)) can be used to provide us a guideline for selection of the skeleton points (see e.g. Roy et al., 2015). It is suggested that we select more skeleton points from the region of the tuning parameters whose corresponding SE relative to the Bayes Factor is relatively large than others.

### Soil data

This data set comes from a study of the association between soil characteristics and forest diversity in the Appalachian mountains of North Carolina (Bondell and Reich, 2008). The response of interest is the number of different plant species found in a plot. The covariates are 15 soil characteristics as follows: (1) % base saturation, (2) sum cations, (3) CEC, (4) calcium, (5) magnesium, (6) potassium, (7) sodium, (8) phosphorous, (9) copper, (10) zinc, (11) manganese, (12) humic matter, (13) density, (14) soil pH, and (15) exchangeable acidity. A more detailed description can be obtained from (Bondell and Reich, 2008). In this study, we have only 20 samples. In this data set several predictors are very highly correlated (absolute correlation $> 0.90$). For example, we see predictor variables 1 to 5 are all very strongly correlated. The correlation between sodium and phosphorous is also very high and there is a strong negative correlation between soil PH and exchangeable acidity. Due to strong correlation structure among the covariates, it would be very useful to group together the variables that are strongly correlated. Therefore lasso and all Bayesian counter parts of lasso which cannot take into account the underlying correlation structure will give sub-optimal result. In Table 4 we report leave one out cross-validation MSE. Under the presence of strong correlation among the predictor variables our GIS-EN resulted in lowest MSE among all. Comparing with all lassos our GIS-EN lowered the CV-error by more than 10%. On the other hand in comparison to the existing elastic-net and EM-EN our GIS-EN improved the accuracy by around 4%. It is important to mention here that although in



