Journal of the Royal Statistical Society
SERIES B Statistical Methodology
J. R. Statist. Soc. B (2014)

The Bayesian bridge

Nicholas G. Polson
University of Chicago, USA

and James G. Scott and Jesse Windle
University of Texas at Austin, USA

[Received October 2011. Final revision June 2013]

Summary. We propose the Bayesian bridge estimator for regularized regression and classification. Two key mixture representations for the Bayesian bridge model are developed: a scale mixture of normal distributions with respect to an α-stable random variable; a mixture of Bartlett–Fejer kernels (or triangle densities) with respect to a two-component mixture of gamma random variables. Both lead to Markov chain Monte Carlo methods for posterior simulation, and these methods turn out to have complementary domains of maximum efficiency. The first representation is a well-known result due to West and is the better choice for collinear design matrices. The second representation is new and is more efficient for orthogonal problems, largely because it avoids the need to deal with exponentially tilted stable random variables. It also provides insight into the multimodality of the joint posterior distribution, which is a feature of the bridge model that is notably absent under ridge or lasso-type priors. We prove a theorem that extends this representation to a wider class of densities representable as scale mixtures of beta distributions, and we provide an explicit inversion formula for the mixing distribution. The connections with slice sampling and scale mixtures of normal distributions are explored. On the practical side, 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 show that the Markov chain Monte Carlo algorithm for fitting the bridge model exhibits excellent mixing properties, particularly for the global scale parameter. This makes for a favourable contrast with analogous Markov chain Monte Carlo algorithms for other sparse Bayesian models. All methods described in this paper are implemented in the R package BayesBridge. An extensive set of simulation results is provided in two on-line supplemental files.

Keywords: Bayesian methods; Bridge estimator; Data augmentation; Prior distributions; Sparsity

1. Introduction

1.1. Penalized likelihood and the Bayesian bridge

This paper develops the Bayesian analogue of the bridge estimator in regression, where $y = X\beta + \varepsilon$ for unknown $\beta = (\beta_1, \ldots, \beta_p)'$. Given $\alpha \in (0, 1]$ 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

Address for correspondence: James G. Scott, Department of Information, Risk and Operations Management, University of Texas at Austin, 1 University Station B6500, Austin, TX 78712, USA.
E-mail: James.Scott@mccombs.utexas.edu

 1369–7412/14/76000

can be found in Frank and Friedman (1993), with recent papers focusing on model selection asymptotics and on computational strategies for fitting the model (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. Specifically, we treat $p(\beta|y) \propto \exp\{-Q_y(\beta)\}$ as a posterior distribution having the minimizer of equation (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):

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

Rather than minimizing equation (1), we sample from the joint posterior distribution for β and the model hyperparameters.

1.2. Relationship with previous work

Our paper emphasizes several interesting features of the Bayesian approach to estimating the bridge regression model.

1.2.1. Comparison with 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 that is 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 estimator satisfies the oracle property (Fan and Li, 2001; Huang *et al.*, 2008). Although this property *per se* is of no 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)), such that

$$\lim_{y \to \pm\infty} \frac{\mathrm{d}}{\mathrm{d}y} \log\{m(y)\} = 0, \qquad m(y) = \int_{\Theta} p(y|\theta)\, p(\theta)\, \mathrm{d}\theta.$$

This property is desirable in sparse situations, as it avoids the overshrinkage of large coefficients even in the presence of sparsity (Polson and Scott, 2011a; Griffin and Brown, 2010).

1.2.2. Comparison with classical bridge estimator

Both the classical and the Bayesian approaches to bridge estimation must confront a significant practical difficulty: exploring and summarizing a multimodal surface in high dimensional Euclidean space. We argue that this feature of the problem recommends 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) reported computational difficulties with local modes in attempting to minimize equation (1). Our sampling-based approach, although not immune to the problem of local modes, seems effective at exploring the whole likelihood surface. As Section 2 will show, there are good reasons for expecting this to be so, based on the structure of the data augmentation strategy that we

pursue. In this respect, Markov chain Monte Carlo (MCMC) sampling behaves like a simulated annealing algorithm that never cools.

In addition, previous researchers have emphasized three other points about penalized likelihood rules that will echo in the examples of Section 5. First, one must choose a penalty parameter ν. Doing so via cross-validation, as is common practice, ignores uncertainty in the penalty parameter. The Bayesian approach is to average over uncertainty in the posterior distribution, under some default prior for the global variance component τ^2 (e.g. Gelman (2006)). 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 equation (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, we wish 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) gave realistic examples where the 'Bayesian lasso' outperforms its classical counterpart, both in prediction and in estimation. Similar conclusions were 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 posterior distribution in the Bayesian model, marginal over τ and σ. In the cases that we study, the Bayesian approach leads to a better estimator.

Third, a fully Bayesian approach can often lead to substantive conclusions that are different from a traditional penalized likelihood analysis, particularly regarding which components of β are important in predicting y. For example, Hans (2010) produced several examples where the classical lasso estimator aggressively zeros 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 diabetes data set (see, for example, Efron *et al.* (2004)). 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.

1.2.3. Comparison with 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 favourable 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.* (2013).

In most of these models, the primary difficulty is the mixing rate of the MCMC algorithm that is 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 introduced by the latent variables is large. Section 3.3 of Hans (2009) contains an informative discussion of this point. We have also included an on-line supplement to the paper that extensively documents the poor mixing behaviour of Gibbs samplers within this realm.

In light of these difficulties, the empirically observed mixing rate of our MCMC approach is a pleasant surprise. For example, Fig. 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-

