Bayesian Analysis (2010) **5**, Number 1, pp. 151–170

The Bayesian Elastic Net

Qing Li* and Nan Lin†

Abstract. Elastic net (Zou and Hastie 2005) is a flexible regularization and variable selection method that uses a mixture of L_1 and L_2 penalties. It is particularly useful when there are much more predictors than the sample size. This paper proposes a Bayesian method to solve the elastic net model using a Gibbs sampler. While the marginal posterior mode of the regression coefficients is equivalent to estimates given by the non-Bayesian elastic net, the Bayesian elastic net has two major advantages. Firstly, as a Bayesian method, the distributional results on the estimates are straightforward, making the statistical inference easier. Secondly, it chooses the two penalty parameters simultaneously, avoiding the "double shrinkage problem" in the elastic net method. Real data examples and simulation studies show that the Bayesian elastic net behaves comparably in prediction accuracy but performs better in variable selection.

Keywords: Bayesian analysis, elastic net, Gibbs sampler, regularization, variable selection.

1 Introduction

Regression regularization methods are developed to carry out parameter estimation and variable selection simultaneously. Tibshirani (1996) proposed the lasso estimator which estimates the linear regression coefficients through an L_1-penalized least squares criterion. Due to the nature of the L_1-penalty, the lasso often yields solutions with some components being exactly 0. The lasso estimates can be solved efficiently by the LARS algorithm proposed by Efron et al. (2004), in which the order of computation load is the same as that of a single ordinary least squares (OLS) fit. While demonstrating promising performance for many problems, the lasso estimator does have some shortcomings. Zou and Hastie (2005) emphasized three inherent drawbacks of the lasso estimator. Firstly, due to the nature of the convex optimization problem, the lasso method cannot select more predictors than the sample size. But in practice there are often studies that involve much more predictors than the sample size, e.g. microarray data analysis (Guyon et al. 2002). Secondly, when there is some group structure among the predictors, the lasso estimator usually selects only one predictor from a group while ignoring others. Thirdly, when the predictors are highly correlated, the lasso estimator performs unsatisfactorily. Zou and Hastie (2005) proposed the elastic net (EN) estimator to achieve improved performance in these cases. The EN estimator can also be viewed as a penalized least squares method where the penalty term is a convex combination of the lasso penalty and the ridge penalty. Simulations and biological data examples showed that the EN

*Department of Mathematics, Washington University in St. Louis, St. Louis, MO, mailto:qli@math.wustl.edu

†Department of Mathematics, Washington University in St. Louis, St. Louis, MO, mailto:nlin@math.wustl.edu

 DOI:10.1214/10-BA506

estimator does well in the three scenarios where the lasso estimator is not proper. A brief review of EN is as follows.

Suppose that we have n observations and p predictors. Denote the response variable by $\mathbf{y} = (y_1, \ldots, y_n)^T$ and the design matrix by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, where $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, \ldots, p$. Apply a linear transformation if necessary, we assume that

$$\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1, \text{ for } \ j = 1, \ldots, p.$$

For any fixed nonnegative penalty parameters λ_1 and λ_2, the EN loss function is defined as

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2,$$

where

$$\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|, \quad \text{and } \|\boldsymbol{\beta}\|_2 = \left(\sum_{j=1}^{p} \beta_j^2\right)^{\frac{1}{2}}.$$

Then the naïve EN estimator $\hat{\boldsymbol{\beta}}_{\text{EN}}$ is defined as $\hat{\boldsymbol{\beta}}_{\text{EN}} = \underset{\boldsymbol{\beta}}{\text{argmin}} L(\lambda_1, \lambda_2, \boldsymbol{\beta})$. The name "naïve" comes from the fact that $\hat{\boldsymbol{\beta}}_{\text{EN}}$ suffers from the "double shrinkage problem" (Zou and Hastie 2005). To correct the "double shrinkage problem", $(1+\lambda_2)\hat{\boldsymbol{\beta}}_{\text{EN}}$ instead of $\hat{\boldsymbol{\beta}}_{\text{EN}}$ is used as the EN estimator. $\hat{\boldsymbol{\beta}}_{\text{EN}}$ can be solved efficiently through the LARS-EN algorithm (Zou and Hastie 2005). However, there are two limitations with the EN method:

1. The LARS-EN algorithm only computes estimates of the regression coefficients, but doing statistical inference for the coefficients is difficult.

2. As the LARS-EN assumes given L_1-penalty coefficient λ_1 and L_2-penalty coefficient λ_2, the tuning parameters λ_1 and λ_2 are selected by cross-validation (Hastie et al. 2009). To avoid intensive computation, a grid of values for λ_2 is first specified. For example, Zou and Hastie (2005) used $(0, 0.01, 0.1, 1, 10, 100, 1000)$. For each λ_2, a 10-fold cross-validation is then used to choose λ_1. This cross-validation procedure selects λ_1 and λ_2 sequentially instead of simultaneously and causes the "double shrinkage problem".

The first limitation is common for most regression regularization methods. A number of people have considered the standard error estimation for the lasso estimator. Tibshirani (1996) suggested using ridge regression approximation or the bootstrap method. The problem for the ridge regression approximation is that the standard error estimate would be 0 when the parameter estimate is 0. The sandwich estimator proposed by Fan and Li (2001) suffers the same problem. Knight and Fu (2000) studied the bootstrap method for standard error estimation and pointed out that the bootstrap estimates are asymptotically biased when some true parameters are 0 or close to 0. Osborne et al. (2000) gave an alternative approximation that leads to better standard error estimates,

but it is based on the implicit assumption that the estimators are approximately linear, which holds only for very small λ_1. All these studies aforementioned are from a frequentist's point of view. In contrast, statistical inference for the Bayesian methods are more straightforward. The Bayesian lasso (BL) (Park and Casella 2008; Hans 2009a,b) not only gives the interval estimation but also provided the posterior distribution for the lasso estimator. BL assigns a double exponential prior on $\boldsymbol{\beta}$, which is essentially a scale mixture of normals where the mixing is through an exponential distribution. Some related works or extensions of BL include the horseshoe estimator by Carvalho et al. (2008) where the mixing is through a half-Cauchy distribution and the normal-gamma prior by Griffin and Brown (2009) where the mixing is through a gamma distribution. The latter is further applied and explored by Scheipl and Kneib (2009).

In this paper, we propose a Bayesian analysis of the EN problem and solve the problem using a Gibbs sampler. As an automatic byproduct of the Markov chain Monte Carlo procedure, the inference on the estimates is straightforward. In addition, our method chooses the penalty parameters λ_1 and λ_2 simultaneously by a Monte Carlo EM algorithm proposed by Casella (2001), thus avoids the "double shrinkage problem". Section 2 describes the model and the Gibbs sampler, as well as criteria for variable selection. Section 3 presents some simulation studies to compare the Bayesian elastic net (BEN) method with some other regression regularization methods. Sections 4 and 5 check the performance of BEN on some real data examples. Some discussions are provided in Section 6. An appendix contains technical proofs and derivations .

