Biometrika (2009), **96**, 4, *pp*. 835–845
doi: 10.1093/biomet/asp047
Advance Access publication 24 September 2009

Bayesian lasso regression

BY CHRIS HANS

Department of Statistics, The Ohio State University, Columbus, Ohio 43210, U.S.A.
hans@stat.osu.edu

SUMMARY

The lasso estimate for linear regression corresponds to a posterior mode when independent, double-exponential prior distributions are placed on the regression coefficients. This paper introduces new aspects of the broader Bayesian treatment of lasso regression. A direct characterization of the regression coefficients' posterior distribution is provided, and computation and inference under this characterization is shown to be straightforward. Emphasis is placed on point estimation using the posterior mean, which facilitates prediction of future observations via the posterior predictive distribution. It is shown that the standard lasso prediction method does not necessarily agree with model-based, Bayesian predictions. A new Gibbs sampler for Bayesian lasso regression is introduced.

Some key words: Double-exponential distribution; Gibbs sampler; L_1 penalty; Laplace distribution; Markov chain Monte Carlo; Posterior predictive distribution; Regularization.

1. INTRODUCTION

The lasso of Tibshirani (1996) has become a widely used alternative to ordinary least squares for parameter estimation in regression problems. Its popularity is due in part to a key feature of the procedure: shrinkage of the vector of regression coefficients toward zero with the possibility of setting some coefficients identically equal to zero, resulting in a simultaneous estimation and variable selection procedure. The lasso is a form of penalized least squares that minimizes the residual sum of squares while controlling the L_1-norm of the coefficient vector β:

$$\hat{\beta}_{\mathrm{L}} = \mathrm{argmin}_{\beta}(y - X\beta)^{\mathrm{T}}(y - X\beta) \; + \; \lambda\|\beta\|_1, \tag{1}$$

where $\lambda \geqslant 0$ determines the amount of shrinkage. The case $\lambda = 0$ results in $\hat{\beta}_{\mathrm{L}} = \hat{\beta}_{\mathrm{OLS}}$, the ordinary least-squares estimate, and λ sufficiently large shrinks $\hat{\beta}_{\mathrm{L}}$ to zero. The modified least angle regression algorithm (Efron et al., 2004) provides an efficient method for computing the entire path of lasso estimates as a function of λ.

The lasso has a Bayesian interpretation. Tibshirani (1996) notes that the lasso estimate can be viewed as the mode of the posterior distribution of β, $\hat{\beta}_{\mathrm{L}} = \mathrm{argmax}_{\beta} p(\beta \mid y, \sigma^2, \tau)$, when independent, double-exponential prior distributions are placed on the p regression coefficients,

$$p(\beta \mid \tau) = (\tau/2)^p \exp(-\tau\|\beta\|_1), \tag{2}$$

and the likelihood component is taken to be the normal linear regression model $p(y \mid \beta, \sigma^2) = \mathrm{N}(y \mid X\beta, \sigma^2 I_n)$. For any fixed values $\sigma^2 > 0$ and $\tau > 0$, the posterior mode of β is the lasso estimate with penalty $\lambda = 2\tau\sigma^2$. The double-exponential prior distribution can be represented as a scale mixture of normal distributions (Andrews & Mallows, 1974; West, 1987) and is not new to Bayesian inference. Its Bayesian robustness properties, especially in the normal mean problem,

have been examined by Spiegelhalter (1977), Pericchi & Walley (1991), Pericchi & Smith (1992) and Mitchell (1994). Particular emphasis has been placed on studying the effect of the prior on posterior moments.

The first explicit treatment of Bayesian lasso regression was provided by Park & Casella (2008). The earlier work of Fernández & Steel (2000) considered prior (2) as a special case in a general Bayesian regression modelling framework but did not make specific connections to the lasso procedure. Both used the scale mixture of normals representation of the double-exponential distribution to create a hierarchical formulation of the model by introducing a vector of latent scale variables θ. They obtain samples from the joint posterior distribution $p(\beta, \theta \mid y, \sigma^2, \tau)$ via a data-augmentation Gibbs sampler (Tanner & Wong, 1987) that iteratively samples from the full conditional distributions $p(\beta \mid \theta, y, \sigma^2, \tau)$ and $p(\theta \mid \beta, y, \sigma^2, \tau)$. Carlin & Polson (1991) and Carlin et al. (1992) show that, in the intercept-only, normal-mean setting, the full conditional for θ is a generalized inverse Gaussian distribution, from which samples can be easily obtained. Fernández & Steel (2000) and Park & Casella (2008) provide a similar result for the general regression setting. Park & Casella (2008) extend the Bayesian lasso regression model to account for uncertainty in the hyperparameters by placing prior distributions on σ^2 and τ^2. They obtain point estimates of the regression coefficients using the median of the posterior distribution, but they do not address prediction of future observations.

This paper presents several new contributions to the Bayesian treatment of lasso regression. A new, direct characterization of the posterior distribution $p(\beta \mid y, \sigma^2, \tau)$ is introduced, along with a discussion about estimation and prediction under the lasso from a model-based perspective. The Bayesian connection to the lasso is examined, with particular attention paid to the problem of predicting future observations via the posterior predictive distribution. The direct characterization of the posterior is shown to facilitate sampling from the posterior distribution via two new Gibbs samplers that do not require the use of latent variables.

