The Bayesian Bridge

Nicholas G. Polson

University of Chicago

James G. Scott

University of Texas at Austin

First Version: July 2011
This Version: August 2011

Abstract

We develop the Bayesian bridge estimator for regularized regression and classification. We focus on two key 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). We show how it can be exploited to yield an efficient, parallelizable EM algorithm for calculating the joint posterior mode—even under certain non-Gaussian likelihoods, such as those that arise in logistic or quantile regression. The second representation is new. Its main virtue is that avoids the need to deal with exponentially tilted stable random variables, and therefore leads to a much simpler MCMC scheme than the scale mixture representation. It also provides insight into the multimodality of the joint posterior distribution, which is a notable feature of the Bayesian bridge model that is absent under more-traditional ridge or lasso-type priors (Park and Casella, 2008; Hans, 2009). Finally, we show how our approach can be extended to a wider class of non-convex regularization penalties based on Polya-type characteristic functions, including the double-Pareto model of Armagan et al. (2010). We compare the performance of the Bayesian bridge model to its classical cousin across a variety of data sets, both simulated and real.

*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.
1 Introduction

1.1 Penalized likelihood and the Bayesian bridge

This paper studies the Bayesian analogue of the classical bridge estimator for regression and classification problems. The bridge is most easily understood in the multiple regression model, where

\[ y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I), \]  

for \( y = (y_1, \ldots, y_n)' \) and some unknown vector of coefficients \( \beta = (\beta_1, \ldots, \beta_p)' \). To compute the bridge estimator, solve for \( \hat{\beta} \) as

\[ \hat{\beta}_{(\alpha, \nu)} = \arg \min Q(\beta), \quad \text{where} \quad Q(\beta) = \frac{1}{2}||y - X\beta||^2 + \nu \sum_{j=1}^{p} |\beta_j|^\alpha \]  

for specific choices of \( \alpha \in (0, 1) \) and regularization parameter \( \nu \in \mathbb{R}^+ \). The name comes from the fact that the objective function in (2) bridges an entire class of concave penalties, with the combinatorially hard \( \ell^0 \) (or best-subset-selection) penalty at one end, and the convex \( \ell^1 \) (or lasso) penalty at the other.

As many previous authors have pointed out, the bridge naturally yield sparse estimates for \( \beta \), in the sense that some components of \( \hat{\beta}_{(\alpha, \nu)} \) may be explicitly zeroed out. The solution to (2) thus simultaneously accomplishes the goals of estimation and model selection. Most of the recent work on this estimator concerns either model-selection asymptotics or computation—this latter issue being especially important, in light of the well-documented challenges posed by non-convex optimization problems like (2). Useful references on these topics include the papers by [Frank and Friedman (1993), Huang et al. (2008), and Mazumder et al. (2009)].

Our paper differs from this line of work in adopting a Bayesian approach to bridge estimation. Specifically, we treat \( p(\beta \mid y) \propto e^{-Q(\beta)} \) as a posterior probability distribution which has the solution to (2) as its global mode. This posterior arises in assuming a Gaussian likelihood for \( y \), as in (1), along with a prior for \( \beta \) that decomposes as a product of exponential-power priors (Box and Tiao, 1973):

\[ p(\beta \mid \alpha, \nu) \propto \prod_{j=1}^{p} \exp(-|\beta_j/\tau|^\alpha), \quad \tau = \nu^{-1/\alpha}. \]  

In pursuing this approach, we follow along the same lines as well-known Bayesian treatments of other classical regularization penalties, such as the ridge (Lindley and Smith, 1972), lasso (Park and Casella, 2008; Hans, 2009, 2010), and elastic net (Li and Lin, 2010; Hans, 2011). Yet there are several unique features of the Bayesian-bridge model that distinguish our paper from these other parallel Bayesian treatments of...
classical penalty functions. We are especially interested in two of these features, one statistical and the other practical.

First, from a statistical standpoint, the most interesting feature of the bridge penalty is its concavity. This is both a blessing and a curse. When the underlying signal is sparse, and when some further regularity conditions are met, the bridge penalty—precisely by virtue of its concavity—is known to dominate the lasso and ridge penalties when judged by 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: namely, the property of yielding a redescending score function for the predictive distribution of $y$ (Pericchi and Smith, 1992). This property is a blessing, in the sense that it avoids overshrinkage of large regression coefficients even in the presence of extreme sparsity (Polson and Scott, 2011a). Yet it is also a curse, in that it produces multimodality—and all its attendant computational difficulties—in the joint posterior distribution for $\beta$. Indeed, we will argue that the multimodality of the objective function in (2) is one of the strongest arguments for pursuing a fully Bayes approach to bridge regression, for the simple reason that it can be misleading to summarize a highly multimodal surface in terms of a single point estimate, no matter how appealingly sparse that estimate may be.

Second, from a practical standpoint, sampling from the full posterior under the Bayesian bridge model is very challenging—much more so than in the ridge, lasso, or elastic net models. MCMC sampling in these models, and in many similar ones (e.g. Griffin and Brown, 2010) relies upon representing the implied prior distribution for $\beta_j$ as a scale mixture of normals. The exponential-power prior in (3) is known to lie within this class of models (West, 1987). Yet the mixing distribution that arises in the conditional posterior is that of an exponentially tilted alpha-stable random variable. This is notoriously difficult to handle, owing to the lack of a closed-form expression for the density function. The same MCMC approach that works for the lasso is therefore quite difficult to apply in the case of the bridge. This fact was recognized by Armagan (2009), who proposed using variational methods to perform approximate Bayesian inference while sidestepping the challenges associated with stable distributions.

1.2 Goals of the paper

In dealing with these challenges, this paper makes three specific contributions to the literature on Bayesian regularized regression.
1. We prove the following novel mixture representation for the Bayesian bridge:

\[
(y \mid \beta, \sigma^2) \sim N(X\beta, \sigma^2 I) \\
p(\beta_j \mid \tau, \omega_j, \alpha) = \frac{1}{\tau \omega_j^{1/\alpha}} \cdot \left\{ 1 - \left| \frac{\beta_j}{\tau \omega_j^{1/\alpha}} \right| \right\} _+ \\
(\omega_j \mid \alpha) \sim \frac{1+\alpha}{2} \cdot \text{Ga}(2 + 1/\alpha, 1) + \frac{1-\alpha}{2} \cdot \text{Ga}(1 + 1/\alpha, 1).
\]

This representation, which recovers \(e^{-Q(\beta)}\) as the marginal posterior in \(\beta\), is depicted in Figure 1 and explained in detail in Section 3. It is based not upon the theory of normal scale mixtures, but upon mixtures of Bartlett–Fejer kernels (which are special cases of triangle densities). It leads to a simple MCMC that avoids the need to deal with alpha-stable distributions, and—since it involves directly working with a bimodal mixing distribution—is capable of easily and naturally hopping between modes in the joint posterior. This solves one of the major outstanding difficulties of working with the bridge objective function. We also extend this representation to encompass a wider class of non-convex regularization penalties based on Polya-type characteristic functions, including recent proposals based on Pareto-type models (Armagan et al., 2010).

