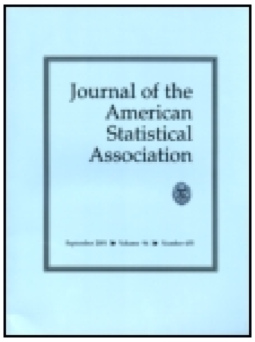

Journal of the
American
Statistical
Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Elastic Net Regression Modeling With the Orthant Normal Prior

Chris Hans

Elastic Net Regression Modeling With the Orthant Normal Prior

Chris HANS

The elastic net procedure is a form of regularized optimization for linear regression that provides a bridge between ridge regression and the lasso. The estimate that it produces can be viewed as a Bayesian posterior mode under a prior distribution implied by the form of the elastic net penalty. This article broadens the scope of the Bayesian connection by providing a complete characterization of a class of prior distributions that generate the elastic net estimate as the posterior mode. The resulting model-based framework allows for methodology that moves beyond exclusive use of the posterior mode by considering inference based on the full posterior distribution. Two characterizations of the class of prior distributions are introduced: a properly normalized, direct characterization, which is shown to be conjugate for linear regression models, and an alternate representation as a scale mixture of normal distributions. Prior distributions are proposed for the regularization parameters, resulting in an infinite mixture of elastic net regression models that allows for adaptive, data-based shrinkage of the regression coefficients. Posterior inference is easily achieved using Markov chain Monte Carlo (MCMC) methods. Uncertainty about model specification is addressed from a Bayesian perspective by assigning prior probabilities to all possible models. Corresponding computational approaches are described. Software for implementing the MCMC methods described in this article, written in `C++` with an `R` package interface, is available at *http://www.stat.osu.edu/~hans/software/*.

KEY WORDS: Bayesian regression; Lasso; Model uncertainty; Regularization; Scale mixtures; Shrinkage; Variable selection.

1. INTRODUCTION

The elastic net (Zou and Hastie 2005) is a regularization procedure for linear regression that also performs variable selection. Regularized optimization as a form of estimation has attracted broad interest; the goal of such procedures is generally to improve on predictions based on ordinary least squares by shrinking parameter estimates toward zero. Certain regularization methods allow some parameters to be set equal to 0, providing a method for identifying important variables. Parameter estimates for linear regression in this framework are typically

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda J(\boldsymbol{\beta}) \qquad (1)$$

for some nonnegative penalty function J and regularization parameter $\lambda \geq 0$.

The elastic net is defined in two stages. Letting $|\boldsymbol{\beta}|^2 = \sum_{j=1}^{p} \beta_j^2$ and $|\boldsymbol{\beta}|_1 = \sum_{j=1}^{p} |\beta_j|$, a "naïve" estimate $\hat{\boldsymbol{\beta}}_{\mathrm{N}}$ is first found via (1) with $J(\boldsymbol{\beta}) = \alpha|\boldsymbol{\beta}|^2 + (1-\alpha)|\boldsymbol{\beta}|_1$. The regularization parameter λ is the sum of two nonnegative penalties, $\lambda = \lambda_1 + \lambda_2$, so that $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$. The final elastic net estimate is taken to be a rescaled version of the naïve estimate, $\hat{\boldsymbol{\beta}}_{\mathrm{E}} = (1+\lambda_2)\hat{\boldsymbol{\beta}}_{\mathrm{N}}$. The scaling was introduced by Zou and Hastie (2005) to reduce perceived overshrinkage in the naïve estimate. By using a combination of L_1- and L_2-norm penalization, the elastic net is designed to provide shrinkage similar to that of ridge regression while at the same time providing lasso-like variable selection.

Regression parameter estimates based on regularization procedures can often be interpreted as the mode of a Bayesian posterior distribution when the likelihood component is the normal linear regression model $p(\mathbf{y}|\boldsymbol{\beta}) = \mathrm{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. Tibshirani (1996) showed that the lasso estimate corresponds to a posterior mode when the prior for $\boldsymbol{\beta}$ is a product of independent double-exponential distributions. The ridge regression estimate can be viewed as a posterior mode when a product of independent normal priors is used for $\boldsymbol{\beta}$ (Jeffreys 1961). Fu (1998) described the corresponding prior distribution that results in the bridge regression estimate. In the case of the elastic net, Zou and Hastie (2005) noted that the elastic net penalty can be expressed as the unnormalized prior density function

$$p(\boldsymbol{\beta}|\lambda, \alpha) \propto \exp\bigl[-\lambda\{\alpha|\boldsymbol{\beta}|^2 + (1-\alpha)|\boldsymbol{\beta}|_1\}\bigr]. \qquad (2)$$

In this article we broaden the Bayesian connection to the elastic net procedure by providing a properly normalized, complete characterization of prior (2). This places the elastic net in the context of a model-based framework where point estimation, prediction, and model uncertainty can be addressed from a Bayesian perspective. The posterior mode does not play a central role in the Bayesian paradigm; rather, inference about $\boldsymbol{\beta}$ is based on the entire posterior distribution $p(\boldsymbol{\beta}|\mathbf{y})$, the prediction of future observations $\tilde{\mathbf{y}}$ is based on the posterior predictive distribution $p(\tilde{\mathbf{y}}|\mathbf{y})$, and uncertainty about the specification of the regression model is addressed via the posterior distribution over the model space. A key advantage of casting elastic net regression in a Bayesian framework is that uncertainty about the regularization parameters α and λ can be incorporated into the model through a prior distribution. Integrating over this uncertainty essentially creates an infinite mixture of elastic net regression models, allowing for adaptive, data-based shrinkage of the regression coefficients. This approach results in more realistic statements of estimation and prediction uncertainty compared with approaches that condition on particular values of α and λ. The normalizing constant for prior (2) is needed to implement this approach, and this is introduced in Section 2.

A key property of the elastic net regularization procedure (1) is the fact that it is possible for some elements of $\boldsymbol{\beta}$ to be set to 0, providing a form of variable selection. Uncertainty about

Chris Hans is Assistant Professor, Department of Statistics, The Ohio State University, Columbus, OH 43210-1247 (E-mail: *hans@stat.osu.edu*). The author thanks the Editor, Associate Editor, and two reviewers for helpful comments. This work was partially supported by National Science Foundation grant DMS-0706948.

© 2011 American Statistical Association
Journal of the American Statistical Association
December 2011, Vol. 106, No. 496, Theory and Methods
DOI: 10.1198/jasa.2011.tm09241

model specification is addressed in the Bayesian framework by explicitly specifying a prior distribution over the possible models and then computing posterior probabilities for each model. Inference can then be made conditionally on a chosen model or by averaging over the posterior distribution of the model space. We introduce the modeling and computational details for addressing Bayesian elastic net model uncertainty in Section 5.

Bayesian shrinkage priors for normal mean estimation have been studied extensively, and shrinkage priors for general regression problems have recently been of interest, due in part to the advent of the regularization procedures discussed earlier. Fernández and Steel (2000) considered a large class of scale mixture of normal distributions for use in regression. Park and Casella (2008) focused on one special case, the double-exponential prior distribution, and made connections to the lasso procedure. Yi and Xu (2008) provided an application to genetics data. Hans (2009) also considered the double-exponential prior, but framed the model outside of the the scale-mixture setting. Griffin and Brown (2007, 2010) generalized the scale-mixture representation of the double-exponential distribution by placing a more flexible mixing distribution on the scale parameter. Polson and Scott (2011a, 2011b) recently developed a very rich class of shrinkage priors. The approach we introduce in this article has connections to each of these settings but arises from a unique prior distribution that results in a posterior with different properties than those considered in the related work. We also address regression model uncertainty and consider model-averaged predictions, which were not addressed in detail in the aforementioned related work.

The purpose of this article is not to simply propose another regularization/shrinkage estimator for regression; rather, more generally, it is to make explicit the Bayesian connection to the elastic net procedure and to develop the tools required for inference in this setting. The core elements of Bayesian elastic net regression—the prior and posterior distributions—are introduced in Section 2, where a properly normalized, direct characterization of (2) is provided. The prior is shown to be conjugate, resulting in a direct characterization of the posterior distribution of $\boldsymbol{\beta}$. The prior is also shown to be representable as a scale mixture of normal distributions. This second representation allows for connections to be made to other Bayesian regression shrinkage priors. A natural set of prior distributions for σ^2, α, and λ is proposed in Section 2, and Bayesian estimation and prediction are discussed. Section 3 describes Markov chain Monte Carlo (MCMC) approaches to computation for this model and shows how both representations of the prior can be used to construct simple Gibbs samplers. A Bayesian treatment of uncertainty in model specification for elastic net regression is described in Section 5, and related MCMC methods are introduced. Section 6 contains several simulation examples that demonstrate the predictive effectiveness of the Bayesian elastic net model, and Section 7 provides a high-dimensional example that underscores the usefulness of model-averaged prediction.