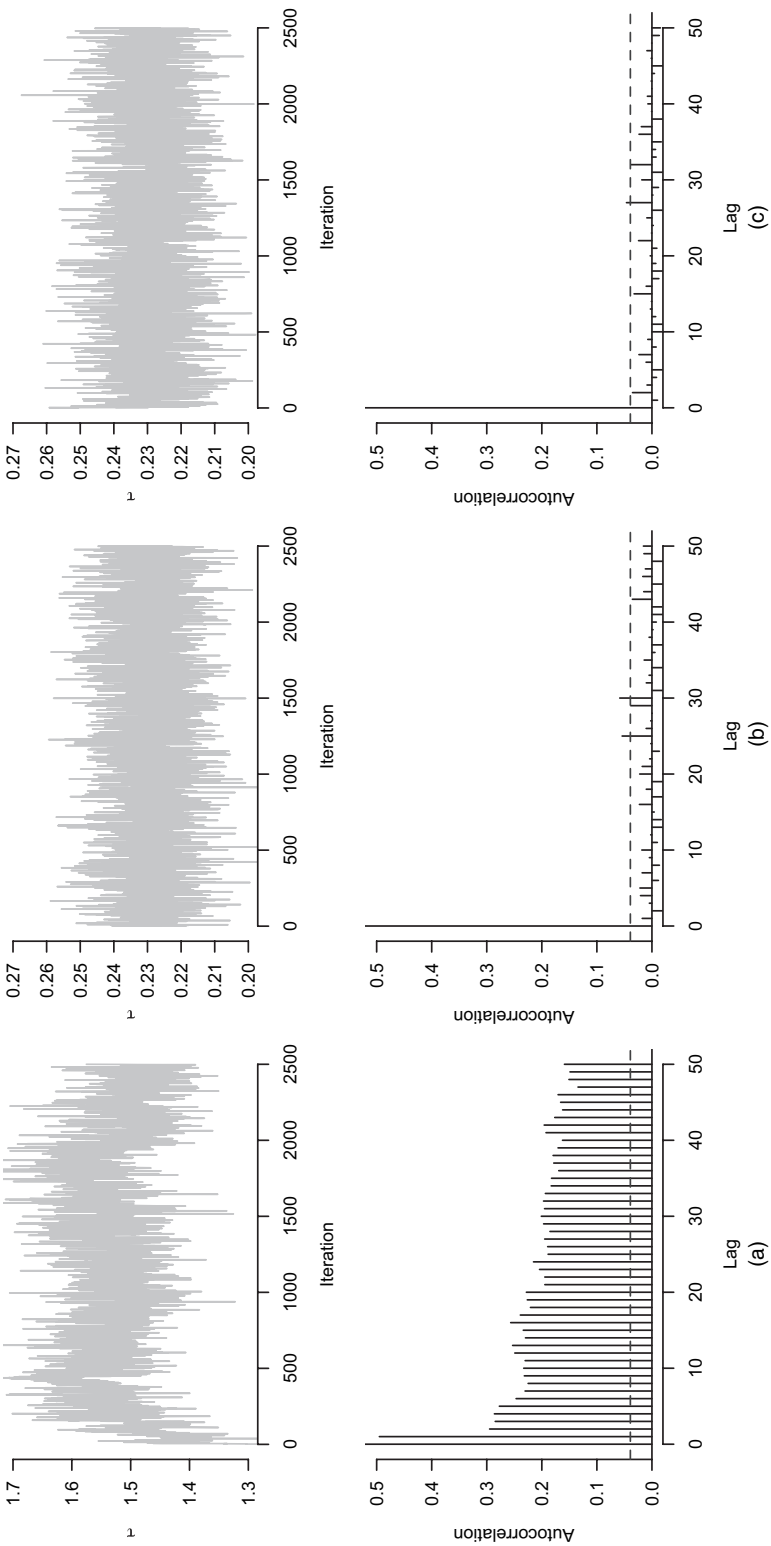

Fig. 1. Comparison of the simulation histories for τ, the global scale parameter, by using MCMC sampling for the bridge and the horseshoe on a 1000-dimensional orthogonal regression problem with $n = 1100$ observations (there were 100 non-zero entries in β simulated from a t_4-distribution, and 900 0s; 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): (a) horseshoe (with parameter expansion): (b) bridge (using Bartlett–Fejer kernels); (c) bridge (using normal mixtures)

variable orthogonal regression problem with 900 zero entries in β. (See the on-line supplement for details.) The plots show the first 2500 iterations of the sampler, starting from $\tau = 1$. Autocorrelation under the horseshoe MCMC sampler is much more severe. (Though these results are not shown here equally large differences emerge when comparing the simulation histories of the local scale parameters under each method.)

1.3. Computational approach

We would argue that the Bayesian bridge model is an interesting object for study on the basis of all three of these comparisons. There are, however, corresponding disadvantages. In particular, posterior inference for the Bayesian bridge is more challenging than in most other Bayesian models of this type, where MCMC sampling relies on representing the implied prior distribution for β_j as a scale mixture of normal distributions. The exponential power prior in equation (2) is known to lie within the normal–scale mixture class (West, 1987). Yet the mixing distribution that arises in the conditional posterior is that of an exponentially tilted α-stable random variable. This complicates matters, owing to the lack of a closed form expression for the density function. This fact was recognized by Armagan (2009), who proposed the use of 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, 3 and 4. The R package `BayesBridge`, which is freely available on line (Windle *et al.*, 2012), implements all methods and experiments that are 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 quite complicated, but seems robust, and leads to generally good performance (see the empirical results in the on-line supplement). Given current technology, it appears to be the best method for sampling the bridge model when the design matrix exhibits strong collinearity.

We also propose a second approach that is more efficient than the mixture-of-normals MCMC method 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 result based on the Schoenberg–Williamson theorem for n-monotone densities:

$$(y|\beta, \sigma^2) \sim N(X\beta, \sigma^2 I),$$

$$p(\beta_j|\tau, \omega_j, \alpha) = \frac{1}{\tau\omega_j^{1/\alpha}} \left\{ 1 - \left| \frac{\beta_j}{\tau\omega_j^{1/\alpha}} \right| \right\}_+, \tag{3}$$

$$(\omega_j|\alpha) \sim \frac{1+\alpha}{2} \mathrm{Ga}\left(2 + \frac{1}{\alpha}, 1\right) + \frac{1-\alpha}{2} \mathrm{Ga}\left(1 + \frac{1}{\alpha}, 1\right). \tag{4}$$

This scale mixture of triangles, or Bartlett–Fejer kernels, recovers $\exp\{-Q_y(\beta)\}$ as the marginal posterior in β. The mixing distribution is depicted in Fig. 2 and explained in detail in Section 2.2. It leads to a simple MCMC algorithm that avoids the need to deal with α-stable distributions and can 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.

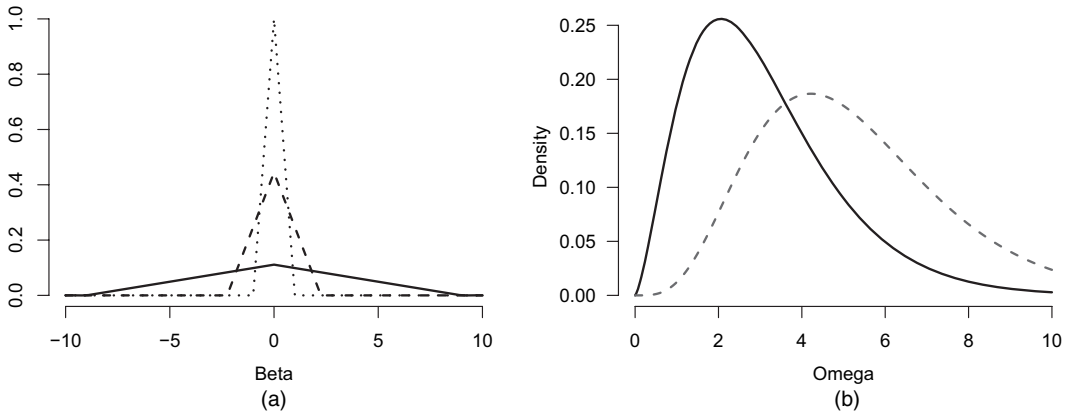

Fig. 2. (a) Triangular densities, or normalized Bartlett–Fejer kernels, of different widths ($\alpha = 0.5$) (· · · · · ·, $\omega = 1.0$; – – –, $\omega = 1.5$; ———, $\omega = 3.0$) and (b) two examples of mixing distributions for ω_j that give rise to exponential power marginals for β_j in conjunction with the Bartlett–Fejer kernel (———, $\alpha = 0.75$; – – –, $\alpha = 0.25$)

