Generalized Beta Mixtures of Gaussians

Artin Armagan
Dept. of Statistical Science
Duke University
Durham, NC 27708
artin@stat.duke.edu

David B. Dunson
Dept. of Statistical Science
Duke University
Durham, NC 27708
dunson@stat.duke.edu

Merlise Clyde
Dept. of Statistical Science
Duke University
Durham, NC 27708
clyde@stat.duke.edu

Abstract

In recent years, a rich variety of shrinkage priors have been proposed that have great promise in addressing massive regression problems. In general, these new priors can be expressed as scale mixtures of normals, but have more complex forms and better properties than traditional Cauchy and double exponential priors. We first propose a new class of normal scale mixtures through a novel generalized beta distribution that encompasses many interesting priors as special cases. This encompassing framework should prove useful in comparing competing priors, considering properties and revealing close connections. We then develop a class of variational Bayes approximations through the new hierarchy presented that will scale more efficiently to the types of truly massive data sets that are now encountered routinely.

1 Introduction

Penalized likelihood estimation has evolved into a major area of research, with ℓ_1[22] and other regularization penalties now used routinely in a rich variety of domains. Often minimizing a loss function subject to a regularization penalty leads to an estimator that has a Bayesian interpretation as the mode of a posterior distribution [8, 11, 1, 2], with different prior distributions inducing different penalties. For example, it is well known that Gaussian priors induce ℓ_2 penalties, while double exponential priors induce ℓ_1 penalties [8, 19, 13, 1]. Viewing massive-dimensional parameter learning and prediction problems from a Bayesian perspective naturally leads one to design new priors that have substantial advantages over the simple normal or double exponential choices and that induce rich new families of penalties. For example, in high-dimensional settings it is often appealing to have a prior that is concentrated at zero, favoring strong shrinkage of small signals and potentially a sparse estimator, while having heavy tails to avoid over-shrinkage of the larger signals. The Gaussian and double exponential priors are insufficiently flexible in having a single scale parameter and relatively light tails; in order to shrink many small signals strongly towards zero, the double exponential must be concentrated near zero and hence will over-shrink signals not close to zero. This phenomenon has motivated a rich variety of new priors such as the *normal-exponential-gamma*, the *horseshoe* and the *generalized double Pareto* [11, 14, 1, 6, 20, 7, 12, 2].

An alternative and widely applied Bayesian framework relies on variable selection priors and Bayesian model selection/averaging [18, 9, 16, 15]. Under such approaches the prior is a mixture of a mass at zero, corresponding to the coefficients to be set equal to zero and hence excluded from the model, and a continuous distribution, providing a prior for the size of the non-zero signals. This paradigm is very appealing in fully accounting for uncertainty in parameter learning and the unknown sparsity structure through a probabilistic framework. One obtains a posterior distribution over the model space corresponding to all possible subsets of predictors, and one can use this posterior for model-averaged predictions that take into account uncertainty in subset selection and to obtain marginal inclusion probabilities for each predictor providing a weight of evidence that a specific signal is non-zero allowing for uncertainty in the other signals to be included. Unfortunately,

the computational complexity is exponential in the number of candidate predictors (2^p with p the number of predictors).

Some recently proposed continuous shrinkage priors may be considered competitors to the conventional mixture priors [15, 6, 7, 12] yielding computationally attractive alternatives to Bayesian model averaging. Continuous shrinkage priors lead to several advantages. The ones represented as scale mixtures of Gaussians allow conjugate block updating of the regression coefficients in linear models and hence lead to substantial improvements in Markov chain Monte Carlo (MCMC) efficiency through more rapid mixing and convergence rates. Under certain conditions these will also yield sparse estimates, if desired, via maximum a posteriori (MAP) estimation and approximate inferences via variational approaches [17, 24, 5, 8, 11, 1, 2].

The class of priors that we consider in this paper encompasses many interesting priors as special cases and reveals interesting connections among different hierarchical formulations. Exploiting an equivalent conjugate hierarchy of this class of priors, we develop a class of variational Bayes approximations that can scale up to truly massive data sets. This conjugate hierarchy also allows for conjugate modeling of some previously proposed priors which have some rather complex yet advantageous forms and facilitates straightforward computation via Gibbs sampling. We also argue intuitively that by adjusting a global shrinkage parameter that controls the overall sparsity level, we may control the number of non-zero parameters to be estimated, enhancing results, if there is an underlying sparse structure. This global shrinkage parameter is inherent to the structure of the priors we discuss as in [6, 7] with close connections to the conventional variable selection priors.

2 Background

We provide a brief background on shrinkage priors focusing primarily on the priors studied by [6, 7] and [11, 12] as well as the Strawderman-Berger (SB) prior [7]. These priors possess some very appealing properties in contrast to the double exponential prior which leads to the Bayesian lasso [19, 13]. They may be much heavier-tailed, biasing large signals less drastically while shrinking noise-like signals heavily towards zero. In particular, the priors by [6, 7], along with the Strawderman-Berger prior [7], have a very interesting and intuitive representation later given in (2), yet, are not formed in a conjugate manner potentially leading to analytical and computational complexity.