2. We use the Lévy representation of Polson and Scott (2011b) to derive an EM algorithm based on Krylov-subspace methods for calculating the posterior mode.
under the Bayesian bridge model. This algorithm employs a block updating scheme—a sharp and useful departure from traditional coordinate-descent algorithms, which are known to encounter difficulties in highly multi-modal situations. Moreover, the algorithm is easily parallelized for use in ultra-high-dimensional data sets (see Algorithm 4).

3. We extend this EM algorithm to provide estimates of the posterior mode under certain non-Gaussian likelihoods, including bridge logistic regression and bridge quantile regression.

1.3 Relationship with previous work

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. We have already mentioned the Bayesian ridge, lasso, and elastic net models, described in papers by Lindley and Smith (1972), Park and Casella (2008), Li and Lin (2010), and Hans (2009, 2010, 2011). Several others in this class include 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 (2005); the normal/gamma and normal/inverse-Gaussian (Caron and Doucet, 2008; Griffin and Brown, 2010); the horseshoe prior of Carvalho et al. (2010); the hypergeometric inverted–beta model of Polson and Scott (2010); and the double-Pareto model of Armagan et al. (2010).

Since the bridge penalty is also a normal scale-mixture prior, our work naturally fits within this larger body of literature on Bayesian regularized regression. Yet our paper relies far less upon the theory of normal scale mixtures, as in the treatment of Armagan (2009), and instead uses the Bartlett-kernel mixture of Equation (4) and (5) in deriving an MCMC sampling scheme for the encompassing model. This difference will be elaborated further in Section 3.

These authors repeatedly emphasize two points that will echo in the examples we present in Section 6. First, the solution to the penalized-likelihood problem in (2) may produce a sparse estimator, but this estimate is provably suboptimal (in a Bayes-risk sense) with respect to many traditional loss functions. If, for example, one wishes either to estimate $\beta$ or to predict future values of $y$ under squared-error loss, then the optimal choice is to use the posterior mean, usually estimated via MCMC, in lieu of the mode. This conclusion is no mere tempest in a theoretical teapot. Indeed, both Park and Casella (2008) and Hans (2009) give realistic examples where the lasso posterior mean significantly outperforms the posterior mode, 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 mode and the mode of the joint distribution in the Bayesian model, marginal over over $\tau$
and σ. (A collection of R scripts implementing all of our numerical experiments is available as a supplemental file.)

Second, a fully Bayesian approach can often lead to very different 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 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.

2 The usual scale-mixture representation

The traditional construction of the exponential-power prior involves the following mixture:

\[ k \exp(-|t|^\alpha) = \int_{0}^{\infty} \phi(t \mid 0, \lambda^{-1}) \lambda^{-1/2} f_{\alpha/2}(\lambda), \]  

where \( k \) is a normalization constant; \( \phi(z \mid m, v) \) is the normal density function with mean \( m \) and variance \( v \); and \( f_{b}(t) \) is the density of a positive alpha-stable random variable with index of stability \( b \). [West (1987)] proves the result when \( 1 < \alpha < 2 \), while [Polson and Scott (2011b)] appeal to an argument involving moment-generating functions of Lévy processes to extend the result for all \( 0 < \alpha < 2 \). This involves treating the unnormalized density directly as the moment-generating function of a subordinator (or strictly increasing Lévy process), evaluated at \( t^2/2 \):

\[ e^{-|t|^\alpha} = \mathbb{E}(e^{-\lambda t^2/2}), \]  

where the expectation is with respect to \( f_{\alpha/2}(\lambda) \). This avoids the extra term of \( \lambda^{-1/2} \) and greatly simplifies the analysis.

Similar mixture representations have been exploited to yield simple MCMC algorithms for a wide variety of models, including many of those referenced in the introduction. [Armagan (2009)] also uses (6) to derive a variational-EM approach for approximate Bayesian analysis.

Unlike these other models, however, the mixture representation for the bridge is less than hospitable to full Bayes analysis via MCMC. To see the difficulty, observe
that (7) leads to the following latent-variable representation of the joint posterior:

\[ p(\beta, \Lambda | y) = C \exp \left( -\nu^{2/\alpha} \beta' \Lambda \beta - \frac{1}{2\sigma^2} \beta' X' X \beta + \beta' \sigma^{-2} \beta' X' y \right) \prod_{j=1}^{p} f_{\alpha/2}(\lambda_j) \]

\[ = C \exp \left\{ -\frac{1}{2} \beta' \left( \sigma^{-2} X' X + 2\nu^{2/\alpha} \Lambda \right) \beta + \beta' \sigma^{-2} \beta' X' y \right\} \prod_{j=1}^{p} f_{\alpha/2}(\lambda_j), \tag{8} \]

where \( \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_j) \). The conditional posterior of \( \lambda_j \) given \( \beta_j \) is then an exponentially tilted stable random variable, leading to a Gibbs sampler for \( p(\beta, \Lambda | y) \) involving two steps.

1. Generate \( p(\beta | \Lambda, y) \). This is a multivariate normal draw with

\[ (\beta | \Lambda, \nu, y) \sim \mathcal{N} \left( (\sigma^{-2} X' X + 2\nu^{2/\alpha} \Lambda)^{-1} \cdot \sigma^{-2} X' y, \sigma^2(\sigma^{-2} X' X + 2\nu^{2/\alpha} \Lambda)^{-1} \right). \]

2. Generate \( p(\Lambda | \beta, \nu, y) \) component-wise, with each component having an exponentially tilted stable conditional distribution:

\[ p(\lambda_j | \beta_j) = \frac{e^{-\nu^{2/\alpha} |\beta_j|^2 \lambda_j} f_{\alpha/2}(\lambda_j)}{\mathbb{E} \left( e^{-\nu^{2/\alpha} |\beta_j|^2 \lambda_j} \right)}. \]

The difficulty here is that neither the prior density \( f_{\alpha/2}(\lambda_j) \), nor the posterior density just given, are known in closed form, and can be only be given explicitly in terms of integral or series representations. Any efficient sampling scheme must therefore avoid evaluating both the prior and posterior density functions for \( \lambda_j \).

