Stat Comput (2010) 20: 221–229
DOI 10.1007/s11222-009-9160-9

Model uncertainty and variable selection in Bayesian lasso regression

Chris Hans

Received: 27 July 2009 / Accepted: 9 November 2009 / Published online: 25 November 2009

Abstract While Bayesian analogues of lasso regression have become popular, comparatively little has been said about formal treatments of model uncertainty in such settings. This paper describes methods that can be used to evaluate the posterior distribution over the space of all possible regression models for Bayesian lasso regression. Access to the model space posterior distribution is necessary if model-averaged inference—e.g., model-averaged prediction and calculation of posterior variable inclusion probabilities—is desired. The key element of all such inference is the ability to evaluate the marginal likelihood of the data under a given regression model, which has so far proved difficult for the Bayesian lasso. This paper describes how the marginal likelihood can be accurately computed when the number of predictors in the model is not too large, allowing for model space enumeration when the total number of possible predictors is modest. In cases where the total number of possible predictors is large, a simple Markov chain Monte Carlo approach for sampling the model space posterior is provided. This Gibbs sampling approach is similar in spirit to the stochastic search variable selection methods that have become one of the main tools for addressing Bayesian regression model uncertainty, and the adaption of these methods to the Bayesian lasso is shown to be straightforward.

Keywords Double-exponential distribution · Gibbs sampler · Marginal likelihood · MCMC · Model averaging · Orthant-normal distribution · SSVS

C. Hans (✉)
Department of Statistics, The Ohio State University, Columbus, OH 43210, USA
e-mail: hans@stat.osu.edu

1 Introduction

The purpose of this paper is to introduce analytic and computational approaches for handling model uncertainty under the Bayesian lasso regression model. The "Bayesian lasso" (Park and Casella 2008; Hans 2009) typically refers to use of the double-exponential shrinkage prior for the p-vector of regression coefficients β in the normal linear regression model $(y \mid \beta, \sigma^2) \sim \mathrm{N}(X\beta, \sigma^2 I)$, where y is an n-vector of observations and X is an $n \times p$ matrix of predictor variables. Shrinkage priors play an important role in Bayesian regression modeling, especially when the number of possible predictor variables is large. Within the class of shrinkage priors for β, scale mixtures of normal distributions (Andrews and Mallows 1974; West 1987) have received extensive attention. The trivial one-component mixture prior for β was used by Raiffa and Schlaifer (1961), and Zellner and Siow (1980) considered a Cauchy prior, which can be represented as a mixture. A variety of scale-mixture priors for β have been investigated more recently by Fernández and Steel (2000), Griffin and Brown (2005, 2007, 2009), Liang et al. (2008) and Carvalho et al. (2008).

The particular scale-mixture of normals shrinkage prior for β that underlies Bayesian lasso regression is the double-exponential distribution. Study and use of the double-exponential prior distribution in regression problems have become popular in part due to connections to the lasso procedure of Tibshirani (1996): the posterior mode under the double-exponential prior is equivalent to the lasso estimate, $\hat{\beta}$. A salient feature of $\hat{\beta}$ is that it is possible that $\hat{\beta}_j = 0$ for each $j = 1, \ldots, p$, providing a method for identifying important predictor variables and improving on prediction when p is large. This "variable selection" property, though, is *ad hoc* from a Bayesian perspective. Under the absolutely continuous double-exponential prior distribution, the prior probability of the event $\{\beta_j = 0\}$ is zero, and so the posterior probability of such an event must also be zero. In order for posterior inferences about events such as $\{\beta_j = 0\}$ to be coherent, prior probability mass must be allocated to these events.

In the Bayesian setting, placing prior mass on the events $\{\beta_j = 0\}$ is akin to assigning a prior distribution to the space of regression models that are to be considered. Using standard notation, let γ be a p-vector where $\gamma_j = 1$ if predictor variable x_j is included in the regression model and $\gamma_j = 0$ otherwise. If the prior probability of a particular model is $\pi(\gamma)$, the posterior probability of the regression model is

$$\pi(\gamma \mid y) = \frac{m_\gamma(y)\pi(\gamma)}{\sum_{\gamma' \in \Gamma} m_{\gamma'}(y)\pi(\gamma')}, \tag{1}$$

where Γ is the collection of all possible models and

$$m_\gamma(y) = \int \mathrm{N}(y \mid X_\gamma \beta_\gamma, \sigma^2 I_n)\, \pi(\beta_\gamma, \sigma^2 \mid \gamma)\, d\beta_\gamma\, d\sigma^2 \tag{2}$$

is the marginal likelihood of the observed data under model γ. In this notation $\mathrm{N}(y \mid \cdot, \cdot)$ is the density function for a normal random variable evaluated at y, X_γ is the matrix of predictor variables corresponding to model γ, β_γ is the vector of regression coefficients corresponding to model γ and $\pi(\beta_\gamma, \sigma^2 \mid \gamma)$ is the prior distribution for the parameters in model γ. Important predictor variables can be identified by examining the marginal posterior inclusion probabilities $\pi(\gamma_j = 1 \mid y)$. If selection of one model for prediction is desired, Barbieri and Berger (2004) provide conditions under which the median probability model—defined to be the model containing all variables with $\pi(\gamma_j = 1 \mid y) \geq 1/2$—provides Bayes-optimal predictions, and suggest that their approach might be successful even when the conditions are not met. A commonly used approach is to base predictions on the highest posterior probability model, however such predictions are not always Bayes optimal. Other methods for choosing variables or models can be constructed using decision theoretic arguments (e.g. Bernardo and Smith 2000, Chap. 6). Knowledge of (or access to) the posterior distribution $\pi(\gamma \mid y)$ is required for all of these approaches.