[6, 7] propose a useful class of priors for the estimation of multiple means. Suppose a p-dimensional vector $\mathbf{y}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{I})$ is observed. The independent hierarchical prior for θ_j is given by

$$\theta_j|\tau_j \sim \mathcal{N}(0, \tau_j), \;\; \tau_j^{1/2} \sim \mathcal{C}^+(0, \phi^{1/2}), \tag{1}$$

for $j = 1, \ldots, p$, where $\mathcal{N}(\mu, \nu)$ denotes a normal distribution with mean μ and variance ν and $\mathcal{C}^+(0, s)$ denotes a half-Cauchy distribution on $\Re^+$ with scale parameter s. With an appropriate transformation $\rho_j = 1/(1 + \tau_j)$, this hierarchy also can be represented as

$$\theta_j|\rho_j \sim \mathcal{N}(0, 1/\rho_j - 1), \;\; \pi(\rho_j|\phi) \propto \rho_j^{-1/2}(1 - \rho_j)^{-1/2}\frac{1}{1 + (\phi - 1)\rho_j}. \tag{2}$$

A special case where $\phi = 1$ leads to $\rho_j \sim \mathcal{B}(1/2, 1/2)$ (beta distribution) where the name of the prior arises, *horseshoe* (HS) [6, 7]. Here ρ_js are referred to as the *shrinkage coefficients* as they determine the magnitude with which θ_js are pulled toward zero. A prior of the form $\rho_j \sim \mathcal{B}(1/2, 1/2)$ is natural to consider in the estimation of a signal θ_j as this yields a very desirable behavior both at the tails and in the neighborhood of zero. That is, the resulting prior has heavy-tails as well as being unbounded at zero which creates a strong pull towards zero for those values close to zero. [7] further discuss priors of the form $\rho_j \sim \mathcal{B}(a, b)$ for $a > 0$, $b > 0$ to elaborate more on their focus on the choice $a = b = 1/2$. A similar formulation dates back to [21]. [7] refer to the prior of the form $\rho_j \sim \mathcal{B}(1, 1/2)$ as the Strawderman-Berger prior due to [21] and [4]. The same hierarchical prior is also referred to as the quasi-Cauchy prior in [16]. Hence, the tail behavior of the Strawderman-Berger prior remains similar to the horseshoe (when $\phi = 1$), while the behavior around the origin changes. The hierarchy in (2) is much more intuitive than the one in (1) as it explicitly reveals the behavior of the resulting marginal prior on θ_j. This intuitive representation makes these hierarchical priors interesting despite their relatively complex forms. On the other hand, what the prior in (1) or (2) lacks is a more trivial hierarchy that yields recognizable conditional posteriors in linear models.

[11, 12] consider the normal-exponential-gamma (NEG) and normal-gamma (NG) priors respectively which are formed in a conjugate manner yet lack the intuition the Strawderman-Berger and horseshoe priors provide in terms of the behavior of the density around the origin and at the tails. Hence the implementation of these priors may be more user-friendly but they are very implicit in how they behave. In what follows we will see that these two forms are not far from one another. In fact, we may unite these two distinct hierarchical formulations under the same class of priors through a generalized beta distribution and the proposed equivalence of hierarchies in the following section. This is rather important to be able to compare the behavior of priors emerging from different hierarchical formulations. Furthermore, this equivalence in the hierarchies will allow for a straightforward Gibbs sampling update in posterior inference, as well as making variational approximations possible in linear models.

3 Equivalence of Hierarchies via a Generalized Beta Distribution

In this section we propose a generalization of the beta distribution to form a flexible class of scale mixtures of normals with very appealing behavior. We then formulate our hierarchical prior in a conjugate manner and reveal similarities and connections to the priors given in [16, 11, 12, 6, 7]. As the name *generalized beta* has previously been used, we refer to our generalization as the *three-parameter beta* (TPB) distribution.

In the forthcoming text $\Gamma(.)$ denotes the gamma function, $\mathcal{G}(\mu,\nu)$ denotes a gamma distribution with shape and rate parameters μ and ν, $\mathcal{W}(\nu,S)$ denotes a Wishart distribution with ν degrees of freedom and scale matrix S, $\mathcal{U}(\alpha_1,\alpha_2)$ denotes a uniform distribution over (α_1,α_2), $\mathcal{GIG}(\mu,\nu,\xi)$ denotes a generalized inverse Gaussian distribution with density function $(\nu/\xi)^{\mu/2}\{2\mathrm{K}_\mu(\sqrt{\nu\xi})\}^{-1}x^{\mu-1}\exp\{(\nu x+\xi/x)/2\}$, and $\mathrm{K}_\mu(.)$ is a modified Bessel function of the second kind.

Definition 1. *The three-parameter beta (TPB) distribution for a random variable X is defined by the density function*

$$f(x;a,b,\phi)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\phi^b x^{b-1}(1-x)^{a-1}\left\{1+(\phi-1)x\right\}^{-(a+b)}, \tag{3}$$

for $0<x<1$, $a>0$, $b>0$ and $\phi>0$ and is denoted by $\mathcal{TPB}(a,b,\phi)$.

It can be easily shown by a change of variable $x=1/(y+1)$ that the above density integrates to 1. The kth moment of the TPB distribution is given by

$$\mathbb{E}(X^k)=\frac{\Gamma(a+b)\Gamma(b+k)}{\Gamma(b)\Gamma(a+b+k)}{}_2\mathrm{F}_1(a+b,b+k;a+b+k;1-\phi) \tag{4}$$

where ${}_2\mathrm{F}_1$ denotes the hypergeometric function. In fact it can be shown that TPB is a subclass of Gauss hypergeometric (GH) distribution proposed in [3] and the compound confluent hypergeometric (CCH) distribution proposed in [10].