Godsill (2000) studied a similar problem and gives three options for sampling from this class of distributions. The first two use the alpha-stable prior \( f_{\alpha/2}(\lambda_j) \) as a proposal distribution within a rejection sampler and a Metropolis-Hastings sampler, respectively. The third option involves an alternative accept/reject algorithm within an overall slice sampler. Unfortunately, as Godsill (2000) writes, all three methods have significant practical limitations: “The problem is that all three schemes rely on sampling from the prior distribution . . . which may not be well attuned to the posterior distribution itself. Hence rejection rates and convergence rates can be very poor in application to both real and simulated data.” The paper goes on to recommend an alternative approach involving a power-series expansions of the tails of the stable distribution for use in practical applications. But this must also be tuned to different configurations of the prior vis-a-vis the region where the likelihood is largest. Devroye (1996) describes an alternative method for sampling from exponentially tilted stables. But we have encountered difficulties in implementing this method, since it requires evaluating the hyperbolic cosine function for sometimes-extreme arguments, leading to numerical instability.
Despite these difficulties, the Lévy representation of the bridge penalty is still very useful: as we will show, it leads to fast, efficient EM algorithms for calculating posterior modes. These algorithms do not require evaluating the density of a stable random variable; offer a parallelizable block updating scheme (unlike traditional coordinate-descent methods); and can be extended to many non-Gaussian likelihoods. This class of algorithms is described in detail in Section 5.

3 Mixtures of Bartlett–Fejer kernels

We now describe an alternative approach for sampling from the Bayesian-bridge posterior distribution. This approach involves a counter-intuitive step: abandoning the normal scale-mixture framework entirely, in favor of a representation that uses mixtures of Bartlett–Fejer kernels. Compared with the above approach involving stable distributions, the Bartlett-kernel approach has three crucial advantages within the context of MCMC sampling: it is simple to program; it requires no additional tuning; and it generalizes to a much wider class of models that can be sampled using a similar scheme. It therefore represents a promising new approach for generating regularization penalties with tractable Bayesian analogues. The representation is also highly intuitive, in that it allows one to see precisely how two salient features of the bridge posterior—its non-differentiable point at zero, and its multimodality—arise from the prior.

We now describe this approach in further detail. The following theorem describes a class of priors that can be represented as mixtures of Bartlett–Fejer kernels where the mixing distribution is explicitly known. It relies upon the useful fact that certain characteristic functions of Polya type can be reinterpreted as symmetric density functions on the real line.

**Theorem 3.1.** Let \( f(t) \) be a function that is symmetric about the origin; integrable, convex, and twice-differentiable on \((0, \infty)\); and for which \( f(0) = 1 \), and \( \lim_{t \to \infty} f(\beta) = 0 \). Let \( k = \left\{ 2 \int_0^\infty f(t) \, dt \right\}^{-1} \) denote the normalizing constant that makes \( f(t) \) a density on the real line. Then \( f \) is the following mixture of Bartlett–Fejer kernels:

\[
k f(t) = \int_0^\infty \frac{1}{s} \left\{ 1 - \left| \frac{t}{s} \right| \right\} a_+ s^2 f''(s) \, ds ,
\]

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

As an immediate corollary, we have the following mixture representation for the exponential power density.

**Corollary 3.2.** The exponential power density, \( \alpha \in (0, 1] \), is mixture of Bartlett–Fejer kernels.
Fejer kernels, where
\[
\frac{\alpha}{2\pi \Gamma(1/\alpha)} \exp(-|\beta/\tau|^\alpha) = \int_0^\infty \frac{1}{\tau \omega^{1/\alpha}} \left\{ 1 - \left| \frac{\beta}{\tau \omega^{1/\alpha}} \right| \right\}^+ \, p(\omega | \alpha) \, d\omega \quad (10)
\]

\[
p(\omega | \alpha) = \frac{1 + \alpha}{2} c_1 \omega^{1+1/\alpha} e^{-\omega} + \frac{1 - \alpha}{2} c_2 \omega^{1/\alpha} e^{-\omega}, \quad (11)
\]

where \(c_1\) and \(c_2\) are the normalizing constants of their respective gamma densities.

The Bayesian bridge prior can therefore be written using the simple two-component mixture of gamma random variables from Equations (4) and (5) and Figure 1. Note that this family includes the Bayesian lasso of Park and Casella (2008) and Hans (2009), which is seen to be a simple gamma mixture of triangle densities (since the second mixture component drops out when \(\alpha = 1\)).

4 MCMC sampling for the Bayesian bridge

4.1 Sampling \(\beta\) and the latent variables

To see why the representation in (4)–(5) leads to such a simple algorithm for posterior sampling, consider the joint distribution for \(\beta\) and the latent \(\omega_j\)'s:

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

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

\[
p(\beta, \Omega, u | \tau, y) \propto \exp \left( -\frac{1}{2\sigma^2} \beta'X'X\beta + \frac{1}{\sigma^2} \beta'X'y \right) \prod_{j=1}^p p(\omega_j | \alpha) \prod_{j=1}^p \left( 0 \leq u_j \leq 1 - \frac{|\beta_j|}{\tau \omega_j^{1/\alpha}} \right),
\]

where \(\mathbb{I}(\cdot)\) is the indicator function. 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 \(\omega_j\). This will make inverting the slice region for \(\omega_j\) far easier.

Applying Corollary 3.2, if we marginalize out both the slice variables and the latent \(\omega_j\)'s, we recover the Bayesian bridge posterior distribution,

\[
p(\beta | y) = C \exp \left( -\frac{1}{2\sigma^2} \|y - X\beta\|^2 - \sum_{j=1}^p |\beta_j/\tau|^\alpha \right) .
\]
We can invert the slice region in (13) by defining \((a_j, b_j)\) as

\[
|\beta_j| \leq \tau^{-1}(1-u_j)\omega_{j}^{1/\alpha} = b_j \quad \text{and} \quad \omega_j \geq \left(\frac{\tau|\beta_j|}{1-u_j}\right)^\alpha = a_j.
\]

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

1. Generate \((u_j | \beta_j, \omega_j) \sim \text{Unif}\left(0, 1 - \tau|\beta_j|\omega_{j}^{1/\alpha}\right)\).
2. Generate each \(\omega_j\) from a mixture of truncated gammas, as described below.
3. Generate \(\beta\) from a truncated multivariate normal proportional to

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

where \(\hat{\beta}\) indicates the least-squares estimate for \(\beta\). This step can be done component-wise in Gibbs fashion. See Devroye (1986), Geweke (1991), and Chopin (2011) for fast rejection-sampling schemes for one-dimensional truncated normals.