(a) Age

(b) Sex

(c) Body mass index

(d) Mean arterial pressure

(e) Total cholesterol

(f) Low density lipoprotein

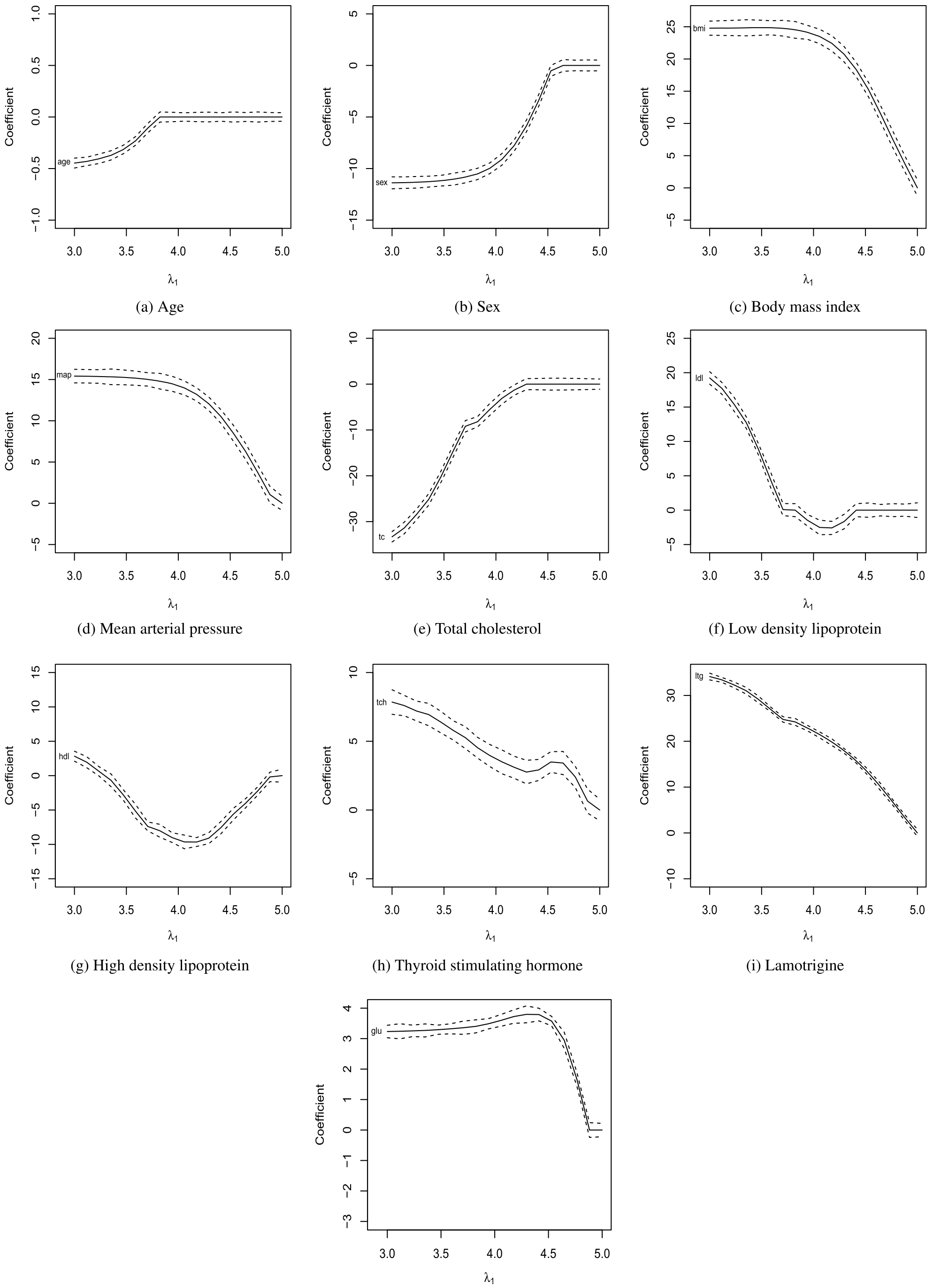
(g) High density lipoprotein

(h) Thyroid stimulating hormone

(i) Lamotrigine

(j) Glucose

Figure 2: Diabetes Data: 95% Confidence Interval for GIS-lasso Path.



(a) Age

(b) Sex

(c) Body mass index

(d) Mean arterial pressure

(e) Total cholesterol

(f) Low density lipoprotein

(g) High density lipoprotein

(h) Thyroid stimulating hormone

(i) Lamotrigine

(j) Glucose

Figure 3: Diabetes Data: 95% Confidence Interval for GIS-EN Path ($\lambda_2 = 12.65$).

comparison GIS-EN and EM-EN both seems to be equally performing well, but EM-EN is extremely slow to implement and thus its unusable in any practical application. On the other hand our GIS-EN is producing the same accurate result 15 times faster. In Figure 4 we list the solution paths for our GIS-lasso and GIS-EN. From the solution paths (Figures 4(a) and (b)) we can identify that GIS-lasso picked up 9 variables as important where as GIS-EN has selected only 6 as important. The 9 selected variables by our GIS-lasso are soil pH, copper, exchangeable acidity, potassium, density, humic Matter, calcium, zinc, and phosphorous. On the other hand the 6 selected variables by our GIS-EN are soil pH, copper, exchangeable acidity, potassium, density and humic Matter.

