The Bayesian Bridge

Nicholas G. Polson
*University of Chicago**

James G. Scott
Jesse Windle
University of Texas at Austin

First Version: July 2011
This Version: October 2012

Abstract

We develop the Bayesian bridge estimator for regularized regression and classification. We focus on two distinct mixture representations for the prior distribution that give rise to the Bayesian bridge model: (1) a scale mixture of normals with respect to an alpha-stable random variable; and (2) a mixture of Bartlett–Fejer kernels (or triangle densities) with respect to a two-component mixture of gamma random variables. The first representation is a well known result due to West (1987), and is the more efficient choice for collinear design matrices. The second representation is new, and is more efficient for orthogonal problems, largely by avoiding the need to deal with exponentially tilted stable random variables. It also provides insight into the multimodality of the joint posterior distribution, a feature of the bridge model that is notably absent under ridge or lasso-type priors (Park and Casella, 2008; Hans, 2009). We find that the Bayesian bridge model outperforms its classical cousin in estimation and prediction across a variety of data sets, both simulated and real. We also prove a theorem that extends the approach to a wider class of regularization penalties that can be represented as scale mixtures of betas, and provide an explicit inversion formula for the mixing distribution. Finally, we show that the MCMC for fitting the bridge model has the striking property of generating nearly independent draws for the global scale parameter. This makes it far more efficient than analogous MCMC algorithms for fitting other sparse Bayesian models.

*Polson is Professor of Econometrics and Statistics at the Chicago Booth School of Business. email: ngp@chicagobooth.edu. Scott is Assistant Professor of Statistics at the University of Texas at Austin. email: James.Scott@mccombs.utexas.edu. Windle is a Ph.D student at the University of Texas at Austin. email: jwindle@ices.utexas.edu.

1 Introduction

1.1 Penalized likelihood and the Bayesian bridge

This paper studies the Bayesian analogue of the bridge estimator in regression, where $y = X\beta + \epsilon$ for some unknown vector $\beta = (\beta_1, \ldots, \beta_p)'$. Given choices of $\alpha \in (0, 1]$ and $\nu \in \mathbb{R}^+$, the bridge estimator $\hat{\beta}$ is the minimizer of

$$Q_y(\beta) = \frac{1}{2}||y - X\beta||^2 + \nu \sum_{j=1}^{p} |\beta_j|^\alpha \,. \tag{1}$$

This bridges a class of shrinkage and selection operators, with the best-subset-selection penalty at one end, and the ℓ^1 (or lasso) penalty at the other. An early reference to this class of models can be found in Frank and Friedman (1993), with recent papers focusing on model-selection asymptotics, along with strategies for actually computing the estimator (Huang et al., 2008; Zou and Li, 2008; Mazumder et al., 2011).

Our approach differs from this line of work in adopting a Bayesian perspective on bridge estimation. Specifically, we treat $p(\beta \mid y) \propto \exp\{-Q_y(\beta)\}$ as a posterior distribution having the minimizer of (1) as its global mode. This posterior arises in assuming a Gaussian likelihood for y, along with a prior for β that decomposes as a product of independent exponential-power priors (Box and Tiao, 1973):

$$p(\beta \mid \alpha, \nu) \propto \prod_{j=1}^{p} \exp(-|\beta_j/\tau|^\alpha) \;,\; \tau = \nu^{-1/\alpha} \,. \tag{2}$$

Rather than minimizing (1), we proceed by constructing a Markov chain having the joint posterior for β as its stationary distribution.

1.2 Relationship with previous work

Our paper emphasizes several interesting features of the Bayesian approach to bridge estimation. We summarize these features here, grouping them into three main categories.

Versus the Bayesian ridge and lasso priors. There is a large literature on Bayesian versions of classical estimators related to the exponential-power family, including the ridge (Lindley and Smith, 1972), lasso (Park and Casella, 2008; Hans, 2009, 2010), and elastic net (Li and Lin, 2010; Hans, 2011). Yet the bridge penalty has a crucial feature not shared by these other approaches: it is concave over $(0, \infty)$. From a Bayesian perspective, this implies that the prior for β has heavier-than-exponential tails. As a result, when the underlying signal is sparse, and when further regularity conditions are met, the bridge penalty dominates the lasso and ridge according to a classical criterion known as the oracle property (Fan and Li, 2001; Huang et al., 2008). Although the oracle property *per se* is of no particular relevance to a Bayesian treatment of the problem, it does correspond to a feature of certain prior distributions that Bayesians have long found important: the property of yielding a redescending score function for the marginal distribution of y (e.g.

Pericchi and Smith, 1992). This property is highly desirable in sparse situations, as it avoids the overshrinkage of large regression coefficients even in the presence of many zeros (Polson and Scott, 2011a).

Versus the classical bridge estimator. Both the classical and Bayesian approaches to bridge estimation must confront a significant practical difficulty: exploring and summarizing a multimodal surface in high-dimensional Euclidean space. In our view, multimodality is one of the strongest arguments for pursuing a full Bayes approach. For one thing, it is misleading to summarize a multimodal surface in terms of a single point estimate, no matter how appealingly sparse that estimate may be. Moreover, Mazumder et al. (2011) report serious computational difficulties with getting stuck in local modes in attempting to minimize (1). Our sampling-based approach, while not immune to this difficulty, seems very effective at exploring the whole space. (As Section 2 will show, there are very good reasons for expecting this to be the case, based on the structure of the data-augmentation strategy we pursue.) In this respect, MCMC behaves like a simulated annealing algorithm that never cools.

In addition, previous authors have emphasized three other points about penalized-likelihood rules that will echo in the examples we present in Section 4. First, one must choose a penalty parameter ν. In the classical setting this can be done via cross validation, which usually yields reasonable results. Yet this ignores uncertainty in the penalty parameter, which may be considerable. We are able to handle this in a principled way by averaging over uncertainty in the posterior distribution, under some default prior for the global variance component τ^2 (e.g. Gelman, 2006; Polson and Scott, 2012b). In the case of the bridge estimator, this logic may also be extended to the concavity parameter α, for which even less prior information is typically available.

Second, the minimizer of (1) may produce a sparse estimator, but this estimate is provably suboptimal, in a Bayes-risk sense, with respect to most traditional loss functions. If, for example, one wishes either to estimate β or to predict future values of y under squared-error loss, then the optimal solution is the posterior mean, not the mode. Both Park and Casella (2008) and Hans (2009) give realistic examples where the "Bayesian lasso" significantly outperforms its classical counterpart, both in prediction and in estimation. Similar conclusions are reached by Efron (2009) in a parallel context. Our own examples provide evidence of the practical differences that arise on real data sets—not merely between the mean and the mode, but also between the classical bridge solution and the mode of the joint distribution in the Bayesian model, marginal over over τ and σ. In the cases we study, the Bayesian approach leads to lower risk, often dramatically so.

