Generalized double Pareto shrinkage

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

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

Jaeyong Lee
Department of Statistics
Seoul National University
Seoul, 151-747, Korea
leejyc@gmail.com

Abstract

We propose a generalized double Pareto prior for Bayesian shrinkage estimation and inferences in linear models. The prior can be obtained via a scale mixture of Laplace or normal distributions, while forming a bridge between the Laplace and Normal-Jeffreys' priors. While it has a spike at zero like the Laplace density, it also has a Student-t-like tail behavior. Bayesian computation is straightforward via a simple Gibbs sampling algorithm. We investigate the properties of the maximum a posteriori estimator, as many are interested in sparse solutions, reveal connections with some well-established regularization procedures and show some asymptotic results. The performance of the prior is tested through simulations.

Key words: Heavy tails; High-dimensional data; Lasso; Maximum a posteriori estimation; Relevance vector machine; Robust prior; Shrinkage estimation.

1 Introduction

There has been a great deal of work in shrinkage estimation and simultaneous variable selection in the frequentist framework. The Lasso of Tibshirani (1996) has drawn much attention to the area, particularly after the introduction of LARS (Efron et al., 2004) due to its superb computational performance. There is a rich literature analyzing the Lasso and related approaches (Fu, 1998; Knight and Fu, 2000; Fan and Li, 2001; Yuan and Lin, 2005; Zhao and Yu, 2006; Zou, 2006; Zou and Li, 2008), with a number of articles considering asymptotic properties.

Bayesian approaches to the same problem became popular with the works of Tipping (2001) and Figueiredo (2003). By expressing Student-t priors for basis coefficients as scale mixtures of normals (West, 1987) and relying on type II maximum likelihood estimation (Berger, 1985), Tipping (2001) developed the relevance vector machine for sparse estimation in kernel regression. However, sparsity comes with the price of forfeiting propriety of the posterior by driving the degrees of freedom and the scale parameter of the Student-t distribution towards zero. This yielded the so-called Normal-Jeffreys' prior on the parameters, $p(\theta) \propto 1/|\theta|$. Figueiredo (2003) proposed an expectation-maximization algorithm for maximum a posteriori estimation under Laplace and Normal-Jeffreys' priors, with estimates under the Laplace corresponding to the Lasso. The Normal-Jeffreys' prior leads to substantially improved performance due to the property of strongly shrinking small coefficients to zero while minimally shrinking large coefficients due to

the heavy tails. Although interpretable as posterior modes, these estimators and other penalized likelihood estima-
tors do not correspond to Bayes estimators under a reasonable choice of loss function, and hence lack a fully Bayes
justification.

A Bayesian Lasso was proposed by Park and Casella (2008) and Hans (2009). However, these procedures inherit
the problem of over-shrinking large coefficients due to the relatively light tails of the Laplace prior. Strawderman-
Berger priors (Strawderman, 1971; Berger, 1980) have some desirable properties yet lack an analytic form. Recently
proposed priors have been designed to have high density near zero and heavy tails without the impropriety problem
of Normal-Jeffreys. The horseshoe prior of Carvalho et al. (2009, 2010) is induced through a carefully-specified
mixture of normals, leading to desirable properties, such as an infinite spike at zero and very heavy tails. They studied
sparse shrinkage estimation properties of the horseshoe in a normal means problem. Griffin and Brown (2007, 2010)
proposed an alternative class of hierarchical priors for shrinkage, which has some similarities to the prior we propose,
but lacks the simple analytic form facilitating the study of some properties.

There is a need for alternative shrinkage priors that lead to sparse point estimates if desired, do not over-shrink
coefficients that are not close to zero, facilitate straightforward computation even in large p cases, and result in a
joint posterior distribution that does a good job in quantifying uncertainty. We propose the generalized double Pareto
prior which independently finds mention in Cevher (2009). It has a simple analytic form, yields a proper posterior
and possesses appealing properties, including a spike at zero, Student-t-like tails, and a simple characterization as a
scale mixture of normals leading to a straightforward Gibbs sampler for posterior inferences. We consider both fully
Bayesian and frequentist penalized likelihood approaches based on this prior. We show that the induced penalty in
the regularization framework yields a consistent thresholding rule having the continuity property in the orthogonal
case, with a simple Expectation-Maximization algorithm described for sparse estimation in non-orthogonal cases.
Similarities to Cevher (2009) are very limited and the contributions beyond these are (i) a formal introduction of a
generalized Pareto density thresholded and folded at zero as a shrinkage prior in Bayesian analysis, (ii) the scale
mixture representation of the generalized double Pareto given in Proposition 1 which is central to our work, (iii) its
connection to the Laplace and Normal-Jeffreys' priors as limiting cases given in Proposition 2, (iv) the resulting fully
conditional posteriors in a linear regression setting along with a simple Gibbs sampling procedure, (v) a discussion on
the hyper-parameters α and η and their treatment along with the incorporation of a griddy sampling scheme into the
Gibbs sampler, (vi) a detailed analysis of the induced penalty by the generalized double Pareto prior and the properties
of the resulting thresholding rule, (vii) an explicit analytic form for the maximum a posteriori estimator in orthogonal
cases, (viii) consistency of the resulting thresholding rule with a diverging number of parameters in orthogonal cases,
(ix) an expectation-maximization procedure to obtain the maximum a posteriori estimate in non-orthogonal cases
using the normal mixture representation given in Section 5.1, and finally (x) the one-step estimator (Zou and Li,
2008) resulting from the Laplace mixture representation and its oracle properties given in Section 5.2 revealing the
connection of the resulting procedure to the adaptive Lasso (Zou, 2006).

2 Generalized Double Pareto Prior

The generalized double Pareto density is given by

$$f(\theta|\xi,\alpha) = \frac{1}{2\xi}\left(1 + \frac{|\theta|}{\alpha\xi}\right)^{-(1+\alpha)},$$ (1)

where $\xi > 0$ is a scale parameter and $\alpha > 0$ is a shape parameter. In contrast to (1), the generalized Pareto density
of Pickands (1975) is parametrized in terms of a location parameter $\mu \in \Re$, a scale parameter $\xi > 0$, and a shape
parameter $\alpha \in \Re$ as follows,

$$f(\theta\,|\,\xi,\alpha,\mu) = \frac{1}{\xi}\left(1 + \frac{\theta - \mu}{\alpha\xi}\right)^{-(1+\alpha)},$$ (2)

with $\theta \geq \mu$ for $\alpha > 0$ and $\mu \leq \theta \leq \mu - \xi\alpha$ for $\alpha < 0$. The mean and variance for the generalized Pareto distribution is
respectively given by $E(\theta) = \mu + \xi/(1 - 1/\alpha)$ for $\alpha \notin [0,1]$ and $\text{Var}(\theta) = \xi^2(1 - 1/\alpha)^{-2}(1 - 2/\alpha)^{-1}$ for $\alpha \notin [0,2]$.
If we let $\mu = 0$, (2) becomes an exponential density as $\alpha \to \infty$ with mean ξ and variance ξ^2.

To modify the generalized Pareto density to be appropriate as a shrinkage prior, we let $\mu = 0$ and reflect the positive part about the origin assuming $\alpha > 0$. This leads to a density that is symmetric about zero. The mean and variance for the generalized double Pareto distribution is respectively given by $E(\theta) = 0$ for $\alpha > 1$ and $\text{Var}(\theta) = 2\xi^2\alpha^2(\alpha - 1)^{-1}(\alpha - 2)^{-1}$ for $\alpha > 2$. The dispersion is controlled by ξ and α, with α controlling the tail heaviness and $\alpha = 1$ corresponding to Cauchy-like tails and no finite moments.

Figure 1 compares density (1) to Cauchy and Laplace densities in the special case in which $\xi = \alpha = 1$, so that $f(\theta) = 1/\{2(1 + |\theta|)^2\}$. We refer to this form as the standard double Pareto. Near zero the standard double Pareto resembles the Laplace density, suggesting similar sparse shrinkage properties of small coefficients in maximum a posteriori estimation. It also has Cauchy-like tails which is appealing in avoiding over-shrinkage away from the origin. This is illustrated in Figure 1(a). Figure 1(b) illustrates how density (1) changes for different values of ξ and α.

Prior (1) can be represented as a scale mixture of normal distributions leading to computational simplifications. As shorthand notation, let $\theta \sim \text{GDP}(\xi, \alpha)$ denote that θ follows density (1).

Proposition 1. *Let $\theta \sim N(0, \tau)$, $\tau \sim Exp(\lambda^2/2)$ and $\lambda \sim Ga(\alpha, \eta)$ where $\alpha > 0$ and $\eta > 0$. The resulting marginal density for θ is $GDP(\xi = \eta/\alpha, \alpha)$.*

Proposition 1 reveals a relationship between prior (1) and the prior of Griffin and Brown (2007), with the difference being that Griffin and Brown (2007) place a mixing distribution on λ^2 leading to a marginal with no simple analytic form. Proposition 2 shows that prior (1) forms a bridge between Laplace and Normal-Jeffreys' priors.