2. THE LASSO POSTERIOR DISTRIBUTION

2·1. *Direct characterization*

A version of the Bayesian lasso regression model is

$$p(y \mid \beta, \sigma^2, \tau) = \mathrm{N}(y \mid X\beta, \sigma^2 I_n),$$
$$p(\beta \mid \tau, \sigma^2) = \left(\frac{\tau}{2\sigma}\right)^p \exp(\tau\sigma^{-1}\|\beta\|_1). \tag{3}$$

The expression $\mathrm{N}(t \mid m, S)$ represents the density function, evaluated at t, of a multivariate normal random variable with mean m and covariance matrix S. Assume that the observed data y and the columns of the $n \times p$ matrix X have been mean centred so that an intercept is not included in the model. Prior (3), introduced by Park & Casella (2008), is used instead of the usual lasso prior (2); the penalty parameter is now scaled by the square root of the error variance. Prior (3) retains the property that, for fixed τ and σ^2, the mode of $p(\beta \mid y, \sigma^2, \tau)$ is the lasso estimate with penalty parameter $\lambda = 2\tau\sigma$. The parameters σ^2 and τ are considered for the moment to be known, however this assumption is later relaxed.

A direct characterization of the posterior distribution $p(\beta \mid y, \sigma^2, \tau)$ that does not require the inclusion of latent variables is constructed as follows. Let $\mathcal{Z} = \{-1, 1\}^p$ represent the set of all 2^p possible p-vectors whose elements are ± 1. For any vector $z \in \mathcal{Z}$, let $\mathcal{O}_z \subset R^p$ represent the corresponding orthant: if $\beta \in \mathcal{O}_z$, then $\beta_j \geqslant 0$ if $z_j = 1$ and $\beta_j < 0$ if $z_j = -1$. Write the density

function for the orthant-truncated normal distribution and its associated orthant integrals as

$$\mathrm{N}^{[z]}(\beta \mid m, S) \equiv \frac{\mathrm{N}(\beta \mid m, S)}{\mathrm{P}(z, m, S)} 1(\beta \in \mathcal{O}_z), \qquad \mathrm{P}(z, m, S) = \int_{\mathcal{O}_z} \mathrm{N}(t \mid m, S)\, dt.$$

Applying Bayes' theorem to the lasso regression model, the posterior distribution is orthant-wise Gaussian,

$$p(\beta \mid y, \sigma^2, \tau) = \sum_{z \in \mathcal{Z}} \omega_z \mathrm{N}^{[z]}(\beta \mid \mu_z, \Sigma), \tag{4}$$

i.e. a collection of 2^p different normal distributions that are each restricted to a different orthant. The covariance structure in each of the 2^p orthants is the same, $\Sigma = \sigma^2 (X^{\mathrm{T}} X)^{-1}$, whereas the location parameters depend on the orthants: $\mu_z = \hat{\beta}_{\mathrm{OLS}} - \tau \sigma^{-1} \Sigma z$, where $\hat{\beta}_{\mathrm{OLS}} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y$. The normalized weight for each orthant is

$$\omega_z = \left\{ \frac{\mathrm{P}(z, \mu_z, \Sigma)}{\mathrm{N}(0 \mid \mu_z, \Sigma)} \right\} \Bigg/ \left\{ \sum_{z \in \mathcal{Z}} \frac{\mathrm{P}(z, \mu_z, \Sigma)}{\mathrm{N}(0 \mid \mu_z, \Sigma)} \right\}.$$

If the classic lasso prior (2) is used in place of (3), the posterior distribution has the same form; the only change is that the location parameters become $\mu_z = \hat{\beta}_{\mathrm{OLS}} - \tau \Sigma z$.

If good routines are available for computing the multivariate normal orthant integrals, the posterior distribution (4) can be integrated or sampled from when p is small, allowing for straightforward posterior inference and estimation of the posterior mean for fixed values of σ^2 and τ. For moderate to large p, posterior inference is easily accomplished via the Markov chain Monte Carlo methods introduced in § 3.

The model can be extended to account for uncertainty about the parameters σ^2 and τ by assigning the independent prior distributions $\sigma^{-2} \sim \mathrm{Ga}(a, b)$ and $\tau \sim \mathrm{Ga}(r, s)$. This is similar to the approach taken by Park & Casella (2008); they assign a Gamma prior distribution to τ^2 rather than τ, which leads to conditional conjugacy in their data-augmentation Gibbs sampler.

2·2. *Posterior-based estimation and prediction*

The Bayes rule for point estimation for fixed σ^2 and τ under a given loss function $l(b, \beta)$ is the estimator $\hat{b}$ that minimizes the expected posterior loss $\int l(b, \beta) p(\beta \mid y, \sigma^2, \tau) d\beta$. The posterior mean and median, for example, are the Bayes rules under squared-error loss and absolute-error loss, respectively. The lasso estimate of the regression coefficients corresponds to the mode of the posterior distribution of β, and so the lasso procedure (1) is sometimes referred to as a Bayesian procedure due to the claim that the posterior mode is the Bayes rule under the zero-one loss function: $l(b, \beta) = 0$ if $b = \beta$ and $l(b, \beta) = 1$ otherwise. Strictly speaking, when $\beta \in R^p$, this interpretation is not correct: under this loss function, the expected posterior loss is equal to one for all $b \in R^p$ and so any point $b \in R^p$ is a suitable $\hat{b}$. The posterior mode can, though, be interpreted as the limit of a sequence of Bayes rules. Consider the sequence of loss functions $l_\varepsilon(b, \beta) = 1 - 1\{\beta \in B_{\varepsilon(b)}\}$, where the indicator function equals one if an ε-ball centred at b contains β and is zero otherwise. In the limit as $\varepsilon \to 0$, the sequence of Bayes-optimal estimators $\hat{b}_\varepsilon$ converges to the posterior mode (Bernardo & Smith, 2000, pp. 257–258). The extent to which such a construction truly represents one's loss for a given decision problem is, of course, up to the individual.