The two major difficulties that arise when addressing regression model uncertainty in this manner are the evaluation of the integral in (2) for a given prior $\pi(\beta_\gamma, \sigma^2 \mid \gamma)$ and the ability to compute the summation in (1) for even moderately sized p. Both issues are a concern for the Bayesian lasso, and various approximations and alternative approaches have been used to avoid them. For example, Yuan and Lin (2005) avoid evaluating all 2^p marginal likelihoods in a search for the highest posterior probability model under their formulation of the Bayesian lasso by carefully restricting the parameters of their model to lie in a particular hyperplane. This allowed them to focus on a smaller subset of possible regression models, the marginal likelihoods of which they approximated using a Laplace approximation. While this approach provides a quick method for finding a high probability model, it does not address broader questions related to model uncertainty: because emphasis is placed on finding a single model, functionals of the model space posterior distribution—e.g., model-averaged predictions and variable inclusion probabilities—cannot be evaluated. Additionally, the restrictions placed on key parameters preclude a full Bayesian treatment of the model under this approach.

This paper introduces the tools that are required for addressing model uncertainty for Bayesian lasso regression. After a review of the Bayesian lasso regression model in Sect. 2, Sect. 3.1 describes how the marginal likelihood can be accurately evaluated when the model size is not too large, allowing for enumeration of the model space posterior distribution (1) when the total number of predictors p is modest. When p is large and we wish to consider models with moderate to large numbers of predictors, direct calculation of (1) and (2) will not be feasible. For these cases, a simple Markov chain Monte Carlo (MCMC) method for providing samples from (1) is described in Sects. 3.3–3.4. Model-averaged inference and prediction can be easily accomplished using the output from this Gibbs sampler. All computations described in this paper were implemented on a Mac Pro running Mac OS 10.6.1 with 8 GB of memory and dual 2.66 GHz quad-core Intel Xeon processors. The model space enumeration methods described in Sect. 3.1 were implemented in R (R Development Core Team 2009), and the MCMC methods described in Sects. 3.3 and 3.4 were implemented in the C++ programming language.

2 The Bayesian lasso regression model

In this section we briefly review the Bayesian lasso regression model in order to provide the proper context for the new material on model uncertainty presented in Sect. 3. Park and Casella (2008) consider the Bayesian lasso regression model

$$y \mid \beta, \sigma^2 \sim \mathrm{N}(X\beta, \sigma^2 I_n),$$

$$\beta_j \mid \sigma^2, \tau \overset{\text{iid}}{\sim} \mathrm{DE}(\tau/\sigma), \qquad j = 1, \ldots, p,$$

where $\mathrm{DE}(\tau/\sigma)$ is the double-exponential distribution with density function

$$p(\beta_j \mid \tau, \sigma^2) = \frac{\tau}{2\sigma} e^{-\tau|\beta_j|/\sigma}. \tag{3}$$

Throughout, we assume that y and the columns of X have been demeaned, and so an intercept term is not included in the model (although one could be easily accommodated).

For now we assume that σ^2 and τ are known values, however this will be relaxed later. For a given model γ with k_γ predictor variables, the key to assessing model uncertainty is the ability to evaluate

$$\begin{aligned} m_\gamma(y \mid \sigma^2, \tau) &= \int \mathrm{N}(y \mid X_\gamma \beta_\gamma, \sigma^2 I_n)\pi(\beta_\gamma \mid \sigma^2, \tau) d\beta_\gamma \\ &= \int (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}(y - X_\gamma\beta_\gamma)^T(y - X_\gamma\beta_\gamma)} \\ &\quad \times \left(\frac{\tau}{2\sigma}\right)^{k_\gamma} e^{-\tau\|\beta_\gamma\|_1/\sigma}\, d\beta_\gamma, \end{aligned} \tag{4}$$

where $\|\beta\|_1$ is the L_1-norm of β.

Most of the study and application of the Bayesian lasso regression model (Fernández and Steel 2000; Park and Casella 2008; Yi and Xu 2008) has focused on the scale mixture of normals representation of the double-exponential distribution, where latent variables are used to create a hierarchical representation of the prior distribution. This formulation of the model admits a simple Gibbs sampler for obtaining draws from the posterior distribution of β for a fixed model, however it does not lead to a simple expression for the marginal likelihood. Rather than working with the scale-mixture representation, we consider the direct representation of the posterior distribution of β provided by Hans (2009), which will facilitate both calculation of the marginal likelihood (Sect. 3.1) and computational approaches for addressing model uncertainty (Sect. 3.3).

By breaking the density function for the double-exponential distribution (3) into separate positive and negative components, Hans (2009) shows that for a given set of $p \le n$ predictor variables, the posterior distribution of β is a mixture of orthant-specific normal distributions:

$$\pi(\beta \mid \sigma^2, \tau, y) = \sum_{z \in \mathcal{Z}_p} \omega_z\, \mathrm{N}^{[z]}(\beta \mid \mu_z\,,\, \sigma^2(X^TX)^{-1}). \tag{5}$$

The sum is taken over the set $\mathcal{Z}_p = \{-1, 1\}^p$ which represents the 2^p orthants of $\mathbb{R}^p$. The orthant corresponding to a given $z \in \mathcal{Z}_p$ is defined to be $\mathcal{O}_z = R_{z_1} \times \cdots \times R_{z_p}$ where R_{z_j} is $[0, \infty)$ if $z_j = 1$ and is $(-\infty, 0)$ if $z_j = -1$. Each term in the sum contains a normalized density function for a normal distribution restricted to lie in a particular orthant:

$$\mathrm{N}^{[z]}(\beta \mid m, S) \equiv \frac{\mathrm{N}(\beta \mid m, S)}{\mathrm{P}(z, m, S)} \mathbf{1}(\beta \in \mathcal{O}_z),$$

$$\text{where } \mathrm{P}(z, m, s) = \int_{\mathcal{O}_z} \mathrm{N}(t \mid m, S) dt.$$