The density functions of GH and CCH distributions are given by

$$f_{\mathrm{GH}}(x;a,b,r,\zeta) = \frac{x^{b-1}(1-x)^{a-1}(1+\zeta x)^{-r}}{\mathrm{B}(b,a){}_2\mathrm{F}_1(r,b;a+b;-\zeta)}, \tag{5}$$

$$f_{\mathrm{CCH}}(x;a,b,r,s,\nu,\theta) = \frac{\nu^b x^{b-1}(1-x)^{a-1}(\theta+(1-\theta)\nu x)^{-r}}{\mathrm{B}(b,a)\exp(-s/\nu)\Phi_1(a,r,a+b,s/\nu,1-\theta)}, \tag{6}$$

for $0<x<1$ and $0<x<1/\nu$, respectively, where $\mathrm{B}(b,a)=\Gamma(a)\Gamma(b)/\Gamma(a+b)$ denotes the beta function and Φ_1 is the degenerate hypergeometric function of two variables [10]. Letting $\zeta=\phi-1$, $r=a+b$ and noting that ${}_2\mathrm{F}_1(a+b,b;a+b;1-\phi)=\phi^{-b}$, (5) becomes a TPB density. Also note that (6) becomes (5) for $s=1$, $\nu=1$ and $\zeta=(1-\theta)/\theta$ [10].

[20] considered an alternative special case of the CCH distribution for the shrinkage coefficients, ρ_j, by letting $\nu=r=1$ in (6). [20] refer to this special case as the hypergeometric-beta (HB) distribution. TPB and HB generalize the beta distribution in two distinct directions, with one practical advantage of the TPB being that it allows for a straightforward conjugate hierarchy leading to potentially substantial analytical and computational gains.

Now we move onto the hierarchical modeling of a flexible class of shrinkage priors for the estimation of a potentially sparse p-vector. Suppose a p-dimensional vector $\mathbf{y}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \mathbf{I})$ is observed where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ is of interest. Now we define a shrinkage prior that is obtained by mixing a normal distribution over its scale parameter with the TPB distribution.

Definition 2. *The TPB normal scale mixture representation for the distribution of random variable θ_j is given by*

$$\theta_j|\rho_j \sim \mathcal{N}(0, 1/\rho_j - 1), \quad \rho_j \sim \mathcal{TPB}(a, b, \phi), \tag{7}$$

where $a > 0$, $b > 0$ and $\phi > 0$. The resulting marginal distribution on θ_j is denoted by $\mathcal{TPBN}(a, b, \phi)$.

Figure 1 illustrates the density on ρ_j for varying values of a, b and ϕ. Note that the special case for $a = b = 1/2$ in Figure 1(a) gives the horseshoe prior. Also when $a = \phi = 1$ and $b = 1/2$, this representation yields the Strawderman-Berger prior. For a fixed value of ϕ, smaller a values yield a density on θ_j that is more peaked at zero, while smaller values of b yield a density on θ_j that is heavier tailed. For fixed values of a and b, decreasing ϕ shifts the mass of the density on ρ_j from left to right, suggesting more support for stronger shrinkage. That said, the density assigned in the neighborhood of $\theta_j = 0$ increases while making the overall density lighter-tailed. We next propose the equivalence of three hierarchical representations revealing a wide class of priors encompassing many of those mentioned earlier.

Proposition 1. *If $\theta_j \sim \mathcal{TPBN}(a, b, \phi)$, then*
1) $\theta_j \sim \mathcal{N}(0, \tau_j)$, $\tau_j \sim \mathcal{G}(a, \lambda_j)$ and $\lambda_j \sim \mathcal{G}(b, \phi)$.
2) $\theta_j \sim \mathcal{N}(0, \tau_j)$, $\pi(\tau_j) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\phi^{-a}\tau^{a-1}(1 + \tau_j/\phi)^{-(a+b)}$ which implies that $\tau_j\phi \sim \beta'(a, b)$, the inverted beta distribution with parameters a and b.

The equivalence given in Proposition 1 is significant as it makes the work in Section 4 possible under the TPB normal scale mixtures as well as further revealing connections among previously proposed shrinkage priors. It provides a rich class of priors leading to great flexibility in terms of the induced shrinkage and makes it clear that this new class of priors can be considered simultaneous extensions to the work by [11, 12] and [6, 7]. It is worth mentioning that the hierarchical prior(s) given in Proposition 1 are different than the approach taken by [12] in how we handle the mixing. In particular, the first hierarchy presented in Proposition 1 is identical to the NG prior up to the first stage mixing. While fixing the values of a and b, we further mix over λ_j (rather than a global λ) and further over ϕ if desired as will be discussed later. ϕ acts as a global shrinkage parameter in the hierarchy. On the other hand, [12] choose to further mix over a and a global λ while fixing the values of b and ϕ. By doing so, they forfeit a complete conjugate structure and an explicit control over the tail behavior of $\pi(\theta_j)$.

As a direct corollary to Proposition 1, we observe a possible equivalence between the SB and the NEG priors.