Given a point estimate $\hat{\beta}$, it is useful if predictions of future observations $\tilde{y}$ at new values $\tilde{X}$ can be made via $\tilde{X}\hat{\beta}$. In a Bayesian setting, prediction of future observations $\tilde{y}$ at new values $\tilde{X}$ is based on the posterior predictive distribution $p(\tilde{y} \mid y, \sigma^2, \tau) = \int p(\tilde{y} \mid y, \beta, \sigma^2, \tau) p(\beta \mid y, \sigma^2, \tau) d\beta$. For a given loss function, predictions are made using the estimator that minimizes expected

posterior predictive loss. Under squared-error loss, predictions are made via the mean of the posterior predictive distribution. For this regression model, the posterior predictive mean is $E(\tilde{y} \mid y, \sigma^2, \tau) = \tilde{X}E(\beta \mid y, \sigma^2, \tau)$, which indicates that the posterior mean facilitates both point estimation and prediction. This is not the case under zero-one loss: the mode of the posterior predictive distribution is not necessarily $\tilde{X}\hat{\beta}_{\mathrm{L}}$, the standard lasso prediction. A simple example where this is the case is provided in § 2·3. If predictions were to be made under a limiting sequence of zero-one loss functions, numerical optimization procedures would be needed to maximize $p(\tilde{y} \mid y, \sigma^2, \tau)$. It is unclear if the standard lasso can be justified as a Bayesian predictive procedure.

2·3. *The univariate case*

Pericchi & Smith (1992) introduced the following representation of the posterior distribution of β in the univariate, intercept-only, normal-mean setting. Here the representation is extended to the general regression setting. For fixed σ^2 and τ, the univariate posterior is

$$p(\beta \mid y, \sigma^2, \tau) = w\, \mathrm{N}^{-}(\beta \mid \mu_-, \upsilon^2) \;+\; (1-w)\mathrm{N}^{+}(\beta \mid \mu_+, \upsilon^2),$$

where N^{-} and N^{+} correspond to $\mathrm{N}^{[z]}$ for $z = -1$ and $z = 1$, respectively. The common scale term is $\upsilon^2 = \sigma^2(x^{\mathrm{T}}x)^{-1}$, and the two location parameters are $\mu_+ = \hat{\beta}_{\mathrm{OLS}} - \tau\sigma^{-1}\upsilon^2$ and $\mu_- = \hat{\beta}_{\mathrm{OLS}} + \tau\sigma^{-1}\upsilon^2$. The weight is

$$w = \frac{\Phi\left(\frac{-\mu_-}{\upsilon}\right)/\mathrm{N}(0 \mid \mu_-, \upsilon^2)}{\Phi\left(\frac{-\mu_-}{\upsilon}\right)/\mathrm{N}(0 \mid \mu_-, \upsilon^2) \;+\; \Phi\left(\frac{\mu_+}{\upsilon}\right)/\mathrm{N}(0 \mid \mu_+, \upsilon^2)},$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. The posterior mean can be expressed as a function of the least-squares estimate, $E(\beta \mid y, \sigma^2, \tau) = \hat{\beta}_{\mathrm{OLS}} + \tau\sigma^{-1}\upsilon^2\{w - (1-w)\}$, denoted from now on as $\hat{\beta}_{\mathrm{B}}$, the Bayes estimator under squared-error loss. Because $-1 \leqslant w - (1-w) \leqslant 1$, the effect of the prior on the posterior mean relative to the least-squares estimate is bounded, $|\hat{\beta}_{\mathrm{B}} - \hat{\beta}_{\mathrm{OLS}}| \leqslant \tau\sigma(x^{\mathrm{T}}x)^{-1}$, and the bound is controlled by the amount of penalization. Pericchi & Smith (1992) provide related results for the nonregression, normal-mean setting.

Five univariate posteriors are displayed in Fig. 1(a), corresponding to values of the penalty parameter in half-unit increments from $\tau = 0{\cdot}5$ to $\tau = 2{\cdot}5$ when the observed data are such that $\hat{\beta}_{\mathrm{OLS}} = 1{\cdot}96$. Here σ^2 is assumed to be one and the predictor variables are such that $x^{\mathrm{T}}x = 1$. The posteriors corresponding to $\tau = 2$ and $\tau = 2{\cdot}5$ are penalized enough that their modes occur at zero. The case $\tau = 2$ provides an interesting comparison between the posterior mean and mode. Enough penalization is included in this case so that $\hat{\beta}_{\mathrm{L}} = 0$, however there is appreciable posterior mass away from zero, e.g. $\mathrm{pr}(\beta > 1 \mid y, \sigma^2, \tau) = 0{\cdot}258$, due to the asymmetric nature of the posterior. The posterior mean is $\hat{\beta}_{\mathrm{B}} = 0{\cdot}617$. While any single-point summary of a skewed distribution can be misleading, the posterior mean in this case captures the nontrivial posterior mass to the positive side of zero, an important feature of the distribution masked by the posterior mode.

As described in § 2·2, predictive inference for a new observation $\tilde{y}$ at a single point $\tilde{x}$ is done via the posterior predictive distribution, which for the univariate case is

$$p(\tilde{y} \mid y, \sigma^2, \tau) = w\left\{\frac{\Phi\left(\frac{-\tilde{\mu}_-}{\tilde{\upsilon}}\right)}{\Phi\left(\frac{-\mu_-}{\upsilon}\right)}\mathrm{N}(\tilde{y} \mid \tilde{x}\mu_-, \tilde{\sigma}^2)\right\} + (1-w)\left\{\frac{\Phi\left(\frac{\tilde{\mu}_+}{\tilde{\upsilon}}\right)}{\Phi\left(\frac{\mu_+}{\upsilon}\right)}\mathrm{N}(\tilde{y} \mid \tilde{x}\mu_+, \tilde{\sigma}^2)\right\}, \tag{5}$$