2 The Bayesian elastic net

2.1 Model hierarchy and prior distributions

Consider the linear model $E(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$, where we assume that the response variables follow the normal distribution conditionally, i.e.

$$\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n).$$

We assume that all analysis hereafter is conditional on $\mathbf{X}$. Zou and Hastie (2005) pointed out that, under these assumptions, solving the EN problem is equivalent to finding the marginal posterior mode of $\boldsymbol{\beta} \mid \mathbf{y}$ when the prior distribution of $\boldsymbol{\beta}$ is given by

$$\pi(\boldsymbol{\beta}) \propto \exp\left\{-\lambda_1 \|\boldsymbol{\beta}\|_1 - \lambda_2 \|\boldsymbol{\beta}\|_2^2\right\}, \tag{1}$$

a compromise between Gaussian and Laplacian priors. Specifically, the conditional posterior distribution has the probability density function (pdf)

$$\begin{aligned} f(\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}) &\propto \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \lambda_1 \|\boldsymbol{\beta}\|_1 - \lambda_2 \|\boldsymbol{\beta}\|_2^2\right\}, \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + (2\sigma^2\lambda_1)\|\boldsymbol{\beta}\|_1 + (2\sigma^2\lambda_2)\|\boldsymbol{\beta}\|_2^2\right]\right\}. \end{aligned}$$

However, under (1), neither the conditional posterior mode of $\boldsymbol{\beta} \mid \sigma^2, \mathbf{y}$ nor the marginal posterior mode of $\boldsymbol{\beta} \mid \mathbf{y}$ would be equivalent to the EN estimator $\hat{\boldsymbol{\beta}}_{\text{EN}}$ unless the analysis is conditional on σ^2 or σ^2 is given a point-mass prior. Instead we propose a prior for $\boldsymbol{\beta}$, which is conditional on σ^2, as

$$\pi(\boldsymbol{\beta} \mid \sigma^2) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(\lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2\right)\right\}. \tag{2}$$

A noninformative prior is then assigned for σ^2, i.e. $\pi(\sigma^2) \propto 1/\sigma^2$. In this setup, the marginal posterior distribution for $\boldsymbol{\beta} \mid \mathbf{y}$ has the pdf

$$f(\boldsymbol{\beta} \mid \mathbf{y}) = \int_0^\infty \frac{C(\lambda_1, \lambda_2, \sigma^2)}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2}{2\sigma^2}\right\} \pi(\sigma^2) d\sigma^2, \tag{3}$$

where $C(\lambda_1, \lambda_2, \sigma^2)$ is the normalizing constant later mentioned in Lemma 1. From (3), we can see that the EN estimator $\hat{\boldsymbol{\beta}}_{\text{EN}}$ maximizes the integrand for each σ^2, and thus equals the marginal posterior mode of $\boldsymbol{\beta} \mid \mathbf{y}$. This conditional prior specification is also used by Park and Casella (2008) for the sake of accelerating the convergence of the Gibbs sampler in that paper.

Based on the discussion above, we have the following hierarchial model.

$$\begin{aligned} \mathbf{y} \mid \boldsymbol{\beta}, \sigma^2 &\sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \\ \boldsymbol{\beta} \mid \sigma^2 &\sim \exp\left\{-\frac{1}{2\sigma^2}\left(\lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2\right)\right\}, \\ \sigma^2 &\sim \frac{1}{\sigma^2}. \end{aligned} \tag{4}$$

To better understand this prior specification, it is informative to see how the hyperparameters (λ_1, λ_2) affect the shape of the prior. Figure 1 shows the prior density function (2) with different values of λ_1 and λ_2, assuming $\sigma^2 = 1$ and $p = 1$. Three different choices of (λ_1, λ_2) are considered, (5,0), (1,5) and (0,5). These three choices correspond to the lasso prior (Park and Casella 2008), the EN prior and the ridge prior, respectively. It is worth noting that the lasso prior and EN prior are not differentiable at 0, but the ridge prior is. This is the essential difference between sparse model priors and non-sparse model priors, as pointed out by Fan and Li (2001). To yield sparse solutions that are continuous in the data, the prior must be non-differentiable at 0. Furthermore, from the plot we can see that, the EN prior is less "sharp" than the lasso prior. In fact, it is a compromise between the lasso prior and the ridge prior. Qualitatively speaking this suggests that the EN method is not so "aggressive" as the lasso method in terms of predictor exclusions.

2.2 The full conditional distributions and the Gibbs sampler

Computationally, it is difficult to solve the model (4) directly through a Gibbs sampler because the absolute values $|\beta_j|$'s in the prior would yield unfamiliar full conditional distributions. However, the following lemma suggests solving the problem by introducing another hierarchy to the model.

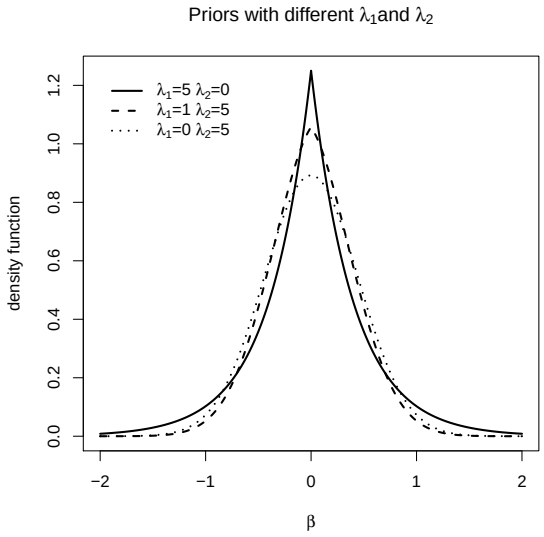

Figure 1: A comparison of prior (2) with different λ_1's and λ_2's.

Lemma 1. *The density $\pi(\boldsymbol{\beta} \mid \sigma^2)$ in (2) can be written as*

$$C(\lambda_1, \lambda_2, \sigma^2) \prod_{j=1}^{p} \int_1^\infty \sqrt{\frac{t}{t-1}} \exp\left\{-\frac{\beta_j^2}{2}\left(\frac{\lambda_2}{\sigma^2}\frac{t}{t-1}\right)\right\} t^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\frac{\lambda_1^2}{4\lambda_2}t\right) dt,$$

where $C(\lambda_1, \lambda_2, \sigma^2)$ is the normalizing constant.

The proof for this lemma is in the Appendix. This lemma implies that we can treat $\beta_j \mid \sigma^2$ as a mixture of normal distributions, $N(0, \sigma^2(t-1)/(\lambda_2 t))$, where the mixing distribution is over the variance $\sigma^2(t-1)/(\lambda_2 t)$ and is given by a truncated gamma distribution with shape parameter $1/2$, scale parameter $8\lambda_2\sigma^2/\lambda_1^2$ and support $(1, \infty)$, $\text{TG}\left(1/2, 8\lambda_2\sigma^2/\lambda_1^2, (1, \infty)\right)$.