Proposition 2. *Given the representation in Proposition 1, $\theta \sim GDP(\xi = \eta/\alpha, \alpha)$ implies*

1. *$f(\theta) \propto 1/|\theta|$ for $\alpha = 0$ and $\eta = 0$,*

2. *$f(\theta|\lambda') = (\lambda'/2)\exp(-\lambda'|\theta|)$ for $\alpha \to \infty$, $\alpha/\eta = \lambda'$ and $0 < \lambda' < \infty$.*

Proof. For the first item, setting $\alpha = \eta = 0$ implies placing a Jeffreys' prior on λ, $p(\lambda) \propto 1/\lambda$. Integration over λ yields $p(\tau) \propto 1/\tau$ which implies the Normal-Jeffreys' prior on θ. For the second item, notice that $p(\lambda) = \delta(\lambda - \lambda')$, where $\delta(.)$ denotes the Dirac delta function, since $\lim_{\alpha \to \infty} \lim_{\alpha/\eta \to \lambda'} E(\lambda) = \lambda'$ and $\lim_{\alpha \to \infty} \lim_{\alpha/\eta \to \lambda'} \text{Var}(\lambda) = 0$. Thus, $\int_0^\infty (\lambda/2)\exp(-\lambda|\theta|)\delta(d\lambda) = (\lambda'/2)\exp(-\lambda'|\theta|)$. $\square$

3 Bayesian Inference

3.1 Posterior Computation

Consider the linear regression model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is an n-dimensional vector of responses, $\mathbf{X}$ is the $n \times p$ design matrix and $\boldsymbol{\varepsilon} \sim N\left(0, \sigma^2\mathbf{I}_n\right)$. Letting $\beta_j|\sigma \sim \text{GDP}(\xi = \sigma\eta/\alpha, \alpha)$ independently for $j = 1, \ldots, p$,

$$\pi(\boldsymbol{\beta}|\sigma) = \prod_{j=1}^p \frac{1}{2\sigma\eta/\alpha}\left(1 + \frac{1}{\alpha}\frac{|\beta_j|}{\sigma\eta/\alpha}\right)^{-(\alpha+1)}. \tag{3}$$

From Proposition 1 this prior is equivalent to $\beta_j|\sigma \sim N(0, \sigma^2\tau_j)$, with $\tau_j \sim \text{Exp}(\lambda_j^2/2)$ and $\lambda_j \sim \text{Ga}(\alpha, \eta)$. We place the Jeffreys' prior on the error variance, $\pi(\sigma) \propto 1/\sigma$.

Using the scale mixture of normals representation, we obtain a simple data augmentation Gibbs sampler having the following conditional posteriors: $(\boldsymbol{\beta}|\sigma^2, \mathbf{T}, \mathbf{y}) \sim N\{(\mathbf{X}'\mathbf{X} + \mathbf{T}^{-1})^{-1}\mathbf{X}'\mathbf{y}, \sigma^2\left(\mathbf{X}'\mathbf{X} + \mathbf{T}^{-1}\right)^{-1}\}$, $(\sigma^2|\boldsymbol{\beta}, \mathbf{T}, \mathbf{y}) \sim$ IG$\{(n + p)/2, (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}'\mathbf{T}^{-1}\boldsymbol{\beta})/2\}$, $(\lambda_j|\beta_j, \sigma^2) \sim$ Ga$(\alpha + 1, |\beta_j|/\sigma + \eta)$, $(\tau_j^{-1}|\beta_j, \lambda_j, \sigma^2) \sim$ Inv-Gauss$\{\mu = (\lambda_j^2\sigma^2/\beta_j^2)^{1/2}, \rho = \lambda^2\}$ where $\mathbf{T} = \text{diag}(\tau_1, ..., \tau_p)$ and Inv-Gauss denotes the inverse Gaussian distribution with location and scale parameters μ and ρ. In our experience, this Gibbs sampler is efficient having fast rates of convergence and mixing.
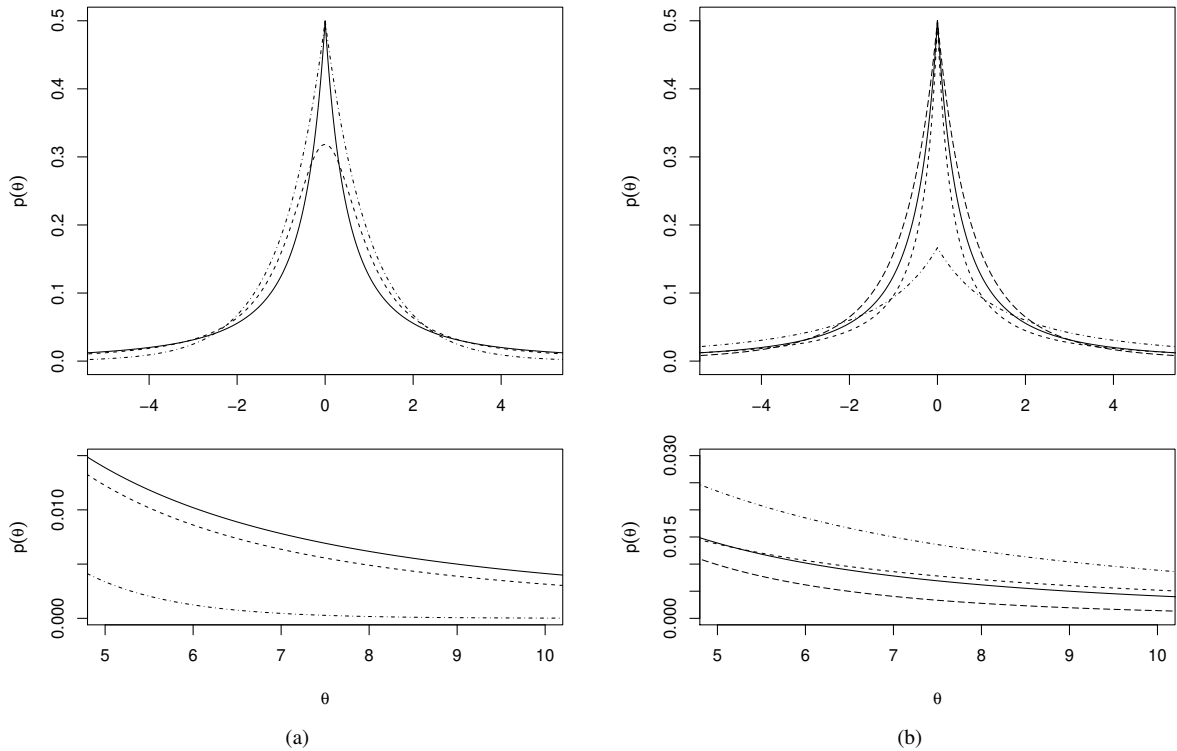
Figure 1: (a) Probability density functions for standard double Pareto (solid line), standard Cauchy (dashed line) and Laplace (dot-dash line) ($\lambda = 1$) distributions. (b) Probability density functions for the generalized double Pareto with (ξ, α) values of $(1, 1)$ (solid line), $(0.5, 1)$ (dashed line), $(1, 3)$ (long-dashed line), and $(3, 1)$ (dot-dash line).

3.2 Hyper-prior Specification and Computation

As α grows the density becomes lighter tailed, more peaked and the variance becomes smaller while as η grows the density becomes flatter and the variance increases. Hence if we increase α, we may cause unwanted bias for large signals, though causing stronger shrinkage for noise-like signals; if we increase η we may lose the ability to shrink noise-like signals also causing less bias for large signals; and finally, If we increase α and η at the same rate, the variance remains constant but the tails become lighter converging to a Laplace density in the limit. This can lead to over-shrinking of the coefficients that are away from zero. Given that the columns of $\mathbf{X}$ are scaled to be of unit length, as a typical default specification for the hyper-parameters, one can let $\alpha = \eta = 1$ (thus $\xi = \sigma$) in (3). This choice leads to Cauchy-like tail behavior, which is well-known to have desirable Bayesian robustness properties.

To further motivate our default choice, we assess the behavior of the prior shrinkage factor $\kappa = 1/(1+\tau) \in (0,1)$ where $\theta \sim \mathrm{N}(0, \tau)$ is the parameter of interest (Carvalho et al., 2010). As $\kappa \to 0$, the prior does not impose any shrinkage while as $\kappa \to 1$ it has a strong pull towards zero. The generalized double Pareto distribution implies a prior $p(\kappa)$ on κ upon integration over λ in Proposition 1. For the standard double Pareto, this is given by

$$p(\kappa) = \frac{1}{2(1-\kappa)^2} \left[\frac{\sqrt{\pi} \exp\left\{ \frac{\kappa}{2(1-\kappa)} \right\} \mathrm{Erfc}\left\{ \sqrt{\frac{\kappa}{2(1-\kappa)}} \right\}}{\sqrt{2\kappa(1-\kappa)}} - 1 \right],$$

where $\mathrm{Erfc}(.)$ denotes the complementary error function. In Figure 2, we compare $p(\kappa)$ under the standard double
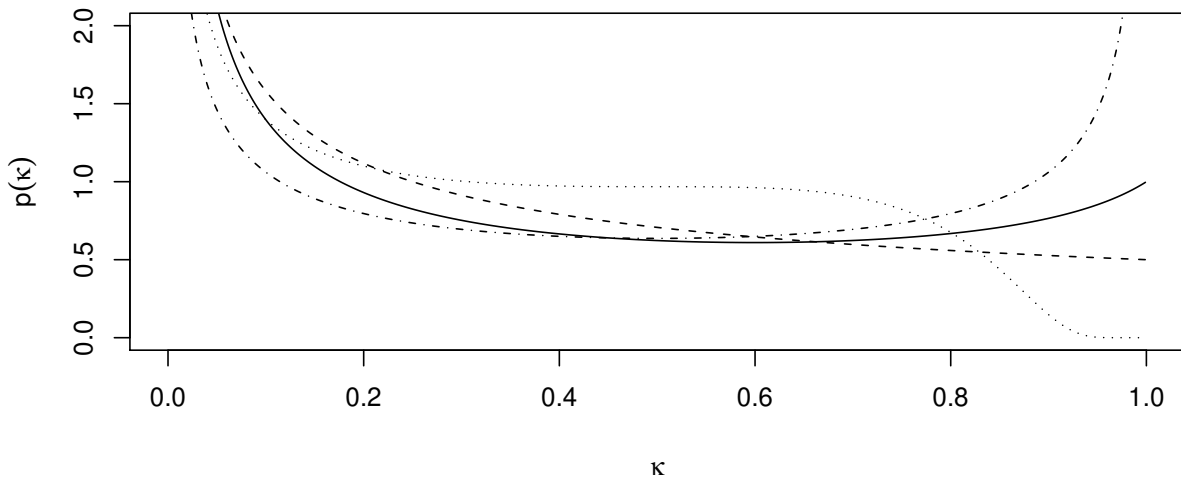
4

Figure 2: Prior density of κ implied by the standard double Pareto prior (solid line), Strawderman–Berger prior (dashed line), horseshoe prior (dot-dash line) and standard Cauchy prior (dotted line).