2. ELASTIC NET PRIOR AND POSTERIOR DISTRIBUTIONS

2.1 Direct Representation

The likelihood considered throughout this article is $p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = \mathrm{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, where $\mathbf{X}$ is an $n \times p$ matrix of predictor variables and $\boldsymbol{\beta}$ is a p-vector of regression coefficients. The columns of $\mathbf{X}$ and the vector $\mathbf{y}$ are assumed to be demeaned, and so the model does not include an intercept, although one could be easily accommodated. Consider a modified version of prior (2),

$$p(\boldsymbol{\beta}|\alpha, \lambda, \sigma^2) \propto \exp\left[-\frac{\lambda}{2\sigma^2}\{\alpha|\boldsymbol{\beta}|^2 + (1-\alpha)|\boldsymbol{\beta}|_1\}\right], \tag{3}$$

where the penalty λ is now scaled by $2\sigma^2$. Under this representation, for fixed values of σ^2 and α, the posterior mode under prior (3) will be the naïve elastic net estimate with overall penalty λ. This prior behaves like a normal distribution when $\alpha \approx 1$ and like a double-exponential distribution when $\alpha \approx 0$. Integration of (3) shows that the normalizing constant is available in closed form up to the evaluation of the univariate standard normal cdf. The properly normalized prior density function is

$$p(\boldsymbol{\beta}|\lambda, \alpha, \sigma^2) = \prod_{j=1}^{p}\left\{(0.5)\cdot \mathrm{N}^{-}\left(\beta_j \middle| \frac{1-\alpha}{2\alpha}, \frac{\sigma^2}{\lambda\alpha}\right)\right.$$
$$\left. + (0.5)\cdot \mathrm{N}^{+}\left(\beta_j \middle| -\frac{1-\alpha}{2\alpha}, \frac{\sigma^2}{\lambda\alpha}\right)\right\}, \tag{4}$$

where N^{-} and N^{+} are properly normalized density functions for truncated normal distributions,

$$\mathrm{N}^{+}(t|m, s^2) \equiv \frac{\mathrm{N}(t|m, s^2)}{\Phi(m/s)}\mathbf{1}(t \geq 0)$$

and

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

and Φ is the univariate standard normal cdf. The location parameter for the positive (negative) component in (4) will always be negative (positive), and so the prior says that β_j will always be "in the tails" of a normal distribution. [See also Pericchi and Smith (1992), who blended two halves of two normal distributions together to obtain a double-exponential prior.]

The prior also can be written suggestively as follows. Letting $\mathcal{Z} = \{-1, 1\}^p$ be the collection of all possible p-vectors with elements ± 1, for any vector $\mathbf{z} \in \mathcal{Z}$ let $\mathcal{O}_\mathrm{z} \subset \mathbb{R}^p$ be the corresponding orthant. If $\boldsymbol{\beta} \in \mathcal{O}_\mathrm{z}$, then $\beta_j \geq 0$ for $z_j = 1$ and $\beta_j < 0$ for $z_j = -1$. The prior (4) then can be rewritten as

$$p(\boldsymbol{\beta}|\lambda, \alpha, \sigma^2) = 2^{-p}\Phi\left(\frac{\alpha - 1}{2\sigma\sqrt{\alpha/\lambda}}\right)^{-p}$$
$$\times \sum_{\mathrm{z}\in\mathcal{Z}} \mathrm{N}\left(\boldsymbol{\beta}\middle|\frac{\alpha-1}{2\alpha}\mathbf{z}, \frac{\sigma^2}{\lambda\alpha}\mathbf{I}_p\right)\mathbf{1}(\boldsymbol{\beta} \in \mathcal{O}_\mathrm{z}).$$

From this, it is clear that the prior is piecewise normal, where each piece is defined over a separate orthant, yielding an "orthant normal" prior. Writing the prior in terms of λ_1 and λ_2 gives

$$p(\boldsymbol{\beta}|\lambda_1, \lambda_2, \sigma^2) = 2^{-p}\Phi\left(\frac{-\lambda_1}{2\sigma\sqrt{\lambda_2}}\right)^{-p}$$
$$\times \sum_{\mathrm{z}\in\mathcal{Z}} \mathrm{N}\left(\boldsymbol{\beta}\middle|-\frac{\lambda_1}{2\lambda_2}\mathbf{z}, \frac{\sigma^2}{\lambda_2}I_p\right)\mathbf{1}(\boldsymbol{\beta} \in \mathcal{O}_\mathrm{z}). \tag{5}$$

We use the (λ_1, λ_2) formulation from this point on unless specified otherwise.

Combining prior (5) with the regression model likelihood via Bayes' theorem yields the posterior distribution

$$p(\boldsymbol{\beta}|\mathbf{y}, \lambda_1, \lambda_2, \sigma^2) = \sum_{\mathbf{z}\in\mathcal{Z}} \omega_{\mathbf{z}} \mathrm{N}^{[\mathbf{z}]}(\boldsymbol{\beta}|\boldsymbol{\mu}_{\mathbf{z}}, \sigma^2\mathbf{R}), \qquad (6)$$

a weighted sum of 2^p orthant-truncated normal distributions,

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

where

$$\mathrm{P}(\mathbf{z}, \mathbf{m}, \mathbf{S}) = \int_{\mathcal{O}_{\mathbf{z}}} \mathrm{N}(\mathbf{t}|\mathbf{m}, \mathbf{S})\, d\mathbf{t},$$

is a multivariate normal orthant integral. The posterior, like the prior, is multivariate piecewise normal, with each component defined on a separate orthant. The parameters of the posterior distribution are

$$\mathbf{R} = (\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{I}_p)^{-1} \qquad \text{and} \qquad \boldsymbol{\mu}_{\mathbf{z}} = \hat{\boldsymbol{\beta}}_{\mathbf{R}} - \frac{\lambda_1}{2}\mathbf{R}\mathbf{z},$$

where $\hat{\boldsymbol{\beta}}_{\mathbf{R}} = \mathbf{R}\mathbf{X}^T\mathbf{y}$ is the ridge regression estimate with penalty λ_2. The final pieces of the posterior are the weights for each orthant,

$$\omega_{\mathbf{z}} = \omega^{-1}\frac{\mathrm{P}(\mathbf{z}, \boldsymbol{\mu}_{\mathbf{z}}, \sigma^2\mathbf{R})}{\mathrm{N}(0|\boldsymbol{\mu}_{\mathbf{z}}, \sigma^2\mathbf{R})}, \quad \text{where}$$
$$\omega = \sum_{\mathbf{z}\in\mathcal{Z}} \frac{\mathrm{P}(\mathbf{z}, \boldsymbol{\mu}_{\mathbf{z}}, \sigma^2\mathbf{R})}{\mathrm{N}(0|\boldsymbol{\mu}_{\mathbf{z}}, \sigma^2\mathbf{R})}. \qquad (7)$$

2.2 Mixture Representation

Scale mixtures of normal distributions have been used extensively in Bayesian modeling (e.g., Andrews and Mallows 1974; West 1987; Carlin and Polson 1991; Fernández and Steel 2000; Liang et al. 2008; Carvalho, Polson, and Scott 2009; Polson and Scott 2011a, 2011b). In particular, for lasso regression, Figueiredo (2003) and Park and Casella (2008) used a scale-mixture of normals representation of the double-exponential distribution to facilitate posterior inference via the EM algorithm (Dempster, Laird, and Rubin 1977) and a data-augmentation Gibbs sampler (Tanner and Wong 1987), respectively.

Theorem 1. The orthant normal distribution (5) is equivalent to a product of independent scale mixtures of normal distributions,

$$p(\boldsymbol{\beta}|\sigma^2, \lambda_1, \lambda_2) = \prod_{j=1}^{p} \int_0^1 \mathrm{N}\left(\beta_j \middle| 0, \frac{\sigma^2}{\lambda_2}(1-\tau_j)\right)$$
$$\times\, \mathrm{IG}_{(0,1)}\left(\tau_j \middle| \frac{1}{2}, \frac{1}{2}\left(\frac{\lambda_1}{2\sigma\sqrt{\lambda_2}}\right)^2\right) d\tau_j,$$

where $\mathrm{IG}_{(0,1)}$ is the inverse gamma distribution restricted to the interval $(0, 1)$.

A proof is provided in the Appendix. This result was reported independently by Li and Lin (2010). By introducing latent variables $\tau_1, \ldots, \tau_p$ and using the notation $\mathbf{S}_{\boldsymbol{\tau}} = \mathrm{diag}(1 - \tau_j)$, the prior can be written hierarchically as

$$p(\boldsymbol{\beta}|\boldsymbol{\tau}, \sigma^2, \lambda_2) = \mathrm{N}\left(\boldsymbol{\beta}\middle| 0, \frac{\sigma^2}{\lambda_2}\mathbf{S}_{\boldsymbol{\tau}}\right),$$
$$p(\boldsymbol{\tau}|\sigma^2, \lambda_1, \lambda_2) = \prod_{j=1}^{p} \mathrm{IG}_{(0,1)}\left(\tau_j \middle| \frac{1}{2}, \frac{1}{2}\left(\frac{\lambda_1}{2\sigma\sqrt{\lambda_2}}\right)^2\right)$$
$$= 2^{-p}\Phi\left(\frac{-\lambda_1}{2\sigma\sqrt{\lambda_2}}\right)^{-p}\left(\frac{\lambda_1^2}{8\pi\sigma^2\lambda_2}\right)^{p/2}$$
$$\times \exp\left(-\frac{\lambda_1^2\sum_{j=1}^{p}\tau_j^{-1}}{8\sigma^2\lambda_2}\right)$$
$$\times \prod_{j=1}^{p}\tau_j^{-3/2}\mathbf{1}(0 < \tau_j < 1).$$

The hierarchical representation is useful because $p(\boldsymbol{\beta}|\boldsymbol{\tau}, \sigma^2, \lambda_2)$ is a product of independent, unrestricted normal distributions. Accordingly, the conditional posterior distribution is multivariate normal: $p(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\tau}, \sigma^2, \lambda_1, \lambda_2) = \mathrm{N}(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_{\mathbf{R}_{\boldsymbol{\tau}}}, \sigma^2\mathbf{R}_{\boldsymbol{\tau}})$, where $\mathbf{R}_{\boldsymbol{\tau}} = (\mathbf{X}^T\mathbf{X} + \lambda_2\mathbf{S}_{\boldsymbol{\tau}}^{-1})^{-1}$ and the mean vector is the ridge-like estimate $\hat{\boldsymbol{\beta}}_{\mathbf{R}_{\boldsymbol{\tau}}} = \mathbf{R}_{\boldsymbol{\tau}}\mathbf{X}^T\mathbf{y}$. This hierarchical representation is exploited by a data augmentation Gibbs sampler in Section 3.2.