By Lemma 1, we have the following hierarchical model.

$$\begin{aligned}
\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2 &\sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n), \\
\boldsymbol{\beta} \mid \boldsymbol{\tau}, \sigma^2 &\sim \prod_{j=1}^{p} N\left(0, \left(\frac{\lambda_2}{\sigma^2}\frac{\tau_j}{\tau_j - 1}\right)^{-1}\right), \\
\boldsymbol{\tau} \mid \sigma^2 &\sim \prod_{j=1}^{p} \text{TG}\left(\frac{1}{2}, \frac{8\lambda_2\sigma^2}{\lambda_1^2}, (1, \infty)\right), \\
\sigma^2 &\sim \frac{1}{\sigma^2}.
\end{aligned}$$

After introducing $\boldsymbol{\tau}$, the model becomes computationally easier to solve since the full conditional distributions now are

$$\begin{aligned}
\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2, \boldsymbol{\tau} &\sim N\left(\mathbf{A}^{-1}\mathbf{X}^T\mathbf{y}, \sigma^2\mathbf{A}^{-1}\right), \\
&\quad \text{where } \mathbf{A} = \mathbf{X}^T\mathbf{X} + \lambda_2 \text{diag}\left(\frac{\tau_1}{\tau_1 - 1}, \cdots, \frac{\tau_p}{\tau_p - 1}\right), \\
(\boldsymbol{\tau} - \mathbf{1}_p) \mid \mathbf{y}, \sigma^2, \boldsymbol{\beta} &\sim \prod_{j=1}^{p} \text{GIG}\left(\lambda = \frac{1}{2}, \psi = \frac{\lambda_1}{4\lambda_2\sigma^2}, \chi = \frac{\lambda_2\beta_j^2}{\sigma^2}\right), \qquad (5) \\
\sigma^2 \mid \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\tau} &\sim \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+p+1} \left\{\Gamma_U\left(\frac{1}{2}, \frac{\lambda_1^2}{8\sigma^2\lambda_2}\right)\right\}^{-p} \times \\
&\quad \exp\left[-\frac{1}{2\sigma^2}\left\{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \sum_{j=1}^{p} \frac{\tau_j}{\tau_j - 1}\beta_j^2 + \frac{\lambda_1^2}{4\lambda_2}\sum_{j=1}^{p}\tau_j\right\}\right], \qquad (6)
\end{aligned}$$

where $\mathbf{1}_p$ is a p-dimensional vector of 1's. $\Gamma_U(\alpha, x) = \int_x^\infty t^{\alpha-1}e^{-t}dt$ is the upper incomplete gamma function (Armido and Alfred 1986). $\text{GIG}(\lambda, \psi, \chi)$ denotes the generalized inverse Gaussian distribution (Jϕrgensen 1982), whose pdf is

$$f(x \mid \lambda, \chi, \psi) = \frac{(\psi/\chi)^{\lambda/2}}{2K_\lambda(\sqrt{\psi\chi})} x^{\lambda-1} \exp\left\{-\frac{1}{2}(\chi x^{-1} + \psi x)\right\}, \qquad (7)$$

for $x > 0$, where $K_\lambda(\cdot)$ is the modified Bessel function of the third kind with order λ.

2.3 Sampling from the full conditional distributions

It is straightforward to sample from $\boldsymbol{\beta} \mid \mathbf{y}, \sigma^2, \boldsymbol{\tau}$ which follows the multivariate normal distribution. Details on sampling from (5) and (4) are as follows.

To sample generalized inverse Gaussian random numbers for $(\boldsymbol{\tau} - \mathbf{1}_p) \mid \mathbf{y}, \sigma^2, \boldsymbol{\beta}$, we may use the `rgig`() function in the "HyperbolicDist" package (Scott 2008) in `R` (R Development Core Team 2005) which implements the algorithm in Dagpunar (1989). This algorithm is based on the ratio method together with rejection sampling on the minimal enclosing rectangle. However, when the product of parameters $\chi\psi$ in (7) is small, the rejection rate could be extremely high, and the algorithm becomes very inefficient. To expedite our algorithm, we consider the full conditional of some transformation of $\boldsymbol{\tau} - \mathbf{1}_p$. It is easy to show that $1/(\tau_j - 1) \mid \mathbf{y}, \sigma^2, \boldsymbol{\beta}$, $j = 1, \ldots, p$ are independent and follow the inverse Gaussian distribution (Chhikara and Folks 1988) with $\mu = \sqrt{\lambda_1}/(2\lambda_2|\beta_j|)$, $\lambda = \lambda_1/(4\lambda_2\sigma^2)$. The pdf of an inverse Gaussian distribution is

$$f(x \mid \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left\{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right\}.$$

Sampling from an inverse Gaussian distribution uses a more efficient algorithm based on multiple roots selection (Michael et al. 1976) and is much faster than the `rgig`() function.

Sampling from $\sigma^2 \mid \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\tau}$ could be done by the acceptance-rejection algorithm. Denote the function of σ^2 on the right-hand side of (4) as $f(\sigma^2)$. Then by the definition of incomplete gamma functions,

$$f(\sigma^2) \le \Gamma\left(\frac{1}{2}\right)^{-p}\left(\frac{1}{\sigma^2}\right)^{a+1}\exp\left\{-\frac{1}{\sigma^2}b\right\} = \frac{\Gamma(a)\Gamma\left(\frac{1}{2}\right)^{-p}}{b^a}h(\sigma^2),$$

where $h(\cdot)$ is the pdf for inverse-gamma(a, b) and

$$a = \frac{n}{2} + p,\ b = \frac{1}{2}\left[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_2\sum_{j=1}^{p}\frac{\tau_j}{\tau_j - 1}\beta_j^2 + \frac{\lambda_1^2}{4\lambda_2}\sum_{j=1}^{p}\tau_j\right].$$

To get σ^2 from $f(\sigma^2)$, we first generate a candidate Z from h and a u from uniform(0,1), and then accept Z if $u \le \Gamma\left(\frac{1}{2}\right)^p b^a f(Z)/\Gamma(a)h(Z)$ or, equivalently, if $\log(u) \le p\log\left(\Gamma\left(\frac{1}{2}\right)\right) - p\log\Gamma_U\left(\frac{1}{2}, \frac{\lambda_1^2}{8Z\lambda_2}\right)$.

2.4 Choosing the penalty parameters by empirical Bayes

As suggested by Park and Casella (2008), we use empirical Bayes estimates for the penalty parameters λ_1 and λ_2, which are given by maximizing the data marginal likelihood. Specifically, by treating the parameters $\boldsymbol{\beta}, \boldsymbol{\tau}$ and σ^2 as missing data, we can use the Monte Carlo EM algorithm in Casella (2001) to estimate λ_1 and λ_2. In each step of the Monte Carlo EM algorithm, any conditional expectations that are hard to compute are substituted by Monte Carlo estimates. The details of the algorithm are described in the Appendix.

2.5 Variable selection

In lasso and EN, some regression coefficient estimates are forced to be zero, so variable selection is straightforward. However, the Bayesian approaches need some ad hoc treatment for variable selection.

In general, variable selection can be viewed as a hypothesis testing problem. Since the posterior probability of $H_0 : \beta_j = 0$, $j = 1, \ldots, p$ would always be 0, a common practice is to consider the posterior probability in some neighborhood $[-r_j, r_j]$ of 0, for some constants $r_j > 0$, by viewing $H_0 : \beta_j = 0$ as an approximation to $H_0 : \beta_j \in [-r_j, r_j]$ (Berger 1993). The choice of r_j's usually demands expert knowledge about the data. In this paper, instead of choosing an r_j for each β_j, $j = 1, \ldots, p$, we circumvent by suggesting two automatic criteria for variable selection.

The first criterion is called the credible interval criterion. A predictor $\mathbf{x}_j$ is excluded if the credible interval of β_j covers 0 and is retained otherwise. Simulation results show that 95% credible intervals are usually too wide and most predictors would consequently be excluded. Some empirical guidance on choosing the appropriate level is discussed in detail in Section 3.