Third, a fully Bayesian approach can often lead to different substantive conclusions than a traditional penalized-likelihood analysis, particularly regarding which components of β are important in predicting y. For example, Hans (2010) produces several examples where the classical lasso estimator aggressively zeroes out components of β for which, according to a full Bayes analysis, there is quite a large amount of posterior uncertainty regarding their size. This is echoed in our analysis of the classic data set on diabetes in Pima Indians. This is not to suggest that one conclusion is right, and the other wrong, in any specific

setting—merely that the two conclusions can be quite different, and that practitioners are well served by having both at hand.

Versus other sparsity-inducing priors in Bayesian regression analysis. Within the broader class of regularized estimators in high-dimensional regression, there has been widespread interest in cases where the penalty function corresponds to a normal scale mixture. Many estimators in this class share the favorable sparsity-inducing property (i.e. heavy tails) of the Bayesian bridge model. This includes the relevance vector machine of Tipping (2001); the normal/Jeffreys model of Figueiredo (2003) and Bae and Mallick (2004); the normal/exponential-gamma model of Griffin and Brown (2012); the normal/gamma and normal/inverse-Gaussian (Caron and Doucet, 2008; Griffin and Brown, 2010); the horseshoe prior of Carvalho et al. (2010); and the double-Pareto model of Armagan et al. (2012).

In virtually all of these models, the primary difficulty is the mixing rate of the MCMC used to sample from the joint posterior for β. Most MCMC approaches in this realm use latent variables to make sampling convenient. But this can lead to poor mixing rates, especially in cases where the fraction of "missing information"—that is, the information in the conditional distribution for β introduced by the latent variables—is large. Section 3.3 of the paper by Hans (2009) contains an informative discussion of this point. We have also included an online supplement to the manuscript that extensively documents the mixing behavior of Gibbs samplers within this realm.

In light of these difficulties, it comes as something of a surprise that the Bayesian bridge model leads to an MCMC strategy with an excellent mixing rate. There are actually two such approaches, both of which have the remarkable property of generating nearly independent draws for τ, the global scale parameter. For example, Figure 1 compares the performance of our bridge MCMC versus the best known Gibbs sampler for fitting the horseshoe prior (Carvalho et al., 2010) on a 1000-variable orthogonal regression problem with 900 zero entries in β. (See the supplement for details.) The plots show the first 2500 iterations of the sampler, starting from $\tau = 1$. There is a dramatic difference in the effective sampling rate for τ, which controls the overall level of sparsity in the estimate of β. (Though these results are not shown here, equally striking differences emerge when comparing the simulation histories of the local scale parameters under each method.)

1.3 Computational approach

Thus we would summarize the potential advantages of the Bayesian bridge as follows. It leads to richer model summaries, superior performance in estimation and prediction, and better uncertainty quantification compared to the classical bridge. It is better at handling sparsity than the Bayesian lasso. And it leads to an MCMC with superior mixing compared to other heavy-tailed, sparsity-inducing priors widely used in Bayesian inference.

These advantages, however, do not come for free. In particular, posterior inference for the Bayesian bridge is more challenging than in most other Bayesian models of this type, where MCMC sampling relies upon representing the implied prior distribution for β_j as a scale mixture of normals. The exponential-power prior in (2) is known to lie within the normal-scale mixture class (West, 1987). Yet the mixing distribution that arises in the

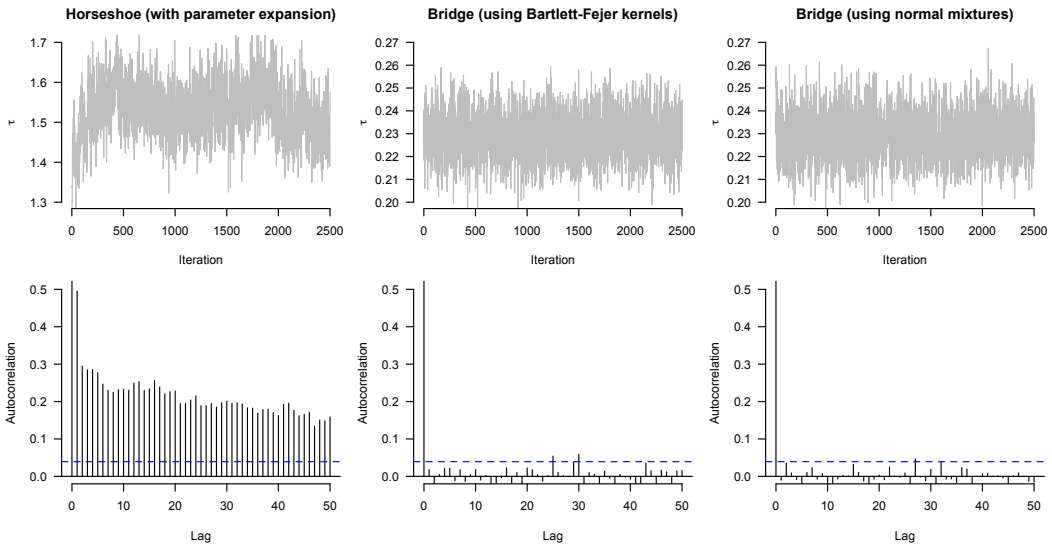

Figure 1: Comparison of the simulation histories for τ, the global scale parameter, using MCMC for the bridge and the horseshoe on a 1000-dimensional orthogonal regression problem with $n = 1100$ observations. There were 100 non-zero entries in β simulated from a t_4 distribution, and 900 zeros. Because the priors have different functional forms, the τ parameters in each model have a comparable role but not a comparable scale, which accounts for the difference between the vertical axes.

conditional posterior is that of an exponentially tilted alpha-stable random variable. This complicates matters, due to the lack of a closed-form expression for the density function. This fact was recognized by Armagan (2009), who proposed using variational methods to perform approximate Bayesian inference.

These issues can be overcome in two ways. We outline our computational strategy here, and provide further details in Sections 2 and 3. The R package `BayesBridge`, freely available online, implements all methods and experiments described in this paper.

The first approach is to work directly with normal mixtures of stable distributions, using rejection sampling or some other all-purpose algorithm within the context of a Gibbs sampler. Some early proposals for sampling stable distributions can be found in Devroye (1996) and Godsill (2000). Neither of these proved to be sufficiently robust in our early implementations of the method. But a referee pointed us to a much more recent algorithm from Devroye (2009). The method is somewhat complicated, but seems very robust, and leads to generally excellent performance (see the empirical results in the online supplement). Given current technology, it appears to be the best method for sampling the bridge model when the design matrix exhibits strong collinearity.