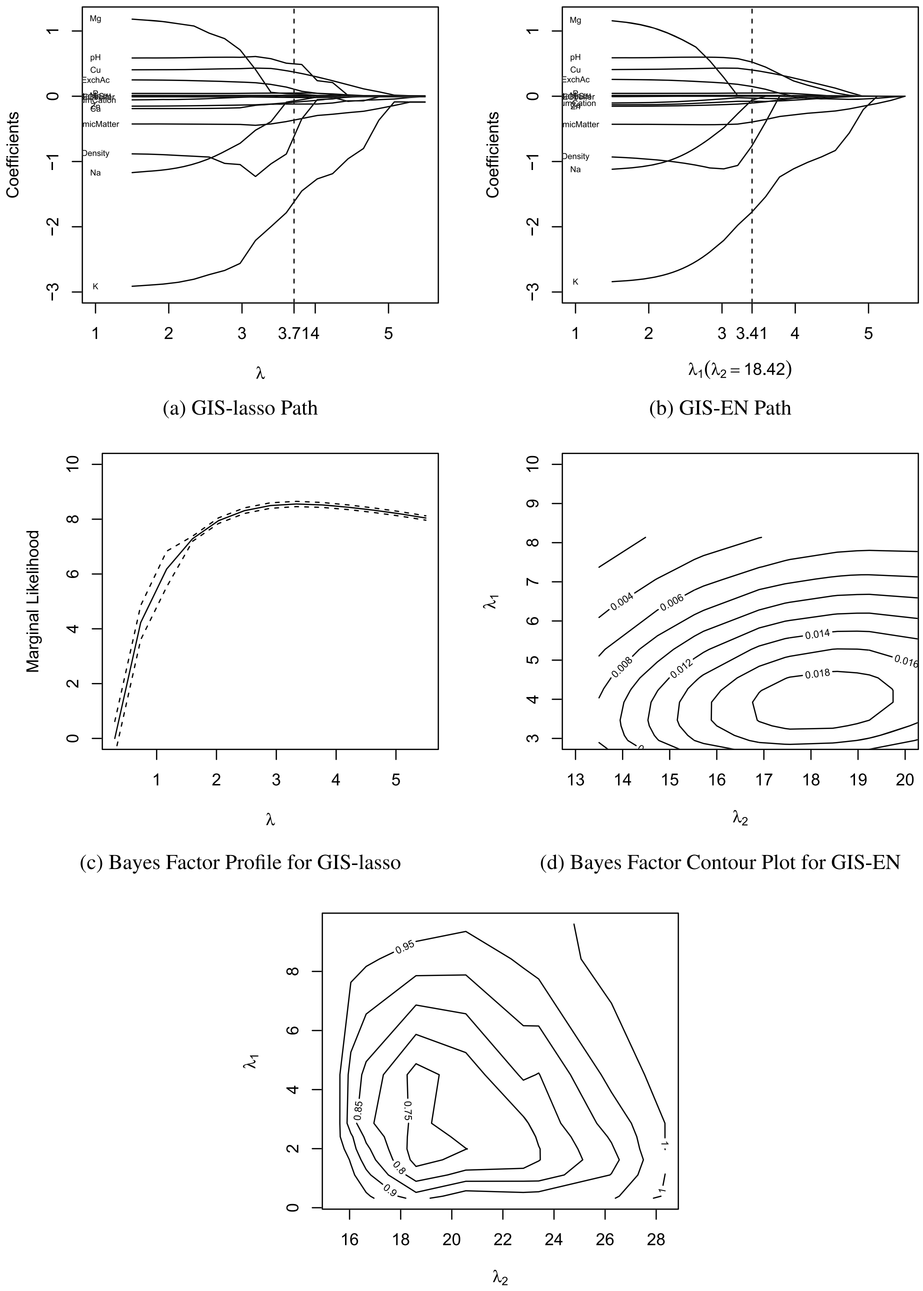
| Method | Test Set MSE | Tuning Parameter Estimates |
|---|---|---|
| lasso | 0.39 | $\lambda = 0.31$ |
| FB-lasso | 0.41 | $\lambda = 3.77$ |
| EM-lasso | 0.46 | $\lambda = 3.14$ |
| GIS-lasso | 0.43 | $\lambda = 3.72$ |
| EN | 0.31 | $\lambda_1 = 0.42, \lambda_1 = 1.77$ |
| FB-EN | 0.32 | $\lambda_1 = 2.11, \lambda_2 = 19.56$ |
| EM-EN | 0.27 | $\lambda_1 = 3.32, \lambda_2 = 17.99$ |
| GIS-EN | 0.27 | $\lambda_1 = 3.41, \lambda_2 = 18.40$ |

Table 4: Soil Data.

## 5 Discussions

Variable selection plays a fundamental role in modern statistical modeling. Classical approaches to deal with the variable selection problems are through regularization methods. These methods minimize the residual sum of squares subject to an imposed penalty. These methods are closely related to Bayesian methods as often the estimate given by a regularization method is indeed the posterior mode of a Bayesian model. The estimation of penalty parameters, although it is very important, can be tricky. In this article, an empirical Bayes methodology based on efficient importance sampling schemes is used for estimating the penalty parameters as well as the full solution paths for Bayesian lasso and EN. The Gibbs sampler for Bayesian EN is proved to be geometrically ergodic, which allows users to calculate asymptotically valid standard errors for the posterior estimates.

In many applications, interaction exists among the variables. In recent years, a variety of models is proposed in the literature to deal with the presence of highly correlated predictors (see e.g. Bondell and Reich, 2008; Liu et al., 2014). As a possible avenue for future work, it would be interesting to apply our proposed methods for estimating the penalty parameters and solution paths of these other grouped lasso methods. The GIS method can also be applied while using other shrinkage priors such as global local priors (Polson and Scott, 2010) (e.g. the horseshoe prior (Carvalho et al., 2010)) and non local priors (Johnson and Rossell, 2012).



(a) GIS-lasso Path

(b) GIS-EN Path

(c) Bayes Factor Profile for GIS-lasso

(d) Bayes Factor Contour Plot for GIS-EN

(e) SE Relative to BF Estimates for GIS-EN

Figure 4: Soil Data.