The second criterion is called the scaled neighborhood criterion. We consider the posterior probability in $\left[-\sqrt{\text{var}(\beta_j \mid \mathbf{y})}, \sqrt{\text{var}(\beta_j \mid \mathbf{y})}\right]$. A predictor is excluded if the posterior probability exceeds a certain probability threshold and is retained otherwise. We would explore the choice of the probability threshold in the next section.

3 Simulation studies

We carry out Monte Carlo simulations to compare the performance of BEN, BL, EN and lasso in prediction accuracy and variable selection. For EN, instead of using the penalty parameters (λ_1, λ_2), it is equivalent to use (s, λ_2), where $s = \|\boldsymbol{\beta}\|_1/\|\boldsymbol{\beta}_{\text{OLS}}\|_1$ with $\boldsymbol{\beta}_{\text{OLS}}$ being the OLS estimate (Zou and Hastie 2005). Similarly, s instead of λ_1 is used in lasso.

The data are simulated from the following linear model,

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

Each simulated sample was partitioned into a training set, a validation set and a testing set. The validation set is used to select s for lasso and (s, λ_2) for EN. After the penalty parameter is selected, we combine the training set and the validation set together to estimate $\boldsymbol{\beta}$. For Bayesian methods, we directly combine the training and validation sets together for estimation.

For the first two simulation studies, we simulate 50 data sets, each of which has 20 observations for the training set, 20 for the validation set, and 200 for the testing set. In Simulation 1, we set $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\sigma^2 = 9$. The design matrix $\mathbf{X}$ is generated from the multivariate normal distribution with mean 0, variance 1 and pairwise correlations between $\mathbf{x}_i$ and $\mathbf{x}_j$ equal to $0.5^{|i-j|}$ for all i and j. In Simulation 2, we set $\boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)^T$, and leave other setups the same as in Simulation 1. In Simulation 3, we first generate Z_1, Z_2 and Z_3 independently from $N(0, 1)$. Then let $\mathbf{x}_i = Z_1 + \epsilon_i$, $i = 1, \ldots, 5$, $\mathbf{x}_i = Z_2 + \epsilon_i$, $i = 6, \ldots, 10$, $\mathbf{x}_i = Z_3 + \epsilon_i$, $i = 11, \ldots, 15$ and $\mathbf{x}_i \sim N(0, 1)$, $i = 16, \ldots, 30$, where $\epsilon_i \sim N(0, 0.01)$, $i = 1, \ldots, 15$. We perform 50 simulations, in each of which we have a training set of size 100, a validation set of size 100 and a testing set of size 400. The parameters are set as $\sigma^2 = 225$ and

$$\boldsymbol{\beta} = (\underbrace{3, \ldots, 3}_{5}, \underbrace{3, \ldots, 3}_{5}, \underbrace{3, \ldots, 3}_{5}, \underbrace{0, \ldots, 0}_{15})^T.$$

In Simulation 4, we set the sizes of the training set and the validation set both to 200, while leaving other setups the same as in Simulation 3. In Simulation 5, we set the sizes of the training set and the validation set both to 20, and the true parameter value to

$$\boldsymbol{\beta} = (\underbrace{3, \ldots, 3}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{3, \ldots, 3}_{10}),$$

while leaving other setups the same as in Simulation 3.

3.1 The variable selection accuracy

Denote by α the level in the credible interval criterion and by p the probability threshold in the scaled neighborhood criterion. The variable selection accuracy of the Bayesian methods depends on α and p. In order to understand how the choices affect the variable selection result, we can draw the receiver operating characteristic (ROC) curve and the power curve for different α's and p's. The ROC curve is derived by plotting the correct inclusion rate, i.e. sensitivity, against the false inclusion rate, i.e. 1 - specificity, along different α's or p's. The power curve is similar to the ROC curve, but with α or p as the horizontal axis. The ROC curves and power curves for the five simulation studies are in Figures 2 and 3. For Simulation 2, as there cannot be any false inclusions, only the power curve is plotted. In these plots, BEN + CI and BEN + SN stand for the BEN method with the credible interval criterion and the BEN method with the scaled neighborhood criterion, respectively. The same is for BL + CI and BL + SN.

Figures 2 and 3 show that the two variable selection criteria give similar ROC curves for both BEN and BL but the credible interval criterion tends to have higher sensitivity for the same α or p. This implies that the scaled neighborhood criterion has higher specificity for the same α or p. So depending on the importance of sensitivity and specificity of a particular study, the researcher may choose the appropriate variable selection criterion. In the usual hypothesis testing setup where specificity is more valued, the scaled neighborhood criterion would be preferred. Secondly, the power curves show that BEN and EN generally have higher sensitivity than BL and lasso, i.e. they are not as "aggressive" in excluding predictors. This validates the observation in the discussion of Figure 1 in Section 2.1. Although EN and BEN also give sparse solutions, they tend to give less sparse ones than those given by lasso and BL. Thirdly, from the power curves for the five simulation studies, we can see that $\alpha = 0.5$ and $p = 0.5$ would generally guarantee that BEN and BL have similar or higher power than the non-Bayesian methods. Based on this observation, together with the discussion in Section 2.5, we would suggest using $\alpha = 0.5$ and $p = 0.5$ in practice. In principle, using a higher level α or a threshold value p would result in a higher sensitivity but a lower specificity. We show in Table 1 some detailed results on the variable selection for the first two simulation studies. Let $N(\beta_j), j = 1, \ldots, 8$, denote the frequency of exclusions for the predictor $\mathbf{x}_j$. For Simulation 1, the smaller $N(\beta_j)$ for $j = 1, 2, 5$ and the larger $N(\beta_j)$ for $j = 3, 4, 6, 7, 8$, the better the method performs. For Simulation 2, we want all $N(\beta_j)$ to be as small as possible. As there are too many predictors in Simulations 3, 4 and 5 to put in a table, we do not show those results.

3.2 The prediction accuracy

We compare the prediction accuracy of the four methods using the median of the prediction mean-squared errors (MMSE). Also, 500 bootstrap resampling on the 50 mean-squared errors are drawn to give the standard error for the MMSE. The results are summarized in Table 2.

Table 2 shows that BL has a better prediction accuracy than other methods in most