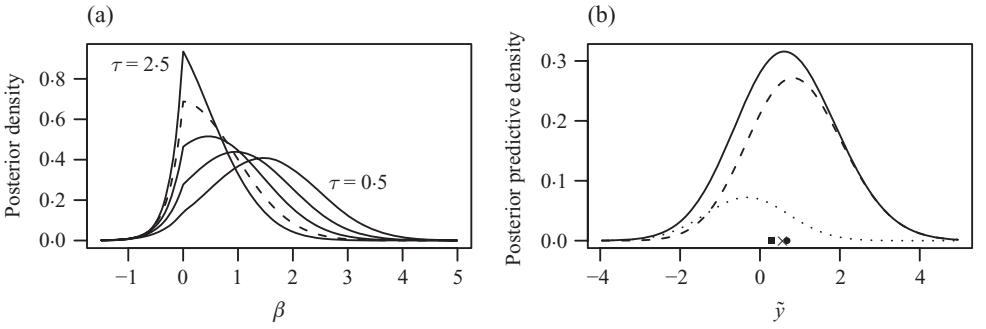

Fig. 1. (a) Univariate posterior densities for five values of τ are shown; the dashed line is $\tau = 2$. (b) The solid line depicts a posterior predictive distribution; the dashed and dotted lines are the weighted density functions in (5). The circle and $\times$ are the posterior predictive mean and mode, respectively. The square is the standard lasso prediction.

where $\tilde{\upsilon}^2 = \upsilon^2/\{1 + \tilde{x}^2/(x^{\mathrm{T}}x)\}$, $\tilde{\sigma}^2 = \sigma^2\{1 + \tilde{x}^2/(x^{\mathrm{T}}x)\}$ and

$$\tilde{\mu}_+ = \left(\frac{\tilde{x}\tilde{y} + x^{\mathrm{T}}y - \sigma\tau}{\tilde{x}^2 + x^{\mathrm{T}}x}\right), \quad \tilde{\mu}_- = \left(\frac{\tilde{x}\tilde{y} + x^{\mathrm{T}}y + \sigma\tau}{\tilde{x}^2 + x^{\mathrm{T}}x}\right).$$

The quantities $\tilde{\mu}_+$ and $\tilde{\mu}_-$ are functions of $\tilde{y}$, and the quantities in curly braces in (5) are properly normalized density functions. The mean of the posterior predictive distribution is $\tilde{x}\hat{\beta}_{\mathrm{B}}$, providing a simple method for prediction under squared-error loss.

For example, consider a dataset where $\hat{\beta}_{\mathrm{OLS}} = 1{\cdot}3$, $x^{\mathrm{T}}x = 1$, we know $\sigma^2 = 1$, we set $\tau = 1$ and we wish to predict at the single point $\tilde{x} = 1$. A plot of (5) in this setting is depicted in Fig. 1(b). The posterior predictive distribution is less skewed than the posterior distribution of β, but the skewness of both distributions requires careful attention when implementing prediction. The posterior mean of the regression coefficient is $\hat{\beta}_{\mathrm{B}} = 0{\cdot}6788$ and so the prediction under squared-error loss is $\tilde{x}\hat{\beta}_{\mathrm{B}} = 0{\cdot}6788$. Numerical optimization methods are required if predictions are to be made based on the posterior predictive mode, which is approximately 0·6070. In contrast, the lasso estimate is $\hat{\beta}_{\mathrm{L}} = 0{\cdot}3$. A prediction based on $\tilde{x}\hat{\beta}_{\mathrm{L}}$ is much smaller in this example than a prediction based on the posterior predictive mode. A similar phenomenon occurs if estimation and prediction are performed under absolute-error loss. The median of the posterior predictive distribution is approximately 0·6563, whereas the posterior median of β is 0·6025.

3. Posterior inference via Gibbs sampling

3·1. *The standard Gibbs sampler*

The most straightforward approach to Gibbs sampling for Bayesian lasso regression is to update each parameter one at a time, conditionally on all other parameters. Prior (3) is assumed unless otherwise noted. The full conditional distributions, with details given below, are

$$p(\beta_j \mid \beta_{-j}, \sigma^2, \tau, y) = \phi_j \mathrm{N}^+\left(\beta_j \mid \mu_{j\cdot}^+, \omega_{jj}^{-1}\right) + (1 - \phi_j)\mathrm{N}^-\left(\beta_j \mid \mu_{j\cdot}^-, \omega_{jj}^{-1}\right), \tag{6}$$

$$p(\sigma^2 \mid \beta, \tau, y) \propto (\sigma^2)^{-(a^*+1)} \exp(-b^*/\sigma^2 - \tau\|\beta\|_1/\sigma), \tag{7}$$

$$p(\tau \mid \beta, \sigma^2, y) = \mathrm{Ga}(\tau \mid p + r, \sigma^{-1}\|\beta\|_1 + s). \tag{8}$$

In (7), $a^* = (n+p)/2 + a$ and $b^* = (y - X\beta)^{\mathrm{T}}(y - X\beta)/2 + b$. The parameters in (6) are

$$\mu_{j\cdot}^{+} = \hat{\beta}_{\mathrm{OLS},j} + \left\{ \sum_{i \neq j} (\hat{\beta}_{\mathrm{OLS},i} - \beta_i)(\omega_{ij}/\omega_{jj}) \right\} + (-\tau\sigma^{-1}\omega_{jj}^{-1}), \tag{9}$$

where ω_{ij} is the ijth element of $\Omega = \Sigma^{-1}$; the expression for $\mu_{j\cdot}^{-}$ is similar, with $-\tau\sigma^{-1}\omega_{jj}^{-1}$ replaced with $\tau\sigma^{-1}\omega_{jj}^{-1}$. The weights are

$$\phi_j = \left\{ \frac{\Phi(\mu_{j\cdot}^{+}\sqrt{\omega_{jj}})}{\mathrm{N}(0 \mid \mu_{j\cdot}^{+}, \omega_{jj}^{-1})} \right\} \Big/ \left\{ \frac{\Phi(\mu_{j\cdot}^{+}\sqrt{\omega_{jj}})}{\mathrm{N}(0 \mid \mu_{j\cdot}^{+}, \omega_{jj}^{-1})} + \frac{\Phi(-\mu_{j\cdot}^{-}\sqrt{\omega_{jj}})}{\mathrm{N}(0 \mid \mu_{j\cdot}^{-}, \omega_{jj}^{-1})} \right\}.$$