The location vector for each term in the sum depends on the orthant: $\mu_z = \hat{\beta}_{\text{OLS}} - \tau\sigma(X^TX)^{-1}z$, where $\hat{\beta}_{\text{OLS}}$ is the least-squares estimate $(X^TX)^{-1}X^Ty$. Each term in (5) also contains a weight,

$$\omega_z = \omega^{-1} \frac{\mathrm{P}(z, \mu_z, \sigma^2(X^TX)^{-1})}{\mathrm{N}(0 \mid \mu_z,\ \sigma^2(X^TX)^{-1})},$$

$$\text{where } \omega = \sum_{z \in \mathcal{Z}_p} \frac{\mathrm{P}(z, \mu_z, \sigma^2(X^TX)^{-1})}{\mathrm{N}(0 \mid \mu_z,\ \sigma^2(X^TX)^{-1})},$$

which makes (5) a properly normalized density function.

When $p > n$ one cannot represent the density function of the posterior distribution as in (5). In this case, the surface of the likelihood (as a function of β) will be flat in a $p - n$ dimensional subspace, meaning that on this subspace the posterior distribution will have exponential tails (due to the prior distribution). Along any direction that does not lie in this subspace, the posterior will have normal tails, as it does in (5). While this complicates the writing of an expression for the posterior density function, it will not cause problems for addressing model uncertainty via the computational methods described in Sects. 3.3 and 3.4.

3 Addressing model uncertainty

3.1 Marginal likelihood

The results of Hans (2009)—which provide expression (5) for the posterior distribution of β for a given model—are extended in this section to provide a new, simple expression for the marginal likelihood. Breaking integral (4) into a sum of integrals over each orthant reveals that the marginal likelihood for a particular model γ is

$$m_\gamma(y \mid \sigma^2, \tau) = \omega_\gamma \left(\frac{\tau}{2\sigma}\right)^{k_\gamma} \mathrm{N}(y \mid 0, \sigma^2 I_n), \tag{6}$$

where $k_\gamma = \sum_{l=1}^p \gamma_l$ is the number of variables included in model γ and ω_γ is the same as ω but computed using only those predictor variables in model γ:

$$\omega_\gamma = \sum_{z \in \mathcal{Z}_{k_\gamma}} \frac{\mathrm{P}(z, (X_\gamma^TX_\gamma)^{-1}(X_\gamma^Ty - \tau\sigma z), \sigma^2(X_\gamma^TX_\gamma)^{-1})}{\mathrm{N}(0 \mid (X_\gamma^TX_\gamma)^{-1}(X_\gamma^Ty - \tau\sigma z), \sigma^2(X_\gamma^TX_\gamma)^{-1})}. \tag{7}$$

When $k_\gamma = 0$, ω_γ is defined to be one. As in Sect. 2, expression (6) only holds if $k_\gamma \le n$. Computing the marginal likelihood for a model γ is easy if the model size is not too large: computing ω_γ for a model of size k_γ only requires evaluating 2^{k_γ} k_γ-dimensional multivariate normal orthant integrals, which can be evaluated numerically with high accuracy (e.g. Genz 1992). If k_γ is large, computation of ω_γ is difficult. Increasing the value of k_γ increases both (i) the number of integrals that must be evaluated and

Table 1 Posterior marginal inclusion probabilities for the diabetes data example. "ML" refers to direct computation of $m_\gamma(y \mid \sigma^2, \tau)$ or $m_\gamma(y \mid \tau)$ using equation (6); "MCMC" refers to estimation of the probabilities based on 1,500,000 iterations of the appropriate Gibbs sampler. When the parameters are not fixed, they have priors $\pi(\sigma^2) = \sigma^{-2}$, $\tau \sim \text{Gamma}(1, 1)$ and $\rho \sim \text{Unif}(0, 1)$. The horizontal lines in the table separate different models. Run times are provided in minutes; the run time for * was 5.62 hours

Fixed Parameters	Method	AGE	SEX	BP	S1	S2	S3	S4	S6	Time
$\sigma^2 = 1$, $\tau = 4.25$,	ML	.192	.776	.983	.519	.372	.696	.402	.251	8.57
$\rho = 0.5$	MCMC	.192	.775	.983	.560	.372	.695	.401	.251	1.32
$\sigma^2 = .492$, $\tau = 4.25$,	ML	.191	.991	1.000	.658	.435	.797	.473	.307	8.19
$\rho = 0.5$	MCMC	.191	.991	1.000	.658	.436	.796	.473	.307	1.39
$\tau = 4.25$, $\rho = 0.5$	ML	.191	.987	1.000	.650	.432	.795	.470	.304	*
	MCMC	.191	.990	1.000	.660	.435	.793	.476	.307	1.51
$\rho = 0.5$	MCMC	.130	.989	1.000	.659	.429	.696	.414	.217	1.52
	MCMC	.381	.995	1.000	.816	.658	.781	.651	.503	1.72

(ii) the dimension of the integrals. The increase associated with (i) eventually results in unreasonably long compute times, while the increase associated with (ii) eventually results in inaccurate estimates of each integral and, consequently, an inaccurate estimate of ω_γ. If the total number of predictors, p, is not too large, all 2^p posterior model probabilities $\pi(\gamma \mid \sigma^2, \tau, y)$ can be accurately computed for a given prior $\pi(\gamma)$, allowing for model averaged inference. The "large p" scenario is considered in Sect. 3.3.