Pareto, Strawderman-Berger, horseshoe and Cauchy priors. The priors behave similarly for $\kappa \approx 0$, implying similar tail behavior. The behavior of $p(\kappa)$ for $\kappa \approx 1$ governs the strength of shrinkage of small signals. As $\kappa \to 1$, $p(\kappa)$ tends towards zero for the Cauchy implying weak shrinkage, while $p(\kappa)$ is unbounded for the horseshoe suggesting very strong pull towards zero for small signals. The Strawderman-Berger and standard double Pareto priors are a compromise between these extremes, with $p(\kappa)$ being bounded for $\kappa \to 1$ in both cases. The standard double Pareto assigns higher density close to one than the Strawderman-Berger prior and has the advantage of a simple analytic form and a conjugate hierarchy over the Strawderman-Berger and horseshoe priors.

As an alternative we recommend choosing hyper-priors to allow the data to inform about the values of α and η, with $p(\alpha) = 1/(1+\alpha)^2$ and $p(\eta) = 1/(1+\eta)^2$ to correspond to generalized Pareto hyper-priors with location 0, scale 1 and shape 1. The median value of the resulting distribution for α and η is one, centering it at the default choices suggested earlier while the mean and variance do not exist.

For sampling purposes let $a = 1/(1+\alpha)$ and $e = 1/(1+\eta)$. These transformations suggest a uniform prior on a and e in $(0,1)$ given the generalized Pareto priors on α and η. Consequently, the conditional posteriors for a and e are

$$p(a|\boldsymbol{\beta},\eta) \quad \propto \quad \left(\frac{1-a}{a}\right)^p \prod_{j=1}^{p}\left(1+\frac{|\beta_j|}{\sigma\eta}\right)^{-1/a},$$

$$p(e|\boldsymbol{\beta},\alpha) \quad \propto \quad \left(\frac{e}{1-e}\right)^p \prod_{j=1}^{p}\left\{1+e\frac{|\beta_j|}{\sigma(1-e)}\right\}^{-(\alpha+1)}.$$

We propose the following embedded griddy Gibbs (Ritter and Tanner, 1992) sampling scheme:

   i. Form a grid of m points $a^{(1)},...,a^{(m)}$ in the interval $(0,1)$,

   ii. Calculate $w^{(k)} = p(a^{(k)}|\boldsymbol{\beta},\eta)$,

   iii. Normalize the weights, $w_N^{(k)} = w^{(k)}/\sum_{k=1}^{m} w^{(k)}$,

iv. Draw a sample from the set $\{a^{(1)}, ..., a^{(m)}\}$ with probabilities $\{w_N^{(1)}, ..., w_N^{(m)}\}$ and set $\alpha = 1/a - 1$ to be used at the current iteration of the Gibbs sampler.

Repeat the same procedure for e and obtain a random draw for η. We also experiment with fixing η as 1 and $\sqrt{\alpha + 1}$, explaining the latter choice in the following section. In these cases, the prior variance of $\boldsymbol{\beta}|\sigma^2$ is determined by α.

In what follows we establish the ties between the Bayesian approach we have taken and some frequentist regularization approaches. The simple analytic structure of the generalized double Pareto prior allows for the following analyses while its hierarchical formulation leads to straight-forward computation.

4 Sparse Maximum a Posteriori Estimation

The generalized double Pareto distribution can be used not only as a prior in a Bayesian analysis but also to induce a sparsity-favoring penalty in regularized least squares,

$$\tilde{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \left\{ \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} p(|\beta_j|) \right\}, \tag{4}$$

where $\mathbf{X}$ is initially assumed to have orthonormal columns and $p(.)$ denotes the penalty function implied by the prior on the regression coefficients. Following Fan and Li (2001), let $\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ and denote the minimization problem in (4) for a component of $\boldsymbol{\beta}$ as

$$\tilde{\beta}_j = \arg \min_{\beta_j} \left\{ \frac{1}{2} \left(\hat{\beta}_j - \beta_j \right)^2 + \sigma^2 p(|\beta_j|) \right\}, \tag{5}$$

with the penalty function implied by (3), $p(|\beta_j|) = (\alpha + 1) \log (\sigma\eta + |\beta_j|)$.

From Fan and Li (2001), a good penalty function should result in an estimator that is (i) nearly unbiased when the true unknown parameter is large, (ii) a thresholding rule, which automatically sets small estimated coefficients to zero to reduce model complexity, and (iii) continuous in data z to avoid instability in model prediction. In the following, we show that the penalty function induced by prior (3) has these three properties.

4.1 Near-unbiasedness

The first order derivative of (5) with respect to β_j is $\mathrm{sgn}(\beta_j)\{|\beta_j| + \sigma^2 p'(|\beta_j|)\} - \hat{\beta}_j = \mathrm{sgn}(\beta_j)\{|\beta_j| + \sigma^2(\alpha + 1)/(\sigma\eta + |\beta_j|)\} - \hat{\beta}_j$, where $p'(|\beta_j|) = \partial p(|\beta_j|)/\partial |\beta_j|$ is the term causing bias in estimation. Although it is appealing to introduce bias in small coefficients to reduce the mean squared error and model complexity, it is also desirable to limit the shrinkage of large coefficients with $p'(|\beta_j|) \to 0$ as $|\beta_j| \to \infty$. In addition, it is desirable for $p'(|\beta_j|)$ to approach zero rapidly, implying shrinkage and the associated introduction of bias rapidly decreases as coefficients get further away from zero. In fact, the rate of convergence of $p'(|\beta_j|)$ to zero is of the same order under generalized double Pareto and Normal-Jeffreys' priors, with $\lim_{|\beta_j|\to\infty}\{(\alpha + 1)/(\sigma\eta + |\beta_j|)\}/\{1/|\beta_j|\} = \alpha + 1$. As α controls the tail heaviness in the generalized double Pareto prior, with lighter tails for larger values of α, convergence of the ratio to $(\alpha + 1)$ is intuitive. In the case of Lasso, the bias, $p'(|\beta_j|)$, remains constant regardless of $|\beta_j|$, which can also be observed in Figure 3(b).

4.2 Sparsity

As noted in Fan and Li (2001), a sufficient condition for the resulting estimator to be a thresholding rule is that the minimum of the function $|\beta_j| + \sigma^2 p'(|\beta_j|)$ is positive.

Proposition 3. *Given the formulation in Proposition 1, prior (3) implies a penalty yielding an estimator that is a thresholding rule if $\eta < 2\sqrt{\alpha + 1}$.*

This result is obtained by finding the minimum of $|\beta_j| + \sigma^2 p'(|\beta_j|)$ and setting it greater than zero. The thresholding is a direct consequence of the fact that when $|\hat{\beta}_j| < \min_{\beta_j}\{|\beta_j| + \sigma^2(\alpha+1)/(\sigma\eta + |\beta_j|)\}$ – which requires that $\min_{\beta_j}\{|\beta_j| + \sigma^2 p'(|\beta_j|)\} > 0$ – the derivative of (5) is positive for all positive β_j and negative for all negative β_j. In this case, the penalized least squares estimator is zero. When $|\hat{\beta}_j| > \min_{\beta_j}\{|\beta_j| + \sigma^2(\alpha+1)/(\sigma\eta + |\beta_j|)\}$, two roots may exist, the larger one being the penalized least squares estimator. To elaborate more on this, the root(s) may exist for $\text{sgn}(\beta_j)\{|\beta_j| + \sigma^2 p'(|\beta_j|)\} - \hat{\beta}_j = 0$ only when $|\hat{\beta}_j| > \min_{\beta_j}\{|\beta_j| + \sigma^2 p'(|\beta_j|)\}$. A helpful illustration is given in Figure 3 of Fan and Li (2001).

4.3   Continuity

Continuity in data is an important property of an estimator to avoid instabilities in prediction. As defined by Breiman (1996), "a regularization procedure is unstable if a small change in data can make large changes in the regularized estimator". Discontinuities in the thresholding rule may result in inclusion or dismissal of a signal with minor changes in the data used (see Figure 3(b)). Hard-thresholding – or namely the "usual" variable selection – is an unstable procedure, while ridge or Lasso estimates are considered to be stable. The penalty yielded by the Normal-Jeffreys' prior (the log penalty) mimics the behavior of the ℓ_γ penalty as $\gamma \to 0$ where ℓ_γ is the γ-norm of a vector for $\gamma > 0$. This close relation can also be observed in Figure 3(b), again by looking at the discontinuities of hard-thresholding and the Normal-Jeffreys' prior. This problem is remedied with the use of prior (3).

A necessary and sufficient condition for continuity is that the minimum of the function $|\beta_j| + \sigma^2 p'(|\beta_j|)$ is obtained at zero (Fan and Li, 2001). For our prior, the minimum of this function is obtained at $|\beta_j| = \sigma(\sqrt{\alpha+1} - \eta)$. Therefore $\eta = \sqrt{\alpha+1}$ will yield an estimator with this property.