Corollary 1. *If $a = 1$ in Proposition 1, then TPBN $\equiv$ NEG. If $(a, b, \phi) = (1, 1/2, 1)$ in Proposition 1, then TPBN $\equiv$ SB $\equiv$ NEG.*

An interesting, yet expected, observation on Proposition 1 is that a half-Cauchy prior can be represented as a scale mixture of gamma distributions, i.e. if $\tau_j \sim \mathcal{G}(1/2, \lambda_j)$ and $\lambda_j \sim \mathcal{G}(1/2, \phi)$, then $\tau_j^{1/2} \sim \mathcal{C}^+(0, \phi^{1/2})$. This makes sense as $\tau^{1/2}|\lambda_j$ has a half-Normal distribution and the mixing distribution on the precision parameter is gamma with shape parameter $1/2$.

[7] further place a half-Cauchy prior on $\phi^{1/2}$ to complete the hierarchy. The aforementioned observation helps us formulate the complete hierarchy proposed in [7] in a conjugate manner. This should bring analytical and computational advantages as well as making the application of the procedure much easier for the average user without the need for a relatively more complex sampling scheme.

Corollary 2. *If $\theta_j \sim \mathcal{N}(0, \tau_j)$, $\tau_j^{1/2} \sim \mathcal{C}^+(0, \phi^{1/2})$ and $\phi^{1/2} \sim \mathcal{C}^+(0, 1)$, then $\theta_j \sim \mathcal{TPBN}(1/2, 1/2, \phi)$, $\phi \sim \mathcal{G}(1/2, \omega)$ and $\omega \sim \mathcal{G}(1/2, 1)$.*

Hence disregarding the different treatments of the higher-level hyper-parameters, we have shown that the class of priors given in Definition 1 unites the priors in [16, 11, 12, 6, 7] under one family and reveals their close connections through the equivalence of hierarchies given in Proposition 1. The first hierarchy in Proposition 1 makes much of the work possible in the following sections.

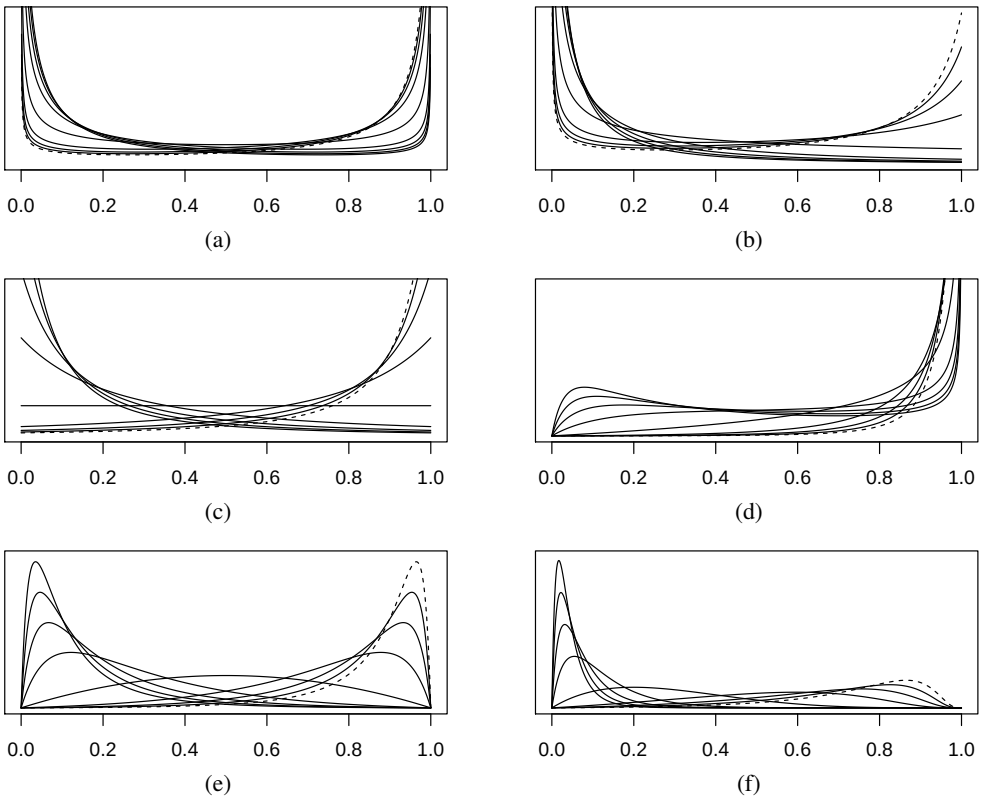


Figure 1: $(a, b) = \{(1/2, 1/2), (1, 1/2), (1, 1), (1/2, 2), (2, 2), (5, 2)\}$ for (a)-(f) respectively. $\phi = \{1/10, 1/9, 1/8, 1/7, 1/6, 1/5, 1/4, 1/3, 1/2, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ considered for all pairs of a and b. The line corresponding to the lowest value of ϕ is drawn with a dashed line.

4 Estimation and Posterior Inference in Regression Models

4.1 Fully Bayes and Approximate Inference

Consider the linear regression model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$, where $\mathbf{y}$ is an n-dimensional vector of responses, $\mathbf{X}$ is the $n \times p$ design matrix and $\boldsymbol{\epsilon}$ is an n-dimensional vector of independent residuals which are normally distributed, $\mathcal{N}(0, \sigma^2\mathbf{I}_n)$ with variance σ^2.