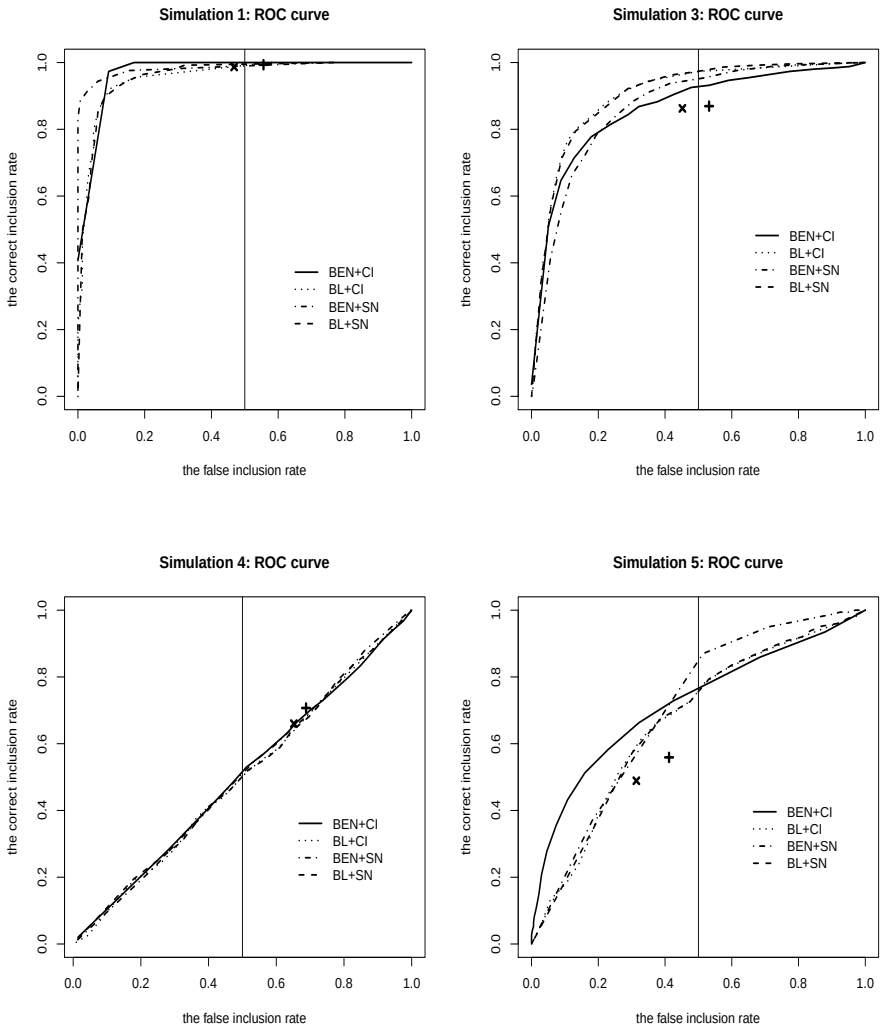

Figure 2: The ROC curves for the four simulation studies. BEN + CI (solid line), BL + CI (dotted line), BEN + SN (dot-dashed line), BL + SN (dashed line), EN (+) and lasso (×).

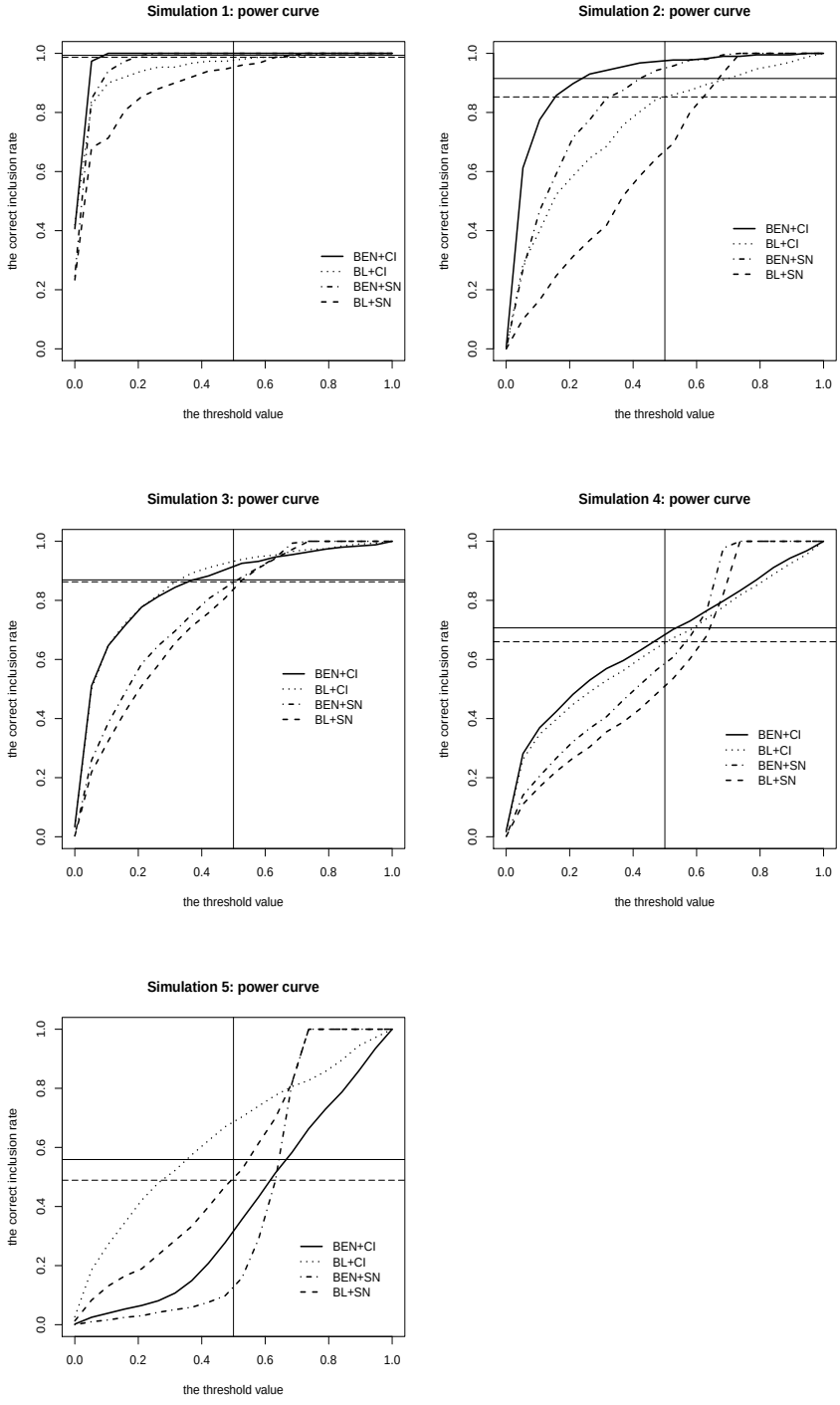

Figure 3: The power curves for the five simulation studies. BEN + CI (solid line), BL + CI (dotted line), BEN + SN (dot-dashed line), BL + SN (dashed line), EN (horizontal solid line) and lasso (horizontal dashed line).

Method	N(β_1)	N(β_2)	N(β_3)	N(β_4)	N(β_5)	N(β_6)	N(β_7)	N(β_8)
BEN + CI	0	0	9	13	0	22	30	26
BEN + SN	0	0	16	19	0	30	37	41
EN	0	1	12	10	0	22	27	31
BL + CI	0	3	26	31	0	27	27	25
BL + SN	0	7	40	41	0	38	36	36
lasso	0	2	22	24	0	23	24	25

(a) Simulation 1

Method	N(β_1)	N(β_2)	N(β_3)	N(β_4)	N(β_5)	N(β_6)	N(β_7)	N(β_8)
BEN + CI	2	0	1	0	0	1	2	3
BEN + SN	3	1	2	0	0	2	4	7
EN	9	1	2	1	0	1	5	8
BL + CI	6	6	8	8	10	6	7	10
BL + SN	13	15	16	19	22	14	14	18
lasso	6	3	5	5	5	5	5	9

(b) Simulation 2

Table 1: Comparison of the four methods (BEN, BL, EN, and lasso) on variable selection accuracy with $\alpha = 0.5$ and $p = 0.5$ for the first two simulation studies.

simulation studies. However, for small samples under a less sparse model (Simulation 5), BEN behaves the best. Figure 2 also shows that the ROC curve for variable selection of BEN dominates that of BL in that the correct inclusion rate is always higher at any given false inclusion rates, although BL has a higher power (Figure 3). Secondly, when the true model structure is relatively simple (Simulations 1 and 2), the four methods perform comparably in prediction accuracy. But when the true model has a complex structure (Simulations 3, 4 and 5), BEN and BL outperform EN and lasso significantly. By comparing BEN and BL in Simulation 3 and EN and lasso in Simulation 4 (the bolded numbers in Table 2), we can see that even with only half as many data, the Bayesian methods perform better than the non-Bayesian methods in prediction accuracy. One possible reason is that complicated models would result in highly variational estimators, and the Bayesian methods use prior information to integrate across uncertainty to reduce the variance, which leads to a smaller mean squared error. In this sense, the Bayesian methods furnish better results with less data when the true model is complicated. Furthermore, with the sample size doubled from Simulation 3 to Simulation 4, the MMSE of the Bayesian methods decreases about 15 while that of the non-Bayesian methods decreases about 40.