As an example, consider the diabetes dataset of Efron et al. (2004), which consists of measurements on $n = 442$ patients with diabetes. The response variable—a one-year measure of disease progression—is to be predicted using the information in $p = 10$ clinical covariates. The predictor variables x_j and the outcome variables y have been demeaned and then standardized to have unit sample variance. The marginal likelihoods $m_\gamma(y \mid \sigma^2, \tau)$ for all 1,024 models were computed twice, once while fixing $\sigma^2 = 1$ (an estimate of the residual variance under the null model) and once while fixing $\sigma^2 = 0.492$ (an estimate of the residual variance under the full model). The fixed value $\tau = 4.25$ was used in all calculations and was chosen because it represents a reasonable amount of penalization for this dataset. The marginal likelihoods were then converted into posterior probabilities $\pi(\gamma \mid y, \sigma^2, \tau)$, where the prior distribution on the model space was specified so that $\gamma_j \overset{\text{iid}}{\sim} \text{Bernoulli}(\rho = 0.5)$. Under this prior, all models are *a priori* equally likely.

The purpose of this example is not to attempt a sophisticated analysis of the data, but rather to illustrate that these calculations are feasible for a dataset with $p = 10$. The posterior probabilities computed here will be compared, in Sect. 3.3, to an MCMC approach designed to estimate the same probabilities. Agreement of the probabilities computed using both approaches is an indication that accurate calculations can be made using either method.

The first and third rows of Table 1 show the posterior variable inclusion probabilities $\pi(\gamma_j = 1 \mid y, \sigma^2, \tau)$ for eight of the ten variables. The two missing variables, BMI and S5, had inclusion probabilities of approximately 1.000 across all rows of the table, and so they are not displayed. The time required to compute all 1,024 marginal likelihoods was just over eight minutes, a reasonable amount of time for a one-time computation. Roughly speaking, under both fixed values of σ^2, there is strong evidence that BMI, BP and S5 are important predictors, moderate evidence for S3 and S1, and weaker evidence for S4, S2, S6 and AGE. There is also fairly strong evidence that SEX is an important predictor, especially when σ^2 is fixed at 0.492. Except for AGE, the variable inclusion probabilities are all higher under the model where $\sigma^2 = 0.492$, which is not surprising as this value of σ^2 is an estimate of residual variation under the full model.

In general, it is not desirable to condition on a particular value of σ^2. Different values of σ^2 represent different amounts of residual variation, which in turn can correspond to different regions of the model space. When σ^2 is unknown and given a prior distribution $\pi(\sigma^2)$, the marginal likelihood becomes $m_\gamma(y \mid \tau) = \int m_\gamma(y \mid \sigma^2, \tau)\pi(\sigma^2)d\sigma^2$. Due to the complicated way in which σ^2 appears in (6) through the term ω_γ, an analytic solution to this integral for standard choices of $\pi(\sigma^2)$ is not obvious. Fortunately, the integral is one dimensional, and standard techniques can be used to evaluate the integral numerically. Such techniques typically require repeated evaluations of $m_\gamma(y \mid \sigma^2, \tau)$ at many different values of σ^2, which is feasible if k_γ is not too large. Applying this approach to the diabetes dataset using the prior $\pi(\sigma^2) \propto \sigma^{-2}$ (propriety of the posterior was shown by Park and Casella 2008), the calculation of all 1,024 marginal likelihoods took over five hours. The fifth row of Table 1 shows the resulting posterior marginal inclusion probabilities. The results will be compared, in Sect. 3.4, to an

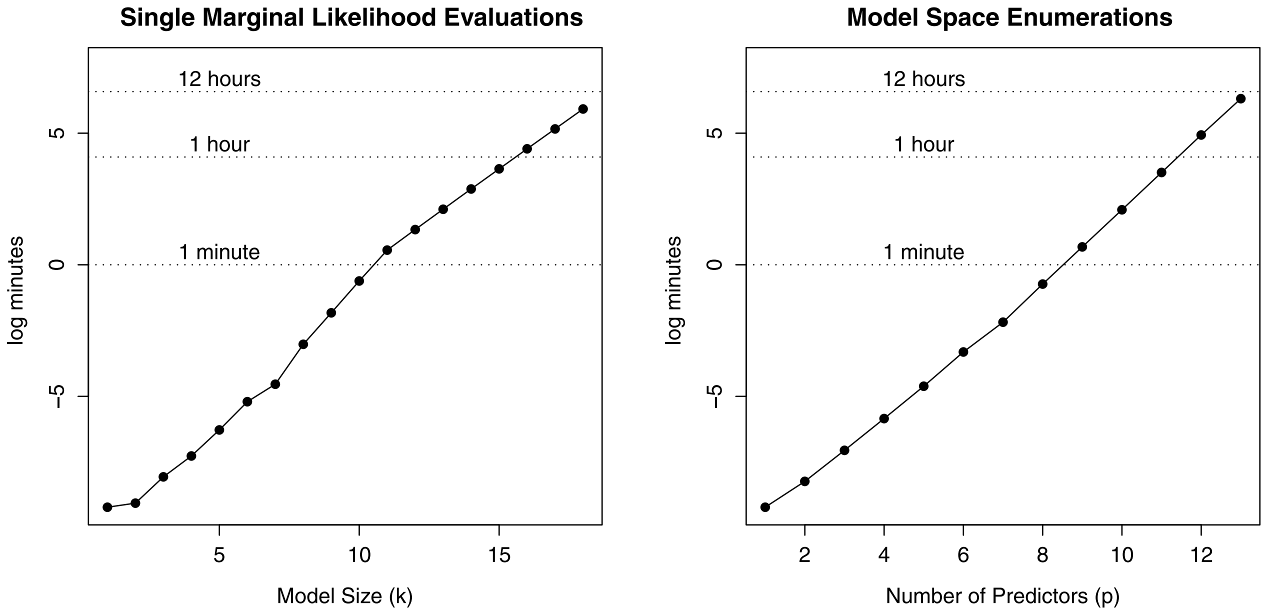

Fig. 1 The *left panel* displays the logarithm of the time required (in minutes) to compute the marginal likelihood for models of sizes $k = 1, \dots, 18$. The *right panel* displays the logarithm of the time required (in minutes) to enumerate the model space as the total number of candidate predictor variables increases from $p = 1$ to $p = 13$