Proposition 4. *Given the formulation in Proposition 1, a subfamily of prior (1) with $\eta = \sqrt{\alpha+1}$ implies a penalty function that yields an estimator with the continuity property.*

In this particular case, the penalized likelihood estimator is set to zero if $|\hat{\beta}_j| < \sigma\sqrt{\alpha+1}$. When $|\hat{\beta}_j| > \sigma\sqrt{\alpha+1}$,

$$\tilde{\beta}_j = \begin{cases} \frac{\hat{\beta}_j - \sigma\sqrt{\alpha+1} + \{\hat{\beta}_j^2 + 2\hat{\beta}_j\sigma\sqrt{\alpha+1} - 3\sigma^2(\alpha+1)\}^{1/2}}{2} & \hat{\beta}_j > 0, \\ \frac{\hat{\beta}_j + \sigma\sqrt{\alpha+1} - \{\hat{\beta}_j^2 - 2\hat{\beta}_j\sigma\sqrt{\alpha+1} - 3\sigma^2(\alpha+1)\}^{1/2}}{2} & \hat{\beta}_j < 0. \end{cases} \tag{6}$$

As can be observed in Figure 3(a), ensuring continuity by letting $\eta = \sqrt{\alpha+1}$ in prior (3) creates a trade-off between sparsity and tail-robustness. As the thresholding region becomes wider, the larger values are penalized further, yet not nearly at the level of Lasso. To achieve a similar thresholding rule to the Normal-Jeffreys' prior, we must pick $\alpha = 3$, which induces a lighter tailed distribution than a Cauchy distribution. We may reduce α to 1 and make the tail behavior similar to that of a Cauchy distribution, however, the thresholding region now is reduced to $\pm\sqrt{2}\sigma$. Choosing similar tail behavior to the Normal-Jeffreys' prior by letting $\alpha \to 0$, leading to an improper prior, induces a thresholding region of $\pm\sigma$.

4.4   Consistency in Estimation

We investigate the estimation consistency of the implied thresholding rule under orthogonal designs with a diverging number of parameters. Such designs are common in problems such as multiple mean estimation, wavelet smoothing, and principal component regression. Consider the following model which results in a multiple mean estimation (or orthogonal $\mathbf{X}$ such that $\mathbf{x}'_j\mathbf{x}_j = n$ where $\mathbf{x}_j$ is the jth column of $\mathbf{X}$) setting.

$$\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}^*_n + \frac{\sigma}{\sqrt{n}}\mathbf{Z}_n, \tag{7}$$

where $\hat{\boldsymbol{\beta}}_n$, $\boldsymbol{\beta}^*_n$ are p_n dimensional vectors, and $\mathbf{Z}_n$ is a p_n dimensional multivariate standard normally distributed random variable. Index n denotes sequences that change with n. Here $\boldsymbol{\beta}^*_n$ denotes the true unknown mean/coefficient vector and $\hat{\boldsymbol{\beta}}_n$ denotes the maximum likelihood estimator. Here $\boldsymbol{\beta}^*_n$ is assumed to have only a finite number, r, of nonzero elements. Hence, we let the number of zero elements grow with n.
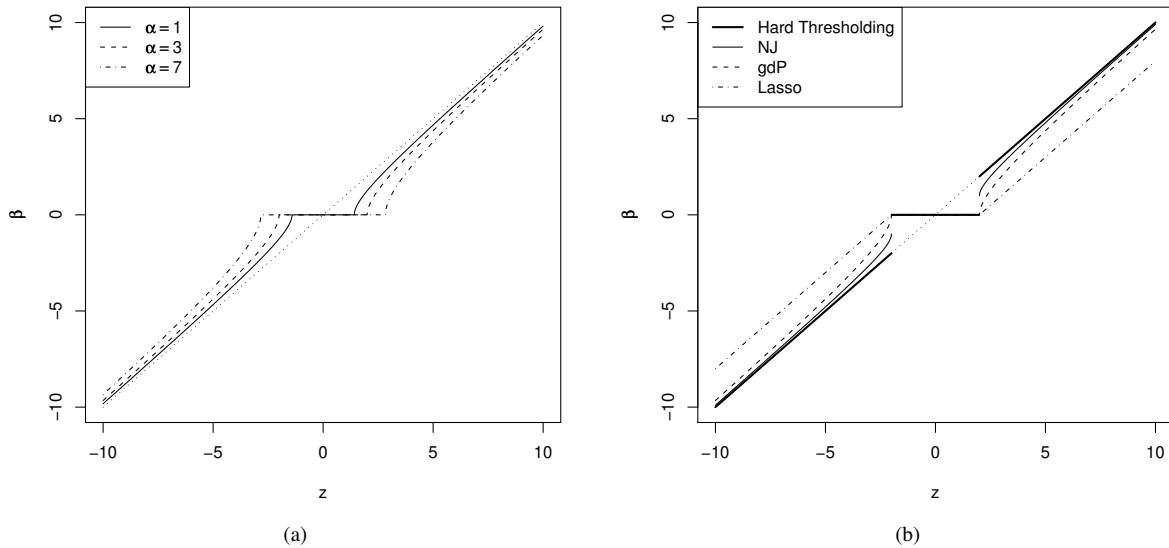
Figure 3: Thresholding functions for (a) generalized double Pareto prior with $\eta = \sqrt{\alpha + 1}$, $\alpha = \{1, 3, 7\}$, (b) Hard thresholding, Normal-Jeffreys' prior, generalized double Pareto prior with $\eta = 2$, $\alpha = 3$ and Lasso with $\sigma = 1$.

The maximum a posteriori estimator under prior (3) (with the continuity property) is given by (6) where σ is replaced by $\sigma/\sqrt{n}$, and $\alpha'_n = \alpha_n + 1$.

Theorem 1. *Let $\alpha'_n \to \infty$, $p_n \to \infty$, $\alpha'_n/n \to 0$ and $\sigma < \infty$. Given the model in (7) and the estimator $\tilde{\boldsymbol{\beta}}$ in (6),*

$$E\|\boldsymbol{\beta}^*_n - \tilde{\boldsymbol{\beta}}_n\|^2 = \mathcal{O}\left\{ \frac{p_n \sqrt{\alpha'_n} \exp(-\alpha'_n/2)}{n} \right\} + \mathcal{O}\left(\sqrt{\frac{\alpha'_n}{n}} \right). \tag{8}$$

The proof is deferred to the Appendix.

Corollary 1. *If $p_n = n$ and $\alpha'_n = \log n$, $E\|\boldsymbol{\beta}^*_n - \tilde{\boldsymbol{\beta}}_n\|^2 = \mathcal{O}(\sqrt{\log n/n})$.*

Proof. This result is obtained by letting $p_n = n$, equating the two terms on the right-hand side of (8), so that they will both have the same rate, and then solving for α'_n. □

Thus, the threshold for this procedure becomes $\sigma\sqrt{\log n}$. The multiplier of the standard error in the threshold $\sqrt{\log n}$ has a striking similarity to that arising in the so called "universal thresholding", $\sqrt{2 \log n}$.

Remark 1. *Here the rate of the estimator is repressed by the bias caused in the non-thresholded elements. Recall that as α'_n increases, the bias increases as well. Thus, there occurs a trade-off between the rate of the estimator and the continuity property we attained.*

Therefore, using the penalty implied by the generalized double Pareto prior and choosing an appropriate α'_n we may achieve consistency in estimation with a diverging number of parameters if the true signal has a finite number of nonzero elements. The result given in Corollary 1 is of particular relevance in wavelet smoothing and principle component regression. These properties are relevant to readers interested in sparse estimation and show connections to frequentist approaches.

8

5 Maximum a Posteriori Estimation via Expectation-Maximization

5.1 Exploiting the Normal Mixture Representation

We assume a normal likelihood to formulate the procedure for non-orthogonal linear regression. Estimation is carried out via the Expectation-Maximization algorithm. We first take the expectation of the log-posterior with respect to the conditional posterior distributions of $(\tau_j^{-1}|\beta_j^{(k)}, \lambda_j, \sigma^{2(k)})$ and $(\lambda_j|\beta_j^{(k)}, \sigma^{2(k)})$ at the kth step, and then maximize with respect to β_j and σ^2 yielding the values for the $(k+1)th$ step. Removing the terms of the log-posterior that do not depend on $\boldsymbol{\beta}$ and σ^2, we are left with

$$-\left(\frac{n+p}{2}+1\right)\log\sigma^2 - \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - \sum_{j=1}^p \beta_j^2/\tau_j}{2\sigma^2}.$$

- *E-step*:

$$-\left(\frac{n+p}{2}+1\right)\log\sigma^2 - \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{j=1}^p \beta_j^2 \underbrace{\left\{ \frac{(\alpha+1)\sigma^{2(k)}}{|\beta_j^{(k)}|\left(|\beta_j^{(k)}|+\sigma^{(k)}\eta\right)} \right\}}_{d_j^{(k)}}$$

- *M-step*: Letting $D^{(k)} = \text{diag}\left(d_1^{(k)},...,d_p^{(k)}\right)$, we have

$$\boldsymbol{\beta}^{(k+1)} = \left(\mathbf{X}'\mathbf{X} + D^{(k)}\right)^{-1}\mathbf{X}'\mathbf{y}, \quad \sigma^{2(k+1)} = \frac{\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(k)}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(k)}\right) + \boldsymbol{\beta}^{(k)'}D^{(k)}\boldsymbol{\beta}^{(k)}}{n+p+2}.$$

We refer to this estimator as GDP(MAP).