4 The diabetes example

The data in this example are from Efron et al. (2004), where the predictors are ten baseline variables: age, sex, body mass index (bmi), average blood pressure (map), and

Method	Simulation 1 MMSE(SE)	Simulation 2 MMSE(SE)	Simulation 3 MMSE(SE)	Simulation 4 MMSE(SE)	Simulation 5 MMSE(SE)
BEN	11.99(0.28)	11.39(0.29)	**232.3(3.83)**	215.1(1.95)	335.3(4.17)
EN	11.29(0.53)	11.10(0.29)	281.4(6.80)	**243.0(4.24)**	376.9(9.81)
BL	10.45(0.24)	10.55(0.21)	**227.1(4.20)**	211.4(2.36)	352.6(13.1)
lasso	10.99(0.20)	11.41(0.33)	280.1(8.00)	**243.0(4.34)**	384.9(9.61)

Table 2: Comparison of the four methods (BEN, BL, EN, and lasso) on prediction accuracy.

six blood serum measurements (tc, ldl, hdl, tch, lth, glu). The response of interest is a quantitative measure of disease progression one year after baseline. There are $n = 442$ diabetes patients in the data set. Four different methods are applied to this data set: BEN, BL, EN and lasso. The estimation results are summarized in Figure 4. In this example, the posterior distributions of $\boldsymbol{\beta} \mid \mathbf{y}$ are very close to normal distributions, so posterior medians instead of the posterior modes are reported for BEN and BL.

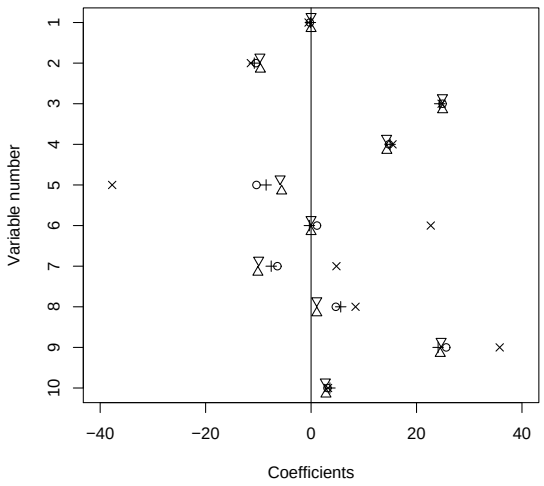

Figure 4: The estimates of the predictor effects for the diabetes data using different methods: posterior median BEN estimates (+), posterior median BL estimates (○), EN estimates (△), lasso estimates (∇) and the OLS estimates (×).

The penalty parameters are selected as $(s, \lambda_2) = (0.78, 0.01)$ for EN and $s = 0.57$ for lasso. Figure 4 shows that these two methods give almost identical estimates of $\boldsymbol{\beta}$ for this data set. For the BEN method, the penalty parameters estimated by the Monte Carlo EM algorithm are $(\lambda_1, \lambda_2) = (0.996, 14.89)$ while for the BL method, λ_1 is estimated as 4.536. These two Bayesian methods also behave similarly for this data set. From Figure

4 we can see that the Bayesian methods (BEN and BL) tend to achieve milder shrinkage compared to the non-Bayesian methods (EN and lasso), i.e. their estimates usually lie between the OLS estimates and the estimates given by EN and lasso.

For the variable selection, if we use the credible interval criterion with level $\alpha = 0.5$, two variables are excluded (for both BEN and BL): age and ldl. These are consistent with the variables excluded by the EN and lasso. If we use the scaled neighborhood criterion with probability threshold $p = 0.5$, then three variables are excluded (for both BEN and BL): age, ldl and tch. Table 3 summarizes the posterior probabilities for the scaled neighborhood criterion. Since a larger posterior probability implies stronger evidence for exclusion, Table 3 shows that BEN and BL give the same ordering for variable exclusion.

predictor	posterior probability	
	BEN	BL
bmi	0.00	0.00
ltg	0.00	0.00
map	0.00	0.00
sex	0.00	0.00
hdl	0.37	0.46
glu	0.42	0.46
tc	0.49	0.48
tch	0.51	0.57
age	0.68	0.69
ldl	0.69	0.71

Table 3: Diabetes example: posterior probabilities in the scaled neighborhood.

5 The prostate example

The data in this example are from a prostate cancer study (Stamey et al. 1989) and were analyzed by Zou and Hastie (2005). The predictors are eight clinical measures: the logarithm of cancer volume, the logarithm of prostate weight, age, the logarithm of the amount of benign prostatic hyperplasia, seminal vesicle invasion, the logarithm of capsular penetration, the Gleason score and the percentage Gleason score 4 or 5. The response of interest is the logarithm of prostate-specific antigen. In total we have 97 observations for the data. The training data has 67 observations and the testing data has 30 observations. The estimates given by different methods are summarized in Figure 5. It shows that estimates given by the four methods are much more different than those in the diabetes data. In particular, some BL estimates are large compared to the corresponding estimates from the other three methods.

The scaled neighborhood criterion with threshold $p = 0.5$ is used for variable selection in BEN and BL. Table 4 gives the posterior probabilities in the scaled neighborhood, and Table 5 gives the variable selection results using four different methods. In contrast to the diabetes example, we see that the variable selection results for BEN and BL are

very different. One possible reason is that the predictors in this data set are more correlated than those in the diabetes data. Table 5 also reports the prediction mean-squared error from the testing data using four different methods. The two Bayesian methods have slightly larger prediction error than the non-Bayesian methods. One possible reason is that the true model structure is not complicated, so that not much variance is reduced by introducing the prior, but the increased bias leads to a larger prediction error.

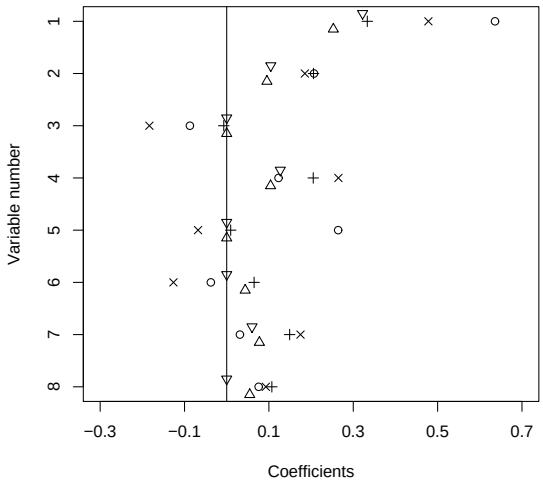

Figure 5: The estimates of the predictor effects for the prostate cancer data using different methods: posterior median BEN estimates (+), posterior median BL estimates (∘), EN estimates (△), lasso estimates (∇) and the OLS estimates (×).