MCMC approach for estimating the same quantities that requires much less computing time.

3.2 Computational limitations

Our ability to compute a single marginal likelihood $m_\gamma(y \mid \sigma^2, \tau)$ in a reasonable amount of time depends on the size of model, as computation of ω_γ requires the evaluation of 2^{k_γ} ratios as in (7). Assuming that the bulk of the computing time is spent computing these ratios, the logarithm of the time required to compute a marginal likelihood should scale approximately linearly in k_γ. The practical extent of this limitation was examined using the diabetes dataset. Starting with the model containing only the predictor AGE, variables were added to the model one at a time, and the time required to compute the marginal likelihood for each model in this sequence was recorded. When the ten variables in the dataset were exhausted, white-noise predictors were added until a total of 18 variables were in the model. The left panel of Fig. 1 displays the logarithm of the time required (in minutes) to make these calculations, which scales approximately linearly in k_γ as expected. It took less than one minute to compute $m_\gamma(y \mid \sigma^2, \tau)$ for models with $k_\gamma \le 10$, forty minutes for a model of size $k_\gamma = 15$ and six hours for a model of size $k_\gamma = 18$. Computation of the multivariate normal probabilities in (7) was performed using the `R` package `mvtnorm` (Genz et al. 2009).

Our ability to enumerate the model space in a reasonable amount of time depends on the total number of predictor variables p. To enumerate the space, 2^p marginal likelihoods must be computed, and for a given model of size k_γ, 2^{k_γ} ratios must be computed as in (7). This means that $\sum_{k=0}^{p} \binom{p}{k} 2^k = 3^p$ ratios must be computed in order to enumerate the model space. Again assuming that computation of these ratios dominates the total time required to compute a marginal likelihood, the logarithm of the time required to enumerate the model space should scale approximately linearly in p. The practical extent of this limitation on model space enumeration was examined in the same way as above by constructing a nested sequences of model spaces for $p = 1, \dots, 13$; this time, instead of computing a single marginal likelihood, all 2^p marginal likelihoods were computed for each p. The right panel of Fig. 1 displays the logarithm of the time required (in minutes) to make these calculations, which is approximately locally linear in p. Model space enumeration can be done in less than a minute for $p \le 8$, in about thirty minutes for $p = 11$, two hours for $p = 12$ and nine hours for $p = 13$. Extrapolating, for this example it would take between one and two days to enumerate a model space with $p = 14$.

3.3 Model uncertainty in higher dimensions

As seen above, computation of the marginal likelihood is intractable in practice for large values of k_γ, and model space enumeration is difficult when p is larger than 12 or 13. Additionally, if $p > n$ we can only use expression (6) to compute the marginal likelihood of models of size $k_\gamma \le n$. We can, however, construct a simple Gibbs sampler which will allow us to obtain samples from the posterior distribution of γ, providing a computational approach for addressing model uncertainty for the Bayesian lasso regression model that can be used for any number of predictor variables. Assuming that $\gamma_j \mid \rho \overset{\text{iid}}{\sim} \text{Bernoulli}(\rho)$, the prior distribution for β_j can be written as the mixture

$$\pi(\beta_j \mid \sigma^2, \tau, \rho) = (1-\rho)\delta_0(\beta_j) + \rho\left(\frac{\tau}{2\sigma}\right)e^{-\tau|\beta_j|/\sigma}, \qquad (8)$$

where $\delta_0(\beta_j)$ is a point mass at zero. The case where a normal distribution is used in place of the double-exponential distribution has been studied extensively; different forms of that prior distribution and various computational treatments for related priors have been considered by George and McCulloch (1993, 1997), Carlin and Chib (1995), Geweke (1996), Smith and Kohn (1996), Raftery et al. (1997) and Kuo and Mallick (1998), among others.

Using the mixture representation (8) of the prior, an iteration of the Gibbs sampler cycles through the full conditional distributions $\beta_j \mid \beta_{-j}, \sigma^2, \tau, y$, $j = 1, \ldots, p$, where β_{-j} contains all elements of β except for β_j. Separating the portion of (8) corresponding to the double-exponential distribution into separate positive and negative components, the full conditional distribution of β_j is revealed to be a mixture with two components:

$$\begin{aligned}\pi(\beta_j &\mid \beta_{-j}, \sigma^2, \tau, y)\\ &= \phi_{0j}\delta_0(\beta_j) + (1-\phi_{0j})\left\{\phi_j \mathrm{N}^+(\beta_j \mid \mu_j^+, s_j^2)\right.\\ &\quad \left. + (1-\phi_j)\mathrm{N}^-(\beta_j \mid \mu_j^-, s_j^2)\right\}. \end{aligned} \qquad (9)$$