2.3 Relationship to the Double-Exponential Prior

In their definition of the elastic net procedure, Zou and Hastie (2005) noted that the penalty term $J(\boldsymbol{\beta}) = \alpha|\boldsymbol{\beta}|^2 + (1-\alpha)|\boldsymbol{\beta}|_1$ provides a continuum of penalties that have ridge regression $(\alpha = 1)$ and the lasso $(\alpha = 0)$ at the two extremes. Thus the orthant normal prior indexes a continuum of priors between a normal prior $(\alpha = 1)$ and a double-exponential prior $(\alpha = 0)$. Regression modeling under the double-exponential prior and connections with the lasso have drawn recent interest (e.g., Park and Casella 2008; Hans 2009). A natural means for comparing the double-exponential prior with the elastic net (orthant normal) prior is through the scale mixture of normals representation. If we consider the scaled penalty term $-(\lambda_2/2)\beta_j^2 - (\lambda_1/2)|\beta_j|$, corresponding to prior (3) with $\sigma^2 = 1$, then when $\lambda_2 = 0$, we obtain the double-exponential prior

$$p(\beta_j|\lambda_1) = \frac{\lambda_1/2}{2}e^{-(\lambda_1/2)|\beta_j|}.$$

Using the results of Andrews and Mallows (1974), this prior can be represented as the scale mixture

$$\beta_j|\theta_j \sim \mathrm{N}(0, \theta_j^2),$$
$$p(\theta_j|\lambda_1) = \frac{\lambda_1^2}{4}\theta_j e^{-\lambda_1^2\theta_j^2/8}, \qquad \theta_j > 0, \qquad (8)$$

so that $\theta_j \sim \mathrm{Weibull}(2, \lambda_1^2/8)$ or, equivalently, $\theta_j^2 \sim \mathrm{Exp}(\lambda_1^2/8)$. Under a suitable transformation, the scale-mixture representation of the orthant normal prior can be written in the same form,

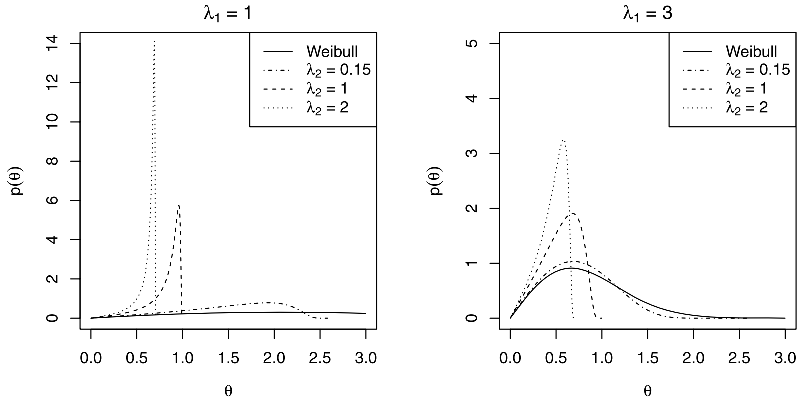

Figure 1. Mixing distributions for the scale parameter θ.

with

$$\beta_j|\theta_j \sim \mathrm{N}(0,\theta_j^2),$$

$$p(\theta_j|\lambda_1,\lambda_2) = \frac{\lambda_1\theta_j(\lambda_2^{-1}-\theta_j^2)^{-3/2}}{\lambda_2\sqrt{8\pi}\,\Phi(-\lambda_1/(2\sqrt{\lambda_2}))} \times \exp\left\{-\frac{\lambda_1^2(\lambda_2^{-1}-\theta_j^2)^{-1}}{8\lambda_2^2}\right\}, \qquad 0<\theta_j<1/\sqrt{\lambda_2}. \tag{9}$$

Figure 1 compares the Weibull mixing distribution with (9) for $\lambda_1 \in \{1,3\}$ and $\lambda_2 \in \{0.15, 1, 2\}$. Small values of θ_j correspond to large amounts of shrinkage. In both (8) and (9), increasing the value of λ_1 smoothly transitions the mixing distribution toward favoring smaller values of θ, inducing more shrinkage. Including the extra parameter λ_2 in (9) also induces more shrinkage, but the resulting shrinkage is different than what would be obtained by simply inreasing λ_1. As λ_2 increases, the distribution of θ favors increasingly smaller values of θ; however, the distribution tends to become more concentrated near the upper boundary of the support of θ at the nonzero value $\lambda_2^{-1/2}$. The dual-action shrinkage inherent in the elastic net prior provides greater flexibility than what is obtained under the double-exponential prior. The simulation results presented in Section 6 indicate that transitioning from the single-parameter double-exponential prior to the two-parameter orthant normal prior can improve predictive performance.

2.4 Bayesian Learning and Adaptive Shrinkage

The posterior distribution introduced above was defined conditionally for fixed values of λ_1, λ_2, and σ^2. One advantage of considering elastic net regression from a Bayesian perspective is that these parameters can be modeled with prior distributions, eliminating the need to pick specific values of the penalty parameters a priori. In non-Bayesian regularization settings, K-fold cross validation is a popular method for determining reasonable values of the penalty parameters. In the case of the elastic net, the fact that there are two penalty parameters requires K-fold cross-validation over a grid of values (λ_1, λ_2). Modeling these parameters directly with prior distributions creates an infinite mixture of elastic net regression models, where the amount of penalization can adapt to the specific dataset being analyzed. This approach also allows for marginal inference on the regression parameter $\boldsymbol{\beta}$. It is important to note that the normalizing constant of the conditional posterior distribution $p(\boldsymbol{\beta}|\lambda_1,\lambda_2,\sigma^2,\mathbf{y})$ contains terms involving λ_1 and λ_2 that do not appear in the traditional regularization setup (1). These terms contain information about the penalty parameters that can be leveraged by a fully Bayesian analysis (see also the discussion of Polson and Scott 2011b).

A natural set of independent prior distributions for the elastic net hyperparameters is $\lambda_1 \sim \mathrm{Gamma}(L,\nu/2)$, $\lambda_2 \sim \mathrm{Gamma}(R,\nu/2)$, $\sigma^2 \sim \mathrm{IG}(a/2, b/2)$. This formulation is equivalent to specifying independent prior distributions $\alpha \sim \mathrm{Beta}(R, L)$ and $\lambda \sim \mathrm{Gamma}(R+L,\nu/2)$. Thus, setting $R=L=1$ provides a uniform prior for α.

3. COMPUTATIONAL APPROACHES

3.1 Basic Gibbs Sampler

Posterior inference is most easily achieved via MCMC simulation methods. The most basic Gibbs sampler updates each component of $\boldsymbol{\beta}$ conditionally on all other components by generating from the full conditional distributions $p(\beta_j|\boldsymbol{\beta}_{-j},\mathbf{y},\sigma^2,\lambda_1,\lambda_2)$. Under the direct characterization of the orthant normal prior, the posterior distribution (6) yields full conditional distributions

$$p(\beta_j|\boldsymbol{\beta}_{-j},\mathbf{y},\lambda_1,\lambda_2,\sigma^2) = (1-\phi_j)\mathrm{N}^-(\beta_j|\mu_j^-,s_j^2) + \phi_j\mathrm{N}^+(\beta_j|\mu_j^+,s_j^2), \tag{10}$$

which, like the prior, are piecewise normal. The two components share common scale parameters $s_j^2 = \sigma^2/(\mathbf{x}_j^T\mathbf{x}_j + \lambda_2)$. Using the shorthand $\hat{\beta}_j$ to represent the jth element of $\hat{\boldsymbol{\beta}}_{\mathbf{R}}$, the location parameter for the positive normal distribution is

$$\mu_j^+ = \hat{\beta}_j + \left\{\sum_{i\neq j}(\hat{\beta}_i - \beta_i)\frac{\mathbf{x}_i^T\mathbf{x}_j}{\mathbf{x}_j^T\mathbf{x}_j+\lambda_2}\right\} + \frac{-\lambda_1}{2(\mathbf{x}_j^T\mathbf{x}_j+\lambda_2)}.$$

The expression for μ_j^- is similar, with $-\lambda_1$ replaced with λ_1. The weight is

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

Sampling from the full conditional distributions is straightforward: the positive or negative component of (10) is chosen by sampling a Bernoulli random variable with probability ϕ_j, followed by a draw from the appropriate truncated normal distribution (see Geweke 1991, for efficient computational methods).

