Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Dirichlet–Laplace Priors for Optimal Shrinkage

Anirban Bhattacharya, Debdeep Pati, Natesh S. Pillai & David B. Dunson

To cite this article: Anirban Bhattacharya, Debdeep Pati, Natesh S. Pillai & David B. Dunson (2015) Dirichlet–Laplace Priors for Optimal Shrinkage, Journal of the American Statistical Association, 110:512, 1479-1490, DOI: 10.1080/01621459.2014.960967

To link to this article: https://doi.org/10.1080/01621459.2014.960967

Dirichlet–Laplace Priors for Optimal Shrinkage

Anirban BHATTACHARYA, Debdeep PATI, Natesh S. PILLAI, and David B. DUNSON

Penalized regression methods, such as L_1 regularization, are routinely used in high-dimensional applications, and there is a rich literature on optimality properties under sparsity assumptions. In the Bayesian paradigm, sparsity is routinely induced through two-component mixture priors having a probability mass at zero, but such priors encounter daunting computational problems in high dimensions. This has motivated continuous shrinkage priors, which can be expressed as global-local scale mixtures of Gaussians, facilitating computation. In contrast to the frequentist literature, little is known about the properties of such priors and the convergence and concentration of the corresponding posterior distribution. In this article, we propose a new class of Dirichlet–Laplace priors, which possess optimal posterior concentration and lead to efficient posterior computation. Finite sample performance of Dirichlet–Laplace priors relative to alternatives is assessed in simulated and real data examples.

KEY WORDS: Bayesian; Convergence rate; High-dimensional; L_1; Lasso; Penalized regression; Regularization; Shrinkage prior.

1. INTRODUCTION

The overwhelming emphasis in the literature on high-dimensional data analysis has been on rapidly producing point estimates with good empirical and theoretical properties. However, in many applications, it is crucial to obtain a realistic characterization of uncertainty in estimates of parameters, functions of parameters, and predictions. Usual frequentist approaches to characterize uncertainty, such as constructing asymptotic confidence regions or using the bootstrap, can break down in high-dimensional settings. For example, in regression when the number of subjects n is equal to or larger than the number of predictors p, one cannot naively appeal to asymptotic normality and resampling from the data may not provide an adequate characterization of uncertainty.

Most penalized estimators correspond to the mode of a Bayesian posterior distribution. For example, lasso/L_1 regularization (Tibshirani 1996) is equivalent to maximum a posteriori (MAP) estimation under a Gaussian linear regression model having a double exponential (Laplace) prior on the coefficients. Given this connection, it is natural to ask whether we can use the entire posterior distribution to provide a probabilistic measure of uncertainty. In addition to providing a characterization of uncertainty, a Bayesian perspective has distinct advantages in terms of tuning parameter choice, allowing key penalty parameters to be marginalized over the posterior distribution instead of relying on cross-validation.

From a frequentist perspective, we would like to be able to choose a default shrinkage prior that leads to similar optimality properties to those shown for L_1 penalization and other approaches. However, instead of showing that a penalized estimator obtains the minimax rate under sparsity assumptions, we would like to show that the entire posterior distribution concentrates at the optimal rate, that is, the posterior probability assigned to a shrinking neighborhood of the true parameter value converges to one, with the neighborhood size proportional to the frequentist minimax rate.

An amazing variety of shrinkage priors has been proposed in the Bayesian literature; however with essentially no theoretical justification in the high-dimensional settings for which they were designed. Ghosal (1999) and Bontemps (2011) provided conditions on the prior for asymptotic normality of linear regression coefficients allowing the number of predictors p to increase with sample size n, with Ghosal (1999) requiring a very slow rate of growth and Bontemps (2011) assuming $p \leq n$. These results required the prior to be sufficiently flat in a neighborhood of the true parameter value, essentially ruling out shrinkage priors. Armagan, Dunson, and Lee (2013) considered shrinkage priors in providing simple sufficient conditions for posterior consistency in linear regression where the number of variables grows slower than the sample size, though no rate of contraction was provided.