The first component is a point mass at zero with corresponding weight

$$\begin{aligned}\phi_{0j} &\equiv \Pr(\beta_j = 0 \mid \beta_{-j}, \sigma^2, \tau, \rho, y)\\ &= \left[1 + \frac{\rho}{1-\rho} \times \frac{\tau}{2\sigma}\left\{\frac{\Phi(\mu_j^+/s_j)}{\mathrm{N}(0 \mid \mu_j^+, s_j^2)}\right.\right.\\ &\quad \left.\left. + \frac{\Phi(-\mu_j^-/s_j)}{\mathrm{N}(0 \mid \mu_j^-, s_j^2)}\right\}\right]^{-1},\end{aligned}$$

where Φ is the univariate standard normal distribution function. The parameters of this weight are $s_j^2 = \sigma^2/(x_j^T x_j)$ and $\mu_j^+ = (x_j^T x_j)^{-1}\{x_j^T(y - X_{-j}\beta_{-j}) - \tau\sigma/(x_j^T x_j)\}$, where X_{-j} is the matrix X with the jth column removed. Some elements of β_{-j} may be equal to zero, allowing for the possibility of a computational speed up when computing $(x_j^T X_{-j})\beta_{-j}$. The parameter μ_j^- is the same as μ_j^+, with the exception that $-\tau\sigma/(x_j^T x_j)$ is replaced with $+\tau\sigma/(x_j^T x_j)$.

The second component of the mixture in (9) has weight $1 - \phi_{0j}$ and is itself a mixture with two components: a normal distribution restricted to be positive (with weight ϕ_j) and a normal distribution restricted to be negative (with weight $1 - \phi_j$). In the notation above, N^+ and N^- are the univariate truncated normal distributions with density functions

$$\mathrm{N}^+(t \mid m, s^2) = \frac{\mathrm{N}(t \mid m, s^2)}{\Phi(m/s)}\mathbf{1}(t > 0) \quad \text{and}$$

$$\mathrm{N}^-(t \mid m, s^2) = \frac{\mathrm{N}(t \mid m, s^2)}{\Phi(-m/s)}\mathbf{1}(t < 0),$$

and the weight in favor of the positive component is

$$\begin{aligned}\phi_j = &\left\{\frac{\Phi(\mu_j^+/s_j)}{\mathrm{N}(0 \mid \mu_j^+, s_j^2)}\right\}\\ &\bigg/ \left\{\frac{\Phi(\mu_j^+/s_j)}{\mathrm{N}(0 \mid \mu_j^+, s_j^2)} + \frac{\Phi(-\mu_j^-/s_j)}{\mathrm{N}(0 \mid \mu_j^-, s_j^2)}\right\}.\end{aligned}$$

An alternate view of (9) is that the full conditional distribution is a three component mixture: a point mass at zero, a positive normal distribution and a negative normal distribution with weights ϕ_{0j}, $(1-\phi_{0j})\phi_j$ and $(1-\phi_{0j})(1-\phi_j)$, respectively, which sum to one.

The computational advantage of this Gibbs sampling approach over direct calculation of the marginal likelihood (6) is evident in the form of the full conditional distribution. Rather than having to compute 2^{k_γ} k_γ-dimensional multivariate normal orthant probabilities, sampling from the full conditional requires computation of only two one-dimensional normal probabilities, which can be done quickly and with high accuracy. Additionally, Rao–Blackwellized estimates of the posterior variable inclusion probabilities are easily obtained by averaging the values of $1 - \phi_{0j}$ across MCMC iterations for each variable.

The Gibbs sampler was implemented for the diabetes dataset under the conditions $\tau = 4.25$ and $\rho = 0.5$. The fixed values $\sigma^2 = 1$ and $\sigma^2 = .492$ were both considered, and two separate runs of length 1,500,000 were performed. The values obtained using the MCMC approach (see Table 1) match very well with those obtained in Sect. 3.1 based on direct calculation of the marginal likelihoods (6), providing confidence that both methods can produce stable and accurate answers. The MCMC approach, however, produced the results in less time.

3.4 MCMC when σ^2, τ and ρ are unknown

As discussed in Sect. 3.1, fixing σ^2 at a particular value is undesirable, especially when multiple models are to be compared. The same is true for τ, the penalty parameter, and ρ, the hyperparameter controlling the number of zeroed-out regression coefficients. In particular, Scott and Berger (2008) show that fixing ρ in seemingly reasonable ways can provide unsatisfactory results as p grows large. A better approach is to assign these parameters prior distributions and add extra steps to the Gibbs sampler to update them from their full conditional distributions.

A common choice of prior for the residual variance is $\sigma^2 \sim \mathrm{IG}(a, b)$, the inverse gamma distribution. The limiting case $a \to 0$ and $b \to 0$ yields the improper prior $\pi(\sigma^2) = \sigma^{-2}$. Under this prior, the density function for the full conditional distribution is

$$\pi(\sigma^2 \mid \beta, \tau, y) \propto (\sigma^2)^{-(a^*+1)} \exp(-b^*/\sigma^2 - \tau\|\beta\|_1/\sigma),$$

where $a^* = (n + k_\gamma)/2 + a$, k_γ is the number of nonzero elements of β and $b^* = (y - X\beta)^T(y - X\beta)/2 + b$. This is not a standard distribution, however samples can be obtained via a rejection sampling method after a suitable transformation has been applied (see Hans 2009, for details). The sixth row of Table 1 displays MCMC estimates of the marginal variable inclusion probabilities for the diabetes data when $\pi(\sigma^2) = \sigma^{-2}$, $\tau = 4.25$ and $\rho = 0.5$. The estimates match fairly well with those in the fifth line, obtained via numerical integration of $m_\gamma(y \mid \sigma^2, \tau)$ in Sect. 3.1. Experience with MCMC for similar models suggests that the MCMC-based estimates should be accurate, and so the discrepancies are likely due to the numerical integration procedure. It is encouraging, though, that when τ and ρ are fixed, marginal likelihoods $m_\gamma(y \mid \tau)$ can be computed with relatively high accuracy without needing to resort to MCMC. Note, however, that the MCMC approach required only 1.5 minutes, compared to 5.62 hours for the model space enumeration using numerical integration.