## Supplementary Material

Supplementary Material for "Selection of Tuning Parameters, Solution Paths and Standard Errors for Bayesian Lassos" (DOI: 10.1214/16-BA1025SUPP; .pdf). The online supplementary materials contain the proofs of lemmas. Also a summary of the steps involved in the estimation of the tuning parameters and the solution paths is given in the supplementary materials.

## References

Andrews, D. F. and Mallows, C. F. (1974). "Scale mixtures of normal distributions." *Journal of the Royal Statistical Society, Series B*, 36: 99–102. MR0359122. 756

Asmussen, S. and Glynn, P. W. (2011). "A new proof of convergence of MCMC via the ergodic theorem." *Statistics and Probability Letters*, 81: 1482–1485. MR2818658. doi: http://dx.doi.org/10.1016/j.spl.2011.05.004. 758

Bondell, H. and Reich, B. (2008). "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with Oscar." *Biometrics*, 64: 115–123. MR2422825. doi: http://dx.doi.org/10.1111/j.1541-0420.2007.00843.x. 754, 764, 768, 770, 773

Buta, E. and Doss, H. (2011). "Computational approaches for empirical Bayes methods and Bayesian sensitivity analysis." *The Annals of Statistics*, 39: 2658–2685. MR2906882. doi: http://dx.doi.org/10.1214/11-AOS913. 762

Carvalho, C., Polson, N., and Scott, J. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97: 465–480. MR2650751. doi: http://dx.doi.org/10.1093/biomet/asq017. 773

Doss, H. (2010). "Estimation of large families of Bayes factors from Markov chain output." *Statistica Sinica*, 20: 537–560. MR2682629. 761, 762

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). "Least angle regression." *The Annals of Statistics*, 32: 407–499. MR2060166. doi: http://dx.doi.org/10.1214/009053604000000067. 768

Fan, J. and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle property." *Journal of the American Statistical Association*, 96: 1348–1360. MR1946581. doi: http://dx.doi.org/10.1198/016214501753382273. 755

Flegal, J. M. and Jones, G. L. (2010). "Batch means and spectral variance estimators in Markov chain Monte Carlo." *The Annals of Statistics*, 38: 1034–1070. MR2604704. doi: http://dx.doi.org/10.1214/09-AOS735. 759

Geyer, C. J. (1994). "Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo." Technical report 568, School of Statistics, University of Minnesota. 761, 762

Geyer, C. J. (1996). *Markov Chain Monte Carlo in Practice*, chapter Estimation and optimization of functions, 241–258. Boca Raton, FL: Chapman and Hall/CRC Press. MR1397966. doi: http://dx.doi.org/10.1007/978-1-4899-4485-6. 761

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics*, 12: 55–67. 753

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer, New York, USA. MR3100153. doi: http://dx.doi.org/10.1007/978-1-4614-7138-7. 753

Johnson, V. E. and Rossell, D. (2012). "Bayesian model selection in high-dimensional settings." *Journal of the American Statistical Association*, 107: 649–660. MR2980074. doi: http://dx.doi.org/10.1080/01621459.2012.682536. 773

Jones, G. L. and Hobert, J. P. (2001). "Honest exploration of intractable probability distributions via Markov chain Monte Carlo." *Statistical Science*, 16: 312–34. MR1888447. doi: http://dx.doi.org/10.1214/ss/1015346317. 759

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90: 773–795. MR3363402. doi: http://dx.doi.org/10.1080/01621459.1995.10476572. 765

Khare, K. and Hobert, J. P. (2013). "Geometric ergodicity of Bayesian lasso." *Electronic Journal of Statistics*, 7: 2150–2163. MR3104915. doi: http://dx.doi.org/10.1214/13-EJS841. 755

Knight, K. and Fu, W. (2000). "Asymptotics for lasso-type estimators." *The Annals of Statistics*, 28: 1356–1378. MR1805787. doi: http://dx.doi.org/10.1214/aos/1015957397. 755

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). "Penalized regression, standard errors, and Bayesian lassos." *Bayesian Analysis*, 5: 369–412. MR2719657. doi: http://dx.doi.org/10.1214/10-BA607. 754, 755, 757, 764, 765