There is also a second, novel approach that turns out to be more efficient than the mixture-of-normals MCMC when the design matrix is orthogonal, or nearly so. Specifically, we appeal to the following mixture representation, which is a special case of a more general

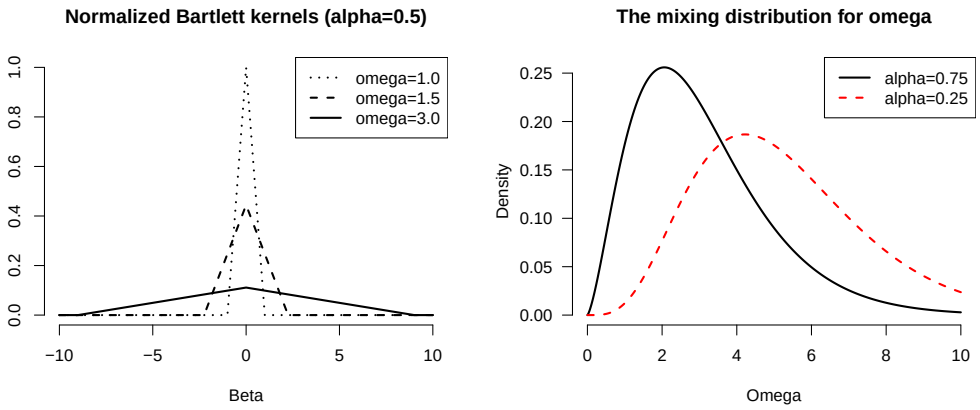

Figure 2: Left: triangular densities, or normalized Bartlett–Fejer kernels, of different widths. Right: two examples of mixing distributions for ω_j that give rise to exponential-power marginals for β_j in conjunction with the Bartlett–Fejer kernel.

result based on the Schoenberg–Williamson theorem for n-monotone densities:

$$
\begin{aligned}
(y \mid \beta, \sigma^2) &\sim \mathrm{N}(X\beta, \sigma^2 I) \\
p(\beta_j \mid \tau, \omega_j, \alpha) &= \frac{1}{\tau\omega_j^{1/\alpha}} \cdot \left\{ 1 - \left| \frac{\beta_j}{\tau\omega_j^{1/\alpha}} \right| \right\}_+ \qquad (3) \\
(\omega_j \mid \alpha) &\sim \frac{1+\alpha}{2} \cdot \mathrm{Ga}(2 + 1/\alpha, 1) + \frac{1-\alpha}{2} \cdot \mathrm{Ga}(1 + 1/\alpha, 1)\,. \qquad (4)
\end{aligned}
$$

This scale mixture of triangles, or Bartlett-Fejer kernels, recovers $e^{-Q_y(\beta)}$ as the marginal posterior in β. The mixing distribution is depicted in Figure 2 and explained in detail in Section 2.2. It leads to a simple MCMC that avoids the need to deal with alpha-stable distributions, and can easily hop between distinct modes in the joint posterior. (This is aided by the fact that the mixing distribution for the local scale ω_j has two distinct components.)

This alleviates one of the major outstanding difficulties of working with the bridge objective function. As Section 2 will show, it is also of significant interest in its own right, and can be used to represent many other penalty functions and likelihoods that arise in high-dimensional inference. Our main theorem leads to an explicit Bayesian representation for any non-convex penalty function whose corresponding density version is proper. This is a very wide class of penalties that can be accommodated via data augmentation.

2 Data augmentation for the bridge model

2.1 As a scale mixture of normals

We begin by discussing the two different data augmentation strategies that facilitate posterior inference for the Bayesian bridge model.

First, there is the mixture-of-normals representation, well known since West (1987).

This can be seen by appealing to Bernstein's theorem, which holds that a function $f(x)$ is completely monotone if and only if it can be represented as a Laplace transform of some distribution function $G(\lambda)$:

$$f(x) = \int_0^\infty e^{-sx} dG(s)\,. \tag{5}$$

To represent the exponential-power prior as a Gaussian mixture for $\alpha \in (0, 2]$, let $x = t^2/2$. We then have

$$\exp(-|t|^\alpha) = \int_0^\infty e^{-st^2/2} g(s)\; ds\,, \tag{6}$$

where $g(s)$ can be identified by recognizing the left-hand side as the Laplace transform, evaluated at $t^2/2$, of a positive alpha-stable random variable with index of stability $\alpha/2$ (also see Polson and Scott, 2012a).

Similar Gaussian representations have been exploited to yield conditionally conjugate MCMC algorithms for a variety of models, such as the lasso and the horseshoe priors. Unfortunately, the case of the bridge is less simple. To see this, consider the joint posterior implied by (1) and (6):

$$\begin{aligned} p(\beta, \Lambda \mid y) &= C \exp\left(-\nu^{2/\alpha}\beta'\Lambda\beta - \frac{1}{2\sigma^2}\beta' X'X\beta + \beta'\sigma^{-2}X'y\right)\prod_{j=1}^p p(\lambda_j) \\ &= C \exp\left\{-\frac{1}{2}\beta'\left(\sigma^{-2}X'X + 2\nu^{2/\alpha}\Lambda\right)\beta + \beta'\sigma^{-2}X'y\right\}\prod_{j=1}^p p(\lambda_j)\,, \end{aligned} \tag{7}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_j)$, and $p(\lambda_j) = \lambda_j^{-1/2} g(\lambda_j)$, g denoting the stable density from the integrand in (6). The conditional posterior of λ_j given β_j is then an exponentially tilted stable random variable,

$$p(\lambda_j \mid \beta_j) = \frac{e^{-\nu^{\frac{2}{\alpha}}|\beta_j|^2\lambda_j} p(\lambda_j)}{\mathbb{E}\left(e^{-\nu^{\frac{2}{\alpha}}|\beta_j|^2\lambda_j}\right)}\,,$$

with the expectation in the denominator taken over the prior. Neither the prior nor posterior for λ_j are known in closed form, and can be only be written explicitly as an infinite series.

2.2 An alternative approach for n-monotone densities

Bernstein's theorem holds for completely monotone density functions, and can be used to construct scale mixtures of normals by evaluating the right-hand side of (5) at $t^2/2$. As we have seen in the case of the bridge, this results in a conditionally Gaussian form for the parameter of interest, but a potentially difficult mixing distribution for the latent variable.

We now construct an alternate data-augmentation scheme that avoids these difficulties. Specifically, consider the class of symmetric density functions $f(x)$ that are n-monotone on $(0, \infty)$ for some integer n: that is, $(-1)^k f^{(k)}(|x|) \geq 0$ for $k = 0, \ldots, n-1$, where $f^{(k)}$ is the kth derivative of f, and $f^{(0)} \equiv f$.