The conditional posterior of the latent \(\omega_j\)'s can be determined as follows. Suppressing subscripts for the moment, it is clear from (13) that

\[
p(\omega | \alpha) = \alpha(\omega e^{-\omega}) + (1-\alpha)e^{-\omega} \quad p(\omega | a, \alpha) = C_a \{\alpha(\omega e^{-\omega}) + (1-\alpha)e^{-\omega}\} \mathbb{I}(\omega \geq a),
\]

where \(a\) comes from inverting the slice region in (13) and \(C_a\) is an appropriate 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

\[
p(\bar{\omega} | a, \alpha) = C_a \{\alpha e^{-a}(a + \bar{\omega})e^{-\bar{\omega}} + (1-\alpha)e^{-a}e^{-\bar{\omega}}\} = \frac{\alpha}{1-\alpha a} \cdot \bar{\omega}e^{-\bar{\omega}} + \frac{1-\alpha}{1-\alpha a} \cdot e^{-\bar{\omega}}.
\]

This is a mixture of gammas, where

\[
(\bar{\omega} | a) \sim \begin{cases} 
\Gamma(1, 1) & \text{with prob } \frac{1-\alpha}{1-\alpha a} \\
\Gamma(2, 1) & \text{with prob } \frac{\alpha}{1-\alpha a}.
\end{cases}
\]

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 \(\beta\) in step 3 is centered at the usual least-squares estimate \(\hat{\beta}\). Only the
Algorithm 1: Gibbs sampling for Bayesian bridge regression

For variable $j = 1, \ldots, p$

1. Sample $(u_j \mid \beta_j, \omega_j) \sim \text{Unif} \left(0, 1 - \tau \mid \beta_j \mid \omega_j^{-1/\alpha} \right)$.

2. Set

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

3. Sample $\omega_j$ in two steps:
   (a) Sample

   $$(\bar{\omega}_j \mid a_j) \sim \left\{ \begin{array}{ll}
   \text{Ga}(1, 1) & \text{with prob } \frac{1 - a(1 + a_j)}{1 - aa_j} \\
   \text{Ga}(2, 1) & \text{with prob } \frac{a}{1 - aa_j}.
   \end{array} \right.$$ 

   (b) Set $\omega_j = \bar{\omega}_j + a_j$.

4. Generate $\beta$ from a truncated multivariate normal proportional to

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

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

Update hyperparameters $(\tau, \sigma, \alpha)$ as required.

Figure 2: A Gibbs-sampling algorithm for the Bayesian bridge model.
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 \(\omega_j\) represents a distinct mode of the marginal posterior for \(\beta_j\). As the examples later will show, this endows the algorithm with the ability to explore various modes of the joint posterior very easily.

We have summarized the algorithm in Figure 2.

### 4.2 Sampling hyperparameters

To update the global scale parameter \(\tau\), we work directly with the exponential-power density, marginalizing out the latent variables \(\{\omega_j, u_j\}\). From (2), observe that the posterior for \(\nu \equiv \tau^{-\alpha}\), given \(\beta\), is conditionally independent of \(y\), and takes the form

\[
p(\nu \mid \beta) \propto \nu^{p/\alpha} \exp\left(-\nu \sum_{j=1}^{p} |\beta_j|^\alpha\right) p(\nu).
\]

Therefore if \(\nu\) has a Gamma\((c, d)\) prior, its conditional posterior will also be a gamma distribution, with hyperparameters \(c^* = c + p/\alpha\) and \(d^* = d + \sum_{j=1}^{p} |\beta_j|^\alpha\). To sample \(\tau\), simply draw \(\nu\) from this gamma distribution, and use the transformation \(\tau = \nu^{-1/\alpha}\). Alternative priors for \(\nu\) can also be considered, in which case the gamma form of the conditional likelihood in \(\nu\) will make for a useful proposal distribution that closely approximates the posterior.

In many cases the concavity parameter \(\alpha\) will be fixed ahead of time to reflect a particular desired shape of the penalty function. But in principle it too can be given a prior \(p(\alpha)\). From (12), the conditional posterior in \(\alpha\) would then be

\[
p(\alpha \mid \Omega, \tau, \beta) \propto \prod_{j=1}^{p} \left\{1 - \left|\frac{\beta_j}{\tau \omega_j^{1/\alpha}}\right|\right\} p(\omega_j \mid \alpha) p(\alpha),
\]

which can be updated using, for example, a random-walk Metropolis sampler.

### 5 Computing MAP estimates

#### 5.1 In the usual regression problem

We now show—first for the usual regression problem, and then for logistic and quantile regression—that the mode of the marginal posterior distribution for \(\beta\) under the Bayesian bridge can be found using a simple EM algorithm, using the tilted, iteratively
reweighted least-squares (TIRLS) algorithm of Polson and Scott (2011c).

We represent the conditional posterior distribution for \( \beta \) in two different ways in order to derive the E and the M steps. First, begin with joint posterior distribution for \( \beta \) and \( \Lambda \) in the normal scale-mixture representation of (8). Taking the natural log and factorizing as a function \( \beta \), we have

\[
\log p(\beta | \Lambda, \tau, y) = c_1(\Lambda, y, \tau) - \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i' \beta)^2 - \frac{1}{2 \tau^2} \sum_{j=1}^{k} \lambda_j \beta_j^2,
\]

where \( c_1 \) does not depend upon \( \beta \).

We treat the above expression as the complete-data log posterior, with \( \Lambda \) as missing data. The E step involves taking the conditional expectation of this expression, given some current guess \( \beta(g) \). Since the complete-data log posterior depends linearly upon each \( \lambda_j \),

\[
E\{\log p(\beta | \Lambda, \tau, y) | \beta(g)\} = -\frac{1}{2} \sum_{i=1}^{n} (y_i - x_i' \beta)^2 - \frac{1}{2 \tau^2} \sum_{j=1}^{k} \hat{\lambda}_j(g) \beta_j^2,
\]

where \( \hat{\lambda}_j(g) = E(\lambda_j | \beta_j(g), \tau) \) is the conditional moment of \( \lambda_j \).

To derive these conditional moments, recall from (7) that the exponential power density can be written as a normal scale mixture:

\[
p(\beta_j | \tau) = \int_0^{\infty} \phi(\beta_j | 0, \tau^2 \lambda_j^{-1}) p(\lambda_j) \ d\lambda_j,
\]

where \( p(\lambda_j) \) is a power-tilted stable distribution.