We place the hierarchical prior given in Proposition 1 on each β_j, i.e. $\beta_j \sim \mathcal{N}(0, \sigma^2\tau_j)$, $\tau_j \sim \mathcal{G}(a, \lambda_j)$, $\lambda_j \sim \mathcal{G}(b, \phi)$. ϕ is used as a global shrinkage parameter common to all β_j, and may be inferred using the data. Thus we follow the hierarchy by letting $\phi \sim \mathcal{G}(1/2, \omega)$, $\omega \sim \mathcal{G}(1/2, 1)$ which implies $\phi^{1/2} \sim \mathcal{C}^+(0, 1)$ that is identical to what was used in [7] at this level of the hierarchy. However, we do not believe at this level in the hierarchy the choice of the prior will have a huge impact on the results. Although treating ϕ as unknown may be reasonable, when there exists some prior knowledge, it is appropriate to fix a ϕ value to reflect our prior belief in terms of underlying sparsity of the coefficient vector. This sounds rather natural as soon as one starts seeing ϕ as a parameter that governs the multiplicity adjustment as discussed in [7]. Note also that here we form the dependence on the error variance at a lower level of hierarchy rather than forming it in the prior of ϕ as done in [7]. If we let $a = b = 1/2$, we will have formulated the hierarchical prior given in [7] in a completely conjugate manner. We also let $\sigma^{-2} \sim \mathcal{G}(c_0/2, d_0/2)$. Under a normal likelihood, an efficient Gibbs sampler may be obtained as the fully conditional posteriors can be extracted: $\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, \sigma^2, \tau_1, \ldots, \tau_p \sim \mathcal{N}(\boldsymbol{\mu}_\beta, \mathbf{V}_\beta)$, $\sigma^{-2}|\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \tau_1, \ldots, \tau_p \sim \mathcal{G}(c^*, d^*)$, $\tau_j|\beta_j, \sigma^2, \lambda_j \sim \mathcal{GIG}(a-1/2, 2\lambda_j, \beta_j^2/\sigma^2)$, $\lambda_j|\tau_j, \phi \sim \mathcal{G}(a+b, \tau_j+\phi)$, $\phi|\lambda_j, \omega \sim \mathcal{G}(pb+1/2, \sum_{j=1}^p \lambda_j+\omega)$, $\omega|\phi \sim \mathcal{G}(1, \phi+1)$, where $\boldsymbol{\mu}_\beta = (\mathbf{X}'\mathbf{X} + \mathbf{T}^{-1})^{-1}\mathbf{X}'\mathbf{y}$, $\mathbf{V}_\beta = \sigma^2(\mathbf{X}'\mathbf{X} + \mathbf{T}^{-1})^{-1}$, $c^* = (n + p + c_0)/2$, $d^* = \{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}'\mathbf{T}^{-1}\boldsymbol{\beta} + d_0\}/2$, $\mathbf{T} = \text{diag}(\tau_1, \ldots, \tau_p)$.

As an alternative to MCMC and Laplace approximations [23], a lower-bound on marginal likelihoods may be obtained via variational methods [17] yielding approximate posterior distributions on the model parameters. Using a similar approach to [5, 1], the approximate marginal posterior distributions of the parameters are given by $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_\beta, \mathbf{V}_\beta)$, $\sigma^{-2} \sim \mathcal{G}(c^*, d^*)$, $\tau_j \sim \mathcal{GIG}(a-1/2, 2\langle\lambda_j\rangle, \langle\sigma^{-2}\rangle\langle\beta_j^2\rangle)$, $\lambda_j \sim \mathcal{G}(a+b, \langle\tau_j\rangle+\langle\phi\rangle)$, $\phi \sim \mathcal{G}(pb+1/2, \langle\omega\rangle+\sum_{j=1}^p\langle\lambda_j\rangle)$, $\omega \sim \mathrm{G}(1, \langle\phi\rangle+1)$, where $\boldsymbol{\mu}_\beta = \langle\boldsymbol{\beta}\rangle = (\mathbf{X}'\mathbf{X}+\mathbf{T}^{-1})^{-1}\mathbf{X}'\mathbf{y}$, $\mathbf{V}_\beta = \langle\sigma^{-2}\rangle^{-1}(\mathbf{X}'\mathbf{X}+\mathbf{T}^{-1})^{-1}$, $\mathbf{T}^{-1} = \mathrm{diag}(\langle\tau_1^{-1}\rangle, \ldots, \langle\tau_p^{-1}\rangle)$, $c^* = (n+p+c_0)/2$, $d^* = (\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\langle\boldsymbol{\beta}\rangle + \sum_{i=1}^n \mathbf{x}_i\langle\boldsymbol{\beta}\boldsymbol{\beta}'\rangle\mathbf{x}_i + \sum_{j=1}^p\langle\beta_j^2\rangle\langle\tau_j^{-1}\rangle + d_0)/2$, $\langle\boldsymbol{\beta}\boldsymbol{\beta}'\rangle = \mathbf{V}_\beta + \langle\boldsymbol{\beta}\rangle\langle\boldsymbol{\beta}'\rangle$, $\langle\sigma^{-2}\rangle = c^*/d^*$, $\langle\lambda_j\rangle = (a+b)/(\langle\tau_j\rangle+\langle\phi\rangle)$, $\langle\phi\rangle = (pb+1/2)/(\langle\omega\rangle+\sum_{j=1}^p\langle\lambda_j\rangle)$, $\langle\omega\rangle = 1/(\langle\phi\rangle+1)$ and