The following result builds on a classic theorem of Schoenberg and Williamson. It establishes that any n-monotone density $f(x)$ may be represented as a scale mixture of

betas, and that we may invert for the mixing distribution using the derivatives of f.

Theorem 2.1. *Let $f(x)$ be a bounded density function that is symmetric about zero and n-monotone over $(0, \infty)$, normalized so that $f(0) = 1$. Let $C = \{2 \int_0^\infty f(t)\ dt\}^{-1}$ denote the normalizing constant that makes $f(x)$ a proper density on the real line. Then f can be represented as the following mixture for any integer k, $1 \leq k \leq n$:*

$$Cf(x) = \int_0^\infty \frac{1}{s}\ k \left(1 - \frac{|x|}{s}\right)_+^{k-1}\ g(s)ds\,, \tag{8}$$

where $a_+ = \max(a, 0)$, and where the mixing density $g(s)$ is

$$g(s) = Ck^{-1} \sum_{j=0}^{k-1} \frac{(-1)^j}{j!} \left\{ j s^j f^{(j)}(s) + s^{j+1} f^{(j+1)}(s) \right\}\,.$$

Crucially, the mixing density in the k-monotone case has only a finite number of terms. Moreover, a function that is completely monotone is also n-monotone for all finite n. Thus the proposition applies to any function for which Bernstein's theorem holds, allowing an arbitrary (presumably convenient) choice of n.

To see the connection between our proposition and Bernstein's theorem, let $u = k/s$. Observe that we obtain the completely monotonic case as k diverges:

$$\begin{aligned} f(x) &\propto \int_0^\infty \left(1 - \frac{ux}{k}\right)_+^{k-1}\ dP(u) \\ &\to \int_0^\infty e^{-sx} d\tilde{P}(s) \end{aligned}$$

for positive x and a suitably defined limiting measure $\tilde{P}(s)$, into which a factor of s has been implicitly absorbed. By evaluating this at $s = t^2/2$, we obtain a scale mixture of normals as a limiting case of a scale mixture of betas. The inversion formula, too, is similar. In particular, for the case of the exponential power kernel, we have

$$\exp\left(-|x|^\alpha\right) = \int_0^\infty e^{-xs} g(s) ds \quad \text{with} \quad g(s) = \sum_{j=1}^\infty (-1)^j \frac{s^{-j\alpha - 1}}{j! \Gamma(-\alpha j)}\,,$$

which clearly parallels the expression given in Proposition 2.1.

Return now to the Bayesian bridge model. The exponential power density is completely monotone on the positive reals, and therefore any value of k may be used in Equation (8). We focus on the choice $k = 2$, which leads to a mixture of Bartlett–Fejer kernels, a special case both of the beta and triangle distributions. The proof involves only simple manipulations, and is omitted.

Corollary 2.2. *Let $f(x)$ be a function that is symmetric about the origin; integrable, convex, and twice-differentiable on $(0, \infty)$; and for which $f(0) = 1$. Let $C = \{2 \int_0^\infty f(t)\ dt\}^{-1}$ denote the normalizing constant that makes $f(x)$ a density on the real line. Then f is the following*

mixture of Bartlett–Fejer kernels:

$$Cf(x) = \int_0^\infty \frac{1}{s} \left\{ 1 - \frac{|t|}{s}| \right\}_+ Cs^2 f''(s)\ ds\,, \tag{9}$$

where $a_+ = \max(a, 0)$.

These have been referred to as Bartlett kernels in econometrics, a usage which appears to originate in a series of papers by Newey and West on robust estimation. They have also been called Fejer densities in probability theory; see Dugué and Girault (1955), who study them in connection with the theory of characteristic functions of Polya type.

Using this corollary, the exponential power density with $\alpha \in (0, 1]$ can be represented in a particularly simple way. To see this, transform $s \to \omega \equiv s^\alpha$ and observe that:

$$\begin{aligned} \frac{1}{2\tau} \exp(-|\beta/\tau|^\alpha|) &= \int_0^\infty \frac{1}{\tau} \left\{ 1 - \left| \frac{\beta}{\tau\omega^{1/\alpha}} \right| \right\}_+ p(\omega \mid \alpha)\ d\omega \\ p(\omega \mid \alpha) &= \alpha\omega e^{-\omega} + (1-\alpha)e^{-\omega}\,. \end{aligned}$$

Simple algebra with the normalizing constants yields a properly normalized mixture of Bartlett–Fejer kernels:

$$\begin{aligned} \frac{\alpha}{2\tau\Gamma(1+1/\alpha)} \exp(-|\beta/\tau|^\alpha|) &= \int_0^\infty \frac{1}{\tau\omega^{1/\alpha}} \left\{ 1 - \left| \frac{\beta}{\tau\omega^{1/\alpha}} \right| \right\}_+ p(\omega \mid \alpha)\ d\omega \\ p(\omega \mid \alpha) &= \frac{1+\alpha}{2} c_1 \omega^{1+1/\alpha} e^{-\omega} + \frac{1-\alpha}{2} c_2 \omega^{1/\alpha} e^{-\omega}\,, \end{aligned}$$

This is a simple two-component mixture of gammas, where c_1 and c_2 are the normalizing constants of each component. The Bayesian lasso is a special case, for which the second mixture component drops out.

2.3 The connection with slice sampling

The above scheme was originally motivated by the potential inefficiencies of working with exponentially tilted stable random variables, and does lead to noticeable improvements in the orthogonal case. Moreover, the representation is very intuitive, in that it allows one to see precisely how two salient features of the bridge posterior arise from the prior. Its nondifferentiable point at zero is reflected directly in the triangular kernel. Its multimodality is reflected in the fact that the conditional posterior for each latent ω_j will have two distinct components.

But the representation is of considerable interest in its own right, quite apart from its application to the bridge model. It leads to MCMC sampling methods that are simple to program, that require no ad-hoc tuning, and that generalize to a very wide class of problems.

The analogy with slice sampling is instructive. In both cases, the basic problem is to sample from a posterior distribution of the form $L(\theta)p(\theta)/Z$, where L is a likelihood, p is a prior, and Z is the normalization constant. For example, if we slice out the prior, we introduce an auxiliary variable u, conditionally uniform on $0 \le u < p(\theta)$, and sample from

the joint distribution

$$\pi(\theta, u) = \mathbb{I}\left(u < p(\theta)\right) L(\theta)/Z\,,$$

where $\mathbb{I}(\cdot)$ is the indicator function. The posterior of interest is then the marginal distribution for θ. The difficulty is that, given u, one needs to be able to calculate the slice region where $p(\theta) > u$. This is often nontrivial. See, for example, Damien et al. (1999), Roberts and Rosenthal (2002), or Neal (2003).