Since \( \phi \) is a normal kernel,

\[
\frac{\partial \phi(\beta_j | \tau^2 \lambda_j^{-1})}{\partial \beta_j} = -\frac{\beta_j \lambda_j}{\tau^2} \phi(\beta_j | 0, \tau^2 \lambda_j^{-1}).
\]

Differentiating under the integral sign, dividing through by \( p(\beta_j | \tau) \), and using the above identity for the inner function, we obtain the following expression:

\[
\frac{\partial}{\partial \beta_j} p(\beta_j | \tau) = -\frac{\beta_j}{\tau^2} E(\lambda_j | \beta_j(g), \tau).
\]

Therefore, the E step is simply to plug

\[
\hat{\lambda}_j^{(g)} = \frac{\tau^2}{\beta_j^{(g)}} \frac{\partial}{\partial \beta_j} p(\beta_j | \tau) = \alpha \tau^{2-\alpha} \beta_j^{(g)-2},
\]

where \( \alpha \) is a shape parameter.
Algorithm 2: EM for Bayesian bridge linear regression

Given a starting value $\beta^{(1)}$ and a threshold $\lambda_{max}$ (e.g. $10^9$):

For iteration $g = 1, 2, \ldots$

E Step: Set
$$\lambda_j^{(g)} := \alpha \tau^{2-\alpha} |\beta_j^{(g)}|^{\alpha-2}.$$ 

For all $\lambda_j > \lambda_{max}$, delete $\lambda_j$ and $\beta_j$ from the model, along with the $j$th column of $X$.

M Step: Solve $A^{(g)} \beta^{(g+1)} = b^{(g)}$, where
$$A^{(g)} := \tau^{-2} \Lambda^{(g)} + X'X$$
$$b^{(g)} := X'y$$
for $\Lambda^{(g)} = \text{diag}(\lambda_1^{(g)}, \ldots, \lambda_p^{(g)})$.

End when the sequence of estimates $\{\beta^{(1)}, \beta^{(2)}, \ldots\}$ has converged.

Figure 3: An EM algorithm for MAP estimation in the Bayesian bridge model.
into (14). Notice that it is unnecessary to know the explicit form of the mixing measure. Only the derivative of the negative log-prior for $\beta_j$ (that is, the exponential power kernel) is required.

To derive the M step, we return to (8). Collecting terms, we can represent the log posterior (up to an additive term not involving $\beta$) as a sum of quadratic forms:

$$
\log p(\beta \mid \Lambda, \tau, y) = -\frac{1}{2\sigma^2} (y - X\beta)' (y - X\beta) - \frac{1}{2\tau^2} \beta' \Lambda \beta.
$$

By using the expressions given in, for example, Lindley and Smith (1972), we arrive at the fact that the M step involves choosing $\beta^{(g+1)}$ as the solution to the following linear system:

$$(X'X + \tau^{-2} \Lambda^{(g)} ) \hat{\beta}^{(g+1)} = X' y.$$

We summarize the resulting EM algorithm in Figure 3. One caveat here is that for $\beta_j \approx 0$, the conditional moment in the E step diverges to infinity, and care must be taken. It is important to emphasize here that infinite values for $\hat{\lambda}_j$ do not reflect poor numerical behavior, but rather are the result of a normally functioning algorithm that has found a sparse solution. The numerical difficulty this poses can be easily handled by starting the algorithm from a value where $\beta$ has no zeros, and then removing $\beta_j$ from the model when it gets within a small numerical threshold of its prior mean (or equivalently, when $\lambda_j > \lambda_{\text{max}}$, e.g. $10^9$. This conveys the added benefit of hastening the matrix computations in the weighted least-squares ($M$) step. Although we have found this approach to work well in practice, it has the disadvantage that a variable cannot re-enter the model once it has been deleted. An alternate approach that avoids this difficulty involves the use of restricted least-squares; for details, see Section 3.2 of Polson and Scott (2011d).

### 5.2 In logistic and quantile regression

A similar EM algorithm can be used to find the MAP estimate for bridge-penalized logistic regression (LR) and quantile regression (QR):

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

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

where binary outcomes are encoded as $\pm 1$, and where $q \in (0, 1)$ is the desired quantile.

To compute these two estimators, we make use of the following identities concern-
Algorithm 3: EM for bridge quantile and logistic regression

Given a starting value $\beta^{(1)}$ and a threshold $\lambda_{\text{max}}$ (e.g. $10^9$):

For iteration $g = 1, 2, \ldots$

E Step: Set

$$
\lambda_j^{(g)} := \alpha \tau^{2-\alpha} |\beta_j^{(g)}|^{\alpha-2}
$$

$$
\gamma_i^{(g)} := \begin{cases} 
\frac{1}{y_i x_i^T \beta^{(g)}} \left\{ \frac{\exp[y_i x_i^T \beta^{(g)}]}{1+\exp[y_i x_i^T \beta^{(g)}]} - \frac{1}{2} \right\} & \text{(logit)} \\
|y_i - x_i^T \beta^{(g)}|^{-1} & \text{(quantile)}
\end{cases}
$$

For all $\lambda_j > \lambda_{\text{max}}$, delete $\lambda_j$ and $\beta_j$ from the model, along with the $j$th column of $X$.

M Step: Solve $A^{(g)} \beta^{(g+1)} = b^{(g)}$, where

$$
A^{(g)} := \begin{cases} 
\tau^{-2} \Lambda^{(g)} + X_i^T \Gamma^{(g)} X_i & \text{(logit)} \\
\tau^{-2} \Lambda^{(g)} + X' \Gamma^{(g)} X & \text{(quantile)}
\end{cases}
$$

$$
b^{(g)} := \begin{cases} 
0.5 X_i' \mathbf{1} & \text{(logit)} \\
X' \left( \Gamma^{(g)} y - [1 - 2q] \mathbf{1} \right) & \text{(quantile)}
\end{cases}
$$

where $\Lambda^{(g)} = \text{diag}(\lambda_1^{(g)}, \ldots, \lambda_p^{(g)})$ and $\Gamma^{(g)} = \text{diag}(\gamma_1^{(g)}, \ldots, \gamma_n^{(g)})$; recalling that binary responses are encoded as $y_i = \pm 1$.

End when the sequence of estimates $\{\beta^{(1)}, \beta^{(2)}, \ldots\}$ has converged.

Figure 4: MAP estimation in bridge quantile and logistic regression.
ing the improper limits of variance–mean Gaussian mixtures:

\[
a^{-1} \exp \left\{ -2c^{-1} \max(a \theta, 0) \right\} = \int_{0}^{\infty} \phi(\theta \mid -av, cv) \, dv
\]

(15)

\[
(1 + \exp\{\theta - \mu\})^{-1} = \int_{0}^{\infty} \phi(\theta \mid \mu - v/2, v) \, p_{pol}(v \mid 0, 1) \, dv,
\]

(16)

where \(\rho_{q}(\theta) = \frac{1}{2} |\theta| + (q - \frac{1}{2}) \theta\) is the check-loss function \cite{Johannes2009, Li2010}; and \(p_{pol}(v \mid 0, 1)\) is the improper limit of the density function of a Polya-distributed random variable. These identities, and the corresponding distributional theory, are described in detail by Polson and Scott \cite{Polson2011c}.