3.2 Data Augmentation Gibbs Sampler

The hierarchical representation of the orthant normal prior distribution can be used to construct a data augmentation Gibbs sampler by including the latent variables $\boldsymbol{\tau}$ in the sampling scheme. Working with the augmented posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\tau} | \mathbf{y}, \sigma^2, \lambda_1, \lambda_2)$, the Gibbs sampler cycles through the full conditional distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$. Sampling in this fashion is advantageous because the full conditional distribution for $\boldsymbol{\beta}$ is $\mathrm{N}(\hat{\boldsymbol{\beta}}_{\mathbf{R}_{\boldsymbol{\tau}}}, \sigma^2 \mathbf{R}_{\boldsymbol{\tau}})$, allowing for a block update.

The τ_j are conditionally independent given $\boldsymbol{\beta}$ and $\mathbf{y}$, and the full conditional for each is

$$\begin{aligned} &p(\tau_j | \mathbf{y}, \boldsymbol{\beta}, \sigma^2, \lambda_1, \lambda_2) \\ &\quad \propto (1-\tau_j)^{-1/2} \tau_j^{-3/2} \\ &\qquad \times \exp\left\{ -\frac{\lambda_2 \beta_j^2}{2\sigma^2(1-\tau_j)} - \frac{\lambda_1^2}{8\sigma^2 \lambda_2 \tau_j} \right\} \mathbf{1}(0 < \tau_j < 1). \end{aligned}$$

Sampling directly from this distribution is tedious; however, under the change of variables $\zeta_j = \tau_j/(1-\tau_j)$, we have

$$\begin{aligned} &p(\zeta_j | \mathbf{y}, \boldsymbol{\beta}, \sigma^2, \lambda_1, \lambda_2) \\ &\quad = \sqrt{\frac{c}{2\pi\zeta^3}} \exp\left\{ -\frac{c(\zeta_j - d_j)^2}{2 d_j^2 \zeta_j} \right\}, \qquad \zeta_j > 0, \end{aligned}$$

where

$$c = \frac{\lambda_1^2}{4\lambda_2 \sigma^2} \qquad \text{and} \qquad d_j = \frac{\lambda_1}{2\lambda_2 |\beta_j|},$$

and so ζ_j has an inverse Gaussian distribution (Tweedie 1957; Folks and Chhikara 1978), which can be easily sampled. This sampler has been reported independently by Li and Lin (2010).

3.3 Updating σ^2 and the Penalty Parameters

The conditional distribution $p(\sigma^2, \lambda_1, \lambda_2 | \mathbf{y}, \boldsymbol{\beta})$ contains the unusual term $\Phi(-\lambda_1/(2\sigma\sqrt{\lambda_2}))^{-p}$, resulting in nonstandard full conditional distributions for σ^2, λ_1, and λ_2. The same is true for $p(\sigma^2, \lambda_1, \lambda_2 | \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\tau})$ when the scale-mixture representation is used. The most straightforward approach to sampling these parameters uses random-walk Metropolis–Hastings updates for $\log\sigma^2$, $\log\lambda_1$, and $\log\lambda_2$ using normal proposals centered at the current values. The standard deviations of these proposals can be easily tuned to provide reasonable acceptance rates.

4. ILLUSTRATING THE ELASTIC NET POSTERIOR

The elastic net posterior distribution is illustrated here using a dataset that has been analyzed previously under the elastic net procedure. The dataset comprises the prostate cancer data of Stamey et al. (1989), which is available in the `R` package `lasso2` (Lokhorst, Venables, and Turlach 2007). The data consist of $p = 8$ clinical variables used to predict the logarithm of a measurement of prostate-specific antigen for each patient ($n = 97$).

4.1 Contextualizing the Posterior Mode

The traditional elastic net estimate corresponds to the mode of the Bayesian conditional posterior distribution $p(\boldsymbol{\beta} | \lambda_1, \lambda_2, \sigma^2, \mathbf{y})$ for fixed values of λ_1, λ_2, and σ^2. In this section we illustrate how this traditional estimate—the posterior mode—is situated in the context of the entire posterior distribution. To make this illustration, we must select values of λ_1, λ_2, and σ^2 on which to condition. Although any values will allow us to make such an illustration, a natural choice is to choose the values $\hat{\lambda}_1$ and $\hat{\lambda}_2$ that would be chosen by someone performing a traditional elastic net analysis, which typically involves choosing these parameters via cross-validation. Choosing the parameters in this way allows us to set the classical answer in the context of the posterior distribution of which it is the mode. Examination of how posterior probability is distributed about the mode will provide an immediate visualization of posterior uncertainty, whereas juxtaposition of the posterior mode with the posterior mean will allow comparison of the traditional estimate with a more typical Bayesian posterior summary. This approach—estimating the parameters via cross-validation and then computing the corresponding conditional posterior distribution—is not suggested as a general method of data analysis, but rather is used here as a device for understanding how the posterior mode relates to the entire posterior distribution.

To choose the values of λ_1 and λ_2 to use in the illustration, we applied the elastic net procedure (1) to the data, and used 10-fold cross-validation to select values $\hat{\lambda}_1 = 1.695$ and $\hat{\lambda}_2 = 0.14$. The corresponding naïve ($\hat{\boldsymbol{\beta}}_{\mathrm{N}}$) and rescaled ($\hat{\boldsymbol{\beta}}_{\mathrm{E}}$) estimates of the regression coefficients are provided in Table 1. The naïve elastic net estimate $\hat{\boldsymbol{\beta}}_{\mathrm{N}}$—which is the same regardless of the value of σ^2 on which we condition—is the mode of the posterior distribution $p(\boldsymbol{\beta} | \hat{\lambda}_1, \hat{\lambda}_2, \sigma^2, \mathbf{y})$. For the illustration, we condition on $\hat{\sigma}^2 = 0.496$, the maximum likelihood estimate of σ^2. Table 1 also provides $\hat{\boldsymbol{\beta}}_{\mathrm{B}}$, the Bayesian estimator of $\boldsymbol{\beta}$ under squared-error loss, which is the mean of the posterior distribution given in (6).

Figure 2 provides a visualization of $\hat{\boldsymbol{\beta}}_{\mathrm{N}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ in the context of the marginal posterior distributions $p(\beta_j | \hat{\lambda}_1, \hat{\lambda}_2, \hat{\sigma}^2, \mathbf{y})$. The marginal posterior density functions were estimated on a fine grid of points via Rao–Blackwellization using the output from the data augmentation Gibbs sampler. At least two interesting features of the posterior distribution are readily apparent. First, the posterior distribution is highly asymmetric, which could not be known by examining only the posterior mode. Regression coefficients that are clearly "signal" (e.g., `lcavol`, `svi`, and `lweight`) have marginal distributions that are roughly sym-

Table 1. Posterior summaries for the prostate cancer data in Section 4. $\hat{\boldsymbol{\beta}}_{\mathrm{B}}$ refers to the posterior mean, $\hat{\boldsymbol{\beta}}_{\mathrm{N}}$ refers to the posterior mode (naïve elastic net estimate), and $\hat{\boldsymbol{\beta}}_{\mathrm{E}} = (1+\lambda_2)\hat{\boldsymbol{\beta}}_{\mathrm{N}}$ is the (rescaled) elastic net estimate

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45
$\hat{\boldsymbol{\beta}}_{\mathrm{B}}$	4.74	1.51	−0.15	0.66	1.95	0.49	0.37	0.54
$\hat{\boldsymbol{\beta}}_{\mathrm{N}}$	5.04	1.50	0	0.47	2.16	0	0	0.56
$\hat{\boldsymbol{\beta}}_{\mathrm{E}}$	5.75	1.71	0	0.54	2.46	0	0	0.64

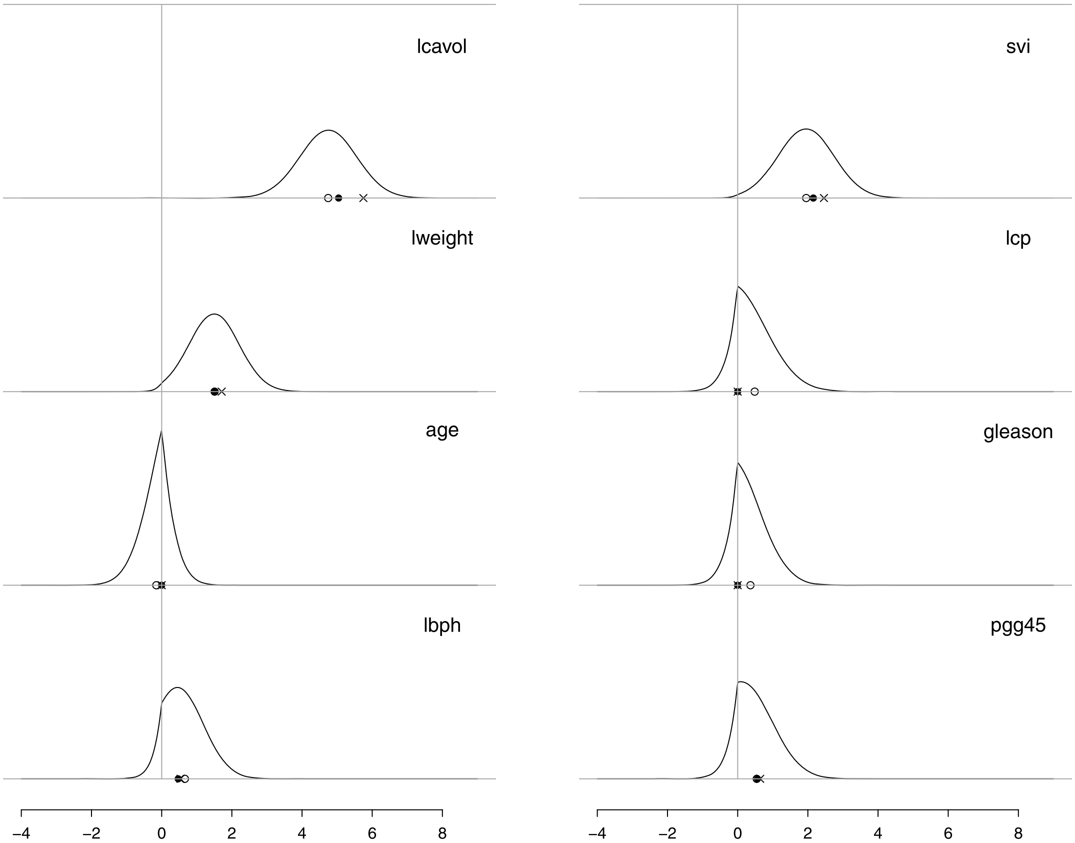

Figure 2. Density estimates of marginal posterior distributions for the prostate data as described in Section 4.1. The solid circles denote the posterior mode $\hat{\boldsymbol{\beta}}_{\mathrm{N}}$, and the open circles denote the posterior mean $\hat{\boldsymbol{\beta}}_{\mathrm{B}}$. The × symbols are the (rescaled) elastic net estimate $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$.