In our data-augmentation approach, the analogous inversion problem is already done. For example, using a mixture of triangles, it reduces to the set where $|\theta| < \omega$. Instead, we must work with a joint distribution, with ω replacing u, given by

$$\pi(\theta, \omega) = \mathbb{I}\left(|\theta| < \omega\right) g(\omega)\left(1 - |\theta|/\omega\right) L(\theta)/Z\,.$$

We have removed the problem of inverting a slice region, at the cost of introducing two new problems. First, we must identify $g(\omega)$ such that we get the appropriate marginal upon integrating out ω. This is where Theorem 2.1 proves useful, as it can be applied to derive the explicit form of g for a wide class of densities. Second, we must sample from the tilted distribution whose density is proportional to $(1 - |\theta|/\omega)\, L(\theta)$. In many cases, $L(\theta)$ itself can be used to construct an envelope in a rejection sampler. In other cases, one may appeal to the algorithm of Stein and Keblis (2009) for simulating the triangle distribution, which can be extended to the case of a triangle times another density.

Of course, the question of whether the slice method or the mixture-of-betas method leads to simpler calculations will be context dependent. Although a long discussion here would lead us astray from our main point, there are clearly many interesting cases where the new approach could prove fruitful. One such example is the type-I extreme value distribution, $p(x) = \exp(-x - e^{-x})$. From Theorem 2.2, we have

$$e^{-x} = \int_0^\infty \left(1 - \frac{|x|}{\omega}\right)_+ e^{-\omega} d\omega\,,$$

and therefore $e^{-e^{-x}}$ can be written as a mixture of gammas:

$$e^{-e^{-x}} = \int_0^\infty \frac{1}{\omega}\left(1 - \frac{e^{-x}}{\omega}\right)_+ \omega e^{-\omega} d\omega\,.$$

3 MCMC sampling for the Bayesian bridge

3.1 Overview of approach

For sampling the Bayesian bridge posterior, we recommend a hybrid computational approach, which we have implemented as the default setting in our `BayesBridge` R package. Due to space constraints, the evidence supporting this recommendation is outlined in an online supplemental file, where we describe the results of an extensive benchmarking study. We briefly summarize our conclusions here.

When the design matrix X exhibits strong collinearity, the normal scale mixture representation is the better choice. In cases where there is interest in fitting many higher-order

interaction terms, the efficiency advantage can be substantial. On the other hand, the Bartlett-Fejer representation is the better choice when the design matrix is orthogonal, usually enjoying an effective sampling rate roughly two to three times that of the Gaussian method. The orthogonal case applies to nonlinear regression problems where the effect of a covariate is expanded in an orthogonal basis. It also has connections with the generalized g-priors for $p > n$ problems discussed in Polson and Scott (2012a).

Once one has a method for sampling exponentially tilted alpha-stable random variables, it is easy to use (7) to generate posterior draws, appealing to standard multivariate normal theory. Thus we omit a discussion of this method in the main manuscript, and focus on the mixture-of-betas approach.

3.2 Sampling β and the latent variables

To see why the representation in (3)–(4) leads to a simple algorithm for posterior sampling, consider the joint distribution for β and the latent ω_j's:

$$p(\beta, \Omega \mid \tau, y) = C \exp\left(-\frac{1}{2\sigma^2}\beta' X' X \beta + \frac{1}{\sigma^2}\beta' X' y\right) \prod_{i=1}^{p} p(\omega_j \mid \alpha) \prod_{i=1}^{p} \left(1 - \frac{|\beta_j|}{\tau \omega_j^{1/\alpha}}\right)_+ . \quad (10)$$

Introduce further slice variables $u_1, \ldots, u_j$. This leads to the joint posterior

$$\begin{aligned} p(\beta, \Omega, u \mid \tau, y) &\propto \exp\left(-\frac{1}{2\sigma^2}\beta' X' X \beta + \frac{1}{\sigma^2}\beta' X' y\right) \\ &\times \prod_{j=1}^{p} p(\omega_j \mid \alpha) \prod_{j=1}^{p} \mathbb{I}\left(0 \le u_j \le 1 - \frac{|\beta_j|}{\tau \omega_j^{1/\alpha}}\right) . \end{aligned} \quad (11)$$

Note that we have implicitly absorbed a factor of $\omega^{1/\alpha}$ from the normalization constant for the Bartlett–Fejer kernel into the gamma conditional for ω_j. This will make inverting the slice region for ω_j far easier.

Applying Corollary 2.2, if we marginalize out both the slice variables and the latent ω_j's, we recover the Bayesian bridge posterior distribution,

$$p(\beta \mid y) = C \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2 - \sum_{j=1}^{p} |\beta_j/\tau|^\alpha\right) .$$

We can invert the slice region in (11) by defining (a_j, b_j) as

$$|\beta_j| \le \tau^{-1}(1 - u_j)\omega_j^{1/\alpha} = b_j \;\text{ and }\; \omega_j \ge \left(\frac{\tau|\beta_j|}{1 - u_j}\right)^\alpha = a_j .$$

This leads us to an exact Gibbs sampler that starts at initial guesses for (β, Ω) and iterates the following steps:

1. Generate $(u_j \mid \beta_j, \omega_j) \sim \text{Unif}\left(0, 1 - \tau|\beta_j|\omega_j^{-1/\alpha}\right)$.
2. Generate each ω_j from a mixture of truncated gammas, as described below.

3. Generate β from a truncated multivariate normal proportional to

$$\mathrm{N}\left(\hat{\beta}, \sigma^2 (X'X)^{-1}\right) \mathbb{I}\left(|\beta_j| \leq b_j \text{ for all } j\right) ,$$

where $\hat{\beta}$ indicates the least-squares estimate for β.

We explored several different methods for simulating from the truncated multivariate normal, ultimately settling on the proposal of Rodriguez-Yam et al. (2004) as the most efficient. The conditional posterior of the latent ω_j's can be determined as follows. Suppressing subscripts for the moment, it is clear from (11) that

$$\begin{aligned} p(\omega \mid \alpha) &= \alpha(\omega e^{-\omega}) + (1-\alpha)e^{-\omega} \\ p(\omega \mid a, \alpha) &= C_a \left\{\alpha(\omega e^{-\omega}) + (1-\alpha)e^{-\omega}\right\} \mathbb{I}\left(\omega \geq a\right) , \end{aligned}$$

where a comes from inverting the slice region in (11) and C_a is the normalization constant.