2. Data augmentation for the bridge model

2.1. As a scale mixture of normal distributions

We begin by discussing the two different data augmentation strategies that facilitate posterior inference for the Bayesian bridge model. The mixture-of-normals representation is well known (West, 1987). It arises from 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 $G(\lambda)$:

$$f(x) = \int_0^\infty \exp(-sx)\,\mathrm{d}G(s). \tag{5}$$

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

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

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

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 that is implied by equations (1) and (6):

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

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 equation (6). The conditional posterior of λ_j given β_j is then an exponentially tilted

stable random variable,

$$p(\lambda_j|\beta_j) = \frac{\exp(-\nu^{2/\alpha}|\beta_j|^2\lambda_j)\, p(\lambda_j)}{\mathbb{E}\{\exp(-\nu^{2/\alpha}|\beta_j|^2\lambda_j)\}},$$

with the expectation in the denominator taken over the prior. Neither the prior nor posterior for λ_j is known in closed form and can 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 normal distributions by evaluating the right-hand side of equation (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 alternative data augmentation scheme. Specifically, consider the class of symmetric density functions $f(x)$ that are n monotone on $(0, \infty)$ for some integer n, i.e. $(-1)^k f^{(k)}(|x|) \geqslant 0$ 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 beta distributions, and that we may invert for the mixing distribution by using the derivatives of f.

Theorem 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 $C = \{2\int_0^\infty f(t)\,\mathrm{d}t\}^{-1}$ 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, $1 \leqslant k \leqslant n$:

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

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

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

Importantly, the mixing density in the k-monotone case has finitely many 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.

Return now to the Bayesian bridge model. The exponential power density for $0 < \alpha \leqslant 1$ is completely monotone on the positive real numbers, and therefore any value of k may be used in equation (8). We focus on the choice $k = 2$. The kernel functions that arise here 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 studied them in connection with the theory of characteristic functions of Polya type. Thus we refer to them as Bartlett–Fejer kernels.

Corollary 1. 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 $C = \{2\int_0^\infty f(t)\,\mathrm{d}t\}^{-1}$ denote the normalizing constant that makes $f(x)$ a density on the real line. Then f is the following mixture

of Bartlett–Fejer kernels:

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

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

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

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

Further algebra yields a properly normalized mixture of Bartlett–Fejer kernels:

$$\frac{\alpha}{2\tau\Gamma(1 + 1/\alpha)} \exp\left(- \left| \frac{\beta}{\tau} \right|^\alpha \right) = \int_0^\infty \frac{1}{\tau\omega^{1/\alpha}} \left\{ 1 - \left| \frac{\beta}{\tau\omega^{1/\alpha}} \right| \right\}_+ p(\omega|\alpha)\, \mathrm{d}\omega$$
$$p(\omega|\alpha) = \frac{1 + \alpha}{2} c_1 \omega^{1+1/\alpha} \exp(-\omega) + \frac{1 - \alpha}{2} c_2 \omega^{1/\alpha} \exp(-\omega).$$

This is a two-component mixture of gamma distributions, 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.

3. Connection with other latent variable methods

The latent variable scheme that is suggested by theorem 1 was originally motivated by the potential inefficiencies of working with exponentially tilted stable random variables and does lead to improvements in the orthogonal case. Here we focus on some potentially interesting features of the representation in its own right, apart from its application to the bridge model.

The analogy with the auxiliary variable method of Damien *et al.* (1999) is instructive. In both cases, the 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 $0 \leqslant u < p(\theta)$, and sample from the joint distribution

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

where $\mathbb{I}(\cdot)$ is the indicator function. The posterior of interest is simply the marginal distribution for θ.

The difficulty is that, given u, we need to be able to calculate the slice region where $p(\theta) > u$. In our data augmentation approach, the analogous inversion problem is already done: it reduces to the set where $|\theta| < \omega$. For example, in the once monotone case, we have

$$\pi(\theta, \omega) = \mathbb{I}(|\theta| < \omega)\, g(\omega)\, L(\theta)/Z,$$

where ω now plays a role similar to that of u. We have removed the problem of inverting a slice region, at the cost of generating ω from a non-uniform distribution $g(\omega)$, which is uniquely identified by theorem 1. Of course, the question of which method leads to simpler calculations will depend on context. We simply point out that there are many cases where inverting a slice region is non-trivial. See Damien *et al.* (1999), Roberts and Rosenthal (1999) or Neal (2003).

Moreover, the once monotone case is just one example of the wider family. It is well known that the efficiency of a latent variable scheme is inversely related to the amount of information about θ that is conveyed by the latent variable. Therefore, all else being equal, one should mimic the target density as closely as possible when choosing a kernel over which to mix. Theorem 1 gives the designer of sampling algorithms wide latitude in this regard. For example, in the twice monotone case, we have

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

For the Bayesian bridge model, this representation is attractive, in that the bimodality of each marginal posterior for β_j is nicely matched by the fact that $g(\omega_j)$ has two distinct components (Fig. 3).

Moving beyond the once monotone case does pose trade-offs. The mixing distribution $g(\omega)$ potentially becomes more complicated, and we must now sample from the tilted distribution with density proportional to $(1 - |\theta|/\omega)^{k-1} L(\theta)$. The difficulty of this step will also depend on context. In the twice monotone representation of the bridge model, it turns out to be straightforward. 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