A reasonable prior distribution for the penalty parameter is $\tau \sim \mathrm{Gamma}(r, s)$. Under this prior, the resulting full conditional distribution is

$$\tau \mid \beta, \sigma^2, y \sim \mathrm{Gamma}(k_\gamma + r, \sigma^{-1}\|\beta\|_1 + s),$$

where again k_γ is the number of nonzero elements of β. The seventh row of Table 1 displays MCMC estimates of the marginal inclusion probabilities for the diabetes data when σ^2 is modeled as above, τ is assigned the Gamma(1, 1) prior distribution, and ρ is fixed at 0.5. The probabilities are, in general, slightly lower when uncertainty about τ is averaged over, suggesting the posterior has moved slightly toward smaller-sized models.

A commonly used prior distribution for the sparsity parameter is $\rho \sim \mathrm{Beta}(g, h)$. See George and McCulloch (1993, 1997), George and Foster (2000), Chipman et al. (2001), Kohn et al. (2001), Ley and Steel (2007), Cui and George (2008) and Scott and Berger (2008) for various treatments and discussion of this important parameter. Under the beta prior distribution, the full conditional distribution is

$$\rho \mid \beta, \sigma^2, \tau, y \sim \mathrm{Beta}(g + k_\gamma, h + p - k_\gamma).$$

The final row of Table 1 displays MCMC estimates of the marginal inclusion probabilities for the diabetes data when $\rho \sim \mathrm{Beta}(1, 1)$ (the uniform distribution) and σ^2 and τ are modeled as above. The posterior probabilities under this more general model are much higher than they were under the previous models, suggesting that the data support larger models than those favored by the prior with $\rho = 0.5$; indeed, the posterior mean of ρ is $\mathrm{E}[\rho \mid y] = 0.732$. The usual Bayesian point estimates for σ^2 and τ under this model are $\mathrm{E}[\sigma^2 \mid y] = 0.493$ and $\mathrm{E}[\tau \mid y] = 2.93$.

3.5 Additional approaches to model space MCMC

The MCMC algorithms described above construct Markov chains over the joint parameter and model spaces, requiring that a sample of β be obtained at each iteration. An alternate approach to regression model space MCMC, described by Smith and Kohn (1996) and George and McCulloch (1997), constructs a Markov chain directly on the model indicator γ by marginalizing over β. Given fixed values of σ^2, τ and ρ, one iteration of this MCMC algorithm for the Bayesian lasso regression model cycles through the full conditional distributions $\gamma_j \mid \gamma_{-j}, \sigma^2, \tau, \rho, y$, where

$$\Pr(\gamma_j = 1 \mid \gamma_{-j}, \sigma^2, \tau, y) = \left(1 + \frac{1-\rho}{\rho} \times \frac{m_{\gamma_0}(y \mid \sigma^2, \tau)}{m_{\gamma_1}(y \mid \sigma^2, \tau)}\right)^{-1},$$

$m_{\gamma_1}(y \mid \sigma^2, \tau)$ is the marginal likelihood for the model with γ_{-j} fixed as indicated in the conditioning statement and $\gamma_j = 1$, and $m_{\gamma_0}(y \mid \sigma^2, \tau)$ is the marginal likelihood for the same model, but with $\gamma_j = 0$. The advantage of constructing a Markov chain in this fashion is that β is never sampled, which may result in faster convergence to the target distribution $\pi(\gamma \mid y, \sigma^2, \tau, \rho)$. The main disadvantage to using this approach in the Bayesian lasso setting is that because computing the marginal likelihood for large models can take an unreasonable amount of time, the time required to complete a cycle of the sampler will increase greatly as the Markov chain transitions to models with large numbers of predictor variables. Constructing a Markov chain over the entire model space in this fashion is not feasible unless the total number of predictor variables, p, is small, in which case model space enumeration can be accomplished as in Sect. 3.1.

It is sometimes the case, especially when p is very large, that investigators are not interested in the entire space of 2^p models and that, instead, interest is on the subspace where model size is restricted to be no larger than a small number k^*. The restriction on the model space can be encoded in the prior via a constraint: $\pi(\gamma) \propto \rho^{k_\gamma}(1-\rho)^{p-k_\gamma}\mathbf{1}(k_\gamma \leq k^*)$. With this cap on model size in place, MCMC over the restricted space is feasible; however, experience with this approach for the Bayesian lasso regression model suggests that even when k^* is small (≈ 6) and p is not too large (≈ 100), the potential gains obtained by not sampling β are outweighed by the increased computational costs.

4 Discussion and extensions

This paper introduced the necessary tools for addressing regression model uncertainty for the Bayesian lasso. The marginal likelihood was seen to be accurately computable when the number of predictor variables is modest, allowing for model space enumeration and model-averaged inferences. When the number of predictor variables is large, it was shown that functionals of the model space posterior distribution could be accurately estimated using a simple Gibbs sampler, the form of which is similar to the stochastic search variable selection (SSVS) samplers that are now widely used for large p regression problems.