We can simulate from this mixture of truncated gammas by defining $\bar{\omega} = \omega - a$, where $\bar{\omega} > 0$. Then $\bar{\omega}$ has density

$$\begin{aligned} p(\bar{\omega}|a, \alpha) &= C_a \left\{\alpha e^{-a}(a + \bar{\omega})e^{-\bar{\omega}} + (1-\alpha)e^{-a}e^{-\bar{\omega}}\right\} \\ &= \frac{\alpha}{1+\alpha a} \cdot \bar{\omega} e^{-\bar{\omega}} + \frac{1-\alpha(1+a)}{1+\alpha a} \cdot e^{-\bar{\omega}} . \end{aligned}$$

This is a mixture of gammas, where

$$(\bar{\omega} \mid a) \sim \begin{cases} \Gamma(1,1) & \text{with prob} \quad \frac{1-\alpha(1+a)}{1+\alpha a} \\ \Gamma(2,1) & \text{with prob} \quad \frac{\alpha}{1+\alpha a} . \end{cases}$$

After sampling $\bar{\omega}$, simply transform back using the fact that $\omega = a + \bar{\omega}$.

This representation has two interesting and intuitive features. First, full conditional for β in step 3 is centered at the usual least-squares estimate $\hat{\beta}$. Only the truncations (b_j) change at each step, which speeds matrix operations. Compare this to the usual scale-mixture representation, which involves inverting a matrix of the form $(X'X + \Lambda)^{-1}$ at every MCMC step.

Second, the mixture-of-gammas form of $p(\omega)$ naturally accounts for the bimodality in the marginal posterior distribution, $p(\beta_j \mid y) = \int p(\beta_j \mid \omega, y) p(\omega_j \mid y) d\omega_j$. Each mixture component of the conditional for ω_j represents a distinct mode of the marginal posterior for β_j. As the examples later will show, this endows the algorithm with the ability to explore various modes of the joint posterior very easily.

3.3 Sampling hyperparameters

To update the global scale parameter τ, we work directly with the exponential-power density, marginalizing out the latent variables $\{\omega_j, u_j\}$. From (1), observe that the posterior for

$\nu \equiv \tau^{-\alpha}$, given β, is conditionally independent of y, and takes the form

$$p(\nu \mid \beta) \propto \nu^{p/\alpha} \exp(-\nu \sum_{j=1}^{p} |\beta_j|^{\alpha}) \; p(\nu) \, .$$

Therefore if ν has a Gamma(c, d) prior, its conditional posterior will also be a gamma distribution, with hyperparameters $c^{\star} = c + p/\alpha$ and $d^{\star} = d + \sum_{j=1}^{p} |\beta_j|^{\alpha}$. To sample τ, simply draw ν from this gamma distribution, and use the transformation $\tau = \nu^{-1/\alpha}$. Alternative priors for ν can also be considered, in which case the gamma form of the conditional likelihood in ν will make for a useful proposal distribution that closely approximates the posterior. As Figure 1 from the introduction shows, the ability to marginalize over the local scales in sampling τ is crucial here in leading to a good mixing rate.

In many cases the concavity parameter α will be fixed ahead of time to reflect a particular desired shape of the penalty function. But it too can be give a prior $p(\alpha)$, most conveniently from the beta family, and can be updated using a random-walk Metropolis sampler.

4 Examples

4.1 Diabetes data

We first explore the Bayesian bridge estimator using the well-known data set on diabetes among Pima Indians, available in the R package `lars` (see, e.g. Efron et al., 2004). The main data set has 10 predictors and 442 observations. Yet even for this relatively information-rich problem, significant differences emerge between the Bayesian and classical methods.

We also fit the Bayesian bridge, using Algorithm 1 and a default Gamma(2,2) prior for ν. We also fit the classical bridge, using generalized cross validation and the EM algorithm from Polson and Scott (2011b). Both the predictor and responses were centered, while the predictors were also re-scaled to have unit variance. At each step of the MCMC for the Bayesian model, we calculated the conditional posterior density for each β_j at a discrete grid of values.

Figure 3 summarizes the results of the two fits, showing both the marginal posterior density and the classical bridge solution for each of the 10 regression coefficients. One notable feature of the problem is the pronounced multimodality in the joint posterior distribution for the Bayesian bridge. Observe, for example, the two distinct modes in the marginal posteriors for the coefficients associated with the TCH and Glucose predictors (and, to a lesser extent, for the HDL and Female predictors). In none of these cases does it seem satisfactory to summarize information about β_j using only a single number, as the classical solution forces one to do.

Second, observe that the classical bridge solution does not coincide with the joint mode of the fully Bayesian posterior distribution. This discrepancy can be attributed to uncertainty in τ and σ, which is ignored in the classical solution. Marginalizing over these hyperparameters leads to a fundamentally different objective function, and therefore a different joint posterior mode.

The difference between the classical mode and the Bayesian mode, moreover, need not