metric about their modes. Coefficients that are less clearly signal have density functions that are noticeably nondifferentiable at 0 and are skewed in the direction of the least squares estimate. A consequence of this skewness is that our choice of estimator will clearly have an impact on inference. For example, the posterior mode of the coefficient for `lcp` is at 0; however, the skewness of the distribution results in a posterior mean of 0.49. The simulations in Section 6 indicate that incorporating the skewness into the predictive procedure by basing prediction on the posterior mean can provide improvements over the usual elastic net procedure.

The second interesting feature is that there is substantial uncertainty about some of the coefficients that have posterior modes at 0. For example, the posterior mode of the coefficients for `gleason` and `lcp` are at 0; however, there is substantial posterior probability assigned to values greater than 1 for both coefficients. This highlights the difficulty of quantifying model uncertainty when only the posterior mode is computed. Bayesian methods for addressing model uncertainty under the orthant normal prior are introduced in Section 5.

4.2 Incorporating Regularization Parameter Uncertainty

An important benefit to adopting a Bayesian approach to elastic net regression is that the penalty parameters can be modeled as described in Section 2.4. Figure 3 displays marginal posterior distributions under the model where the prior on α was taken to be uniform, the prior for σ^2 was specified with $a = 5$ and $b = 1$, and the prior for λ was specified with $\nu = 4/3$. Compared with Section 4.1, where the regularization parameters were fixed, the posterior distribution of the regression coefficients is now much less shrunken toward and spiked near 0.

4.3 Comparison With the Bayesian Lasso Posterior

To assess the impact of adding the L_2-norm term in prior (2), we also fit the Bayesian lasso model of Park and Casella (2008) to the data (dashed lines in Figure 3). The priors for the Bayesian lasso model were chosen so that the prior on σ^2 was the same under both models and the marginal prior variances of the β_j are approximately the same under both models. The variables with large coefficients—`lcavol`, `lweight`, and `svi`—have similar marginal posteriors under the two models. Differences can be seen in the marginal posteriors for variables with small coefficients. For these parameters, the posterior is smoother near 0 under the orthant normal prior than under the double-exponential prior because of inclusion of the λ_2 term. In addition, moderate coefficients are shrunken somewhat less under the orthant normal prior. For example, the distributions of the coefficients for `age` and `pgg45` have more mass away from 0 than they do under the double-exponential prior. This suggests an advantage in estimating small but nonzero coefficients, which will we investigated in a simulation reported in Section 6.

5. MODEL UNCERTAINTY

A key property of the elastic net procedure based on (1) is that it is possible for components of $\hat{\boldsymbol{\beta}}_{\mathrm{N}}$ to equal 0, providing a

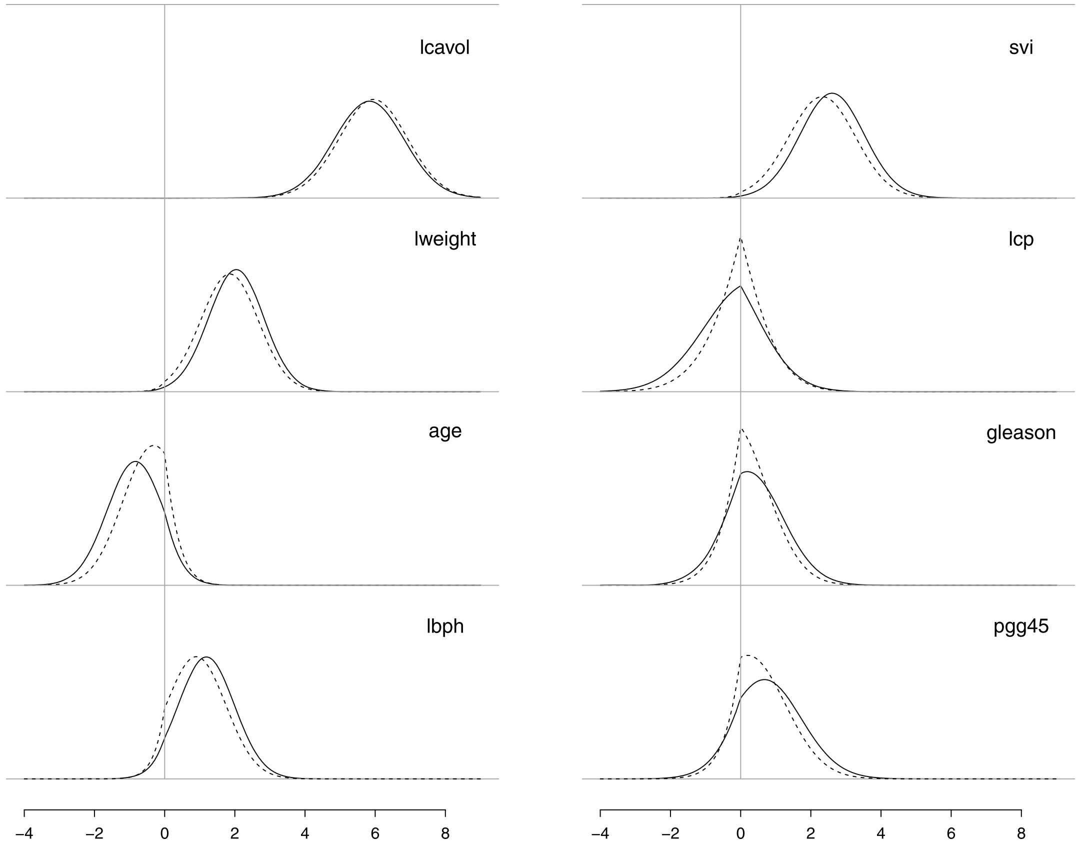

Figure 3. Density estimates of the marginal posterior distributions for the prostate data as described in Section 4.2. The solid lines represent the elastic net posteriors; the dashed lines, the lasso posteriors.

method for variable selection. It is convenient to index models by $\boldsymbol{\gamma}$, an indicator vector of length p such that $\gamma_j = 1$ means that the jth predictor variable is included in the model and $\gamma_j = 0$ means that it is not. From a Bayesian perspective, after a prior probability $p(\boldsymbol{\gamma})$ is assigned to each model, uncertainty about model composition is typically described by the posterior distribution over the model space,

$$p(\boldsymbol{\gamma}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}'\in\Gamma} p(\mathbf{y}|\boldsymbol{\gamma}')p(\boldsymbol{\gamma}')},$$

where Γ is the collection of all 2^p possible regression models. It is convenient to take $p(\boldsymbol{\gamma}|w) = \prod_{j=1}^{p} w^{\gamma_j}(1-w)^{1-\gamma_j}$, $0 < w < 1$, which treats the γ_j exchangeably. Variations on this prior distribution have been considered by George and McCulloch (1993, 1997), George and Foster (2000), Chipman, George, and McCulloch (2001), and Kohn, Smith, and Chan (2001), among others.

For fixed values of σ^2, λ_1, and λ_2, the marginal likelihood for the Bayesian elastic net regression model under the orthant normal prior (5) is

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\gamma}, \sigma^2, \lambda_1, \lambda_2) &= 2^{-k}(2\pi\sigma^2)^{-(n+k)/2}\lambda_2^{k/2}\omega_{\boldsymbol{\gamma}}\,\Phi\left(\frac{-\lambda_1}{2\sigma\sqrt{\lambda_2}}\right)^{-k} \\ &\quad\times \exp\left\{-\frac{\mathbf{y}^T\mathbf{y} + k\lambda_1^2/(4\lambda_2)}{2\sigma^2}\right\}, \end{aligned} \tag{11}$$

where $k = \sum_{j=1}^{p}\gamma_j$ and $\omega_{\boldsymbol{\gamma}}$ is the same as ω defined in Section 2.1, but here using only the predictor variables indicated by $\boldsymbol{\gamma}$. If p is small, then the marginal likelihood can be computed numerically for all 2^p models; if p is even moderately large, then computation of $\omega_{\boldsymbol{\gamma}}$ becomes difficult. [See Hans (2010) for an investigation of these limitations in the related Bayesian lasso regression setting.]