In studying posterior contraction in high-dimensions, several properties of the prior distribution are critical, including the prior concentration around sparse vectors and the implied dimensionality of the prior. Studying these properties of shrinkage priors is challenging due to the lack of exact zeros, with the prior draws being sparse in only an approximate sense. This technical hurdle has prevented any previous results on posterior concentration in high-dimensional settings for shrinkage priors. Investigating these properties is critical not just in studying frequentist optimality properties of Bayesian procedures but for Bayesians in obtaining improved insight into prior elicitation. Without such a technical handle, prior selection remains an art. Our overarching goal is to obtain theory allowing design of novel priors, which are appealing from a Bayesian perspective while having frequentist optimality properties.

Anirban Bhattacharya is Assistant Professor, Department of Statistics, Texas A&M University, College Station, TX 77843 (E-mail: *anirbanb@stat.tamu.edu*), Debdeep Pati is Assistant Professor, Department of Statistics, Florida State University, Tallahassee, FL 32306 (E-mail: *debdeep@stat.fsu.edu*), Natesh S. Pillai is Associate Professor, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: *pillas@fas.harvard.edu*), David B. Dunson is Arts and Sciences Distinguished Professor, Department of Statistical Science, Duke University, Durham, NC 27708 (E-mail: *dunson@stat.duke.edu*) The authors thank Dr. Ismael Castillo and Dr. James Scott for sharing source code. Dr. Anirban Bhattacharya, Dr. Debdeep Pati and Dr. Natesh S. Pillai acknowledge support from the Office of Naval Research (ONR BAA 14-0001). The authors would like to thank ONR Program Officer Predrag Neskovic for his interest in this work. Dr. Pillai is also partially supported by NSF-DMS 1107070. Dr. David B. Dunson is partially supported by the DARPA MSEE program and the grant number R01 ES017240-01 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH).

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

Journal of the American Statistical Association
December 2015, Vol. 110, No. 512, Theory and Methods
DOI: 10.1080/01621459.2014.960967

2. A NEW CLASS OF SHRINKAGE PRIORS

2.1 Bayesian Sparsity Priors in Normal Means Problem

For concreteness, we focus on the normal means problem (Donoho et al. 1992; Johnstone and Silverman 2004; Castillo and van der Vaart 2012); although the methods developed in this article generalize trivially to high-dimensional linear and generalized linear models. In the normal means setting, one aims to estimate an n-dimensional mean (following standard practice in this literature, we use n to denote the dimensionality and it should not be confused with the sample size) based on a single observation corrupted with iid standard normal noise:

$$y_i = \theta_i + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0,1), \quad 1 \le i \le n. \tag{1}$$

Let $l_0[q;n]$ denote the subset of $\mathbb{R}^n$ given by

$$l_0[q;n] = \{\theta \in \mathbb{R}^n : \#(1 \le j \le n : \theta_j \ne 0) \le q\}.$$

For a vector $x \in \mathbb{R}^r$, let $\|x\|_2$ denote its Euclidean norm. If the true mean θ_0 is q_n-sparse, that is, $\theta_0 \in l_0[q_n;n]$, with $q_n = o(n)$, the *squared minimax rate* in estimating θ_0 in ℓ_2 norm is $2q_n \log(n/q_n)(1+o(1))$ (Donoho et al. 1992); that is (given sequences a_n, b_n, we denote $a_n = O(b_n)$ or $a_n \lesssim b_n$ if there exists a global constant C such that $a_n \le Cb_n$ and $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$).

$$\inf_{\hat\theta} \sup_{\theta_0 \in l_0[q_n;n]} E_{\theta_0} \|\hat\theta - \theta_0\|_2^2 \asymp q_n \log(n/q_n). \tag{2}$$

In the above display, E_{θ_0} denotes an expectation with respect to an $\mathrm{N}_n(\theta_0, \mathrm{I}_n)$ density. In the presence of sparsity, one loses a logarithmic factor in the ambient dimension for not knowing the locations of the zeroes. Moreover, (2) implies that one only needs a number of replicates in the order of the sparsity to consistently estimate the mean. Appropriate thresholding/penalized estimators achieve the minimax rate (2); see Castillo and van der Vaart (2012) for detailed references.