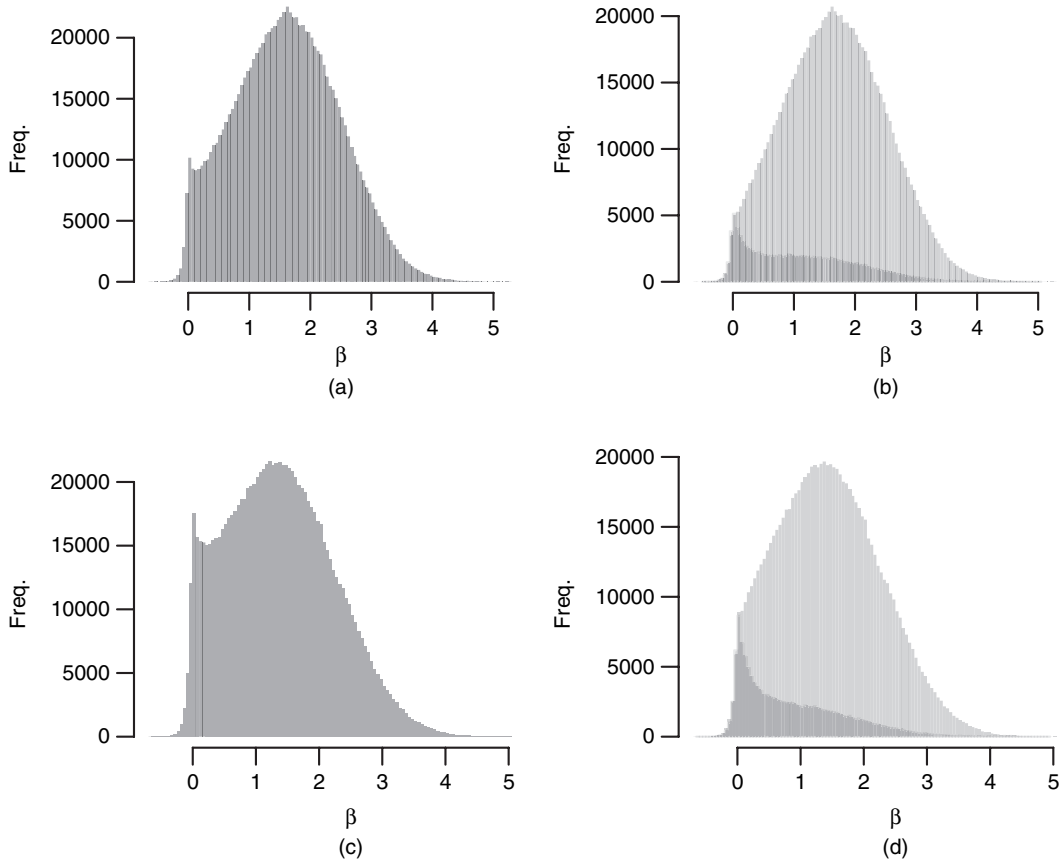

Fig. 3. (a) Marginal draws and (b) draws stratified by the mixture component of the most recent draw for ω_2 in posterior draws for β_2 in the simulated example, and (c), (d) the same as (a) and (b) respectively for β_3, which in this case facilitate the identification of both modes

another density. Other general strategies for sampling tilted densities were alluded to in Devroye (2009).

There is a further connection between our result and the Gaussian mixture representation, which parallels the relationship between the Schoenberg–Williamson and Bernstein theorems. To see this, let $u = k/s$. Observe that we obtain the completely monotonic case as k diverges:

$$f(x) \propto \int_0^\infty \left(1 - \frac{ux}{k}\right)_+^{k-1} \mathrm{d}P(u)$$
$$\to \int_0^\infty \exp(-sx)\, \mathrm{d}\tilde{P}(s)$$

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 normal distributions as a limiting case of a scale mixture of beta distributions. The inversion formula, also, is similar. In particular, for the case of the exponential power kernel, we have

$$\exp(-|x|^\alpha) = \int_0^\infty \exp(-xs)\, g(s)\, \mathrm{d}s \qquad g(s) = \sum_{j=1}^\infty (-1)^j \frac{s^{-j\alpha-1}}{j!\,\Gamma(-\alpha j)},$$

which parallels the expression given in theorem 1.

As a final side point, we observe that there may be further cases where the new approach could prove fruitful. One such example is the type I extreme value distribution, $p(x) = \exp\{-x - \exp(-x)\}$. From corollary 1, we have

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

and therefore $\exp\{\exp(-x)\}$ can be written in terms of a gamma mixing measure:

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

One other approach to slice sampling, which is not pursued here, is the single-variable slice sampler of Neal (2003). This method avoids the difficulty of inverting a slice region at the cost of introducing additional tuning parameters. Our method, in contrast, also avoids the inversion of a slice region, but requires no tuning parameters; the corresponding cost comes in the form of the additional analytical work in identifying the conditional distribution with density proportional to $(1 - |\theta|/\omega)^{k-1} L(\theta)$.

Rather than inverting the slice region explicitly, the single-variable method instead constructs an interval around the current parameter value that contains all or much of the slice, and then samples from the part of this interval containing the slice region. This can be done in many ways that leave the uniform distribution over the slice region invariant, with the two recommended ways referred to as the 'stepping-out' and 'doubling' methods. Each method requires the choice of a tuning parameter w_j that controls how quickly the proposed interval around the current value of β_j is allowed to grow. Neal (2003) described the fundamental tension that arises in the choice of w_j:

> 'We would like this interval to contain as much of the slice as is feasible, so as to allow the new point to differ as much as possible from the old point, but we would also like to avoid intervals that are much larger than the slice, as this will make the subsequent sampling step less efficient'.

The strength of this method is that it can lead to very efficient samplers that require no inversion of a slice region. The weakness is that it requires p tuning parameters for the 'stepping-out'

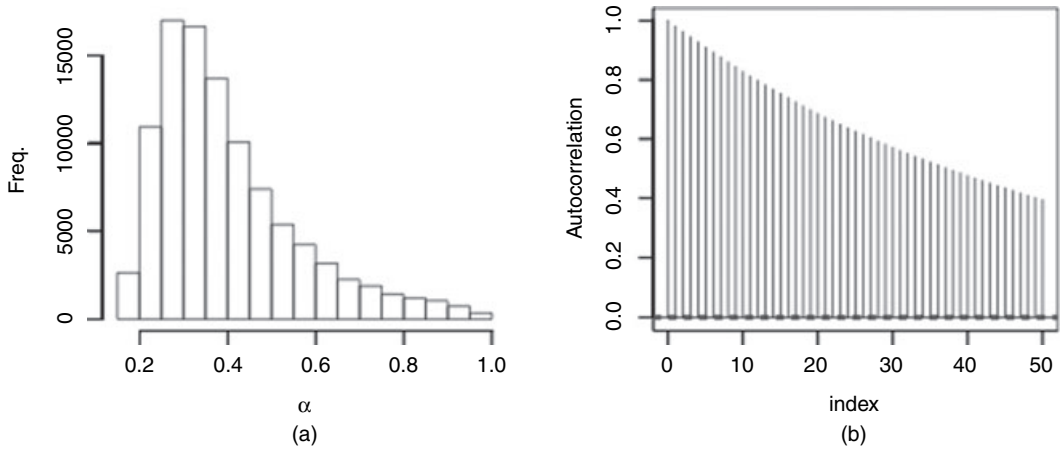

Fig. 4. (a) Posterior distribution of the concavity parameter α for the diabetes data, using a random-walk Metropolis sampler, and (b) auto-correlation function of the posterior draws

part of the algorithm to be maximally efficient in each dimension. In cases where the parameter space is unbounded, these may be difficult to tune optimally. In contrast, updating the concavity parameter α involves a draw of a scalar random variable on a bounded interval, for which the slice sampler is ideally suited. We use the random-walk Metropolis sampler in the next section to update this parameter, as it is adequate for the problem that we consider (Fig. 4). But the slice sampler would probably improve matters for the reasons described in Neal (2003) and is recommended for larger problems.