5.1 Model Uncertainty When p Is Large

When p is large, or when one wishes to integrate over the parameters σ^2, λ_1, and λ_2, a computationally feasible method for addressing model uncertainty is to construct a variable-selection Gibbs sampler in which the parameter $\boldsymbol{\gamma}$ is included in the MCMC algorithm (see, e.g., George and McCulloch 1993, 1997; Geweke 1996; Smith and Kohn 1996). The direct specification of the orthant normal prior can be adapted to incorporate model uncertainty for Bayesian elastic net regression by specifying

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2, \lambda_1, \lambda_2) = \prod_{j=1}^{p}(1-\gamma_j)\delta_0(\beta_j) + \gamma_j\,\mathrm{ON}(\beta_j|\lambda_1, \lambda_2, \sigma^2),$$

where δ_0 is a point mass at zero and “ON” is the density function for the univariate orthant normal prior (4). For $j = 1, \ldots, p$, the Gibbs sampler updates the vector (β_j, γ_j) by first sampling

$$\gamma_j|\mathbf{y}, \boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}_{-j}, \sigma^2, \lambda_1, \lambda_2 \sim \mathrm{Bernoulli}(\phi_{j+}),$$

$$\phi_{j+} = \left[1 + \frac{1-w}{w} \right.$$
$$\times 2\Phi\left(\frac{-\lambda_1}{2\sigma\sqrt{\lambda_2}}\right)(2\pi\sigma^2/\lambda_2)^{1/2}\exp\left(\frac{\lambda_1^2}{8\sigma^2\lambda_2}\right)$$
$$\left. \Big/ \left\{\frac{\Phi(-\mu^-_{\boldsymbol{\gamma}_{-j}}/s_j)}{\mathrm{N}(0|\mu^-_{\boldsymbol{\gamma}_{-j}}, s_j^2)} + \frac{\Phi(\mu^+_{\boldsymbol{\gamma}_{-j}}/s_j)}{\mathrm{N}(0|\mu^+_{\boldsymbol{\gamma}_{-j}}, s_j^2)}\right\}\right]^{-1},$$

where $\mu^+_{\boldsymbol{\gamma}_{-j}} = (\mathbf{x}_j^T\mathbf{x}_j + \lambda_2)^{-1}(\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}_{-j}\boldsymbol{\beta}_{-j}) - \lambda_1/2)$. The expression for $\mu^-_{\boldsymbol{\gamma}_{-j}}$ is similar, with $+\lambda_1/2$ replacing $-\lambda_1/2$. If the sampled value of γ_j is 0, then β_j is set to 0; otherwise, β_j is updated via

$$p(\beta_j|\mathbf{y}, \boldsymbol{\gamma}, \boldsymbol{\beta}_{-j}, \sigma^2, \lambda_1, \lambda_2)$$
$$= (1 - \phi_{\boldsymbol{\gamma}_{-j}})\mathrm{N}^-(\beta_j|\mu^-_{\boldsymbol{\gamma}_{-j}}, s_j^2) + \phi_{\boldsymbol{\gamma}_{-j}}\mathrm{N}^+(\beta_j|\mu^+_{\boldsymbol{\gamma}_{-j}}, s_j^2),$$

where $\phi_{\boldsymbol{\gamma}_{-j}}$ is the same as ϕ_j in Section 3.1, but here is computed using $\mu^+_{\boldsymbol{\gamma}_{-j}}$ and $\mu^-_{\boldsymbol{\gamma}_{-j}}$. The remaining parameters, σ^2, λ_1, and λ_2, are updated as described in Section 3.3, now conditioning on the model indicated by $\boldsymbol{\gamma}$ rather than on the full model.

Although there is no one Bayesian "answer" to the variable selection question, most approaches require either analytical access to or the ability to sample from the model space posterior distribution. The marginal likelihood (11) coupled with the Gibbs sampling tools described earlier provide such access for the Bayesian elastic net regression model. Although we do not consider the question of variable selection directly in this article, we do note that posterior model and variable inclusion probabilities can be computed using the methodology introduced in this section. In Sections 6 and 7 below, we take advantage of the model space posterior distribution to compute model-averaged predictions.

6. SIMULATIONS

Here we use the four simulation settings described by Zou and Hastie (2005) to assess the predictive performance of the Bayesian elastic net model. In each setting, the response data are simulated according to $\mathbf{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, and the predictor variables are simulated according to $\mathbf{X} \sim \mathrm{N}(\mathbf{0}, \mathbf{V})$, where the diagonal elements of $\mathbf{V}$ are 1 and the remaining elements are specified later. We assess the predictive performance via the prediction mean squared error (MSE), $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbf{V}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. In each example, 1000 datasets are simulated.

Example 1. $n = 20$, $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, $\sigma = 3$, and $V_{ij} = 0.5^{|i-j|}$.

Example 2. Same as Example 1, except $\beta_j = 0.85$ for all j.

Example 3. $n = 100$, $\sigma = 15$, $\boldsymbol{\beta} = (0, \ldots, 0, 2, \ldots, 2, 0, \ldots, 0, 2, \ldots, 2)^T$ with 10 repeats in each block, and $V_{ij} = 0.5$ for all $i \neq j$.

Example 4. $n = 100$, $\sigma = 15$, and $\boldsymbol{\beta} = (3, \ldots, 3, 0, \ldots, 0)^T$, where 3 is repeated 15 times and 0 is repeated 25 times. The covariance matrix $\mathbf{V}$ has diagonal elements $V_{ii} = 1.01$ for $i \leq 15$ and $V_{ii} = 1$ for $i > 15$. The off-diagonal ($i \neq j$) elements are $V_{ij} = 1$ if $1 \leq i, j \leq 5$, $6 \leq i, j \leq 10$, $11 \leq i, j \leq 15$, and are 0 otherwise. This creates four blocks. The first three blocks have within-block correlations of approximately 0.99 and between-block correlations of 0, and the fourth block generates independent predictors.

Table 2. Average prediction MSE for the simulation study. The estimators are described in Section 6. The standard errors are based on 500 bootstrap resamplings of the MSE

	Example 1	Example 2	Example 3	Example 4
$\hat{\boldsymbol{\beta}}_{\mathrm{B}}$	3.54 (0.07)	2.90 (0.06)	60.14 (0.51)	60.83 (0.59)
$\hat{\boldsymbol{\beta}}_{\mathrm{BMA}}$	3.34 (0.07)	3.35 (0.06)	51.51 (0.49)	11.56 (0.29)
$\hat{\boldsymbol{\beta}}_{\mathrm{E}}$	3.86 (0.09)	4.04 (0.10)	37.75 (0.38)	16.07 (0.39)
$\hat{\boldsymbol{\beta}}_{\mathrm{BL}}$	3.72 (0.07)	3.52 (0.07)	71.74 (0.64)	81.75 (0.77)
$\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$	6.40 (0.07)	7.20 (0.06)	129.08 (0.53)	104.25 (0.65)

We compared five different estimators $\hat{\boldsymbol{\beta}}$ in each example; the corresponding prediction mean-squared errors are given in Table 2. The first row of the table corresponds to $\hat{\boldsymbol{\beta}}_{\mathrm{B}} \equiv \mathrm{E}[\boldsymbol{\beta}|\mathbf{y}]$, the posterior mean of $\boldsymbol{\beta}$ under the Bayesian elastic net model described in Section 2.4, where the parameters σ^2, λ_1, and λ_2 are no longer fixed. The hyperparameters for this model are taken to be $a = b = L = R = \nu = 1$. The second row of the table corresponds to $\hat{\boldsymbol{\beta}}_{\mathrm{BMA}}$, the model-averaged posterior mean of $\boldsymbol{\beta}$ under the prior described in Section 5 that incorporates uncertainty about model specification in the prior distribution. The prior over the model space is specified with $w \sim \mathrm{Unif}(0, 1)$. The third row of the table corresponds to $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$, the usual rescaled elastic net estimate of Zou and Hastie (2005). For each simulated dataset, the regularization parameters for the elastic net were chosen based on 10-fold cross-validation. The fourth row of the table corresponds to $\hat{\boldsymbol{\beta}}_{\mathrm{BL}}$, the posterior median of $\boldsymbol{\beta}$ under the Bayesian lasso model of Park and Casella (2008). The Bayesian lasso models were fit under the noninformative prior $p(\sigma^2) \propto \sigma^2$ and a Gamma(1, 1) prior on the square of the lasso penalty parameter. Finally, the fifth row of the table corresponds to $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, the least squares estimate of $\boldsymbol{\beta}$ based on a subset of the predictors that was chosen for each simulated dataset via stepwise model selection. The function `step` in `R` was used to select the models based on the Akaike information criterion. Whereas 1000 datasets were simulated for the first four estimators, 5000 datasets were simulated for $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ to obtain comparable standard errors.

In general, the two Bayesian elastic net estimators $\hat{\boldsymbol{\beta}}_{\mathrm{B}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{BMA}}$ perform well, providing prediction MSEs that are comparable to—and in most cases better than—the traditional elastic net estimate $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$. In Example 1, a low-dimensional example where some coefficients are truly 0, the model-averaged Bayesian elastic net estimator $\hat{\boldsymbol{\beta}}_{\mathrm{BMA}}$ performs best. In Example 2, where all coefficients are small but nonzero, the Bayesian elastic net estimators again perform the best, with the edge going to the Bayesian estimator $\hat{\boldsymbol{\beta}}_{\mathrm{B}}$, which does not average over different models. The traditional elastic net estimator $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ performs best in Example 3, a higher-dimensional example where all predictors are correlated with one another and half of the true coefficients are 0. As expected, $\hat{\boldsymbol{\beta}}_{\mathrm{BMA}}$ outperforms $\hat{\boldsymbol{\beta}}_{\mathrm{B}}$ in this example because some of the true coefficients are 0; however, the Bayesian model averaging appears to overshrink the nonzero coefficients, and as a result $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ has the best performance.