5.2 Exploiting the Laplace Mixture Representation and the One-step Estimator

An intuitive relationship to the adaptive Lasso of Zou (2006) and the one-step sparse estimator of Zou and Li (2008) can be seen via the Laplace mixture representation of prior (3) implied in Proposition 1. In the proof of Proposition 1, the integration over $\boldsymbol{\tau}$ leads to a Laplace mixture representation of the prior. As a computationally fast alternative to estimating the exact mode via the above EM algorithm, we can obtain a "one-step estimator" and exploit the LARS algorithm as in Zou and Li (2008). Since the mixing distribution of the Laplace is a known distribution, the required expectation is obtained with ease resulting in the following step $(k+1)$ maximization:

$$\boldsymbol{\beta}^{(k+1)} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \frac{1}{\sigma}\sum_{j=1}^p |\beta_j|\left(\frac{\alpha+1}{|\beta_j^{(k)}|+\sigma\eta}\right) \right\}. \tag{9}$$

The component-specific multiplier on $|\beta_j|$ is obtained from the expectation of λ_j with respect to its conditional posterior distribution, $p(\lambda_j|\beta_j)$. Similar results to (9) are observed by Candes et al. (2008), Cevher (2009) and Garrigues (2009). The one-step estimator is then given by

$$\boldsymbol{\beta}^{(1)} = \arg\min_{\boldsymbol{\beta}} \left\{ (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \alpha^\dagger \sum_{j=1}^p |\beta_j|\left(|\beta_j^{(0)}|+\eta^\dagger\right)^{-1} \right\}, \tag{10}$$

letting $\alpha^\dagger = 2\sigma(\alpha+1)$ and $\eta^\dagger = \sigma\eta$. This estimator resembles the adaptive Lasso. The LARS algorithm can be used to obtain $\boldsymbol{\beta}^{(1)}$ very quickly. We refer to this estimator as GDP(OS).

Assuming the same conditions as in Section 2 of Zou (2006), where $\mathbf{X}'\mathbf{X}/n \to \mathbf{C}$ is a positive definite matrix, we give the following theorem:

Theorem 2. *Let $\boldsymbol{\beta}_n^{(1)}$ denote the GDP(OS) estimator in (10) indexed by n, $\mathcal{A} = \{j : \beta_{nj}^* \neq 0\}$, $\mathcal{A}_n = \{j : \beta_{nj}^{(1)} \neq 0\}$. Suppose that $\alpha_n^\dagger \to \infty$, $\alpha_n^\dagger/\sqrt{n} \to 0$, and $\eta_n^\dagger \sqrt{n} \to c < \infty$. Then $\boldsymbol{\beta}_n^{(1)}$ satisfies the following:*

1. *Consistency in variable selection: $\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1$*

2. *Asymptotic normality: $\sqrt{n}(\boldsymbol{\beta}_{n\mathcal{A}}^{(1)} - \boldsymbol{\beta}_{\mathcal{A}}^*) \xrightarrow{d} N(0, \sigma^2 \mathbf{C}_{\mathcal{A}}^{-1})$, with $\boldsymbol{\beta}_{n\mathcal{A}}^{(1)} = \{\beta_{nj}^{(1)} : j \in \mathcal{A}\}$ denoting the estimates of the non-zero coefficients, $\boldsymbol{\beta}_{n\mathcal{A}}^* = \{\beta_{nj}^* : j \in \mathcal{A}\}$ denoting the values of the non-zero coefficients and $\mathbf{C}_{\mathcal{A}}$ retaining the rows and columns of $\mathbf{C}$ indexed by $\mathcal{A}$.*

The proof is deferred to the Appendix.

Remark 2. *For $\eta^\dagger = 0$, the GDP(OS) solution path for varying $\alpha^\dagger$ is identical to the adaptive Lasso solution path with $\gamma = 1$ (see (4) in Zou (2006)) using identical $\boldsymbol{\beta}^{(0)}$.*

Remark 3. *GDP(OS) forms a bridge between the Lasso and the adaptive Lasso; as $\eta^\dagger \to \infty$ and $\alpha^\dagger/\eta^\dagger \to \lambda^\dagger < \infty$, GDP(OS) gives the Lasso solution with penalty parameter $\lambda^\dagger$.*

We do not experiment with the results obtained in Sections 4 and 5 with the exception of the procedure given in Section 5.1. As mentioned earlier, these results are primarily presented to show the close connections between the Bayesian approach taken and the regularization framework.

6 Experiments

In this section, we compare the proposed estimators to the posterior means obtained under the normal, Laplace and the horseshoe priors as well as the Bayesian model averaged (BMA) estimator. Although our main focus is the posterior mean, we also give the performance of the maximum a posteriori (MAP) estimates using the procedure given in Section 5.1 for those who are interested in sparse estimation. GDP(PM) and GDP(MAP) denote the posterior mean and the MAP estimates respectively under the generalized double Pareto prior. Hyper-parameter values are provided in footnotes of Tables 1 and 2 when fixed in advance and are otherwise treated as random with the priors specified in Section 3.2. The MAP estimates of the regression coefficients are obtained by conditioning on the posterior means of the hyper-parameters when not fixed, i.e we first obtain the posterior means of the hyper-parameters from an initial Bayesian analysis to use them in the calculation of the MAP estimates.

Simulation 1: We generate $n = \{50, 400\}$ observations from $y_i = \mathbf{x}_i' \boldsymbol{\beta}^* + \varepsilon_i$, where x_{ij} are generated as independently and identically distributed standard normals, $\varepsilon_i \sim N(0, \sigma^2)$ and $\sigma = 3$. We use the following five $\boldsymbol{\beta}^*$ configurations:

Model 1: $\boldsymbol{\beta}^* = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)'$
Model 2: $\boldsymbol{\beta}^* = (3, 3, 3, 3, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)'$
Model 3: $\boldsymbol{\beta}^* = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0)'$
Model 4: $\boldsymbol{\beta}^* = (3, 3, 3, 3, 3, 0, 0, 0, 0, 0, 3, 3, 3, 3, 3, 0, 0, 0, 0, 0)'$
Model 5: $\boldsymbol{\beta}^* = (0.85, ..., 0.85)'$
These allow for signal-to-noise ratios of 0.745, 2.236, 1.054, 3.162, 1.267 respectively.

Simulation 2: We let $\text{Cov}(x_j, x_{j'}) = 0.5^{|j-j'|}$. This setup leads to signal-to-noise ratios of 1.112, 3.335, 1.576, 4.729, 2.120.

In our experiments $\mathbf{y}$ and the columns of $\mathbf{X}$ are centered and the columns of $\mathbf{X}$ scaled to have unit length. For the calculation of competing estimators we use `monomvn` (Gramacy, 2010) and `BAS` (Clyde and Littman, 2005; Clyde et al., 2010) packages in `R`. BMA estimator is calculated using the `BAS` package while the rest of the competing estimators are calculated using the `monomvn` package. We mainly follow the default settings provided by the packages. Under the normal prior, the so-called "ridge" parameter is given an inverse Gamma prior with shape and scale parameters 10^{-3}. Under the Laplace prior, as a default choice, the package places a Gamma prior on the "Lasso parameter" λ^2 as given in (6) of Park and Casella (2008) with shape and rate parameters 2 and 0.1 respectively. Under the horseshoe prior, the package uses the hierarchy given in Section 1.1 of Carvalho et al. (2010). For BMA, we use the default settings of the `BAS` package which employs a Zellner-Siow prior given in Section 3.1 of Liang et al. (2008).

100 data sets are generated for each case. In Tables 1 and 2, we report the median model error. Model error is calculated as $(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})'\mathbf{C}(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})$ where $\mathbf{C}$ is the variance-covariance matrix that generated $\mathbf{X}$. The values in the subscripts give the bootstrap standard error of the median model error values obtained. The bootstrap standard error was calculated by generating 500 bootstrap samples from 100 model error values, finding the median model error for each case, and then calculating the standard error for it. Under each model, the best three performances are boldfaced in the tables.

For larger sample sizes BMA outperforms the competing methods in most cases and GDP(MAP) estimator is the second best. This is not surprising as there indeed exists a true underlying sparse model in most of the scenarios considered. Except for Model 5, normal and Laplace priors are outperformed by other methods as expected. The GDP(PM) shows a similar performance to that of horseshoe. Considering that the GDP(PM) and GDP(MAP) calculations are very straightforward and computationally inexpensive due to the simple normal scale mixture representation used, it offers great utility. The ability to use a simple Gibbs sampler (especially when $\alpha = \eta = 1$) makes the procedure very attractive for the average user.

Table 1: Model error comparisons for Simulation 1.