The update for σ^2 requires special attention. A Metropolis–Hastings step could be used, but this would require tuning a proposal distribution. A description of a simple and efficient rejection sampling method for obtaining exact samples from (7) is provided in the Appendix.

While implementation of the standard Gibbs sampler is straightforward, in some cases, especially when predictor variables are highly correlated, autocorrelation in the chain can be high. The usual solution of block-updating is not feasible because $p(\beta \mid \sigma^2, \tau, y)$ is difficult to sample when p is even moderately large. The potential problem of slow convergence is addressed in § 3·2, where a new Gibbs sampler is proposed that uses a reparameterization of the model to reduce autocorrelation. Even in the case of high autocorrelation, accurate estimates of the posterior mean of the regression coefficients can often be obtained under the standard Gibbs sampler via Rao–Blackwellization; at each iteration, the conditional expectation of β_j is simply

$$E(\beta_j \mid \beta_{-j}, \sigma^2, \tau, y) = \phi_j \left\{ \mu_{j\cdot}^{+} + \frac{\mathrm{N}(0 \mid \mu_{j\cdot}^{+}, \omega_{jj}^{-1})}{\Phi(\mu_{j\cdot}^{+}\sqrt{\omega_{jj}})} \right\} + (1 - \phi_j) \left\{ \mu_{j\cdot}^{-} - \frac{\mathrm{N}(0 \mid \mu_{j\cdot}^{-}, \omega_{jj}^{-1})}{\Phi(-\mu_{j\cdot}^{-}\sqrt{\omega_{jj}})} \right\}.$$

If the classic lasso prior (2) is used in place of (3), then (7) and (8) are replaced by

$$p(\sigma^2 \mid \beta, \tau, y) = \mathrm{IG}(\sigma^2 \mid a + n/2\ ,\ b^*),$$
$$p(\tau \mid \beta, \sigma^2, y) = \mathrm{Ga}\,(\tau \mid p + r\ ,\ \|\beta\|_1 + s)\,,$$

where IG denotes the inverse gamma distribution. The only modification required for sampling the β_j is to replace the final quantity in (9) with $(-\tau\omega_{jj}^{-1})$ for $\mu_{j\cdot}^{+}$ and $(\tau\omega_{jj}^{-1})$ for $\mu_{j\cdot}^{-}$.

3·2. *The orthogonalized Gibbs sampler*

A Gibbs sampler that is less sensitive to collinearity in the design matrix can be constructed as follows. Begin by orthogonally diagonalizing $\Sigma = \sigma^2(X^{\mathrm{T}}X)^{-1}$ according to $\sigma^2 H^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}H = \sigma^2\Lambda = \sigma^2\mathrm{diag}(\lambda_j)$, where H is a $p \times p$ matrix such that $H^{\mathrm{T}}H = HH^{\mathrm{T}} = I_p$. Transforming the regression coefficients as $\eta = H^{\mathrm{T}}\beta$,

$$p(\eta \mid y, \sigma^2, \tau) \propto \sum_{z \in \mathcal{Z}} \frac{\mathrm{N}(\eta \mid H^{\mathrm{T}}\mu_z, \sigma^2\Lambda)}{\mathrm{N}(0 \mid \mu_z, \Sigma)} 1(H\eta \in \mathcal{O}_z).$$

Each element of the sum contains a term $\mathrm{N}(\eta \mid H^{\mathrm{T}}\mu_z, \sigma^2\Lambda)1(H\eta \in \mathcal{O}_z)$, a normal distribution with diagonal covariance matrix and support restricted by linear constraints.

While the multivariate distribution of η is difficult to sample directly, the full conditionals are piecewise normal on the intervals $h_{0,j}^* < \cdots < h_{p+1,j}^*$:

$$p(\eta_j \mid \eta_{-j}, \sigma^2, \tau, y) = \sum_{l=0}^{p} \rho_{lj} \{ C_{lj}^{-1} \mathrm{N}(\eta_j \mid \xi_{lj}, \sigma^2\lambda_j)\ 1(h_{l,j}^* \leqslant \eta_j < h_{l+1,j}^*) \}, \tag{10}$$

where $h^*_{0,j} = -\infty$ and $h^*_{p+1,j} = \infty$. The quantity in curly braces is the properly normalized density function for a doubly truncated normal distribution, C_{lj} is part of the norming constant and the ρ_{lj} are weights that sum to one. Details of the parameters in (10) and strategies for sampling are given in the Appendix. After each pass through η in the Gibbs sampler, the original parameters are recovered as $\beta = H\eta$.

3·3. *Comparing samplers*

Each of the three Gibbs samplers for this model has advantages and disadvantages. The data-augmentation approach is simple to implement and performs a block update for the regression coefficients β, but it involves a vector of latent variables which may result in slower mixing of the chain. The standard Gibbs sampler described above is also simple to implement, but its mixing properties are particularly sensitive to correlation among the predictor variables. The orthogonalized sampler is designed to reduce autocorrelation in such cases, but this reduction comes with an increase in both algorithmic complexity and computing time for a fixed number of iterations.