4. Markov chain Monte Carlo sampling for the Bayesian bridge

4.1. Overview of approach

For sampling the Bayesian bridge posterior, we recommend a hybrid computational approach, which we have implemented as the default setting in the `BayesBridge` R package. When the design matrix X exhibits strong collinearity, the normal–scale mixture representation is the better choice. In data sets involving many higher order interaction terms, the efficiency advantage can be substantial. However, when the design matrix is orthogonal, the Bartlett–Fejer representation usually leads to an effective sampling rate that is roughly 2–3 times that of the Gaussian method. Finally, the representation is relevant to cases where one wishes to model a system with an exponential power likelihood, where conditional independence of each observation will mimic the orthogonal case in regression analysis. See, for example, Godsill (2000) for an application in acoustics.

Most of the evidence supporting this recommendation is detailed in an on-line supplemental file. But Table 1 briefly summarizes the absolute and relative speed of the algorithms in two benchmark scenarios:

(a) the diabetes data set with the original design matrix and all pairwise interactions (64 predictors) and
(b) the Boston house price data set with an orthogonalized design matrix (13 predictors).

We measure the efficiency of each algorithm by using the effective sampling rate, which is defined as the effective sample size per second of run time. The effective sample size approximates the number of independent samples that are needed to mimic the performance of the MCMC ergodic average, as measured by the standard error of the sample mean. The effective sampling rate is

Table 1. Summary of performance for the two MCMC strategies in two design matrix scenarios†

Method	*Results for scenario 1 (strong collinearity)*			*Results for scenario 2 (orthogonal)*		
	Time (s)	*Median effective sampling rate*	*Minimum effective sampling rate*	*Time (s)*	*Median effective sampling rate*	*Minimum effective sampling rate*
Triangle	45.1	15	5	1.6	49559	36737
Normal	33.5	1816	536	5.5	17533	13048

†Scenario 1: diabetes data with all pairwise interactions and the design matrix on the original scale. Scenario 2: Boston housing data with an orthogonalized design matrix.

different for each parameter; we report the minimum and the median across all parameters in the model. See the on-line supplement for further details.

The first scenario is highly unfavourable to the mixture-of-triangles approach: the design matrix is highly collinear, and the inefficiency of sampling a constrained multivariate normal distribution is correspondingly severe. In this case, the normal mixture approach is roughly 10 times more efficient. The second scenario is much more favourable to the mixture-of-triangles approach: the design matrix is orthogonal, and there is no longer a need to sample a high dimensional constrained multivariate normal distribution. In this case, the triangle mixture approach is roughly three times more efficient. These broad trends held across other examples. Thus we recommend the triangle mixture approach for orthogonal problems—including principal component regression, basis expansions and when using generalized g-priors (Polson and Scott, 2012)—and the normal mixture approach for other problems.

4.2. Sampling β and the latent variables

We use the method that was described in Devroye (2009) for sampling exponentially tilted α-stable random variables. With this capability in place, it is easy to use equation (7) to generate posterior draws, appealing to standard multivariate normal theory. Thus we omit a long discussion of the normal mixture method and focus on the mixture-of-betas approach.

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

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

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

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

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.

Applying corollary 1, if we marginalize out both the slice variables and the latent ω_js, we recover the Bayesian bridge posterior distribution,

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

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

$$|\beta_j| \leqslant \tau^{-1}(1-u_j)\omega_j^{1/\alpha} = b_j,$$
$$\omega_j \geqslant \left(\frac{|\beta_j/\tau|}{1-u_j}\right)^{\alpha} = a_j.$$

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

Step 1: generate $(u_j|\beta_j, \omega_j) \sim \text{Unif}(0, 1 - |\beta_j/\tau|\omega_j^{-1/\alpha})$.
Step 2: generate each ω_j from a mixture of truncated gamma distributions as described below.
Step 3: update β from a truncated multivariate normal distribution proportional to

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

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

We explored several methods for simulating from the truncated multivariate normal distribution, ultimately settling on the proposal of Rodriguez-Yam *et al.* (2004) as the most efficient. As this method involves running over the individual coefficients, it is important to observe that β cannot be regenerated at each step, but merely updated componentwise. The conditional posterior of the latent ω_js can be determined as follows. Suppressing subscripts for the moment, we may write the compute conditional for ω as

$$p(\omega|\alpha) = \alpha\{\omega \exp(-\omega)\} + (1-\alpha)\exp(-\omega),$$
$$p(\omega|a, \alpha) = C_a[\alpha\{\omega \exp(-\omega)\} + (1-\alpha)\exp(-\omega)\}\, \mathbb{I}(\omega \geqslant a),$$

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

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

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

This is a mixture of gamma distributions, where

$$(\bar{\omega}|a) \sim \begin{cases} \Gamma(1,1) & \text{with probability } \{1-\alpha(1-a)\}/(1+\alpha a), \\ \Gamma(2,1) & \text{with probability } \alpha/(1+\alpha a). \end{cases}$$

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

This representation has two interesting and intuitive features. First, the full conditional for β in step 3 is centred at the usual least squares estimate $\hat{\beta}$. Only the truncations (b_j) change at each step, which eliminates matrix operations.

Second, the mixture-of-gammas form of $p(\omega)$ naturally accounts for the bimodality in the marginal posterior distribution, $p(\beta_j|y) = \int p(\beta_j|\omega, y)\, p(\omega_j|y)\, \mathrm{d}\omega_j$. In some cases, in fact, each mixture component of the conditional for ω_j represents a distinct mode of the marginal posterior for β_j. As Fig. 3 shows, this endows the algorithm with the ability to explore the multiple modes of the joint posterior. These plots come from a simulated data set with $p = 20$ and $n = 200$, with the predictors having pairwise positive correlation of 0.99. We set $\tau = 0.1$ and $\alpha = 0.85$ and ran the mixture-of-triangles MCMC algorithm for 200000 iterations. (These settings were arbitrary but identified the effect most clearly out of the several that we tried.) Fig. 3 shows the posterior draws for two of the coefficients (β_2 and β_3), stratified by the mixture component for the most recent draw for the corresponding ω_j. In the case of β_3, the stratification helps to identify a mode which might easily be missed in a histogram of draws from the marginal posterior.