Example 4 provides another setting where the predictor variables are correlated; however, in this example the (strong) correlation occurs in block patterns. This is a setting in which we would expect the traditional elastic net estimator $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ to perform well, because it was specifically designed to handle groups of highly correlated predictors. In this example, the Bayesian model-averaged elastic net estimator $\hat{\boldsymbol{\beta}}_{\mathrm{BMA}}$ outperforms all other estimators, providing a 28% reduction in prediction MSE over $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$.

Finally, recall that the Bayesian lasso model is obtained as a limiting case of the Bayesian elastic net model as $\lambda_2 \to 0$. The Bayesian lasso prediction errors are dominated by the Bayesian elastic net prediction errors in this simulation study, providing evidence that orthant normal prior allows for more flexible parameter estimation (and thus prediction) compared with its limiting case, the double-exponential prior.

In general, we expect the Bayesian elastic net estimator $\hat{\boldsymbol{\beta}}_{\mathrm{B}}$ to perform particularly well in situations where the true coefficients are small but nonzero. This was suggested in Sections 2.3 and 4.3, where the orthant normal prior was seen to be able to induce shrinkage that is concentrated on small but nonzero values, and this behavior was confirmed in simulation Example 2. We expect the model-averaged estimator $\hat{\boldsymbol{\beta}}_{\mathrm{BMA}}$ to perform well in situations where there is considerable uncertainty about the true model (Example 3 notwithstanding). In the next section we investigate the model-averaged estimator further in a highly correlated, high-dimensional example.

7. HIGH–DIMENSIONAL EXAMPLE

Here we compare the model-averaged Bayesian elastic net estimator $\hat{\boldsymbol{\beta}}_{\mathrm{BMA}}$ with the traditional elastic net estimator $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$ in a high-dimensional example, where the $p = 300$ predictor variables are highly correlated. The example is the cookie dough dataset of Osborne et al. (1984), which also has been analyzed by Brown, Fearn, and Vannucci (2001) and Griffin and Brown (2007), among others. Following Griffin and Brown (2007), we take the dependent variable to be the flour content of each piece of cookie dough; the predictor variables are chosen to be 300 measurements of the near-infrared spectroscopy reflectance spectrum measured at evenly spaced wavelengths between 1202 and 2398 nm. We split the data into a training set containing $n = 39$ samples and a hold-out set with $n = 31$ samples to test the predictive performance of $\hat{\boldsymbol{\beta}}_{\mathrm{BMA}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$. The data in both the training and test sets were mean-centered and standardized to have unit variance. This is a challenging example because the predictor variables are highly correlated and $p > n$. The smallest correlation in the training data is 0.75 and approximately 75% of the pairs of predictor variables have correlation > 0.95. Thus, the setup is similar to simulation Example 3 in Section 6, although here we do not have information on the true amount of sparsity.

We obtained 300,000 samples from the posterior distribution of the Bayesian elastic net model described in Section 5 that accounts for model uncertainty and calculated the model-averaged posterior mean $\hat{\boldsymbol{\beta}}_{\mathrm{BMA}}$. We set the hyperparameters a, b, L, R, and ν to 1, as done in the simulation study. Three different prior distributions for w were considered: (a) $w \sim \mathrm{Beta}(1, 1)$, the uniform distribution; (b) $w \sim \mathrm{Beta}(1, 29)$, corresponding to a prior expected model size of 10 variables, and (c) $w \sim \mathrm{Beta}(1, 99)$, corresponding to a prior expected model size of 3 variables. The penalty parameters for the non-Bayesian elastic net were computed using 5-fold cross-validation on the training data over a grid of values for λ_1 and λ_2.

Estimates of the regression coefficients for the Bayesian model under the three different priors and the estimates under the traditional elastic net are displayed in Figure 4. The traditional elastic net procedure sets most of the coefficients to 0, with spikes at 50 wavelengths that are identified as important. Using these estimated coefficients, the prediction MSE on the test data is 2.466. The Bayesian estimates $\hat{\boldsymbol{\beta}}_{\mathrm{BMA}}$ are averaged over the possible models and so no coefficient is estimated to be exactly equal to 0; however, under the model with the $\mathrm{Beta}(1, 99)$ prior distribution, which penalizes large models heavily a priori, many of the coefficients are estimated to be extremely close to 0. In this case, the coefficient estimates are similar to $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$: mostly 0 with a few spikes in similar regions. The prediction MSE under this Bayesian model is 2.484, essentially identical to the traditional elastic net. Relaxing the prior on the model space slightly to $\mathrm{Beta}(1, 29)$ allows many of the coefficients that were nearly 0 to have slightly larger coefficients (in absolute value); however, the larger coefficients are shrunk slightly to compensate. The prediction MSE is reduced slightly to 2.247.

When the prior over the model space is relaxed to Beta(1, 1)—so that marginally, each predictor has probability 0.5 of being in the model—the model-averaged coefficient estimates look quite different than $\hat{\boldsymbol{\beta}}_{\mathrm{E}}$; the coefficients are allowed to wiggle more around 0, and the contribution to prediction is smoothed out over longer regions of wavelengths. The prediction MSE is reduced to 1.874. This suggests that when there is considerable uncertainty about model specification (e.g., when the predictor variables are all very highly correlated) prediction under the elastic net framework can be improved by averaging over the model space posterior distribution rather than focusing on a single point estimate corresponding to a posterior mode.

8. DISCUSSION

In this article we introduced a complete characterization of a new class of prior distributions—the orthant normal distribution—and showed that it gave rise to the elastic net estimate as the posterior mode. The article provides the tools required for fully Bayesian inference in the elastic net setting. Parameter estimation and prediction based on (possibly model-averaged) posterior means was shown to outperform the traditional elastic net in several simulation and real data examples.

One perceived drawback to Bayesian inference based on the entire posterior distribution is that sampling from the posterior is typically much more computationally intensive compared with finding the posterior mode through numerical optimization. Whereas the posterior mode is available in seconds, it usually takes minutes to obtain sufficient samples from the posterior to achieve accurate Monte Carlo estimates. Sampling times for the Bayesian elastic net models described in this article are not unreasonably long, however. For example, for the high-dimensional dataset described in Section 7 with $p = 300$, obtaining 300,000 samples from the posterior distribution required only slightly more than 15 minutes on a desktop Mac Pro computer.