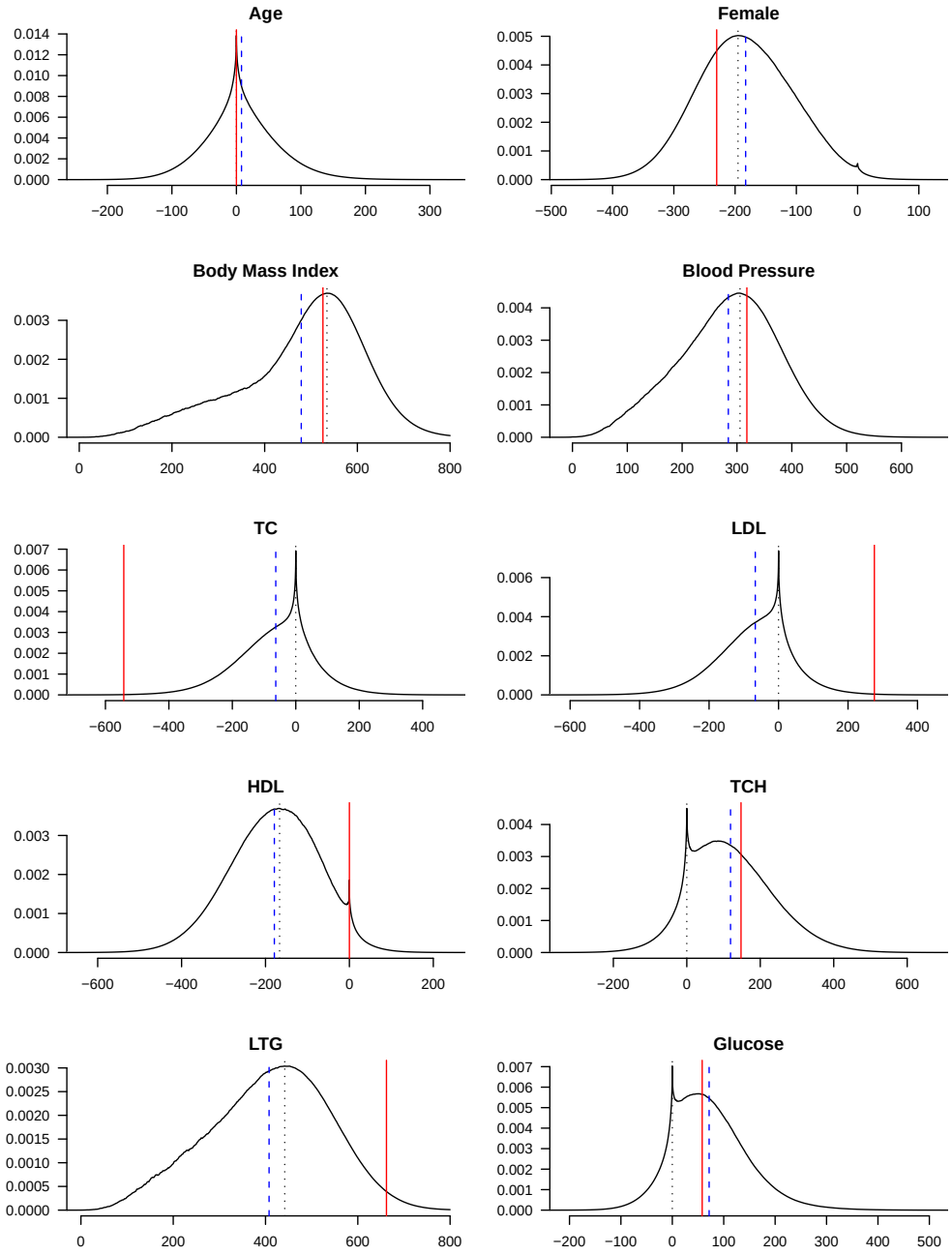

Figure 3: Marginal posterior densities for the marginal effects of 10 predictors in the diabetes data. Solid red line: penalized-likelihood solution with ν chosen by generalized cross validation. Dashed blue line: marginal posterior mean for β_j. Dotted black line: mode of the marginal distribution for β_j under the fully Bayes posterior.

be small. Observe, for example, the middle row in Figure 3, which shows the posterior distributions for the TC and LDL coefficients. These two predictors have a sample correlation of -0.897. The Bayesian solution concentrates in a region of $\mathbb{R}^p$ where neither of these coefficients exerts much of an effect. The classical solution, on the other hand, says that both predictors should be in the model with large coefficients of opposite sign.

It is impossible to say in any objective sense whether TC and HDL are both necessary, or instead are redundant copies of the same unhelpful information. It is highly surprising, however, that such a marked difference would arise between the full Bayes mode and the classical mode, and that this difference would fundamentally alter one's conclusions about two predictors out of ten. (The full Bayes posterior mean is, of course, different yet again.) Clearly an very important role here is played by the decision of whether to account for uncertainty in τ and σ.

4.2 Out-of-sample prediction results

Next, we describe the results from three out-of-sample prediction exercises involving the following benchmark data sets.

Boston housing data: available in the R package `mlbench`. The goal is to predict the median house price for 506 census tracts of Boston from the 1970 census. As covariates, we used the 14 original predictors, plus all interactions and squared terms for quantitative predictors.

Ozone data: available in the R package `mlbench`. The goal is to predict the concentration of ozone in the atmosphere above Los Angeles using various environmental covariates. As covariates, we used the 9 original predictors, plus all interactions and squared terms for quantitative predictors.

NIR Glucose data: available in the R package `chemometrics`. The goal is to predict the concentration of glucose in molecules using data from NIR spectroscopy.

For each data set, we created 100 different train/test splits, using the results from the training data to forecast the test data. For each train/test split we estimated β using least-squares, the classical bridge (using EM), and the Bayesian-bridge posterior mean (using our MCMC method). In all cases we chose $\alpha = 0.5$; centered and standardized the predictors; and centered the response. For the classical bridge estimator, the regularization parameter ν was chosen by generalized cross validation; while for the Bayesian bridge, σ was assigned Jeffreys' prior and ν a default Gamma(2,2) prior.

We measured performance of each method by computing the sum of squared errors in predicting y on the test data set. Details of each data set, along with both the results and the train/test sample sizes used, are in Table 1. In all three cases, the posterior mean estimator outperforms both least squares and the classical bridge estimator.

4.3 Simulated data with correlated design

We conducted three experiments, all with $p = 100$ and $n = 101$, for $\alpha \in \{0.9, 0.7, 0.5\}$. Each experiment involved 250 data sets constructed by: (1) simulating regression coefficients from

Table 1: Average sum of squared errors in predicting hold-out observations for 100 different train/test splits on three real data sets.

				Prediction SSE		
Data set	n	p	train/test	LSE	Bridge	Bayes
Boston housing	506	103	422/84	1288	1147	455
Ozone	203	54	163/40	872	659	415
NIR glucose	40	166	110/56	2791	2980	2375

Table 2: Average sum of squared errors in estimating β for three different batches of 250 simulated data sets.

	LSE	Bridge	Bayes
$\alpha = 0.5$	2254	1611	99
$\alpha = 0.7$	1994	406	225
$\alpha = 0.9$	551	144	85

the exponential power distribution for the given choice of α; (2) simulating correlated design matrices X; and (3) simulating residuals from a Gaussian distribution. In all cases we set $\sigma = \tau = 1$. The rows of each design matrix were simulated from a Gaussian factor model, with covariance matrix $V = BB' + I$ for a 100×10 factor loadings matrix B with independent standard normal entries. As is typical for Gaussian factor models with many fewer factors (10) than ambient dimensions (100), this choice led to marked multi-collinearity among the columns of each simulated X.

For each simulated data set we again estimated β using least squares, the classical bridge, and the Bayesian bridge posterior mean. Performance was assessed by the sum of squared errors in estimating the true value of β. Convergence of both algorithms was assessed by starting from multiple distinct points in $\mathbb{R}^p$ and checking that the final solutions were identical up to machine and/or Monte Carlo precision. As before, for the classical bridge estimator, the regularization parameter ν was chosen by generalized cross validation; while for the Bayesian bridge, σ was assigned Jeffreys' prior and ν a Gamma(2,2) prior.

Table 2 shows the results of these experiments. For all three choices of α, the posterior mean estimator outperforms both least squares and the classical bridge estimator. Sometimes the difference is drastic—such as when $\alpha = 0.5$, where the Bayes estimator outperforms the classical estimator by more than a factor of 16.

5 Discussion

This paper has demonstrated a series of results that allow practitioners to estimate the full joint distribution of regression coefficients under the Bayesian bridge model. Our numerical experiments have shown: (1) that the classical mode, the full Bayes mode, and the full Bayes mean can often lead to very different summaries about the relative importance of

different predictors; and (2) that using the posterior mean offers substantial improvements over the mode when estimating β or making predictions under squared-error loss. Both results parallel the findings of Park and Casella (2008) and Hans (2009) for the Bayesian lasso.