4.3. Sampling hyperparameters

To update the global scale parameter τ, we work directly with the exponential power density, marginalizing out the latent variables $\{\omega_j, u_j\}$. This is a crucial source of efficiency in the bridge MCMC sampler and leads to the favourable mixing that is evident in Fig. 1. From equation (1), observe that the posterior for $\nu \equiv \tau^{-\alpha}$, given β, is conditionally independent of y and takes the form

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

Therefore if ν has a gamma(c, d) prior, its conditional posterior will also be a gamma distribution, with hyperparameters $c^* = c + p/\alpha$ and $d^* = d + \Sigma_{j=1}^{p} |\beta_j|^{\alpha}$. 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 Fig. 1 from Section 1 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 also can be given a prior $p(\alpha)$, most conveniently from the beta family, and can be updated by using a random-walk Metropolis sampler. Fig. 4 shows the results of using this sampler under a uniform prior for α in the well-known data set on blood glucose levels in diabetes patients, which is available in the R package `lars` (Efron *et al.*, 2004). There are only 10 predictors for this problem, and therefore significant uncertainty about the value of α. It is interesting, however, how much the posterior is pulled away from $\alpha = 1$, which corresponds to the lasso prior.

Similar MCMC algorithms can be used to fit Bayesian analogues of bridge-penalized logistic regression and quantile regression:

$$\hat{\beta}_{\mathrm{LR}} = \arg\min_{\beta \in \mathbb{R}^p} \left[\sum_{i=1}^{n} \log\{1 + \exp(-y_i x_i' \beta)\} + \sum_{j=1}^{p} |\beta_j/\tau|^{\alpha} \right],$$

$$\hat{\beta}_{\mathrm{QR}} = \arg\min_{\beta \in \mathbb{R}^p} \left[\sum_{i=1}^{n} \{|y_i - x_i'\beta| + (2q-1)(y_i - x_i'\beta)\} + \sum_{j=1}^{p} |\beta_j/\tau|^{\alpha} \right],$$

where binary outcomes are encoded as ± 1, or where $q \in (0, 1)$ is the desired quantile. To fit these estimators within a Bayesian framework, one introduces a second set of latent variables corresponding to the n individual terms in the likelihood. This allows the quantile regression and logit

likelihoods to be represented as mixtures of Gaussian models with respect to known mixing measures (Polson *et al.*, 2012).

5. Examples

5.1. Diabetes data

We first explore the Bayesian bridge estimator by using the diabetes data, which have already been described. We fit the Bayesian bridge using 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 the responses were centred, whereas the predictors were also rescaled to have unit variance. At each step of the MCMC algorithm for the Bayesian model, we calculated the conditional posterior density for each β_j at a discrete grid of values. It is striking that, even for this relatively information rich problem (10 predictors and 442 observations), significant differences emerge between the Bayesian and classical methods.

Fig. 5 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. For reference, the results of a step-down model fit (starting from the full model) using the Akaike information criterion are also shown. 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 that are associated with the taurocholic acid and glucose predictors (and, to a lesser extent, for the high density lipoprotein and female predictors). In none of these cases does it seem satisfactory to summarize information about β_j by using only a single number, as the classical solution forces us to do.

Second, 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 and shown in Fig. 5(k). 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 be small. Consider the posterior distributions for the total cholesterol and low density lipoprotein coefficients. These two columns of the design matrix have a sample correlation of -0.897. The Bayesian solution concentrates in a region of $\mathbb{R}^p$ where neither of these coefficients exerts a large effect. The classical solution, in contrast, leaves both predictors in the model with large coefficients of opposite sign.

It is surprising that such a marked difference would arise between the full Bayes mode and the classical mode, and that this difference would alter one's substantive conclusions about two predictors out of 10. The full Bayes posterior mean is different yet again. Clearly an important role here is played by the decision of whether, and how, to account for uncertainty in τ and σ.

5.2. Out-of-sample prediction results

Next, we describe the results from three out-of-sample prediction exercises involving three benchmark data sets. First, we used the Boston housing data, which are 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. Second, we used the data on ozone concentration, which are available in the R package `mlbench`. The goal is to predict the concentration of ozone in the atmosphere above Los Angeles by using various environmental covariates. As covariates, we used the nine original predictors, plus all interactions and squared terms for quantitative

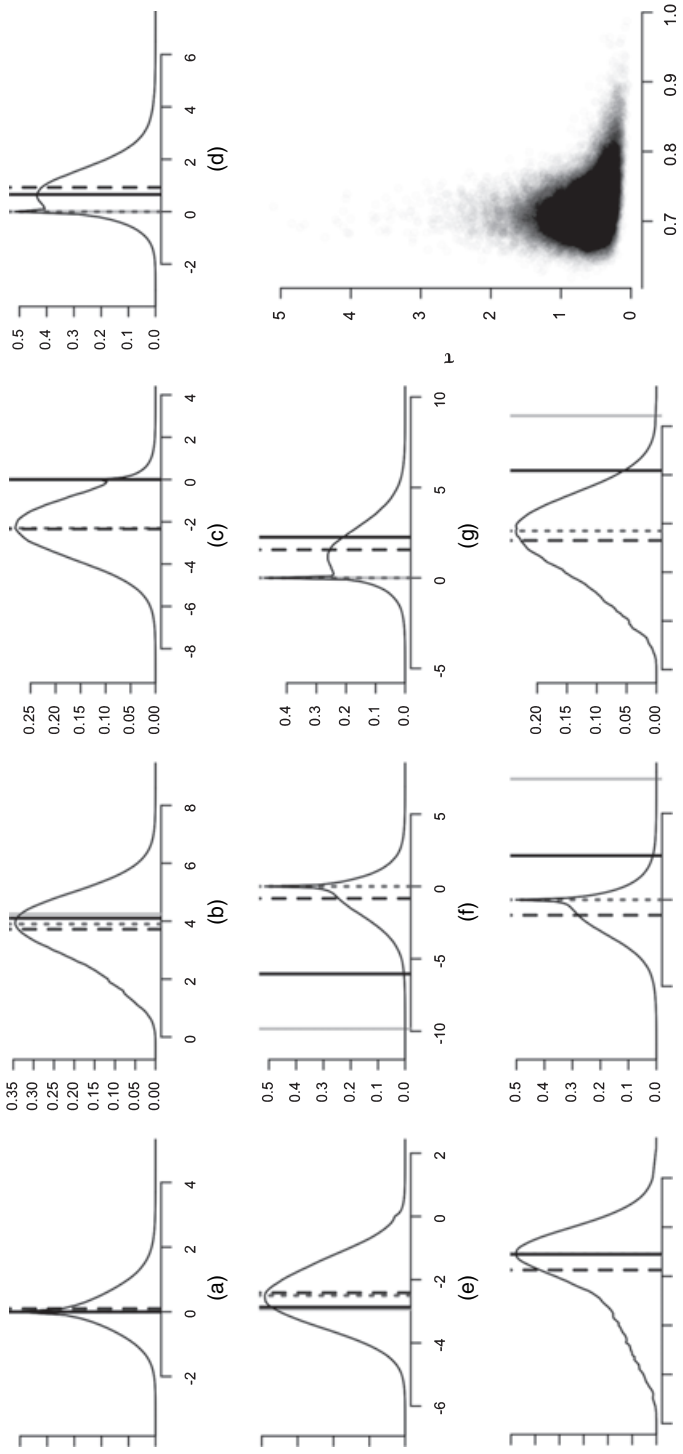

Fig. 5. Marginal posterior densities for the marginal effects of 10 predictors in the diabetes data (|, penalized likelihood solution with ν chosen by generalized cross-validation; $|$, result of stepwise Akaike information criterion selection starting from the full model; $|$, marginal posterior mean for β_j; $\vdots$, mode of the marginal distribution for β_j under the fully Bayes posterior) (all predictors were standardized): (a) age; (b) blood pressure; (c) high density lipoprotein; (d) glucose; (e) female; (f) total cholesterol; (g) taurocholic acid; (h) body mass index; (i) low density lipoprotein; (j) triglyceride level; (k) joint posterior distribution of the scale components τ and σ