$$\begin{aligned}\langle\tau\rangle &= \frac{(\langle\sigma^{-2}\rangle\langle\beta_j^2\rangle)^{1/2}\mathrm{K}_{a+1/2}\left\{(2\langle\lambda_j\rangle\langle\sigma^{-2}\rangle\langle\beta_j^2\rangle)^{1/2}\right\}}{(2\langle\lambda_j\rangle)^{1/2}\mathrm{K}_{a-1/2}\left\{(2\langle\lambda_j\rangle\langle\sigma^{-2}\rangle\langle\beta_j^2\rangle)^{1/2}\right\}}, \\ \langle\tau^{-1}\rangle &= \frac{(2\langle\lambda_j\rangle)^{1/2}\mathrm{K}_{3/2-a}\left\{(2\langle\lambda_j\rangle\langle\sigma^{-2}\rangle\langle\beta_j^2\rangle)^{1/2}\right\}}{(\langle\sigma^{-2}\rangle\langle\beta_j^2\rangle)^{1/2}\mathrm{K}_{1/2-a}\left\{(2\langle\lambda_j\rangle\langle\sigma^{-2}\rangle\langle\beta_j^2\rangle)^{1/2}\right\}}.\end{aligned}$$

This procedure consists of initializing the moments and iterating through them until some convergence criterion is reached. The deterministic nature of these approximations make them attractive as a quick alternative to MCMC.

This conjugate modeling approach we have taken allows for a very straightforward implementation of Strawderman-Berger and horseshoe priors or, more generally, TPB normal scale mixture priors in regression models without the need for a more sophisticated sampling scheme which may ultimately attract more audiences towards the use of these more flexible and carefully defined normal scale mixture priors.

4.2 Sparse Maximum a Posteriori Estimation

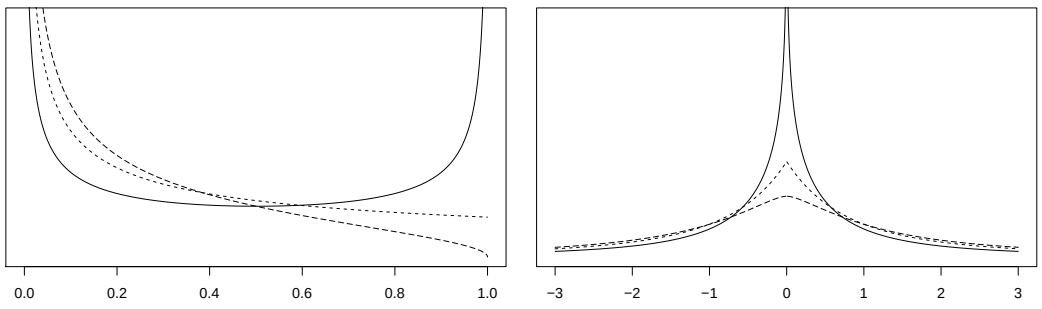
Although not our main focus, many readers are interested in sparse solutions, hence we give the following brief discussion. Given a, b and ϕ, maximum a posteriori (MAP) estimation is rather straightforward via a simple expectation-maximization (EM) procedure. This is accomplished in a similar manner to [8] by obtaining the joint MAP estimates of the error variance and the regression coefficients having taken the expectation with respect to the conditional posterior distribution of τ_j^{-1} using the second hierarchy given in Proposition 1. The kth expectation step then would consist of calculating

$$\langle\tau_j^{-1}\rangle^{(k)} = \frac{\int_0^\infty \tau_j^{a-1/2}(1+\tau_j/\phi)^{-(a+b)}\exp\{-\beta_j^{2(k-1)}/(2\sigma_{(k-1)}^2\tau_j)\}d\tau_j^{-1}}{\int_0^\infty \tau_j^{1/2+a}(1+\tau_j/\phi)^{-(a+b)}\exp\{-\beta_j^{2(k-1)}/(2\sigma_{(k-1)}^2\tau_j)\}d\tau_j^{-1}} \tag{8}$$

where $\beta_j^{2(k-1)}$ and $\sigma_{(k-1)}^2$ denote the modal estimates of the jth component of $\boldsymbol{\beta}$ and the error variance σ^2 at iteration $(k-1)$. The solution to (8) may be expressed in terms of some special function(s) for changing values of a, b and ϕ. $b < 1$ is a good choice as it will keep the tails of the marginal density on β_j heavy. A careful choice of a, on the other hand, is essential to sparse estimation. Admissible values of a for sparse estimation is apparent by the representation in Definition 2, noting that for any $a > 1$, $\pi(\rho_j = 1) = 0$, i.e. β_j may never be shrunk exactly to zero. Hence for sparse estimation, it is essential that $0 < a \leq 1$. Figure 2 (a) and (b) give the prior densities on ρ_j for $b = 1/2$, $\phi = 1$ and $a = \{1/2, 1, 3/2\}$ and the resulting marginal prior densities on β_j. These marginal densities are given by

$$\pi(\beta_j) = \begin{cases} \frac{1}{\sqrt{2}\pi^{3/2}}e^{\beta_j^2/2}\Gamma(0, \beta_j^2/2) & a = 1/2 \\ \frac{1}{\sqrt{2\pi}} - \frac{|\beta_j|}{2}e^{\beta_j^2/2} + \frac{\beta_j}{2}e^{\beta_j^2/2}\mathrm{Erf}(\beta_j/\sqrt{2}) & a = 1 \\ \frac{\sqrt{2}}{\pi^{3/2}}\left\{1 - \frac{1}{2}e^{\beta_j^2/2}\beta_j^2\Gamma(0, \beta_j^2/2)\right\} & a = 3/2 \end{cases}$$

where $\mathrm{Erf}(.)$ denotes the error function and $\Gamma(s, z) = \int_z^\infty t^{s-1}e^{-t}dt$ is the incomplete gamma function. Figure 2 clearly illustrates that while all three cases have very similar tail behavior, their behavior around the origin differ drastically.