Method	Model 1	Model 2	Model 3	Model 4	Model 5
			$n = 50$		
GDP(PM)[a]	$2.659_{0.127}$	$2.170_{0.222}$	$\mathbf{3.963_{0.163}}$	$3.847_{0.167}$	$\mathbf{5.662_{0.257}}$
GDP(PM)[b]	$\mathbf{2.775_{0.153}}$	$2.147_{0.207}$	$4.629_{0.187}$	$3.817_{0.178}$	$6.968_{0.164}$
GDP(PM)[c]	$\mathbf{2.592_{0.109}}$	$2.364_{0.249}$	$\mathbf{4.351_{0.138}}$	$4.001_{0.227}$	$6.540_{0.169}$
GDP(PM)	$2.785_{0.145}$	$2.281_{0.247}$	$4.403_{0.209}$	$4.455_{0.244}$	$\mathbf{6.038_{0.246}}$
GDP(MAP)[a]	$2.884_{0.164}$	$\mathbf{1.498_{0.138}}$	$5.854_{0.261}$	$\mathbf{2.894_{0.157}}$	$10.404_{0.243}$
GDP(MAP)[b]	$4.030_{0.273}$	$\mathbf{1.401_{0.106}}$	$7.099_{0.270}$	$\mathbf{3.017_{0.184}}$	$12.620_{0.155}$
GDP(MAP)[c]	$3.526_{0.168}$	$1.509_{0.127}$	$6.711_{0.248}$	$3.060_{0.165}$	$11.871_{0.169}$
GDP(MAP)	$3.436_{0.258}$	$1.992_{0.292}$	$5.838_{0.197}$	$4.269_{0.272}$	$8.750_{0.364}$
Normal	$3.236_{0.186}$	$5.746_{0.264}$	$\mathbf{3.895_{0.187}}$	$5.515_{0.218}$	$\mathbf{3.940_{0.189}}$
Laplace	$3.208_{0.114}$	$4.024_{0.254}$	$5.109_{0.337}$	$4.875_{0.279}$	$7.324_{0.458}$
Horseshoe	$\mathbf{2.701_{0.128}}$	$2.120_{0.201}$	$4.789_{0.219}$	$3.781_{0.223}$	$7.250_{0.227}$
BMA	$2.760_{0.121}$	$\mathbf{1.438_{0.103}}$	$4.549_{0.187}$	$\mathbf{2.461_{0.179}}$	$7.031_{0.233}$
			$n = 400$		
GDP(PM)[a]	$0.230_{0.016}$	$0.217_{0.010}$	$0.361_{0.018}$	$0.311_{0.017}$	$0.643_{0.039}$
GDP(PM)[b]	$0.219_{0.015}$	$0.228_{0.010}$	$0.365_{0.017}$	$0.338_{0.014}$	$0.591_{0.040}$
GDP(PM)[c]	$0.237_{0.014}$	$0.245_{0.009}$	$0.372_{0.017}$	$0.344_{0.017}$	$0.603_{0.039}$
GDP(PM)	$0.231_{0.013}$	$0.186_{0.010}$	$0.420_{0.020}$	$0.371_{0.016}$	$0.513_{0.032}$
GDP(MAP)[a]	$\mathbf{0.176_{0.014}}$	$\mathbf{0.146_{0.013}}$	$\mathbf{0.313_{0.019}}$	$\mathbf{0.265_{0.017}}$	$0.614_{0.040}$
GDP(MAP)[b]	$\mathbf{0.161_{0.010}}$	$0.153_{0.011}$	$\mathbf{0.313_{0.022}}$	$0.287_{0.016}$	$0.566_{0.037}$
GDP(MAP)[c]	$0.180_{0.014}$	$0.165_{0.011}$	$0.320_{0.022}$	$0.290_{0.015}$	$0.580_{0.038}$
GDP(MAP)	$0.209_{0.018}$	$\mathbf{0.139_{0.013}}$	$0.399_{0.019}$	$0.386_{0.023}$	$\mathbf{0.498_{0.030}}$
Normal	$0.415_{0.016}$	$0.460_{0.024}$	$0.476_{0.025}$	$0.472_{0.020}$	$\mathbf{0.459_{0.027}}$
Laplace	$0.328_{0.013}$	$0.393_{0.022}$	$0.440_{0.017}$	$0.442_{0.020}$	$0.517_{0.032}$
Horseshoe	$0.212_{0.015}$	$0.207_{0.009}$	$0.380_{0.018}$	$0.374_{0.015}$	$0.534_{0.033}$
BMA	$\mathbf{0.156_{0.012}}$	$\mathbf{0.126_{0.016}}$	$\mathbf{0.246_{0.014}}$	$\mathbf{0.242_{0.016}}$	$\mathbf{0.450_{0.021}}$

[a]$\alpha = 1, \eta = 1$; [b]$\eta = 1$; [c]$\eta = \sqrt{\alpha + 1}$

7 Discussion

We proposed a hierarchical prior obtained through a particular scale mixture of normals where the resulting marginal prior has a folded generalized Pareto density thresholded at zero. Although Bayesian model averaging is appealing, it can be argued that allowing parameters to be arbitrarily close to zero instead of exactly equal to zero is more natural. In

Table 2: Model error comparisons for Simulation 2.

	Model 1	Model 2	Model 3	Model 4	Model 5
Method			$n = 50$		
GDP(PM)[a]	$2.123_{0.116}$	$2.149_{0.105}$	$3.205_{0.157}$	$4.213_{0.267}$	$4.440_{0.134}$
GDP(PM)[b]	$\mathbf{1.944_{0.113}}$	$1.997_{0.103}$	$3.260_{0.182}$	$4.176_{0.192}$	$4.651_{0.127}$
GDP(PM)[c]	$\mathbf{1.912_{0.116}}$	$2.154_{0.095}$	$\mathbf{3.117_{0.153}}$	$4.267_{0.218}$	$4.329_{0.124}$
GDP(PM)	$\mathbf{1.948_{0.121}}$	$2.154_{0.125}$	$3.108_{0.121}$	$4.390_{0.207}$	$\mathbf{3.961_{0.149}}$
GDP(MAP)[a]	$2.478_{0.112}$	$1.576_{0.118}$	$4.711_{0.262}$	$\mathbf{3.070_{0.221}}$	$8.654_{0.263}$
GDP(MAP)[b]	$2.594_{0.187}$	$\mathbf{1.494_{0.124}}$	$5.022_{0.245}$	$3.198_{0.226}$	$9.185_{0.231}$
GDP(MAP)[c]	$2.466_{0.186}$	$\mathbf{1.546_{0.104}}$	$4.691_{0.276}$	$3.233_{0.202}$	$8.895_{0.227}$
GDP(MAP)	$2.381_{0.130}$	$1.869_{0.093}$	$3.774_{0.191}$	$4.093_{0.207}$	$5.482_{0.145}$
Normal	$2.352_{0.161}$	$4.165_{0.293}$	$\mathbf{2.739_{0.069}}$	$4.811_{0.238}$	$\mathbf{3.040_{0.177}}$
Laplace	$2.100_{0.146}$	$2.768_{0.171}$	$2.828_{0.119}$	$4.166_{0.259}$	$3.468_{0.154}$
Horseshoe	$1.983_{0.114}$	$2.003_{0.099}$	$3.329_{0.170}$	$4.345_{0.194}$	$4.570_{0.138}$
BMA	$2.408_{0.128}$	$\mathbf{1.346_{0.121}}$	$3.954_{0.124}$	$\mathbf{3.201_{0.259}}$	$6.297_{0.205}$
			$n = 400$		
GDP(PM)[a]	$0.215_{0.010}$	$0.219_{0.011}$	$0.321_{0.015}$	$0.266_{0.013}$	$0.659_{0.037}$
GDP(PM)[b]	$0.205_{0.010}$	$0.230_{0.014}$	$0.330_{0.017}$	$0.287_{0.013}$	$0.595_{0.034}$
GDP(PM)[c]	$0.217_{0.011}$	$0.236_{0.014}$	$0.336_{0.017}$	$0.293_{0.013}$	$0.595_{0.033}$
GDP(PM)	$0.208_{0.014}$	$0.192_{0.010}$	$0.362_{0.020}$	$0.340_{0.014}$	$0.494_{0.030}$
GDP(MAP)[a]	$0.155_{0.010}$	$\mathbf{0.148_{0.012}}$	$\mathbf{0.260_{0.018}}$	$\mathbf{0.227_{0.014}}$	$0.620_{0.039}$
GDP(MAP)[b]	$\mathbf{0.150_{0.009}}$	$0.151_{0.011}$	$\mathbf{0.261_{0.017}}$	$0.248_{0.015}$	$0.571_{0.033}$
GDP(MAP)[c]	$\mathbf{0.151_{0.009}}$	$0.168_{0.010}$	$0.280_{0.016}$	$0.250_{0.014}$	$0.582_{0.034}$
GDP(MAP)	$0.173_{0.012}$	$\mathbf{0.146_{0.012}}$	$0.331_{0.017}$	$0.346_{0.014}$	$\mathbf{0.478_{0.028}}$
Normal	$0.358_{0.013}$	$0.441_{0.024}$	$0.411_{0.019}$	$0.432_{0.013}$	$\mathbf{0.433_{0.026}}$
Laplace	$0.273_{0.013}$	$0.365_{0.019}$	$0.378_{0.020}$	$0.393_{0.013}$	$\mathbf{0.475_{0.028}}$
Horseshoe	$0.198_{0.010}$	$0.211_{0.012}$	$0.333_{0.019}$	$0.332_{0.012}$	$0.535_{0.031}$
BMA	$\mathbf{0.143_{0.010}}$	$\mathbf{0.118_{0.013}}$	$\mathbf{0.231_{0.017}}$	$\mathbf{0.203_{0.014}}$	$0.654_{0.040}$

[a]$\alpha = 1, \eta = 1$; [b]$\eta = 1$; [c]$\eta = \sqrt{\alpha + 1}$

addition, the proposed methods have substantial computational advantages in relying on simple block-updated Gibbs sampling, while BMA requires sampling from a model space with 2^p models. As p increases, it becomes impossible to even visit more than a vanishingly small proportion of the models. Given the simple and fast computation and excellent performance in small sample simulation studies, the generalized double Pareto should be useful as a shrinkage prior in a broad variety of Bayesian hierarchical models, while also suggesting close relationship to frequentist penalized likelihood approaches. The proposed prior can be applied outside of normal linear regression to generalized linear models, shrinkage of basis coefficients in nonparametric regression, and in more complex settings such as factor analysis and nonparametric Bayes modeling.