The diabetes data of Efron et al. (2004) are used here to illustrate the trade-offs between the data-augmentation and orthogonalized samplers, which represent the simplest and most complex of the three sampling methods. The response data consist of $n = 442$ measurements of disease progression along with $p = 10$ predictor variables, several of which are highly correlated. Both samplers were run for 100 000 iterations and the output was used to analyze the subchains for τ and β_5, the latter of which corresponds to a predictor that is highly correlated with other predictors. The lag-one autocorrelation for the β_5 chain was 0·31 under the data-augmentation Gibbs sampler, compared to 0·08 under the orthogonalized sampler; the results for the τ chain were 0·74 and 0·13, respectively. In this sense, the orthogonalized sampler is providing less-correlated samples even though the data-augmentation sampler performs a block update of β.

A per-iteration comparison of the two sampling methods is difficult, however, because the computational cost of the orthogonalized sampler is greater than that of the data-augmentation sampler. To provide a rough comparison of computing speed, both algorithms were programmed in the same environment using as many similar components as possible. The data-augmentation sampler required 6·21 seconds for 100 000 iterations compared to 27·26 seconds for the orthogonalized sampler, an increase of a factor of approximately 4·5. The autocorrelation function plots for the τ subchain indicate that the data-augmentation sampler required a lag of approximately 15 to 20 iterations to obtain samples that could be practically treated as independent, compared to only about two or three iterations for the orthogonalized sampler. If both chains were to be thinned to provide approximately independent samples, the data-augmentation sampler would need to be run for approximately seven times as many iterations as the orthogonalized sampler, resulting in a somewhat longer run time in seconds for the data-augmentation sampler.

Unfortunately, the orthogonalized sampler does not scale as well as the data-augmentation sampler as the number of predictors grows. After adding 40 additional white-noise predictors to this dataset, the orthogonalized sampler's run time was nearly 20 times longer than the data-augmentation sampler's. One strategy for large-p datasets would be to separate β into two blocks: one block corresponding to predictors largely uncorrelated with other predictors, and the other corresponding to predictors exhibiting more complicated correlation structure. The first block could then be updated one at a time and the second block updated via an appropriate orthogonalizing transformation. Such an approach would preserve the autocorrelation reduction without sacrificing too much computing time.

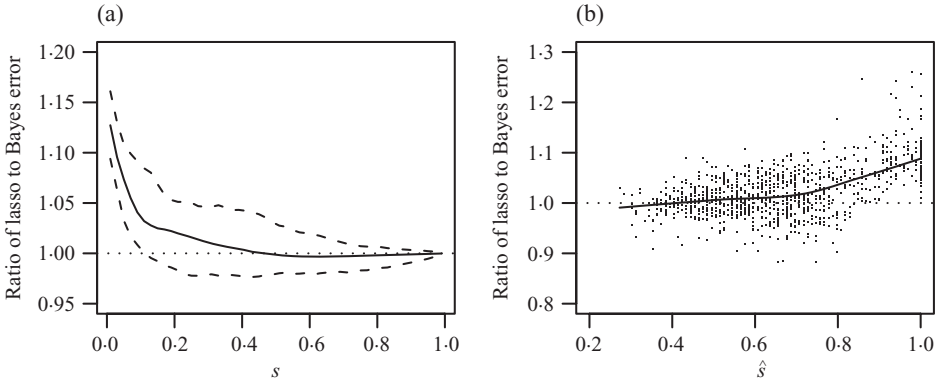

Fig. 2. Ratios of lasso prediction error to Bayes prediction error for examples 1(a) and 2(b). The dashed lines in (a) are the 2·5% and 97·5% empirical quantiles across the 1000 replications. A smoother was used to produce the solid line in (b).

4. EXAMPLES

4·1. *Example 1: Prediction along the solution path*

The diabetes data of Efron et al. (2004), described in § 3·3, are used here to evaluate the predictive performance of both the traditional lasso and the model-based Bayesian version. We randomly split the data into two pieces, generating a training set with $n = 192$ observations and a test set with $m = 250$ observations, and computed the entire lasso solution path for the training data. The solution path is indexed by $s = \|\hat{\beta}_{\mathrm{L}}(\lambda)\|_1 / \|\hat{\beta}_{\mathrm{OLS}}\|_1 \in [0, 1]$, corresponding to the fraction of the L_1 norm of the least-squares estimate represented by the lasso estimate at λ. Large values of λ represent high shrinkage and hence correspond to small values of s. Lasso coefficients were extracted along the path at $l = 1, \ldots, 50$ evenly spaced values in $[0, 1]$, and the corresponding penalty parameters λ_l were determined. The lasso estimate at each value of λ_l corresponds to the mode of the posterior distribution under prior (2) with $\tau = \lambda_l/(2\sigma^2)$, where σ^2 was fixed at an estimate based on a least-squares fit of the full model on the entire dataset. For each of the 50 values λ_l, we estimated the posterior mean based on the training data via the orthogonalized Gibbs sampler. This provided posterior means and modes, $\hat{\beta}_{\mathrm{B},l}$ and $\hat{\beta}_{\mathrm{L},l}$, at each of the 50 values λ_l. The test data were then used to measure the predictive performance at each value λ_l by computing the average test errors $B_l = m^{-1} \sum_{i=1}^{m} (y_{i,\mathrm{test}} - x_{i,\mathrm{test}}^{\mathrm{T}} \hat{\beta}_{\mathrm{B},l})^2$ and $L_l = m^{-1} \sum_{i=1}^{m} (y_{i,\mathrm{test}} - x_{i,\mathrm{test}}^{\mathrm{T}} \hat{\beta}_{\mathrm{L},l})^2$. Because the test errors depend on the random partition of the data, we repeated this procedure $T = 1000$ times. Figure 2(a) shows the average test error ratio $T^{-1} \sum_{t=1}^{\mathrm{T}} L_{lt}/B_{lt}$ at the 50 evenly spaced points in $[0, 1]$. The Bayesian predictions do better on average when a large penalty, corresponding to small s, is used; as the penalization is relaxed, the two methods perform similarly, with the lasso predictions performing marginally better on average for this dataset for large s.