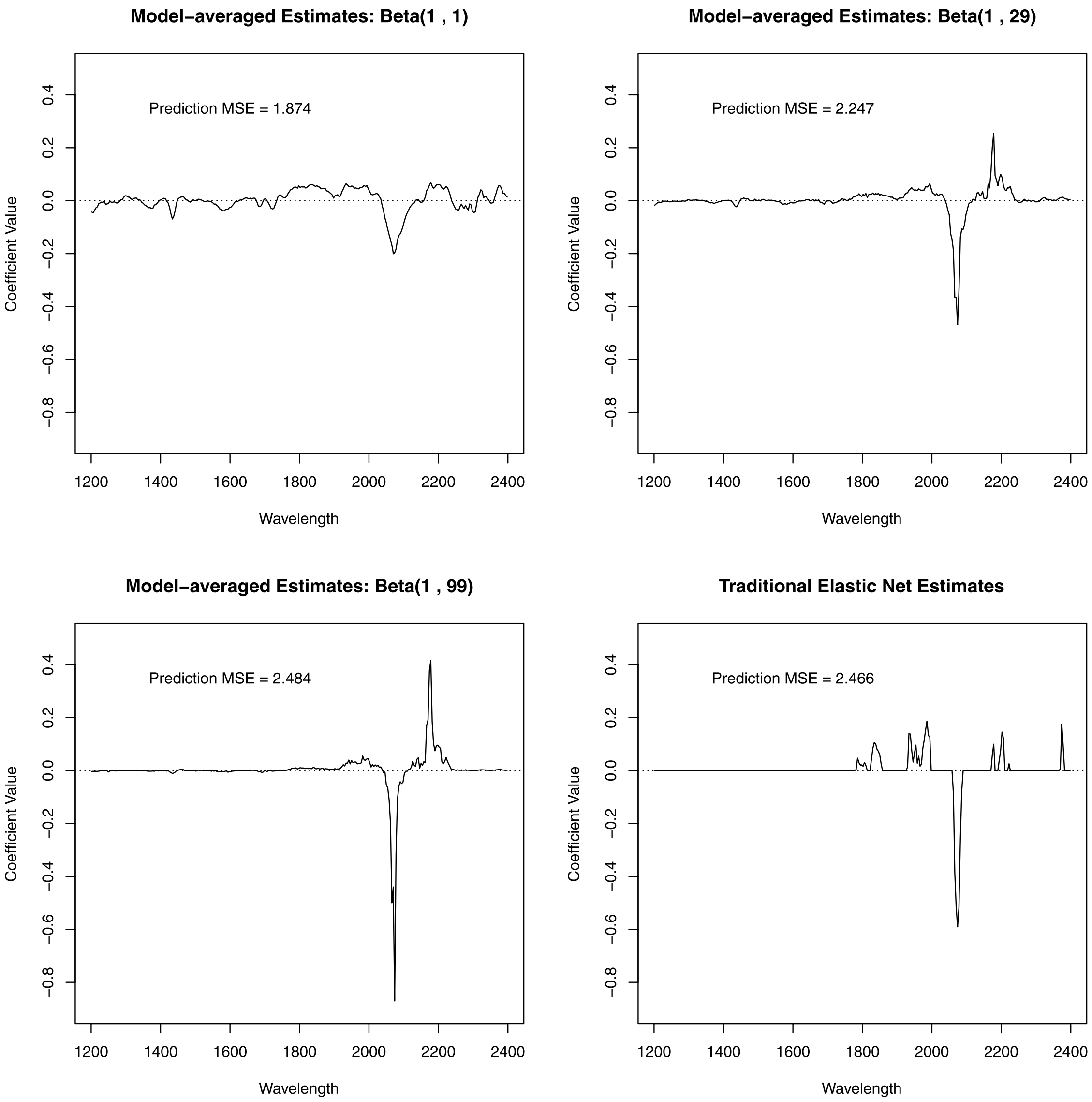

Figure 4. Coefficient estimates for the 300 predictor variables in the cookie dough dataset in Section 7. The panels labeled "Model-averaged Estimates" are $\hat{\boldsymbol{\beta}}_{\text{BMA}}$ under the three difference priors for w indicated in the panel titles. The "Traditional Elastic Net Estimates" are $\hat{\boldsymbol{\beta}}_{\text{E}}$. The prediction MSEs quoted in the plots are for the 31 hold-out test samples.

In any event, the disparity in computational effort must be balanced against the benefits of having access to the entire posterior, including access to the posterior distribution over the space of all possible regression models that is provided in Section 5. Being able to explicitly account for model specification uncertainty—through model averaging or the calculation of posterior model or variable inclusion probabilities—is important in high-dimensional problems as well as in problems with highly correlated predictor variables.

Software for implementing the MCMC methods described in this article, written in `C++` with an `R` package interface, is available at *http://www.stat.osu.edu/~hans/software/*.

APPENDIX: MIXTURE REPRESENTATION

Andrews and Mallows (1974) provided the identity

$$\frac{a}{2}e^{-a|\beta|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}}e^{-\beta^2/(2s)}\,\frac{a^2}{2}e^{-a^2 s/2}\,ds, \qquad a > 0.$$

Letting $a = \lambda_1/(2\sigma^2)$, the orthant normal prior is

$$e^{-1/(2\sigma^2)(\lambda_1|\beta|+\lambda_2\beta^2)} \propto \int_0^\infty s^{-1/2}e^{-1/2(s^{-1}+\lambda_2/\sigma^2)\beta^2}e^{-s\lambda_1^2/(8\sigma^4)}\,ds.$$

After the change of variables $\tau = \sigma^2/(\lambda_2 s + \sigma^2)$, we have

$$\begin{aligned} &e^{-1/(2\sigma^2)(\lambda_1|\beta|+\lambda_2\beta^2)} \\ &\propto \int_0^1 (1-\tau)^{-1/2}e^{-\lambda_2\beta^2/(\sigma^2(1-\tau))}\tau^{-3/2}e^{-\lambda_1^2/(8\sigma^2\lambda_2\tau)}\,d\tau \\ &\propto \int_0^1 \mathrm{N}\left(\beta\,\middle|\,0, \frac{\sigma^2(1-\tau)}{\lambda_2}\right)\tau^{-3/2}e^{-\lambda_1^2/(8\sigma^2\lambda_2\tau)}\,d\tau. \end{aligned}$$

This is an infinite mixture of mean-0 normal distributions, the variances of which range between 0 and σ^2/λ_2. The mixing parameter τ follows the inverse-Gamma distribution restricted to the interval $(0, 1)$, which has density function

$$\mathrm{IG}_{(0,1)}\left(\tau \Big| \frac{1}{2}, \frac{1}{2}\left(\frac{\lambda_1}{2\sigma\sqrt{\lambda_2}}\right)^2\right)$$

$$= \frac{\lambda_1(8\sigma^2\lambda_2\pi)^{-1/2}\tau^{-3/2}\exp\{-\lambda_1^2/(8\sigma^2\lambda_2\tau)\}}{2\Phi(-\lambda_1/(2\sigma\sqrt{\lambda_2}))}\mathbf{1}(0 < \tau < 1).$$

[Received April 2009. Revised March 2011.]

REFERENCES

Andrews, D., and Mallows, C. (1974), "Scale Mixtures of Normal Distributions," *Journal of the Royal Statistical Society, Ser. B*, 36, 99–102. [1385,1392]

Brown, P. J., Fearn, T., and Vannucci, M. (2001), "Bayesian Wavelet Regression on Curves With Application to a Spectroscopic Calibration Problem," *Journal of the American Statistical Association*, 96, 398–408. [1391]

Carlin, B. P., and Polson, N. G. (1991), "Inference for Nonconjugate Bayesian Models Using the Gibbs Sampler," *Canadian Journal of Statistics*, 19, 399–405. [1385]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [1385]

Chipman, H. A., George, E. I., and McCulloch, R. E. (2001), "The Practical Implementation of Bayesian Model Selection" (with discussion), in *Model Selection*, ed. P. Lahiri, Beachwood, OH: IMS, pp. 65–134. [1389]

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 39, 1–38. [1385]

Fernández, C., and Steel, M. (2000), "Bayesian Regression Analysis With Scale Mixtures of Normals," *Econometric Theory*, 16, 80–101. [1384,1385]

Figueiredo, M. (2003), "Adaptive Sparseness for Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 1150–1159. [1385]

Folks, J., and Chhikara, R. (1978), "The Inverse Gaussian Distribution and Its Statistical Application—A Review," *Journal of the Royal Statistical Society, Ser. B*, 40, 263–289. [1387]

Fu, W. (1998), "Penalized Regressions: The Bridge versus the Lasso," *Journal of Computational and Graphical Statistics*, 7, 397–416. [1383]

George, E. I., and Foster, D. P. (2000), "Calibration and Empirical Bayes Variable Selection," *Biometrika*, 87, 731–747. [1389]

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889. [1389]

——— (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373. [1389]

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*, Alexandria, VA: American Statistical Association, pp. 571–578. [1387]

——— (1996), "Variable Selection and Model Comparison in Regression," in *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford, U.K.: Oxford University Press, pp. 609–620. [1389]

Griffin, J., and Brown, P. (2007), "Baysian Adaptive Lassos With Non-Convex Penalization," Working Paper 07-02, CRiSM. [1384,1391]

——— (2010), "Inference With Normal-Gamma Prior Distributions in Regression Problems," *Bayesian Analysis*, 5, 171–188. [1384]

Hans, C. (2009), "Bayesian Lasso Regression," *Biometrika*, 96, 835–845. [1384,1385]

——— (2010), "Model Uncertainty and Variable Selection in Bayesian Lasso Regression," *Statistics and Computing*, 20, 221–229. [1389]

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), London: Oxford University Press. [1383]

Kohn, R., Smith, M., and Chan, D. (2001), "Nonparametric Regression Using Linear Combinations of Basis Functions," *Statistics and Computing*, 11, 313–322. [1389]

Li, Q., and Lin, N. (2010), "The Bayesian Elastic Net," *Bayesian Analysis*, 5, 151–170. [1385,1387]

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. O. (2008), "Mixtures of *g*-Priors for Bayesian Variable Selection," *Journal of the American Statistical Association*, 103, 410–423. [1385]

Lokhorst, J., Venables, B., and Turlach, B. (2007), "lasso2: L1 Constrained Estimation, aka 'lasso'," R package version 1.2-6. [1387]

Osborne, B. G., Fearn, T., Miller, A. R., and Douglas, S. (1984), "Application of Near Infrared Reflectance Spectroscopy to Compositional Analysis of Biscuits and Biscuit Doughs," *Journal of the Science of Food and Agriculture*, 35, 99–105. [1391]

Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [1384,1385,1388,1390]

Pericchi, L., and Smith, A. (1992), "Exact and Approximate Posterior Moments for a Normal Location Parameter," *Journal of the Royal Statistical Society, Ser. B*, 54, 793–804. [1384]

Polson, N. G., and Scott, J. G. (2011a), "Local Shrinkage Rules, Lévy Processes, and Regularized Regression," *Journal of the Royal Statistical Society, Ser. B*, to appear. [1384,1385]

——— (2011b), "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction," in *Bayesian Statistics 9*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxford, U.K.: Oxford University Press. [1384-1386]

Smith, M., and Kohn, R. (1996), "Nonparametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–343. [1389]

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," *Journal of Urology*, 16, 1076–1083. [1387]

Tanner, M., and Wong, W. (1987), "Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540. [1385]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288. [1383]

Tweedie, M. (1957), "Statistical Properties of Inverse Gaussian Distributions. I," *Annals of Mathematical Statistics*, 28, 362–377. [1387]

West, M. (1987), "On Scale Mixtures of Normal Distributions," *Biometrika*, 74, 646–648. [1385]

Yi, N., and Xu, S. (2008), "Bayesian LASSO for Quantitative Trait Loci Mapping," *Genetics*, 179, 1045–1055. [1384]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society, Ser. B*, 67, 301–320. [1383,1385,1390]