Lee, K. H., Chakraborty, S., and Sun, J. (2015). "Survival prediction and variable selection with simultaneous shrinkage and grouping priors." *Statistical Analysis and Data Mining*, 8: 114–127. MR3342983. doi: http://dx.doi.org/10.1002/sam.11266. 757

Li, Q. and Lin, N. (2010). "The Bayesian elastic net." *Bayesian Analysis*, 5: 151–170. MR2596439. doi: http://dx.doi.org/10.1214/10-BA506. 754, 755, 764

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of $g$-priors for Bayesian variable selection." *Journal of the American Statistical Association*, 103: 410–423. MR2420243. doi: http://dx.doi.org/10.1198/016214507000001337. 765

Liu, F., Chakraborty, S., Li, F., Liu, Y., and Lozano, A. C. (2014). "Bayesian regularization via graph Laplacian." *Bayesian Analysis*, 9: 449–474. MR3217003. doi: http://dx.doi.org/10.1214/14-BA860. 773

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. London: Springer Verlag. MR1287609. doi: http://dx.doi.org/10.1007/978-1-4471-3267-7. 758

Newton, M. and Raftery, A. (1994). "Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion)." *Journal of the Royal Statistical Society, Series B*, 56: 3–48. MR1257793. 760

Park, T. and Casella, G. (2008). "The Bayesian lasso." *Journal of the American Statistical Association*, 103: 681–686. MR2524001. doi: http://dx.doi.org/10.1198/016214508000000337. 755, 756, 757, 764, 765

Polson, N. and Scott, J. (2010). "Shrink globally, act locally: sparse Bayesian regularization and prediction." *Bayesian Statistics*, 9: 501–538. MR3204017. doi: http://dx.doi.org/10.1093/acprof:oso/9780199694587.003.0017. 773

Roberts, G. O. and Rosenthal, J. S. (1997). "Geometric ergodicity and hybrid Markov chains." *Electronic Communications in Probability*, 2: 13–25. MR1448322. doi: http://dx.doi.org/10.1214/ECP.v2-981. 759

Rosenthal, J. S. (1995). "Minorization conditions and convergence rates for Markov chain Monte Carlo." *Journal of the American Statistical Association*, 90: 558–566. MR1340509. 760

Roy, V. (2014). "Efficient estimation of the link function parameter in a robust Bayesian binary regression model." *Computational Statistics and Data Analysis*, 73: 87–102. MR3147976. doi: http://dx.doi.org/10.1016/j.csda.2013.11.013. 762

Roy, V. and Chakraborty, S. (2016). "Supplementary material for "Selection of tuning parameters, solution paths and standard errors for Bayesian lassos"." *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/16-BA1025SUPP. 756

Roy, V., Evangelou, E., and Zhu, Z. (2016). "Efficient estimation and prediction for the Bayesian binary spatial model with flexible link functions." *Biometrics*, 72: 289–298. 762, 765

Roy, V. and Hobert, J. P. (2007). "Convergence rates and asymptotic standard errors for MCMC algorithms for Bayesian probit regression." *Journal of the Royal Statistical Society*, Series B, 69: 607–623. MR2370071. doi: http://dx.doi.org/10.1111/j.1467-9868.2007.00602.x. 759

Roy, V., Tan, A., and Flegal, J. (2015). "Estimating standard errors for importance sampling estimators with multiple Markov chains." Technical report, Iowa State University. 755, 761, 762, 763, 770

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society, Series B*, 58: 267–288. MR1379242. 753, 755, 756, 764

Wolpert, R. L. and Schmidler, S. C. (2012). "$\alpha$-stable limit laws for harmonic mean estimators of marginal likelihoods." *Statistica Sinica*, 22: 1233–1251. MR2987490. doi: http://dx.doi.org/10.5705/ss.2010.221. 760

Yuan, M. and Lin, Y. (2006). "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society, Series B*, 68: 49–67. MR2212574. doi: http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x. 754

Zou, H. (2006). "The adaptive lasso and its oracle properties." *Journal of the American Statistical Association*, 101: 1418–1429. MR2279469. doi: http://dx.doi.org/10.1198/016214506000000735. 754

Zou, H. and Hastie, T. (2005). "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society, Series B*, 67: 301–320. MR2137327. doi: http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x. 754, 764

## Acknowledgments

The authors thank an associate editor and an editor for valuable comments that improved the manuscript. The second author thanks the University of Missouri Bioinformatics Consortium for letting him use their Dell EM64T Lewis cluster system. The second author was supported by the NSF Grant DMS-1106717.