(a) (b)

Figure 2: Prior densities of (a) ρ_j and (b) β_j for $a = 1/2$ (solid), $a = 1$ (dashed) and $a = 3/2$ (long dash).

5 Experiments

Throughout this section we use the Jeffreys' prior on the error precision by setting $c_0 = d_0 = 0$. We generate data for two cases, $(n, p) = \{(50, 20), (250, 100)\}$, from $y_i = \mathbf{x}_i'\boldsymbol{\beta}^* + \epsilon_i$, for $i = 1, \ldots, n$ where $\boldsymbol{\beta}^*$ is a p-vector that on average contains $20q$ non-zero elements which are indexed by the set $\mathcal{A} = \{j : \beta_j^* \neq 0\}$ for some random $q \in (0, 1)$. We randomize the procedure in the following manner: (i) $\mathbf{C} \sim \mathcal{W}(p, \mathbf{I}_{p\times p})$, (ii) $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, (iii) $q \sim \mathcal{B}(1, 1)$ for the first and $q \sim \mathcal{B}(1, 4)$ for the second cases, (iv) $I(j \in \mathcal{A}) \sim$ Bernoulli(q) for $j = 1, \ldots, p$ where $I(.)$ denotes the indicator function, (v) for $j \in \mathcal{A}$, $\beta_j \sim \mathcal{U}(0, 6)$ and for $j \notin \mathcal{A}$, $\beta_j = 0$ and finally (vi) $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ where $\sigma \sim \mathcal{U}(0, 6)$. We generated 1000 data sets for each case resulting in a median signal-to-noise ratio of approximately 3.3 and 4.5. We obtain the estimate of the regression coefficients, $\hat{\boldsymbol{\beta}}$, using the variational Bayes procedure and measure the performance by model error which is calculated as $(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})'\mathbf{C}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})$. Figure 3(a) and (b) display the median *relative* model error (RME) values (with their distributions obtained via bootstrapping) which is obtained by dividing the model error observed from our procedures by that of ℓ_1 regularization (lasso) tuned by 10-fold cross-validation. The boxplots in Figure 3(a) and (b) correspond to different (a, b, ϕ) values where C$^+$ signifies that ϕ is treated as unknown with a half-Cauchy prior as given earlier in Section 4.1. It is worth mentioning that we attain a clearly superior performance compared to the lasso, particularly in the second case, despite the fact that the estimator resulting from the variational Bayes procedure is not a thresholding rule. Note that $b = 1$ choice leads to much better performance under Case 2 than Case 1. This is due to the fact that Case 2 involves a much sparser underlying setup on average than Case 1 and that the lighter tails attained by setting $b = 1$ leads to stronger shrinkage.

To give a high dimensional example, we also generate a data set from the model $y_i = \mathbf{x}_i'\boldsymbol{\beta}^* + \epsilon_i$, for $i = 1, \ldots, 100$, where $\boldsymbol{\beta}^*$ is a 10000-dimensional very sparse vector with 10 randomly chosen components set to be 3, $\epsilon_i \sim \mathcal{N}(0, 3^2)$ and $x_{ij} \sim \mathcal{N}(0, 1)$ for $j = 1, \ldots, p$. This $\boldsymbol{\beta}^*$ choice leads to a signal-to-noise ratios of 3.16. For the particular data set we generated, the randomly chosen components of $\boldsymbol{\beta}^*$ to be non-zero were indexed by 1263, 2199, 2421, 4809, 5530, 7483, 7638, 7741, 7891 and 8187. We set $(a, b, \phi) = (1, 1/2, 10^{-4})$ which implies that a priori $\mathbb{P}(\rho_j > 0.5) = 0.99$ placing much more density in the neighborhood of $\rho_j = 1$ (total shrinkage). This choice is due to the fact that $n/p = 0.01$ and to roughly reflect that we do not want any more than 100 predictors in the resulting model. Hence ϕ is used, a priori, to limit the number of predictors in the model in relation to the sample size. Also note that with $a = 1$, the conditional posterior distribution of τ_j^{-1} is reduced to an inverse Gaussian. Since we are adjusting the global shrinkage parameter, ϕ, a priori, and it is chosen such that $\mathbb{P}(\rho_j > 0.5) = 0.99$, whether $a = 1/2$ or $a = 1$ should not matter. We first run the Gibbs sampler for 100000 iterations (2.4 hours on a computer with a 2.8 GHz CPU and 12 Gb of RAM using `Matlab`), discard the first 20000, thin the rest by picking every 5th sample to obtain the posteriors of the parameters. We observed that the chain converged by the 10000th iteration. For comparison purposes, we also ran the variational Bayes procedure using the values from the converged chain as the initial points (80 seconds). Figure 4 gives the posterior means attained by sampling and the variational approximation. The estimates corresponding to the zero elements of

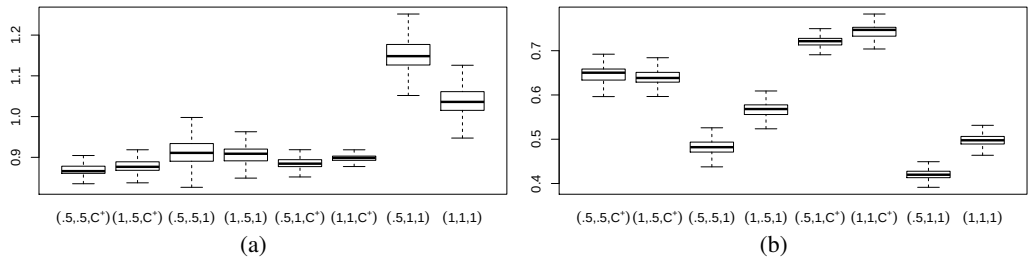


Figure 3: Relative ME at different (a, b, ϕ) values for (a) Case 1 and (b) Case 2.