Appendix

Proof of Theorem 1. Let $\sum_{j=1}^{p_n} E(\beta_{nj}^* - \hat{\beta}_{nj})^2 = \sum_{j=1}^{r} E(\beta_{nj}^* - \hat{\beta}_{nj})^2 + \sum_{j=r+1}^{p_n} E(\hat{\beta}_{nj})^2 = I_1 + I_2$ where $\beta_{nj}^* \neq 0$ for $j = 1, ..., r$ and $\beta_{nj}^* = 0$ for $j = r+1, ..., p_n$.

We first analyze the behavior of I_2. Given the estimator in (6)

$$
\begin{aligned}
I_2 &= \frac{\sigma^2}{4n} \sum_{j=r+1}^{p_n} E\Big\{ I(|Z_{nj}| \geq \sqrt{\alpha'_n}) \\
&\qquad \times \Big[z_{nj} - \mathrm{sgn}\,(z_{nj})\sqrt{\alpha'_n} + \mathrm{sgn}\,(z_{nj})\big\{ z_{nj}^2 + \mathrm{sgn}\,(z_{nj})2\sqrt{\alpha'_n} - 3\alpha'_n \big\}^{1/2} \Big] \Big\}^2 \\
&= \frac{\alpha'_n \sigma^2}{2n} \sum_{j=r+1}^{p_n} \int_{\sqrt{\alpha'_n}}^{\infty} \left[\frac{z_{nj}}{\sqrt{\alpha'_n}} - 1 + \left\{ \left(\frac{z_{nj}}{\sqrt{\alpha'_n}} + 3 \right) \left(\frac{z_{nj}}{\sqrt{\alpha'_n}} - 1 \right) \right\}^{1/2} \right]^2 \phi(z_{nj}) dz_{nj} \\
&\leq \frac{2\sigma^2}{n} \sum_{j=r+1}^{p_n} \int_{\sqrt{\alpha'_n}}^{\infty} z_{nj}^2 \phi(z_{nj}) dz_{nj} \\
&= \frac{2\sigma^2(p_n - r - 1)}{n} \left\{ \frac{\sqrt{\alpha'_n}\exp(-\alpha'_n/2)}{\sqrt{2\pi}} + Q(\sqrt{\alpha'_n}) \right\} \\
&\leq \frac{\sqrt{2}\sigma^2(p_n - r - 1)\sqrt{\alpha'_n}\exp(-\alpha'_n/2)}{\sqrt{\pi}n} \left(1 + \frac{1}{\alpha'_n} \right) \\
&= \mathcal{O}\left\{ \frac{p_n\sqrt{\alpha'_n}\exp(-\alpha'_n/2)}{n} \right\}
\end{aligned}
$$

where $\phi(.)$ and $Q(.)$ denote the density and the tail probability of a standard normal distribution. In the last inequality we make use of $Q(x) \leq \exp(-x^2/2)/(x\sqrt{2\pi})$.

$$
\begin{aligned}
I_1 &= \sum_{j=1}^{r} E\left\{ I\left(\left| \beta_{nj}^* + \frac{\sigma}{\sqrt{n}}Z_{nj} \right| \geq \frac{\sigma}{\sqrt{n}}\sqrt{\alpha'_n} \right) (\beta_{nj}^* - \hat{\beta}_{nj})^2 \right\} \\
&\quad + \sum_{j=1}^{r} E\left\{ I\left(\left| \beta_{nj}^* + \frac{\sigma}{\sqrt{n}}Z_{nj} \right| < \frac{\sigma}{\sqrt{n}}\sqrt{\alpha'_n} \right) (\beta_{nj}^*)^2 \right\} \\
&= J_1 + J_2.
\end{aligned}
$$

Let us first analyze J_2:

$$
\begin{aligned}
J_2 &= \sum_{j=1}^{r} \mathrm{pr}\left(\left| \beta_{nj}^* + \frac{\sigma}{\sqrt{n}}Z_{nj} \right| < \frac{\sigma}{\sqrt{n}}\sqrt{\alpha'_n} \right) (\beta_{nj}^*)^2 \\
&= \sum_{j=1}^{r} P\left(-\sqrt{\alpha'_n} - \beta_{nj}^*\frac{\sqrt{n}}{\sigma} < Z_{nj} < \sqrt{\alpha'_n} - \beta_{nj}^*\frac{\sqrt{n}}{\sigma} \right) (\beta_{nj}^*)^2 \\
&\leq \sum_{j=1}^{r} \mathrm{pr}\left\{ Z_{nj} > -\sqrt{\alpha'_n} + \mathrm{sgn}\,(\beta_{nj}^*)\beta_{nj}^*\frac{\sqrt{n}}{\sigma} \right\} (\beta_{nj}^*)^2 \\
&\leq \frac{1}{\sqrt{2\pi}} \sum_{j=1}^{r} \frac{\exp\left[-\big\{ -\sqrt{\alpha'_n} + \mathrm{sgn}\,(\beta_{nj}^*)\beta_{nj}^*\sqrt{n}/\sigma \big\}^2/2 \right]}{-\sqrt{\alpha'_n} + \mathrm{sgn}\,(\beta_{nj}^*)\beta_{nj}^*\sqrt{n}/\sigma}(\beta_{nj}^*)^2 \\
&= \mathcal{O}\left(\frac{\exp\left[-\frac{n}{2}\big\{ -\sqrt{\alpha'_n/n} + C \big\}^2 \right]}{\sqrt{n}\big\{ -\sqrt{\alpha'_n/n} + C \big\}} \right),
\end{aligned}
$$

13

where $0 < C < \infty$. Let $c_1 = -\sqrt{\alpha_n'} - \beta_{nj}^*\sqrt{n}/\sigma$ and $c_2 = \sqrt{\alpha_n'} - \beta_{nj}^*\sqrt{n}/\sigma$.

$$J_1 = \sum_{j=1}^{r}\left\{\int_{-\infty}^{c_1}\left(\hat{\beta}_{nj} - \beta_{nj}^*\right)^2\phi(z_{nj})dz_{nj} + \int_{c_2}^{\infty}\left(\hat{\beta}_{nj} - \beta_{nj}^*\right)^2\phi(z_{nj})dz_{nj}\right\}$$

Let us consider the case $\beta_{nj}^* > 0$ for a summand of J_1:

$$
\begin{aligned}
J_{1j} &= (\beta_{nj}^*)^2\left(\int_{-\infty}^{c_1}\phi(z_{nj})dz_{nj} + \int_{c_2}^{\infty}\phi(z_{nj})dz_{nj}\right) - 2\beta_{nj}^*\int_{-\infty}^{c_1}\hat{\beta}_{nj}\phi(z_{nj})dz_{nj} \\
&\quad -2\beta_{nj}^*\int_{c_2}^{\infty}\hat{\beta}_{nj}\phi(z_{nj})dz_{nj} + \int_{-\infty}^{c_1}\hat{\beta}_{nj}^2\phi(z_{nj})dz_{nj} + \int_{c_2}^{\infty}\hat{\beta}_{nj}^2\phi(z_{nj})dz_{nj} \\
&\leq (\beta_{nj}^*)^2\{Q(-c_1) + 1 - Q(-c_2)\} - 2\beta_{nj}^*\int_{-\infty}^{c_1}\left(\beta_{nj}^* + z_{nj}\frac{\sigma}{\sqrt{n}}\right)\phi(z_{nj})dz_{nj} \\
&\quad -2\beta_{nj}^*\int_{c_2}^{\infty}\left(\beta_{nj}^* + z_{nj}\frac{\sigma}{\sqrt{n}} - \frac{\sigma}{\sqrt{n}}\sqrt{\alpha_n'}\right)\phi(z_{nj})dz_{nj} + \int_{-\infty}^{c_1}\left(\beta_{nj}^* + z_{nj}\frac{\sigma}{\sqrt{n}}\right)^2\phi(z_{nj})dz_{nj} \\
&\quad + \int_{c_2}^{\infty}\left(\beta_{nj}^* + z_{nj}\frac{\sigma}{\sqrt{n}}\right)^2\phi(z_{nj})dz_{nj} \\
&= 2\beta_{nj}^*\frac{\sigma}{\sqrt{n}}\sqrt{\alpha_n'}\{1 - Q(-c_2)\} + \frac{\sigma^2}{n}\left\{1 + Q(-c_1) - Q(-c_2) + \frac{c_2 e^{-c_2^2/2}}{\sqrt{2\pi}} - \frac{c_1 e^{-c_1^2/2}}{\sqrt{2\pi}}\right\} \quad (11)
\end{aligned}
$$

The slowest term converging to zero in (11) is $2\beta_{nj}^*\sigma\sqrt{\alpha_n'/n}$ which is $\mathcal{O}(\sqrt{\alpha_n'/n})$. Although the derivation is not given, $\beta_{nj}^* < 0$ case can be shown to have the same slowest term. Thus $J_1 = \mathcal{O}(\sqrt{\alpha_n'/n})$, and since J_1 converges at a slower rate than J_2,

$$E\|\boldsymbol{\beta}_n^* - \hat{\boldsymbol{\beta}}_n\|^2 = \mathcal{O}\left\{\frac{p_n\sqrt{\alpha_n'}\exp(-\alpha_n'/2)}{n}\right\} + \mathcal{O}\left(\sqrt{\frac{\alpha_n'}{n}}\right).$$