predictors. Finally, we used the near infrared glucose data, which are available in the R package `chemometrics`. The goal is to predict the concentration of glucose in molecules by using data from near infrared 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 by using least squares, the classical bridge (using the EM algorithm) and the Bayesian bridge posterior mean by using our MCMC method based on stable mixtures. In all cases we centred and standardized the predictors and centred the response. For the classical bridge estimator, the regularization parameter ν was chosen by generalized cross-validation, whereas, for the Bayesian bridge, σ was assigned a Jeffreys prior and ν a default gamma(2,2) prior. We used two different settings for α in the Bayesian bridge: one with α fixed at a default setting of 0.5, and another with α estimated by using a random-walk Metropolis step. R scripts implementing all of these experiments are included as an on-line supplemental file.

We measured performance of each method by computing the sum of squared errors in predicting y on the test data set. The results are in Fig. 6. In all three cases, the posterior mean estimator

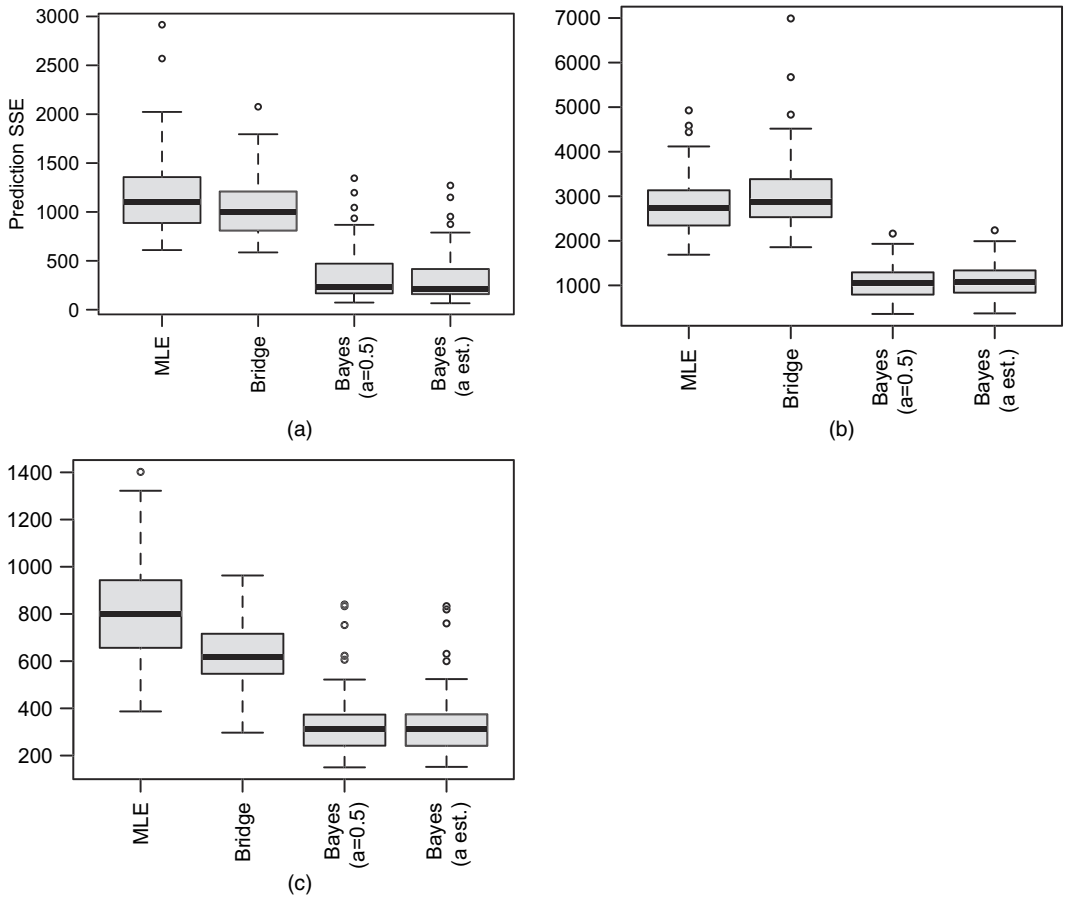

Fig. 6. Boxplots of the sum of squared errors in prediction hold-out data by using four methods for estimating β (maximum likelihood estimation, MLE, the classical bridge with $\alpha = 0.5$ and ν chosen by generalized cross-validation, the Bayesian bridge with $\alpha = 0.5$ and the Bayesian bridge with α estimated under a uniform prior): (a) Boston housing data; (b) near infrared glucose data; (c) ozone data

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

α	*Results for the following methods:*		
	Least squares estimation	*Bridge*	*Bayes bridge*
0.5	2254	1611	99
0.7	1994	406	225
0.9	551	144	85

outperforms both least squares and the classical bridge estimator, and the fixed choice of $\alpha = 0.5$ nearly matched the performance of the model which marginalized over a uniform prior for α.

5.3. 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

(a) simulating regression coefficients from the exponential power distribution for the given choice of α,
(b) simulating correlated design matrices X and
(c) 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 multicollinearity among the columns of each simulated X.

For each simulated data set we again estimated β by 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 Monte Carlo precision. As before, for the classical bridge estimator, the regularization parameter ν was chosen by generalized cross-validation. For the Bayesian bridge, σ was assigned a 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.

6. Discussion

This paper has described a series of tools that allow practitioners to estimate the full joint distribution of regression coefficients under the Bayesian bridge model. Our numerical experiments have shown

(a) that the classical bridge solution, the posterior mode under a joint Bayesian model and the posterior mean can often lead to different summaries about the relative importance of different predictors and
(b) 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 (2008) 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. The evidence suggests that it is the best algorithm in the orthogonal case, but that it suffers from poor mixing when the design matrix is strongly collinear. The chief limiting factor here is our inability to generate samples efficiently from the truncated multivariate normal distribution. Luckily, in this case, the normal mixture method based on the work of Devroye (2009) for sampling exponentially tilted stable random variables performs well. Moreover, both MCMC methods appear to alleviate some of the difficulties that are associated with slow mixing in global–local scale–mixture models described by Hans (2009) and further studied in the on-line supplemental file. They also allow uncertainty about the concavity parameter to be incorporated naturally. Both are implemented in the R package `BayesBridge`, which is available through the Comprehensive R Archive Network. Together, they give practitioners a set of tools for efficiently exploring the bridge model across a wide range of commonly encountered situations.