4·2. *Example 2: Prediction when modelling* λ

The above comparison relied on fixing the penalty parameter over a range of possible values. Several methods for choosing λ have been proposed, including K-fold and generalized cross-validation (Tibshirani, 1996), a C_p-type selection criterion (Efron et al., 2004) and an empirical Bayes approach (Park & Casella, 2008). One advantage of the Bayesian model described in § 2 is that a prior distribution can be placed on τ and inference can be performed after integrating the penalty parameter out of the posterior distribution.

In order to compare predictions under the model using prior (3) with standard lasso predictions using 10-fold crossvalidation, we again randomly split the diabetes data into training and test sets of sizes $n = 192$ and $m = 250$, respectively, and used the Gibbs sampler to estimate the posterior mean based on the training data under a model with hyperparameters $a = b = 0$ and $r = s = 1$. We then computed the prediction error as described in example 1 using the test data. The lasso estimate $\hat{\beta}_{\mathrm{L}}$ was computed using 10-fold crossvalidation on the training data, and the prediction error was calculated using the test data. Because the prediction error for both methods relies on the particular test/training split, we repeated this procedure $T = 10\,000$ times. Comparing the test error across replications, 65·8% of the replications resulted in the lasso procedure having larger test error than the Bayes procedure. Figure 2(b) displays the ratio of test errors L_t/B_t $(t = 1, \ldots, T)$, as a function of the values $\hat{s}_t$ chosen by crossvalidation for each replication. The griddiness is due to the fact that crossvalidation was performed over a grid of values for s. The Bayes procedure is markedly better for the replications where crossvalidation based on the training data selected a large value for s: the adaptive Bayesian penalty induced by modelling both σ^2 and τ was able to induce enough shrinkage in these cases to improve predictions over the crossvalidated lasso.

5. Discussion

The model-based, Bayesian predictions for the two examples presented in § 4 performed, on average, as well as or better than the standard lasso predictions. A similar result was found when simulation examples 1, 2 and 4 from Tibshirani (1996) were performed: the Bayesian approach yielded reductions in average prediction error of 16%, 36% and 19%, respectively, when compared to the standard lasso procedure. One important aspect not addressed in this paper is variable selection. If desired, uncertainty about regression model specification under the double-exponential prior distribution can be incorporated into the analysis by assigning a prior distribution to the model space. The representation of the posterior distribution introduced in § 2·1 can be used to facilitate computation in such a setting. A study of this approach is the subject of current work.

Software written in C++ with an R interface implementing the Markov chain Monte Carlo methods described in § 3 is available at http://www.stat.osu.edu/∼hans/software.html.

Acknowledgement

This research was supported by a grant from the U.S. National Science Foundation. The author would like to thank Steven MacEachern for his helpful comments, as well as two reviewers and the editor for insights that helped improve the clarity of the manuscript.

Appendix 1

Updating σ^2 via rejection sampling

The full conditional (7) for σ^2 is not of a standard form and so requires special attention. Under the transformation $v = \sigma^{-1}$, the full conditional is $p(v \mid \beta, \tau, y) \propto v^{2a^*-1} \exp(-b^* v^2 - \tau v \|\beta\|_1)$. A sufficient condition for this density to be log-concave is $a^* = (n + p)/2 + a > 0{\cdot}5$, which will be met for any dataset if $a \geqslant 0$. A piecewise exponential hull can be created for use as a proposal distribution for rejection sampling for v, which can then be transformed to yield an exact sample from $p(\sigma^2 \mid \beta, \tau, y)$. This approach is similar to that of adaptive rejection sampling (Gilks & Wild, 1992), however because the approximation is usually very good, the adaptive component is not needed. The piecewise exponential hull is created by

finding knot points by deriving a second-order approximation to $\log p(v \mid \beta, \tau, y)$. The mode is

$$\hat{v} = \frac{-\tau\|\beta\|_1 + \left\{\tau^2\|\beta\|_1^2 + 8b^*(2a^* - 1)\right\}^{1/2}}{4b^*},$$

and the curvature at the mode is $|(1 - 2a^*)/\hat{v}^2 - 2b^*|$. Letting $s_{\hat{v}} = |(1 - 2a^*)/\hat{v}^2 - 2b^*|^{-1/2}$, place knots at $\hat{v} + s_{\hat{v}}/2$ and $\hat{v} + ks_{\hat{v}}$ $(k = 1, \ldots K)$. For $v < \hat{v}$, if $\hat{v} - s_{\hat{v}}/2 \leqslant 0$, place a single knot at $\hat{v}/2$. Otherwise, place knots at $\hat{v} - s_{\hat{v}}/2$ and $\hat{v} - ks_{\hat{v}}$ $(k = 1, \ldots, K)$, excluding negative knots. Samples are obtained by taking a proposal from the implied piecewise exponential distribution and then accepting or rejecting with the appropriate probability. In practice, $K = 2$ or 3 tends to work very well; if many rejections are observed, K can be increased adaptively to allow for a better approximation.