These identities allow us to represent both the logistic and quantile-regression objective functions as a conditionally Gaussian model, by introducing a second set of augmentation variables \(\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_n)\) corresponding to the \(n\) individual terms in the likelihood. Details of the argument are presented in Appendix B, with the algorithm itself summarized in Figure 4.

5.3 A conjugate-gradient variation

The bottleneck of Algorithms 2 and 3 lies in solving the linear system \(A^{(g)} \beta^{(g+1)} = b^{(g)}\) for \(\beta^{(g+1)}\), where \(A\) and \(b\) change at every step. Solving this system exactly, however, is inefficient; an approximate solution will suffice in the M step, as long as it leads to an improvement in the observed-data objective function compared to the previous iteration. This will ensure that the sequence \(\{Q(\beta^{(1)}), Q(\beta^{(2)}), \ldots\}\) is monotonically decreasing.

Such an approximate solution can be found using the conjugate-gradient algorithm; see, for example Golub and Van Loan \cite{Golub1996}. We outline this modification in Figure 5. The conjugate-gradient algorithm’s computational bottleneck involves a series matrix-vector multiplications, each of which is \(O(p^2)\). This offers two major advantages.

1. If the algorithm proceeds exactly \(p\) steps, then an exact solution (subject to floating-point error) will be produced in \(O(p^3)\) total operations. But typically far fewer than \(p\) steps are necessary to reach a good approximate solution that improves the observed-data objective function.

2. Many open-source libraries exist that implement parallel matrix-vector multiplication. This fact allows users to implement the algorithm quickly and painlessly in a multi-core processing environment, and can lead to a speed-up that is very nearly linear in the number of cores available. This compares favorably with coordinate-descent algorithms, which cannot be parallelized even in principle.
Algorithm 4: Bridge regression via conjugate gradient
(linear, logit, and quantile)

For iteration $g = 1, 2, \ldots$

E Step: For linear regression with Gaussian error, set $\gamma_i \equiv 1$ and update $\Lambda^{(g)}$ as in Algorithm 2. For logistic and quantile regression, update $\Gamma^{(g)}$ and $\Lambda^{(g)}$.

For all $\lambda_j > \lambda_{\text{max}}$, delete $\lambda_j$ and $\beta_j$ from the model.

M Step: First set $A^{(g)}$ and $b^{(g)}$ as in Algorithm 2 (linear regression) or Algorithm 3 (logistic or quantile regression). Then initialize the following inner-loop terms:

$$
\begin{align*}
\beta^{(g, 0)} &:= \beta^{(g-1)} \\
r^{(g, 0)} &:= b - A^{(g)} \beta^{(g, 0)} \\
d^{(g, 0)} &:= r^{(g)} \\
c^{(g, 0)} &:= A^{(g)} d^{(g, 0)} \\
\alpha^{(g, 0)} &:= \frac{r^{(g, 0)} r^{(g, 0)}}{d^{(g, 0)} c^{(g, 0)}} \\
\Delta^{(g, 0)} &:= \alpha^{(g, 0)} d^{(g, 0)}.
\end{align*}
$$

While $|\Delta^{(g, l)}| > \delta_{\text{min}}$, increment $l$ and set

$$
\begin{align*}
\beta^{(c, l)} &:= \beta^{(c, l-1)} + \Delta^{(g, l-1)} \\
r^{(g, l)} &:= r^{(g, l-1)} - \alpha^{(g, l-1)} c^{(g, l-1)} \\
\gamma^{(g, l)} &:= \frac{r^{(g, l)} r^{(g, l)}}{r^{(g, l-1)} r^{(g, l-1)}} \\
d^{(g, l)} &:= r^{(g, l)} + \gamma^{(g, l)} d^{(g, l-1)} \\
c^{(g, l)} &:= A^{(g)} d^{(g, l)} \\
\alpha^{(g, l)} &:= \frac{r^{(g, 0)} r^{(g, l)}}{d^{(g, l)} c^{(g, l)}} \\
\Delta^{(g, l)} &:= \alpha^{(g, l)} d^{(g, l)}.
\end{align*}
$$

Set $\beta^{(g)} = \beta^{(g, l)}$.

End when the sequence of estimates $\{\beta^{(1)}, \beta^{(2)}, \ldots\}$ has converged.

Figure 5: A conjugate gradient algorithm for bridge regression.
6 Examples

6.1 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 fit both the classical bridge (using Algorithm 2 and generalized cross validation) and the Bayesian bridge (using Algorithm 1 and a default Gamma(2,2) prior for \( \nu \)). 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 \( \beta_j \) at a discrete grid of values.

Figure 6 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 \( \beta_j \) 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 \( \tau \) and \( \sigma \), 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 be small. Observe, for example, the middle row in Figure 6 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 \( \tau \) and \( \sigma \).
Figure 6: Marginal posterior densities for the marginal effects of 10 predictors in the diabetes data. Solid red line: penalized-likelihood solution with $\nu$ chosen by generalized cross validation. Dashed blue line: marginal posterior mean for $\beta_j$. Dotted black line: mode of the marginal distribution for $\beta_j$ under the fully Bayes posterior.
Table 1: Average sum of squared errors in predicting hold-out observations for 100 different train/test splits on three real data sets.

| Data set          | n  | p  | train/test | Prediction SSE |
|-------------------|----|----|------------|----------------|
| Boston housing    | 506| 103| 422/84     | 1288           |
| Ozone             | 203| 54 | 163/40     | 872            |
| NIR glucose       | 40 | 166| 110/56     | 2791           |

6.2 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 \( \beta \) using least-squares, the classical bridge (using Algorithm 2), and the Bayesian-bridge posterior mean (using Algorithm 1). 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 \( \nu \) was chosen by generalized cross validation; while for the Bayesian bridge, \( \sigma \) was assigned Jeffreys’ prior and \( \nu \) a default Gamma(2,2) prior.

We measured performance of each method by computing the sum of squared errors in predictor \( y \) 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.

6.3 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
Table 2: Average sum of squared errors in estimating $\beta$ for three different batches of 250 simulated data sets.

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

coefficients from the exponential power distribution for the given choice of $\alpha$; (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 $100 \times 10$ 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 estimated $\beta$ with least squares, the classical bridge (using Algorithm 2), and the Bayesian bridge posterior mean (using Algorithm 1). Performance was assessed by the sum of squared errors in estimating the true value of $\beta$. 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 $\nu$ was chosen by generalized cross validation; while for the Bayesian bridge, $\sigma$ was assigned Jeffreys’ prior and $\nu$ a Gamma(2,2) prior.