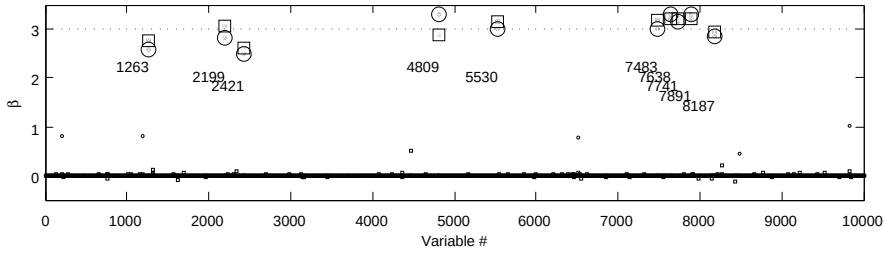


Figure 4: Posterior mean of $\boldsymbol{\beta}$ by sampling (square) and by approximate inference (circle).

$\boldsymbol{\beta}^*$ are plotted with smaller shapes to prevent clutter. We see that in both cases the procedure is able to pick up the larger signals and shrink a significantly large portion of the rest towards zero. The approximate inference results are in accordance with the results from the Gibbs sampler. It should be noted that using a good informed guess on ϕ, rather than treating it as an unknown in this high dimensional setting, improves the performance drastically.

6 Discussion

We conclude that the proposed hierarchical prior formulation constitutes a useful encompassing framework in understanding the behavior of different scale mixtures of normals and connecting them under a broader family of hierarchical priors. While ℓ_1 regularization, or namely lasso, arising from a double exponential prior in the Bayesian framework yields certain computational advantages, it demonstrates much inferior estimation performance relative to the more carefully formulated scale mixtures of normals. The proposed equivalence of the hierarchies in Proposition 1 makes computation much easier for the TPB scale mixtures of normals. As per different choices of hyper-parameters, we recommend that $a \in (0, 1]$ and $b \in (0, 1)$; in particular $(a, b) = \{(1/2, 1/2), (1, 1/2)\}$. These choices guarantee that the resulting prior has a kink at zero, which is essential for sparse estimation, and leads to heavy tails to avoid unnecessary bias in large signals (recall that a choice of $b = 1/2$ will yield Cauchy-like tails). In problems where oracle knowledge on sparsity exists or when $p >> n$, we recommend that ϕ is fixed at a reasonable quantity to reflect an appropriate sparsity constraint as mentioned in Section 5.

Acknowledgments

This work was supported by Award Number R01ES017436 from the National Institute of Environmental Health Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences or the National Institutes of Health.

References

[1] A. Armagan. Variational bridge regression. *JMLR: W&CP*, 5:17–24, 2009.

[2] A. Armagan, D. B. Dunson, and J. Lee. Generalized double Pareto shrinkage. arXiv:1104.0861v2, 2011.

[3] C. Armero and M. J. Bayarri. Prior assessments for prediction in queues. *The Statistician*, 43(1):pp. 139–153, 1994.

[4] J. Berger. A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics*, 8(4):pp. 716–761, 1980.

[5] C. M. Bishop and M. E. Tipping. Variational relevance vector machines. In *UAI '00: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 46–53, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[6] C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. *JMLR: W&CP*, 5, 2009.

[7] C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[8] M. A. T. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1150–1159, 2003.

[9] E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88, 1993.

[10] M. Gordy. A generalization of generalized beta distributions. Finance and Economics Discussion Series 1998-18, Board of Governors of the Federal Reserve System (U.S.), 1998.

[11] J. E. Griffin and P. J. Brown. Bayesian adaptive lassos with non-convex penalization. *Technical Report*, 2007.

[12] J. E. Griffin and P. J. Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

[13] C. Hans. Bayesian lasso regression. *Biometrika*, 96:835–845, 2009.

[14] C. J. Hoggart, J. C. Whittaker, and David J. Balding M. De Iorio. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics*, 4(7), 2008.

[15] H. Ishwaran and J. S. Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):pp. 730–773, 2005.

[16] I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, 32(4):pp. 1594–1649, 2004.

[17] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. *An introduction to variational methods for graphical models*. MIT Press, Cambridge, MA, USA, 1999.

[18] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):pp. 1023–1032, 1988.

[19] T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103:681–686(6), 2008.

[20] N. G. Polson and J. G. Scott. Alternative global-local shrinkage rules using hypergeometric-beta mixtures. Discussion Paper 2009-14, Department of Statistical Science, Duke University, 2009.

[21] W. E. Strawderman. Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, 42(1):pp. 385–388, 1971.

[22] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[23] L. Tierney and J. B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

[24] M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1, 2001.