6 Conclusion and discussion

We propose a Bayesian formulation of the EN problem and the posterior inference can be obtained efficiently using Gibbs sampling. Real data examples and simulation studies show that BEN behaves comparably to EN in prediction accuracy. EN does slightly better than BEN for simple models, but BEN performs significantly better than EN for more complicated models. Simulation studies suggest that BEN outperforms EN in variable selection when coupled with the scaled neighborhood criterion with a proper probability threshold, and BEN gives better prediction accuracy than BL for small samples from less sparse models.

An alternative way of choosing the penalty parameters λ_1 and λ_2 is to introduce hyper-priors on them. This is also mentioned by Park and Casella (2008). Specifically, we may set a gamma prior gamma(a, b) for λ_1^2 and a GIG prior GIG$(1, c, d)$ for λ_2.

predictor	posterior probability	
	BEN	BL
lcavol	0.00	0.00
lweight	0.05	0.17
age	0.68	0.37
lbph	0.06	0.09
svi	0.69	0.62
lcp	0.54	0.63
gleason	0.20	0.41
pgg45	0.36	0.60

Table 4: Prostate cancer example: posterior probabilities in the scaled neighborhood.

method	selected variables	prediction mean-squared error
BEN	locavol, lweight, lbph, gleason, pgg45	0.47
BL	locavol, lweight, age, lbph, gleason	0.44
EN	locavol, lweight, lbph, lcp, gleason, pgg45	0.39
lasso	locavol, lweight, lbph, gleason	0.38

Table 5: Variable selection results and prediction mean-squared error for the prostate cancer example using different methods: BEN, BL, EN and lasso.

Then the resulting conjugacy leads to an easy extension of the Gibbs sampler. This approach does not require updating the penalty parameters after each Markov chain, but would run a single longer chain that updates the penalty parameters in every step. So this approach is much faster computationally. But our experience is that the results depend heavily on the parameters (a, b) and (c, d), which need to be specified by users in advance. However, these parameters are less intuitive to specify than the credible interval level or the probability threshold. In Park and Casella (2008) the specification of parameters in the prior for λ_1 leads to similar results as those given by the Monte Carlo EM method.

Another important desired property for the regression regularization method is the oracle property (Fan and Li 2001). Zou and Zhang (2009) proposed the adaptive elastic net to achieve this goal. Making adjustments to the BEN method to adopt the oracle property deserves some further consideration.

Appendix

1. Proof of Lemma 1

Firstly, note that

$$\exp\left\{-\frac{1}{2\sigma^2}\left(\lambda_1\|\boldsymbol{\beta}\|_1+\lambda_2\|\boldsymbol{\beta}\|_2^2\right)\right\}=\prod_{j=1}^{p}\exp\left\{-\frac{1}{2\sigma^2}\left(\lambda_1|\beta_j|+\lambda_2\beta_j^2\right)\right\}.$$

Secondly, as suggested by Andrews and Mallows (1974), for $a>0$, we have

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

Let $a=\frac{\lambda_1}{2\sigma^2}$ and $z=\beta_j$, and we have

$$\exp\left(-\frac{\lambda_1}{2\sigma^2}|\beta_j|\right)\underset{\sigma^2}{\propto}\int_0^\infty\frac{1}{\sqrt{s}}\exp\left(-\frac{1}{2}\frac{\beta_j^2}{s}\right)\exp\left\{-\frac{1}{2}\left(\frac{\lambda_1}{2\sigma^2}\right)^2s\right\}ds,$$

where $\underset{\sigma^2}{\propto}$ denotes that its two sides differ by a multiplicand involving σ^2 and possibly some constants. Then, we have

$$\begin{aligned}
&\exp\left\{-\frac{1}{2\sigma^2}\left(\lambda_1|\beta_j|+\lambda_2\beta_j^2\right)\right\}\\
&\underset{\sigma^2}{\propto}\int_0^\infty\frac{1}{\sqrt{s}}\exp\left\{-\frac{\beta_j^2}{2}\left(\frac{1}{s}+\frac{\lambda_2}{\sigma^2}\right)\right\}\exp\left\{-\frac{1}{2}\left(\frac{\lambda_1}{2\sigma^2}\right)^2s\right\}ds\\
&=\int_0^\infty\sqrt{\left(\frac{1}{s}+\frac{\lambda_2}{\sigma^2}\right)}\exp\left\{-\frac{\beta_j^2}{2}\left(\frac{1}{s}+\frac{\lambda_2}{\sigma^2}\right)\right\}\frac{1}{\sqrt{1+\frac{\lambda_2}{\sigma^2}s}}\exp\left\{-\frac{1}{2}\left(\frac{\lambda_1}{2\sigma^2}\right)^2s\right\}ds\\
&\underset{\sigma^2}{\propto}\int_1^\infty\sqrt{\frac{t}{t-1}}\exp\left\{-\frac{\beta_j^2}{2}\left(\frac{\lambda_2}{\sigma^2}\frac{t}{t-1}\right)\right\}t^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\frac{\lambda_1^2}{4\lambda_2}t\right)dt.
\end{aligned}$$

The result of Lemma 1 then follows.

2. Choosing the tuning parameters (λ_1,λ_2)

The penalty parameters (λ_1,λ_2) are chosen by the Monte Carlo EM algorithm. By treating $\boldsymbol{\beta},\boldsymbol{\tau},\sigma^2$ as missing data and (λ_1,λ_2) as fixed parameters, the "complete data" likelihood is (after ignoring the constants not involving λ_1 and λ_2)

$$\begin{aligned}
&\lambda_1^p\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+p+1}\left\{\Gamma_U\left(\frac{1}{2},\frac{\lambda_1^2}{8\sigma^2\lambda_2}\right)\right\}^{-p}\prod_{j=1}^{p}\left(\frac{1}{\tau_j-1}\right)^{1/2}\times\\
&\exp\left[-\frac{1}{2\sigma^2}\left\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\lambda_2\sum_{j=1}^{p}\frac{\tau_j}{\tau_j-1}\beta_j^2+\frac{\lambda_1^2}{4\lambda_2}\sum_{j=1}^{p}\tau_j\right\}\right].
\end{aligned}$$

So the log-likelihood is

$$
\begin{aligned}
& p\log(\lambda_1) - \left(\frac{n}{2} + p + 1\right)\log(\sigma^2) - p\log\Gamma_U\left(\frac{1}{2}, \frac{\lambda_1^2}{8\sigma^2\lambda_2}\right) - \frac{1}{2}\sum_{j=1}^{p}\log\left(\tau_j - 1\right) \\
& - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \frac{\lambda_2}{2\sigma^2}\sum_{j=1}^{p}\frac{\tau_j}{\tau_j - 1}\beta_j^2 - \frac{1}{2\sigma^2}\frac{\lambda_1^2}{4\lambda_2}\sum_{j=1}^{p}\tau_j \\
= & \, p\log(\lambda_1) - p\log\Gamma_U\left(\frac{1}{2}, \frac{\lambda_1^2}{8\sigma^2\lambda_2}\right) - \frac{\lambda_2}{2\sigma^2}\sum_{j=1}^{p}\frac{\tau_j}{\tau_j - 1}\beta_j^2 - \frac{1}{2\sigma^2}\frac{\lambda_1^2}{4\lambda_2}\sum_{j=1}^{p}\tau_j \\
& + \text{terms not involving } \lambda_1 \text{ and } \lambda_2.
\end{aligned}
$$