Table 2 shows the results of these experiments. For all three choices of $\alpha$, 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.

7 Generalizations

Our method based on mixtures of Bartlett–Fejer kernels also leads to interesting extensions that encompass other known models. For example, the generalized Linnik $L_{a,b}$ distribution (see [Devroye, 1996]) has density defined by

$$p(\beta \mid a, b) = \frac{1}{(1 + |\beta|^a)^b}.$$  

The Linnik distribution corresponds to $b = 1$. This class of priors nests many special cases including the double Pareto (Armagan et al., 2010), meridian, and Cauchy. One
may simulate from the prior using the identity $\beta \stackrel{D}{=} X^c S_{\alpha,0}$ for $c = b/a$, where $S_{\alpha,0}$ is a symmetric stable and $X$ is a random variable with density $f(x) = e^{-|x|^b}/\Gamma(1 + b^{-1})$. This can also be generalized to the Mittag–Leffler distribution, where in lieu of $S_{\alpha,0}$ one instead uses $S_{\alpha,1} = CS_{\alpha,0}$, with $C$ an extra Cauchy-distributed scale factor.

To see the connection with Bartlett–Fejer kernels, consider the Linnik distribution with kernel $p(t) = (1 + |t|)^{-b}$. Applying Theorem 3.1 the CDF of the mixing distribution is

$$F(t) = 1 - p(t) + tp'(t) = 1 - \frac{(b - 1)t - 1}{(1 + t)^{1+b}}.$$

For the general case,

$$p(t) = \frac{1}{(1 + |t|^a)^b},$$

the conditions of the theorem are met, and with further algebra the mixing measure can be identified as

$$tp''(t) = \frac{abt^{a-1}}{(1 + t^a)^{b+2}} \{ (ab + 1)t^a + (1 - a) \}.$$ 

If we make the transformation $u = t^a$ we have the density

$$g(u) = \frac{ab}{(1 + u)^{b+2}} \{ (ab + 1)u + (1 - a) \}.$$ 

We simulate from this distribution in one line of code directly from the CDF. Indeed, we even know the CDF of the truncated version of this distribution, which is required in the slice sampler:

$$F(u) = 1 - \frac{(1 + ab)b^{-1}}{(1 + u)^b} + \frac{a}{(1 + u)^{b+1}}.$$ 

8 Discussion

This paper has demonstrated a series of results that allow practitioners to estimate both the posterior mode and the full joint distribution 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 $\beta$ or making predictions under squared-error loss. Both results parallel the findings of Park and Casella (2008) and Hans (2009) for the Bayesian lasso. We have also shown how the bridge penalty can be applied both to logistic and quantile regression via a simple EM algorithm. The conjugate-gradient version (Algorithm 4) is also easily parallelized, lending further scalability to the method.
The existence of a second, novel mixture representation for the Bayesian bridge is of particular interest, given the difficulties of the normal scale-mixture approach for this model. It leads to an efficient Gibbs-sampling scheme that—by virtue of working directly with a bimodal mixing measure for each latent scale \( \omega_j \)—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, for example, Hans (2009) and Scott (2010).

A full comparison of our approach with the traditional scale-mixture MCMC would require a fast, robust method for sampling power-tilted or exponentially tilted alpha-stable random variables across a wide variety of input parameters. Constructing such a method, and benchmarking it against the Bartlett-kernel mixture approach, remains an active question for future research.

A Proofs

A.1 Theorem 3.1

Proof. Any function \( f(t) \) meeting the stated conditions is a characteristic function of Polya type. Applying the theorem of Dugué and Girault (1955), we can represent \( f(t) \) as the characteristic function of a known mixture of Fejer–de le Vallee Poussin (FVP) random variables. Specifically, we have the following integral identity for \( f(t) \):

\[
f(t) = \int_0^\infty \left\{ 1 - \left| \frac{t}{s} \right| \right\}_+ dG(s),
\]

where \( G(s) = 1 - f(s) + sf''(s) \) (see also Devroye, 1984). Thus if \( f \) is twice differentiable on \( (0, \infty) \), the unnormalized mixing density is \( G'(s) = sf''(s) \) Thus

\[
kf(t) = \int_0^\infty \left\{ 1 - \left| \frac{t}{s} \right| \right\}_+ ks^2 f''(s) \, ds
\]

\[
= \int_0^\infty \frac{1}{s} \left\{ 1 - \left| \frac{t}{s} \right| \right\}_+ ks^2 f''(s) \, ds,
\]

leaving a mixture with respect to a properly normalized Bartlett–Fejer kernel. \( \square \)
A.2 Corollary 3.2

Proof. Transform \( s \rightarrow \omega \equiv s^\alpha \) and apply Theorem 3.1 to the kernel of the exponential-power density function. This yields

\[
\frac{1}{2\tau} \exp(-|\beta/\tau|) = \int_0^\infty \frac{1}{\tau} \left\{ 1 - \left| \frac{\beta}{\tau \omega^{1/\alpha}} \right| \right\} \cdot p(\omega \mid \alpha) \, d\omega
\]

\[
p(\omega \mid \alpha) = \alpha \omega e^{-\omega} + (1 - \alpha)e^{-\omega}.
\]

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

\[
\frac{\alpha}{2\tau \Gamma(1 + 1/\alpha)} \exp(-|\beta/\tau|) = \int_0^\infty \frac{1}{\tau \omega^{1/\alpha}} \left\{ 1 - \frac{\beta}{\tau \omega^{1/\alpha}} \right\} \cdot p(\omega \mid \alpha) \, d\omega
\]

\[
p(\omega \mid \alpha) = \frac{1 + \alpha}{2} c_1 \omega^{1 + 1/\alpha} e^{-\omega} + \frac{1 - \alpha}{2} c_2 \omega^{1/\alpha} e^{-\omega}.
\]

\[\square\]