Acknowledgements

The authors thank two referees, the Associate Editor and the Joint Editor for their helpful feedback, which has greatly improved the manuscript.

Appendix A: Proof of theorem 1

Let M_n denote the class of n-times monotone functions on $(0, \infty)$. Clearly, for $n \geqslant 2$, $f \in M_n \Rightarrow f \in M_{n-1}$. 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), theorems 1 and 3) states that a necessary and sufficient condition for a function $f(x)$ defined on $(0, \infty)$ to be in M_n is that

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

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

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

Let $s = 1/u$. This yields

$$f(x) = \int_0^\infty (1 - x/s)_+^{n-1} \, \mathrm{d}G(s),$$

$$G(s) = \sum_{j=0}^{n-1} \frac{(-1)^j f^{(j)}(s)}{j!} s^j.$$

Now rewrite the kernel as a scaled beta density to give

$$f(x) = \int_0^\infty \frac{1}{s} n\Big(1 - \frac{x}{s}\Big)_+^{n-1} \frac{s}{n}\, \mathrm{d}G(s).$$

Differentiating the cumulative density function with respect to s and absorbing the factor of s/n into $G(s)$, we conclude that the mixing density is

$$g(s)\, \mathrm{d}s \propto n^{-1} \sum_{j=0}^{n-1} \frac{(-1)^j}{j!} \{ j s^j f^{(j)}(s) + s^{j+1} f^{(j+1)}(s) \}\, \mathrm{d}s,$$

and the result is proven.

References

Armagan, A. (2009) Variational bridge regression. *J. Mach. Learn. Res.*, **5**, 17–24.

Armagan, A., Dunson, D. and Lee, J. (2013) Generalized double Pareto shrinkage. *Statist. Sin.*, **23**, 119–143.

Bae, K. and Mallick, B. (2004) Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, **20**, 3423–3430.

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

Caron, F. and Doucet, A. (2008) Sparse Bayesian nonparametric regression. In *Proc. 25th Int. Conf. Machine Learning*, pp. 88–95. Helsinki: Association for Computing Machinery.

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010) The horseshoe estimator for sparse signals. *Biometrika*, **97**, 465–480.

Damien, P., Wakefield, J. and Walker, S. (1999) Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. R. Statist. Soc.* B, **61**, 331–344.

Devroye, L. (1996) Random variate generation in one line of code. In *Proc. 1996 Winter Simulation Conf.* (eds J. Charnes, D. Morrice, D. Brunner and J. Swain), pp. 265–272. Washington DC: Institute of Electrical and Electronics Engineers Computer Society.

Devroye, L. (2009) On exact simulation algorithms for some distributions related to Jacobi theta functions. *Statist. Probab. Lett.*, **79**, 2251–2259.

Dugué, D. and Girault, M. (1955) Fonctions convexes de Polya. *Publ. Inst. Statist. Univ. Par.*, **4**, 3–10.

Efron, B. (2009) Empirical Bayes estimates for large-scale prediction problems. *J. Am. Statist. Ass.*, **104**, 1015–1028.

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

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Figueiredo, M. (2003) Adaptive sparseness for supervised learning. *IEEE Trans. Pattn Anal. Mach. Intell.*, **25**, 1150–1159.

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

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Baysn Anal.*, **1**, 515–533.

Godsill, S. (2000) Inference in symmetric alpha-stable noise using MCMC and the slice sampler. *Acoust. Spch Signal Process.*, **6**, 3806–3809.

Griffin, J. and Brown, P. (2010) Inference with normal-gamma prior distributions in regression problems. *Baysn Anal.*, **5**, 171–188.

Griffin, J. and Brown, P. (2012) Alternative prior distributions for variable selection with very many more variables than observations. *Aust. New Zeal. J. Statist.*, to be published .

Hans, C. M. (2009) Bayesian lasso regression. *Biometrika*, **96**, 835–845.

Hans, C. M. (2010) Model uncertainty and variable selection in Bayesian lasso regression. *Statist. Comput.*, **20**, 221–229.

Hans, C. M. (2011) Elastic net regression modeling with the orthant normal prior. *J. Am. Statist. Ass.*, **106**, 1383–1393.

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

Li, Q. and Lin, N. (2010) The Bayesian elastic net. *Baysn Anal.*, **5**, 151–170.

Lindley, D. V. and Smith, A. F. M. (1972) Bayes estimates for the linear model (with discussion). *J. R. Statist. Soc.* B, **34**, 1–41.

Mazumder, R., Friedman, J. and Hastie, T. (2011) Sparsenet: coordinate descent with non-convex penalties. *J. Am. Statist. Ass.*, **106**, 1125–1138.

Neal, R. M. (2003) Slice sampling. *Ann. Statist.*, **31**, 705–767.

Park, T. and Casella, G. (2008) The Bayesian lasso. *J. Am. Statist. Ass.*, **103**, 681–686.

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

Polson, N. G. and Scott, J. G. (2011a) Shrink globally, act locally: sparse Bayesian regularization and prediction (with discussion). In *Proc. 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), pp. 501–538. Oxford: Oxford University Press.

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

Polson, N. G. and Scott, J. G. (2012) Local shrinkage rules, Lévy processes and regularized regression. *J. R. Statist. Soc.* B, **74**, 287–311.

Polson, N. G., Scott, J. G. and Windle, J. (2013) Bayesian inference for logistic models using Polya-Gamma latent variables. *J. Am. Statist. Ass.*, to be published, doi 10.1080/01621459.2013.829001.

Roberts, G. O. and Rosenthal, J. S. (1999) Convergence of slice sampler Markov chains. *J. R. Statist. Soc.* B, **61**, 643–660.

Rodriguez-Yam, G., Davis, R. A. and Scharf, L. L. (2004) Efficient Gibbs sampling of truncated multivariate normal with application to constrained linear regression. *Technical Report*. Colorado State University, Fort Collins.

Stein, W. E. and Keblis, M. F. (2009) A new method to simulate the triangular distribution. *Math. Comput. Modlng.*, **49**, 1143–1147.

Tipping, M. (2001) Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, **1**, 211–244.

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

Williamson, R. (1956) Multiply monotone functions and their Laplace transforms. *Duke Math. J.*, **23**, 189–207.

Windle, J., Scott, J. G. and Polson, N. G. (2012) BayesBridge: Bayesian bridge regression. *R Package Version 0.2-1*. (Available from `http://cran.r-project.org/package=BayesBridge`.)

Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, **36**, 1509–1533.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Benchmarking the two MCMC strategies for sampling the Bayesian bridge posterior distribution'

and

'An empirical study of mixing rates in parameter-expanded Gibbs samplers for sparse Bayesian regression models'.