The existence of a second, novel mixture representation for the Bayesian bridge is of particular interest, and suggests many generalizations, some of which we have mentioned. Our main theorem leads to a novel Gibbs-sampling scheme for the bridge that—by virtue of working directly with a two-component mixing measure for each latent scale ω_j—is capable of easily jumping between modes in the joint posterior distribution. It thereby avoids many of the difficulties associated with slow mixing in global-local scale-mixture models described by Hans (2009), and further studied in the online supplemental file. It appears to be the best algorithm in the orthogonal case, but suffers from poor mixing when the design matrix is extremely collinear. Luckily, in this case, the normal-mixture method based on the work of Devroye (2009) for sampling exponentially tilted stable random variables performs well. Both methods are implemented in the R package `BayesBridge`, available through CRAN. Together, they give practioners a set of tools for efficiently exploring the bridge model across a wide range of commonly encountered situations.

A Proofs

A.1 Theorem 2.1

Proof. Let M_n denote the class of n-times monotone functions on $(0, \infty)$. Clearly for $n \geq 2$, $f \in M_n \Rightarrow f \in M_{n-1}$. Thus it is sufficient to prove the proposition for $k = n$. As the density $f(x)$ is symmetric, we consider only positive values of x.

The Schoenberg–Williamson theorem (Williamson, 1956) states that a necessary and sufficient condition for a function $f(x)$ defined on $(0, \infty)$ to be in M_n is that

$$f(x) = \int_0^\infty (1 - ut)_+^{n-1} \, dG(u) \,,$$

for some $G(u)$ that is non-decreasing and bounded below. Moreover, if $G(u) = 0$, the representation is unique, in the sense of being determined at the points of continuity of $G(u)$, and is given by

$$G(u) = \sum_{j=0}^{n-1} \frac{(-1)^j f^{(j)}(1/u)}{j!} \left(\frac{1}{u}\right)^j \,.$$

□

References

A. Armagan. Variational bridge regression. *Journal of Machine Learning Research W&CP*, 5(17–24), 2009.

A. Armagan, D. Dunson, and J. Lee. Generalized double Pareto shrinkage. *Statistica Sinica*, to appear, 2012.

K. Bae and B. Mallick. Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, 20 (18):3423–30, 2004.

G. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, 1973.

F. Caron and A. Doucet. Sparse Bayesian nonparametric regression. In *Proceedings of the 25th International Conference on Machine Learning*, pages 88–95. Association for Computing Machinery, Helsinki, Finland, 2008.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97 (2):465–80, 2010.

P. Damien, J. C. Wakefield, and S. G. Walker. Bayesian nonconjugate and hierarchical models by using auxiliary variables. *J. R. Stat. Soc. Ser. B, Stat. Methodol.*, 61:331–44, 1999.

L. Devroye. Random variate generation in one line of code. In J. Charnes, D. Morrice, D. Brunner, and J. Swain, editors, *Proceedings of the 1996 Winter Simulation Conference*, pages 265–72, 1996.

L. Devroye. On exact simulation algorithms for some distributions related to Jacobi theta functions. *Statistics & Probability Letters*, 79(21):2251–9, 2009.

D. Dugué and M. Girault. Fonctions convexes de Polya. *Publications de l'Institut de Statistique des Universités de Paris*, 4:3–10, 1955.

B. Efron. Empirical Bayes estimates for large-scale prediction problems. *Journal of the American Statistical Association*, 104(487):1015–28, 2009.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32 (2):407–99, 2004.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–60, 2001.

M. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1150–9, 2003.

I. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35(2):109–135, 1993.

A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3): 515–33, 2006.

S. Godsill. Inference in symmetric alpha-stable noise using MCMC and the slice sampler. In *Acoustics, Speech, and Signal Processing*, volume 6, pages 3806–9, 2000.

J. Griffin and P. Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–88, 2010.

J. Griffin and P. Brown. Alternative prior distributions for variable selection with very many more variables than observations. *Australian and New Zealand Journal of Statistics*, 2012. (to appear).

C. M. Hans. Bayesian lasso regression. *Biometrika*, 96(4):835–45, 2009.

C. M. Hans. Model uncertainty and variable selection in Bayesian lasso regression. *Statistics and Computing*, 20:221–9, 2010.

C. M. Hans. Elastic net regression modeling with the orthant normal prior. *Journal of the American Statistical Association*, to appear, 2011.

J. Huang, J. Horowitz, and S. Ma. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics*, 36(2):587–613, 2008.

Q. Li and N. Lin. The Bayesian elastic net. *Bayesian Analysis*, 5(1):151–70, 2010.

D. Lindley and A. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34(1–41), 1972.

R. Mazumder, J. Friedman, and T. Hastie. Sparsenet: coordinate descent with non-convex penalties. *Journal of the American Statistical Association*, 106(495):1125–38, 2011.

R. M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–67, 2003.

T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482): 681–6, 2008.

L. R. Pericchi and A. Smith. Exact and approximate posterior moments for a normal location parameter. *Journal of the Royal Statistical Society (Series B)*, 54(3):793–804, 1992.

N. G. Polson and J. G. Scott. Shrink globally, act locally: sparse Bayesian regularization and prediction (with discussion). In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Proceedings of the 9th Valencia World Meeting on Bayesian Statistics*, pages 501–38. Oxford University Press, 2011a.

N. G. Polson and J. G. Scott. Data augmentation for non-Gaussian regression models using variance-mean mixtures. Technical report, University of Texas at Austin, http://arxiv.org/abs/1103.5407v3, 2011b.

N. G. Polson and J. G. Scott. Local shrinkage rules, Lévy processes, and regularized regression. *Journal of the Royal Statistical Society (Series B)*, 74(2):287–311, 2012a.

N. G. Polson and J. G. Scott. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 2012b.

G. O. Roberts and J. S. Rosenthal. Convergence of slice sampler Markov chains. *Journal of the Royal Statistical Society (Series B)*, 61(3):643–60, 2002.

G. Rodriguez-Yam, R. A. Davis, and L. L. Scharf. Efficient gibbs sampling of truncated multivariate normal with application to constrained linear regression. Columbia, March 2004.

W. E. Stein and M. F. Keblis. A new method to simulate the triangular distribution. *Mathematical and Computer Modeling*, 49(2009):1143–7, 2009.

M. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–44, 2001.

M. West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–8, 1987.

R. Williamson. Multiply monotone functions and their laplace transforms. *Duke Mathematics Journal*, 23 (189–207), 1956.

H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–33, 2008.