So at the kth step of the Monte Carlo EM algorithm, the conditional log-likelihood on $\lambda^{(k-1)} = \left(\lambda_1^{(k-1)}, \lambda_2^{(k-1)}\right)$ and Y is

$$
\begin{aligned}
& Q(\lambda \mid \lambda^{(k-1)}) \\
= & \, p\log(\lambda_1) - pE\left[\log\Gamma_U\left(\frac{1}{2}, \frac{\lambda_1^2}{8\sigma^2\lambda_2}\right) \mid \lambda^{(k-1)}, Y\right] - \frac{\lambda_2}{2}\sum_{j=1}^{p}E\left[\frac{\tau_j}{\tau_j - 1}\frac{\beta_j^2}{\sigma^2} \mid \lambda^{(k-1)}, Y\right] \\
& - \frac{\lambda_1^2}{8\lambda_2}\sum_{j=1}^{p}E\left[\frac{\tau_j}{\sigma^2} \mid \lambda^{(k-1)}, Y\right] + \text{terms not involving } \lambda_1 \text{ and } \lambda_2 \\
= & \, R(\lambda \mid \lambda^{(k-1)}) + \text{terms not involving } \lambda_1 \text{ and } \lambda_2.
\end{aligned}
$$

Then the M-step is to find λ_1 and λ_2 that maximize $R(\lambda \mid \lambda^{(k-1)})$.

To facilitate the maximization problem, we may need the gradient of $R(\lambda \mid \lambda^{(k-1)})$, which is given by

$$
\begin{aligned}
\frac{\partial R}{\partial \lambda_1} = & \, \frac{p}{\lambda_1} + \frac{p\lambda_1}{4\lambda_2}E\left[\left\{\Gamma_U\left(\frac{1}{2}, \frac{\lambda_1^2}{8\sigma^2\lambda_2}\right)\right\}^{-1}\varphi\left(\frac{\lambda_1^2}{8\sigma^2\lambda_2}\right)\frac{1}{\sigma^2} \mid \lambda^{(k-1)}, Y\right] \\
& - \frac{\lambda_1}{4\lambda_2}\sum_{j=1}^{p}E\left[\frac{\tau_j}{\sigma^2} \mid \lambda^{(k-1)}, Y\right], \\
\frac{\partial R}{\partial \lambda_2} = & -\frac{p\lambda_1^2}{8\lambda_2^2}E\left[\left\{\Gamma_U\left(\frac{1}{2}, \frac{\lambda_1^2}{8\sigma^2\lambda_2}\right)\right\}^{-1}\varphi\left(\frac{\lambda_1^2}{8\sigma^2\lambda_2}\right)\frac{1}{\sigma^2} \mid \lambda^{(k-1)}, Y\right] \\
& - \frac{1}{2}\sum_{j=1}^{p}E\left[\frac{\tau_j}{\tau_j - 1}\frac{\beta_j^2}{\sigma^2} \mid \lambda^{(k-1)}, Y\right] + \frac{\lambda_1^2}{8\lambda_2^2}\sum_{j=1}^{p}E\left[\frac{\tau_j}{\sigma^2} \mid \lambda^{(k-1)}, Y\right],
\end{aligned}
$$

where $\varphi(t) = t^{-\frac{1}{2}}e^{-t}$.

References

Andrews, D. F. and Mallows, C. L. (1974). "Scale mixtures of normal distributions." *J. R. Statist. Soc. B*, 36: 99–102. 167

Armido, D. and Alfred, M. (1986). "Computation of the incomplete gamma function ratios and their inverse." *ACM Trans. Math. Soft.*, 12: 377–393. 156

Berger, J. O. (1993). *Statistical Decision Theory and Bayesian Analysis*. Springer, 2nd edition. 157

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2008). "The horseshoe estimator for sparse signals." Discussion Paper 2008-31, Duke University Department of Statistical Science. 153

Casella, G. (2001). "Empirical Bayes Gibbs sampling." *Biostatistics*, 2: 485–500. 153, 157

Chhikara, R. and Folks, L. (1988). *The Inverse Gaussian Distribution Theory: Methodology, and Applications*. Marcel Dekker, Inc, 1st edition. 156

Dagpunar, J. S. (1989). "An easily implemented generalised inverse Gaussian generator." *Commun. Statist. -Simula.*, 18: 703–710. 156

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). "Least angle regression." *Ann. Stat.*, 32: 407–499. 151, 162

Fan, J. and Li, R. (2001). "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." *J. Amer. Statist. Assoc.*, 96: 1348–1360. 152, 154, 166

Griffin, J. E. and Brown, P. J. (2009). "Inference with Normal-Gamma prior distributions in regression problems." Technical report, Institute of Mathematics, Statistics and Actuarial Science, University of Kent. 153

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). "Gene Selection for Cancer Classification Using Support Vector Machines." *Mach. Learn.*, 46: 389–422. 151

Hans, C. (2009a). "Bayesian Lasso Regression." *Biometrika*, 96: 835–845. 153

— (2009b). "Model uncertainty and variable selection in Bayesian Lasso." *Stat. Comput.*. To appear. 153

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2nd edition. 152

Jϕrgensen (1982). "Statistical properties of the generalized inverse Gaussian distribution." *Lecture Notes in Statistics*, 9. 156

Knight, K. and Fu, W. (2000). "Asymptotics for lasso-type estimators." *Ann. Stat.*, 28: 1356–1378. 152

Michael, J. R., Schucany, W. R., and Haas, R. W. (1976). "Generating random variates using transformations with multiple roots." *Am. Stat.*, 30-2: 88–90. 156

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000). "On the LASSO and its dual." *J. Comput. Graph. Stat.*, 9: 319–337. 152

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *J. Amer. Statist. Assoc.*, 103: 681–686. 153, 154, 157, 165, 166

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org 156

Scheipl, F. and Kneib, T. (2009). "Locally adaptive Bayesian P-splines with a Normal-Exponential-Gamma prior." *Computational Statistics & Data Analysis*, 53(10): 3533–3552. 153

Scott, D. (2008). *HyperbolicDist: The hyperbolic distribution*. R package version 0.5-2.
URL http://www.r-project.org 156

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., and Yang, N. (1989). "Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II: radical prostatectomy treated patients." *J. Urol.*, 16: 1076–1083. 164

Tibshirani, R. (1996). "Regression shrinkage and selection via the Lasso." *J. R. Statist. Soc. B*, 58: 267–288. 151, 152

Zou, H. and Hastie, T. (2005). "Regularization and variable selection via the elastic net." *J. R. Statist. Soc. B*, 67: 301–320. 151, 152, 153, 158, 164

Zou, H. and Zhang, H. H. (2009). "On the adaptive elastic-net with a diverging number of parameters." *Ann. Stat.*, 37: 1733–1751. 166