When p is very large, the MCMC algorithms described in Sect. 3.3 will encounter the same problems that the existing SSVS methods encounter. In these cases, the model space is so large that all models cannot be visited by the MCMC algorithm, meaning that estimates of model probabilities $\pi(\gamma \mid y)$ cannot be accurately obtained. It is often the case, though, that the regions of high posterior probability will be visited sufficiently often that estimates of quantities such as $\pi(\gamma_j = 1 \mid y)$ will be reasonably accurate. In this case, the MCMC method described in Sect. 3.3 can be used as a screening method to identify the important predictors.

Computation of the marginal likelihood (6) can be speeded up by noting that each element of the sum that defines ω_γ can be computed independently in parallel. This makes use of the double-exponential prior particularly applicable to parallel computing based stochastic search algorithms such as the SSS approach of Hans et al. (2007) that seek to explore the space of low dimensional regression models. Such methods aim to rapidly explore and catalogue neighborhoods of high probability models that can be used to perform approximate model-averaged inference. The methods introduced in Sect. 3 make it feasible to use the double-exponential prior in conjunction with such search algorithms.

Software for implementing the methods described in Sect. 3, written in C++ with an R package interface, is available at http://www.stat.osu.edu/~hans/software.html.

Acknowledgements This research was supported by National Science Foundation grant DMS-0706948. The author would like to thank the Guest Editor and two anonymous reviewers for comments that improved the paper.

References

Andrews, D., Mallows, C.: Scale mixtures of normal distributions. J. R. Stat. Soc., Ser. B **36**, 99–102 (1974)

Barbieri, M.M., Berger, J.O.: Optimal predictive model selection. Ann. Stat. **32**, 870–897 (2004)

Bernardo, J., Smith, A.: Bayesian Theory. Wiley, New York (2000)

Carlin, B., Chib, S.: Bayesian model choice via Markov chain Monte Carlo methods. J. R. Stat. Soc., Ser. B **57**, 473–484 (1995)

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

Chipman, H.A., George, E.I., McCulloch, R.E.: The practical implementation of Bayesian model selection (with discussion). In: Lahiri, P. (ed.) Model Selection, pp. 65–134. IMS, Beachwood (2001)

Cui, W., George, E.I.: Empirical Bayes vs. fully Bayes variable selection. J. Stat. Plan. Inference **138**, 888–900 (2008)

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

Fernández, C., Steel, M.: Bayesian regression analysis with scale mixtures of normals. Econom. Theory **16**, 80–101 (2000)

Genz, A.: Numerical computation of multivariate normal probabilities. J. Comput. Graph. Stat. **1**, 141–150 (1992)

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Hothorn, T.: mvtnorm: Multivariate Normal and *t* Distributions. R package version 0.9-7 (2009)

George, E.I., Foster, D.P.: Calibration and empirical Bayes variable selection. Biometrika **87**, 731–747 (2000)

George, E.I., McCulloch, R.E.: Variable selection via Gibbs sampling. J. Am. Stat. Assoc. **88**, 881–889 (1993)

George, E.I., McCulloch, R.E.: Approaches for Bayesian variable selection. Stat. Sin. **7**, 339–373 (1997)

Geweke, J.: Variable selection and model comparison in regression. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) Bayesian Statistics 5, pp. 609–620. Oxford Press, London (1996)

Griffin, J., Brown, P.: Alternative prior distributions for variable selection with very many more variables than observations. Tech. Rep., University of Warwick (2005)

Griffin, J., Brown, P.: Bayesian adaptive lassos with non-convex penalization. Tech. Rep., University of Warwick (2007)

Griffin, J., Brown, P.: Inference with Normal-Gamma prior distributions in regression problems. Tech. Rep., University of Warwick (2009)

Hans, C.: Bayesian lasso regression. Biometrika Advance Access published September 24, 2009, doi:10.1093/biomet/asp047

Hans, C., Dobra, A., West, M.: Shotgun stochastic search for "large *p*" regression. J. Am. Stat. Assoc. **102**, 507–516 (2007)

Kohn, R., Smith, M., Chan, D.: Nonparametric regression using linear combinations of basis functions. Stat. Comput. **11**, 313–322 (2001)

Kuo, L., Mallick, B.: Variable selection for regression models. Sankhyā Ser. B **60**, 65–81 (1998)

Ley, E., Steel, M.: On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. Policy Research Working Paper 4238. World Bank (2007)

Liang, F., Paulo, R., Molina, G., Clyde, M., Berger, J.O.: Mixtures of *g*-priors for Bayesian variable selection. J. Am. Stat. Assoc. **103**, 410–423 (2008)

Park, T., Casella, G.: The Bayesian lasso. J. Am. Stat. Assoc. **103**, 681–686 (2008)

R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2009). ISBN 3-900051-07-0

Raftery, A.E., Madigan, D., Hoeting, J.: Bayesian model averaging for linear regression models. J. Am. Stat. Assoc. **92**, 1197–1208 (1997)

Raiffa, H., Schlaifer, R.: Applied Statistical Decision Theory. Graduate School of Business Administration, Harvard University (1961)

Scott, J.G., Berger, J.O.: Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Discussion Paper 2008-10, Duke University Department of Statistical Science (2008)

Smith, M., Kohn, R.: Nonparametric regression using Bayesian variable selection. J. Econom. **75**, 317–343 (1996)

Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc., Ser. B **58**, 267–288 (1996)

West, M.: On scale mixtures of normal distributions. Biometrika **74**, 646–648 (1987)

Yi, N., Xu, S.: Bayesian LASSO for quantitative trait loci mapping. Genetics **179**, 1045–1055 (2008)

Yuan, M., Lin, Y.: Efficient empirical Bayes variable selection and estimation in linear models. J. Am. Stat. Assoc. **100**, 1215–1225 (2005)

Zellner, A., Siow, A.: Posterior odds ratios for selected regression hypotheses. In: Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia, pp. 585–603 (1980)