APPENDIX 2

Details for orthogonalized Gibbs sampling

The full conditional distribution for η_j is

$$p(\eta_j \mid \eta_{-j}, \sigma^2, \tau) \propto \sum_{z \in \mathcal{Z}} \frac{\mathrm{N}(\eta_{-j} \mid (H^{\mathrm{T}}\mu_z)_{-j}, \sigma^2\Lambda_{-j})}{\mathrm{N}(0 \mid \mu_z, \Sigma)} \mathrm{N}(\eta_j \mid (H^{\mathrm{T}}\mu_z)_j, \sigma^2\lambda_j) 1(H\eta \in \mathcal{O}_z),$$

where Λ_{-j} is the matrix obtained by removing row and column j from Λ. Focussing first on the constraints $H\eta \in \mathcal{O}_z$, assume that $H_{ij} \neq 0$ for all i, j. For a given orthant z, the orthant restriction yields linear constraints $\eta_j >< -H_{kj}^{-1} \sum_{i \neq j} H_{ki}\eta_i \equiv H_{kj}^*$ $(k = 1, \ldots, p)$, where $><$ corresponds to $>$ if $\mathrm{sign}(H_{kj}) = z_k$ and $<$ if $\mathrm{sign}(H_{kj}) = -z_k$. If $H_{kj} = 0$ for $k \neq j$, there will be fewer than p linear constraints on η_j and only a few bookkeeping changes need to be made to what is described here. When conditioning on η_{-j}, only $p + 1$ of the 2^p possible values of z will allow the p constraints to be satisfied, and so the sum over $z \in \mathcal{Z}$ can be reduced to a sum over only those $p + 1$ values of z. Refer to these values as $z^{lj} = (z_1^{lj}, \ldots, z_p^{lj})$ $(l = 0, \ldots, p)$. Order the H_{kj}^* as $h_{1,j}^*, \ldots, h_{p,j}^*$, so that $h_{1,j}^* < \cdots < h_{p,j}^*$, with inverse ordering function $d(k)$ such that $h_{k,j}^* = H_{d(k),j}^*$. The boundaries are $h_{0,j}^* = -\infty$ and $h_{p+1,j}^* = \infty$. For $l \in \{0, \ldots, p\}$ and $k = 1, \ldots, p$,

$$z_{d(k)}^{lj} = \begin{cases} -\mathrm{sign}(H_{d(k),j}), & k > l, \\ \mathrm{sign}(H_{d(k),j}), & k \leqslant l \end{cases}$$

$$(H^{\mathrm{T}}\mu_{z^{lj}})_j = \sum_{i=1}^{p} H_{ij}\hat{\beta}_{\mathrm{OLS},i} - \tau\sigma\lambda_j \sum_{i=1}^{p} H_{ij} z_i^{lj} \equiv \xi_{lj}.$$

The normalizing constant C_{lj} in expression (10) is

$$C_{lj} = \Phi\left(\frac{h_{l+1,j}^* - \xi_{l,j}}{\sigma\lambda_j^{1/2}}\right) - \Phi\left(\frac{h_{l,j}^* - \xi_{lj}}{\sigma\lambda_j^{1/2}}\right).$$

The orthant-specific weights in expression (10) are

$$\rho_{lj} \propto C_{lj} \frac{\mathrm{N}(\eta_{-j} \mid (H^{\mathrm{T}}\mu_{z^{lj}})_{-j}, \sigma^2\Lambda_{-j})}{\mathrm{N}(0 \mid \mu_{z^{lj}}, \Sigma)} \propto C_{lj} \exp\left(\frac{\xi_{lj}^2}{2\sigma^2\lambda_j} - \frac{\tau}{\sigma} \sum_{k=1}^{p} z_k^{lj} \sum_{i \neq j} \eta_i H_{ki}\right).$$

When prior (2) is used in place of (3), the expression for ξ_{lj} changes to $\sum_{i=1}^{p} H_{ij}\hat{\beta}_{\mathrm{OLS},i} - \tau\sigma^2\lambda_j \sum_{i=1}^{p} H_{ij} z_i^{lj}$, and the $\tau\sigma^{-1}$ term in the expression for ρ_{lj} is replaced with τ.

Samples are obtained by sampling a component with probability proportional to ρ_{lj} and then sampling from the corresponding truncated normal distribution via the method of Geweke (1991). When all the normal distributions are such that the truncation points are far out in their tails, computation of the ρ_{lj} is numerically unstable. In these cases, a simple rejection sampler, similar to the one described above, is

used. The tails of the normal distributions will be well approximated by exponential distributions, and a piecewise exponential distribution with knot points at the truncation locations serves as a good proposal distribution.

References

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

Bernardo, J. & Smith, A. (2000). *Bayesian Theory*. Chichester: Wiley.

Carlin, B. & Polson, N. (1991). Inference for nonconjugate Bayesian models using the Gibbs sampler. *Can. J. Statist.* **19**, 399–405.

Carlin, B., Polson, N. & Stoffer, D. (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *J. Am. Statist. Assoc.* **87**, 493–500.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407–99.

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

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. In *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pp. 571–8. Alexandria, VA: American Statistical Association.

Gilks, W. & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41**, 337–48.

Mitchell, A. (1994). A note on posterior moments for a normal mean with double-exponential prior. *J. R. Statist. Soc.* B **56**, 605–10.

Park, T. & Casella, G. (2008). The Bayesian lasso. *J. Am. Statist. Assoc.* **103**, 681–6.

Pericchi, L. & Smith, A. (1992). Exact and approximate posterior moments for a normal location parameter. *J. R. Statist. Soc.* B **54**, 793–804.

Pericchi, L. & Walley, P. (1991). Robust Bayesian credible intervals and prior ignorance. *Int. Statist. Rev.* **59**, 1–23.

Spiegelhalter, D. (1977). A test for normality against symmetric alternatives. *Biometrika* **64**, 415–8.

Tanner, M. & Wong, W. (1987). The calculation of posterior densities by data augmentation (with discussion). *J. Am. Statist. Assoc.* **82**, 528–50.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74**, 646–8.

[*Received July* 2008. *Revised March* 2009]