B. EM for logistic and quantile regression

B.1 M Step

Both the logistic and quantile-regression objective functions can be expressed as pseudo posteriors of the form

\[
e^{-Q(\beta)} \propto p(\beta \mid \tau, y) \propto \exp\left\{ -\sum_{i=1}^n f(y_i, x_i' \beta) - \sum_{j=1}^p |\beta_j|/\tau \right\} \quad (17)
\]

\[
\propto \left\{ \prod_{i=1}^n p(z_i \mid x_i' \beta) \right\} \left\{ \prod_{j=1}^k p(\beta_j \mid \tau) \right\}
\]

\[
= p(z \mid \beta) \cdot p(\beta \mid \tau),
\]

where \( z_i = y_i - x_i' \beta \) for quantile regression, or \( z_i = y_i x_i' \beta \) for logistic regression (with the response \( y_i \) coded as \( \pm 1 \)). Moreover, both the pseudo-likelihood \( p(z_i \mid \beta) \) and the prior \( p(\beta_j \mid \tau) \) can be represented as variance-mean mixtures of normals, using the identities in (15) and (16):

\[
p(z_i \mid \beta) = \int_0^\infty \phi(z_i \mid \mu_z + \kappa z \gamma_i^{-1}, \sigma^2 \gamma_i^{-1}) \, dP(\gamma_i) \quad (18)
\]

\[
p(\beta_j \mid \tau) = \int_0^\infty \phi(\beta_j \mid 0, \tau^2 \lambda_j^{-1}) \, dP(\lambda_j). \quad (19)
\]
For logistic regression, $\mu_z = 0$ and $\kappa_z = 1/2$; while for quantile regression $\mu_z = 0$ and $\kappa_z = 1 - 2\cdot q$. (Other likelihoods, such as that corresponding to support-vector machines, can be encoded through different choices of these two constants.) Using this representation, write the complete-data log posterior distribution as

$$
\log p(\beta \mid \gamma, \lambda, \tau, y) = c_0(\gamma, \lambda, y, \tau) - \frac{1}{2} \sum_{i=1}^{n} \gamma_i \left(z_i - \mu_z - \kappa_y \gamma_i^{-1}\right)^2 \\
- \frac{1}{2\tau^2} \sum_{j=1}^{p} \lambda_j (\beta_j - \mu_\beta - \kappa_\beta \lambda_j^{-1})^2
$$

for some constant $c_0$.

First, we derive the M step for the quantile regression likelihood. Plugging in $z_i = y_i - x_i' \beta$ and collecting terms, we can write $\log p(\beta \mid \gamma, \lambda, \tau, y)$ as

$$
- \frac{1}{2} \left(\{y - \mu_z \mathbf{1} - \kappa_z \gamma^{-1}\} - X\beta\right)' \Gamma \left(\{y - \mu_z \mathbf{1} - \kappa_z \gamma^{-1}\} - X\beta\right) - \frac{1}{2\tau^2} (\beta' \Lambda \beta)
$$

This is the log posterior under a normal prior $\beta \sim N(0, \tau^2 \Lambda^{-1})$ and heteroskedastic Gaussian likelihood. Simple algebra then leads to the result.

For logistic regression, let $X_*$ be the matrix with rows $x_i^* = y_i x_i$. The kernel of the conditionally normal likelihood then becomes

$$(X_*\beta - \mu_z \mathbf{1} - \kappa_z \gamma^{-1})' \Gamma (X_*\beta - \mu_z \mathbf{1} - \kappa_z \gamma^{-1}).$$

Hence it is as if we observe the $n$-dimensional “data” vector $\mu_z \mathbf{1} + \kappa_z \gamma^{-1}$ in a regression model having design matrix $X_*$, and we proceed by a similar argument to arrive at the result.

### B.2 E Step

Factorize (20) as a function of $\beta$ to yield

$$
\log p(\beta \mid \gamma, \lambda, \tau, y) = c_1(\gamma, \lambda, y, \tau) - \frac{1}{2} \sum_{i=1}^{n} \gamma_i (z_i - \mu_z)^2 + \kappa_y \sum_{i=1}^{n} (z_i - \mu_z) \\
- \frac{1}{2\tau^2} \sum_{j=1}^{p} \lambda_j (\beta_j - \mu_\beta)^2 + \kappa_\beta \sum_{j=1}^{p} (\beta_j - \mu_\beta), \quad (21)
$$

26
which depends linearly upon $\gamma_i$ and $\lambda_j$. Therefore for the E step, we plug the following conditional moments into the complete-data objective function:

$$\hat{\lambda}_j^{(g)} = \alpha \tau^{2-\alpha} |\beta_j^{(g)}|^{\alpha-2}$$

$$\hat{\gamma}_i^{(g)} = \left\{ \begin{array}{ll} \frac{\gamma_i}{z_i} \left\{ \frac{e^{\gamma_i}}{1+e^{\gamma_i}} - \frac{1}{2} \right\} & \text{for logistic regression} \\
\frac{1}{|y_i - x_i'\beta_j^{(g)}|^{-1}} & \text{for quantile regression.} \end{array} \right.$$  

The conditional moments for $\gamma_i$ can be derived using an argument that is essentially identical to that following Equation (14)—that is, by differentiating the normal kernel under the integral of the mixture representation for the likelihood (c.f. Polson and Scott 2011c).

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. Technical report, Duke University Department of Statistical Science, 2010.

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 *ICML ’08: Proceedings of the 25th international conference on Machine learning*, pages 88–95. ACM, 2008.

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

N. Chopin. Fast simulation of truncated Gaussian distributions. *Statistics and Computing*, 21(2):275–88, 2011.

L. Devroye. Methods for generating random variates with Polya characteristic functions. *Statistics & Probability Letters*, 2:257–61, 1984.

L. Devroye. *Non-uniform random variate generation*. Springer, 1986.

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.

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.

J. Geweke. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In E. Keramidas and S. Kaufman, editors, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 571–8, 1991.

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.

G. Golub and C. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

J. Griffin and P. Brown. Alternative prior distributions for variable selection with very many more variables than observations. Technical report, University of Warwick, 2005.

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

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.

M. Johannes, N. G. Polson, and S. M. Yae. Quantile filtering and learning. Available at SSRN: http://ssrn.com/abstract=1509808, 2009.

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

Q. Li, R. Xi, and N. Lin. Bayesian regularized quantile regression. *Bayesian Analysis*, 5(3):533–56, 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. Technical report, Stanford University, 2009.

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. Large-scale simultaneous testing with hypergeometric inverted-beta priors. Technical report, University of Texas at Austin, http://arxiv.org/abs/1010.5223, 2010.

N. G. Polson and J. G. Scott. Shrink globally, act locally: sparse Bayesian regularization and prediction. In *Proceedings of the 9th Valencia World Meeting on Bayesian Statistics*. Oxford University Press, 2011a.

N. G. Polson and J. G. Scott. Local shrinkage rules, Lévy processes, and regularized regression. *Journal of the Royal Statistical Society (Series B)*, 2011b.

N. G. Polson and J. G. Scott. Sparse bayes estimation in non-gaussian models via data augmentation. Technical report, University of Texas at Austin, http://arxiv.org/abs/1103.5407v1, 2011c.

N. G. Polson and S. Scott. Data augmentation for support vector machines (with discussion). *Bayesian Analysis*, 6(1):1–24, 2011d.

J. G. Scott. Parameter expansion in local-shrinkage models. Technical report, University of Texas at Austin, 2010.

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.