This completes the proof. $\qquad\square$

Proof of Theorem 2. The proof emerges with some modifications to the proof of Theorem 2 in Zou (2006). Here $\boldsymbol{\beta}_n^{(0)}$ denotes the least squares estimator. We first prove asymptotic normality. Let $\boldsymbol{\beta} = \boldsymbol{\beta}_n^* + \mathbf{u}/\sqrt{n}$ and

$$V_n(\mathbf{u}) = \left\{y - \sum_{j=1}^{p}x_j\left(\beta_{nj}^* + \frac{u_j}{\sqrt{n}}\right)\right\}^2 + \alpha_n^\dagger\sum_{j=1}^{p}|\beta_{nj}^* + \frac{u_j}{\sqrt{n}}|\left(|\beta_{nj}^{(0)}| + \eta_n^\dagger\right)^{-1}.$$

Let $\hat{\mathbf{u}}_n = \arg\min V_n(\mathbf{u})$, suggesting $\hat{\mathbf{u}}_n = \sqrt{n}(\boldsymbol{\beta}_n^{(1)} - \boldsymbol{\beta}_n^*)$.

$$V_n(\mathbf{u}) - V_n(\mathbf{0}) = \mathbf{u}'\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)\mathbf{u} - 2\frac{\varepsilon'\mathbf{X}}{\sqrt{n}}\mathbf{u} + \frac{\alpha_n^\dagger}{\sqrt{n}}\sum_{j=1}^{p}\left(|\beta_{nj}^{(0)}| + \eta_n^\dagger\right)^{-1}\sqrt{n}\left(\left|\beta_{nj}^* + \frac{u_j}{\sqrt{n}}\right| - |\beta_{nj}^*|\right)$$

We know that $\mathbf{X}'\mathbf{X}/n \to \mathbf{C}$ and $\varepsilon'\mathbf{X}/\sqrt{n}\overset{d}{\to}\mathbf{W}\overset{d}{=}\mathrm{N}(0, \sigma^2\mathbf{C})$. Now consider the limiting behavior of the third term. If $\beta_{nj}^* \neq 0$, then by the continuous mapping theorem $\{|\beta_{nj}^{(0)}| + \eta_n^\dagger\}^{-1}\overset{p}{\to}\{|\beta_{nj}^*| + \eta_n^\dagger\}^{-1}$ and $\sqrt{n}(|\beta_{nj}^* + u_j/\sqrt{n}| - |\beta_{nj}^*|) \to u_j\mathrm{sgn}(\beta_{nj}^*)$. By Slutsky's theorem $(\alpha_n^\dagger/\sqrt{n})\{|\beta_{nj}^{(0)}| + \eta_n^\dagger\}^{-1}\sqrt{n}(|\beta_{nj}^* + u_j/\sqrt{n}| - |\beta_{nj}^*|)\overset{p}{\to}0$. If $\beta_{nj}^* = 0$, then $\sqrt{n}(|\beta_{nj}^* + u_j/\sqrt{n}| - |\beta_{nj}^*|) = |u_j|$ and $\alpha_n^\dagger\{|\beta_{nj}^{(0)}| + \eta_n^\dagger\}^{-1}/\sqrt{n} = \alpha_n^\dagger/(\sqrt{n}|\beta_{nj}^{(0)}| + \sqrt{n}\eta_n^\dagger)$ where $\sqrt{n}\beta_{nj}^{(0)} = O_p(1)$. Again by Slutsky's theorem

$$V_n(\mathbf{u}) - V_n(0) \overset{d}{\to} \begin{cases} \mathbf{u}_\mathcal{A}'\mathbf{C}_\mathcal{A}\mathbf{u}_\mathcal{A} - 2\mathbf{u}_\mathcal{A}'\mathbf{W}_\mathcal{A} & \text{if } u_j = 0 \text{ for all } j \notin \mathcal{A} \\ \infty & \text{otherwise.} \end{cases}$$

14

$V_n(\mathbf{u}) - V_n(\mathbf{0})$ is convex and the unique minimum of the right hand side is $(\mathbf{C}_{\mathcal{A}}^{-1}\mathbf{W}_{\mathcal{A}}, \mathbf{0})'$. By epiconvergence (Geyer, 1994; Knight and Fu, 2000)

$$\mathbf{u}_{\mathcal{A}}^{(n)} \overset{d}{\to} \mathbf{C}_{\mathcal{A}}^{-1}\mathbf{W}_{\mathcal{A}}, \quad \mathbf{u}_{\mathcal{A}^c}^{(n)} \overset{d}{\to} \mathbf{0}. \tag{12}$$

Since $\mathbf{W}_{\mathcal{A}} \overset{d}{=} \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{C}_{\mathcal{A}})$, this proves the asymptotic normality.

For all $j \in \mathcal{A}$, $\beta_{nj}^{(1)} \overset{p}{\to} \beta_{nj}^*$; thus $\mathrm{pr}(j \in \mathcal{A}_n) \to 1$. Now we show that for all $j' \notin \mathcal{A}$, $\mathrm{pr}(j' \in \mathcal{A}_n) \to 0$. Consider the event $j' \in \mathcal{A}_n$. By the KKT optimality conditions, $2\mathbf{x}_{j'}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n^{(1)}) = \alpha_n^{\dagger}(|\beta_{nj'}^{(0)}| + \eta_n^{\dagger})^{-1}$. We know that $\alpha_n^{\dagger}(|\beta_{nj'}^{(0)}| + \eta_n^{\dagger})^{-1}/\sqrt{n} \overset{p}{\to} \infty$ while

$$\frac{2\mathbf{x}_{j'}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n^{(1)})}{\sqrt{n}} = 2\left\{ \frac{\mathbf{x}_{j'}'\mathbf{X}\sqrt{n}(\boldsymbol{\beta}_n^* - \boldsymbol{\beta}_n^{(1)})}{n} + \frac{\mathbf{x}_{j'}'\boldsymbol{\varepsilon}}{\sqrt{n}} \right\}.$$

By (12) and Slutsky's theorem, we know that both terms in the brackets converge in distribution to some normal suggesting

$$\mathrm{pr}\left(j' \in \mathcal{A}_n\right) \leq \mathrm{pr}\left\{ 2\mathbf{x}_{j'}'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_n^{(1)}\right) = \alpha_n^{\dagger}\left(\left|\beta_{nj'}^{(0)}\right| + \eta_n^{\dagger}\right)^{-1} \right\} \to 0,$$

which proves the consistency part. $\square$

Acknowledgements

References

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

Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer (1985).

Breiman, L. "Heuristics of Instability and Stabilization in Model Selection." *The Annals of Statistics*, 24:2350–2383 (1996).

Candes, E. J., Wakin, M. B., and Boyd, S. P. "Enhancing sparsity by reweighted l1 minimization." *Journal of Fourier Analysis and Applications*, 14:877–905 (2008).

Carvalho, C., Polson, N., and Scott, J. "Handling Sparsity via the Horseshoe." *JMLR: W&CP*, 5 (2009).

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

Cevher, V. "Learning with compressible priors." In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems*, volume 22, 261–269 (2009).

Clyde, M., Ghosh, J., and Littman, M. L. "Bayesian adaptive sampling for variable selection and model averaging." *Journal of Computational and Graphical Statistics* (2010).

Clyde, M. and Littman, M. *Bayesian model averaging using Bayesian adaptive sampling – BAS package manual* (2005).
URL www.stat.duke.edu/ clyde/BAS/BAS-manual.pdf

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. "Least Angle Regression." *The Annals of Statistics*, 32(2):407–499 (2004).

Fan, J. and Li, R. "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." *Journal of the American Statistical Association*, 96(456):1348–1360 (2001).

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

Fu, W. "Penalized regressions: The bridge versus the lasso." *Journal of Computational and Graphical Statistics*, 7:397–416 (1998).

Garrigues, P. J. "Sparse coding models of natural images: Algorithms for efficient inference and learning of higher-order structure." *PhD Thesis, University of California, Berkeley* (2009).

Geyer, C. J. "On the Asymptotics of Constrained M-Estimation." *The Annals of Statistics*, 22(4):1993–2010 (1994).

Gramacy, R. B. *Estimation for multivariate normal and Student-t data with monotone missingness – Monomvn package manual* (2010).
URL http://www.statslab.cam.ac.uk/ bobby/monomvn.html

Griffin, J. E. and Brown, P. J. "Bayesian adaptive lassos with non-convex penalization." *Technical Report* (2007).

—. "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 5(1):171–188 (2010).

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

Knight, K. and Fu, W. "Asymptotics for Lasso-Type Estimators." *The Annals of Statistics*, 28(5):1356–1378 (2000).

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. "Mixtures of *g* priors for Bayesian variable selection." *Journal of the American Statistical Association*, 103(481):410–423 (2008).

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

Pickands, J. "Statistical Inference Using Extreme Order Statistics." *The Annals of Statistics*, 3(1):119–131 (1975).

Ritter, C. and Tanner, M. A. "Facilitating the Gibbs sampler: The Gibbs stopper and the griddy-Gibbs sampler." *Journal of the American Statistical Association*, 97:861–868 (1992).

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

Tibshirani, R. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288 (1996).

Tipping, M. E. "Sparse Bayesian Learning and the Relevance Vector Machine." *Journal of Machine Learning Research*, 1 (2001).

West, M. "On Scale Mixtures of Normal Distributions." *Biometrika*, 74(3):646–648 (1987).

Yuan, M. and Lin, Y. "Efficient Empirical Bayes Variable Selection and Estimation in Linear Models." *Journal of the American Statistical Association*, 100(472):1215–1225 (2005).

Zhao, P. and Yu, B. "On Model Selection Consistency of Lasso." *J. Mach. Learn. Res.*, 7 (2006).

Zou, H. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association*, 101:1418–1429 (2006).

Zou, H. and Li, R. "One-step sparse estimates in nonconcave penalized likelihood models." *The Annals of Statistics*, 36(4):1509–1533 (2008).