For a subset $S \subset \{1,\dots,n\}$, let $|S|$ denote the cardinality of S and define $\theta_S = (\theta_j : j \in S)$ for a vector $\theta \in \mathbb{R}^n$. Denote $\mathrm{supp}(\theta)$ to be the *support* of θ, the subset of $\{1,\dots,n\}$ corresponding to the nonzero entries of θ. For a vector $\theta \in \mathbb{R}^n$, a natural way to incorporate sparsity is to use point mass mixture priors:

$$\theta_j \sim (1-\pi)\delta_0 + \pi g_\theta, \quad j = 1,\dots,n, \tag{3}$$

where $\pi = \Pr(\theta_j \ne 0), \mathbb{E}\{|\mathrm{supp}(\theta)| \mid \pi\} = n\pi$ is the prior guess on model size (sparsity level), and g_θ is an absolutely continuous density on $\mathbb{R}$. These priors are highly appealing in allowing separate control of the level of sparsity and the size of the signal coefficients. If the sparsity parameter π is estimated via empirical Bayes, the posterior median of θ is a minimax-optimal estimator (Johnstone and Silverman 2004), which can adapt to arbitrary sparsity levels as long as $q_n = o(n)$.

In a fully Bayesian framework, it is common to place a beta prior on π, leading to a beta-Bernoulli prior on the model size, which conveys an automatic multiplicity adjustment (Scott and Berger 2010). In a recent article, Castillo and van der Vaart (2012) established that prior (3) with an appropriate beta prior on π and suitable tail conditions on g_θ leads to a minimax optimal rate of *posterior contraction*, that is, the posterior concentrates most of its mass on a ball around θ_0 of squared radius of the order of $q_n \log(n/q_n)$:

$$E_{\theta_0} \mathbb{P}(\|\theta - \theta_0\|_2 < Ms_n \mid y) \to 1, \text{ as } n \to \infty, \tag{4}$$

where $M > 0$ is a constant and $s_n^2 = q_n \log(n/q_n)$. Narisetty et al. (2014) obtained consistency in model selection using point mass mixture priors with appropriate data-driven hyperparameters.

2.2 Global-Local Shrinkage Rules

Although point mass mixture priors are intuitively appealing and possess attractive theoretical properties, posterior sampling requires a stochastic search over an enormous space, leading to slow mixing and convergence (Polson and Scott 2010). Computational issues and consideration that many of the θ_j's may be small but not exactly zero have motivated a rich literature on continuous shrinkage priors; for some flavor refer to Park and Casella (2008), Carvalho, Polson, and Scott (2010), Griffin and Brown (2010), Hans (2011), and Armagan, Dunson, and Lee (2013). Polson and Scott (2010) noted that essentially all such shrinkage priors can be represented as global-local (GL) mixtures of Gaussians,

$$\theta_j \sim \mathrm{N}(0, \psi_j \tau), \quad \psi_j \sim f, \quad \tau \sim g, \tag{5}$$

where τ controls global shrinkage toward the origin while the local scales $\{\psi_j\}$ allow deviations in the degree of shrinkage. If g puts sufficient mass near zero and f is appropriately chosen, GL priors in (5) can intuitively approximate (3) but through a continuous density concentrated near zero with heavy tails.

GL priors potentially have substantial computational advantages over point mass priors, since the normal scale mixture representation allows for conjugate updating of θ and ψ in a block. Moreover, a number of frequentist regularization procedures such as ridge, lasso, bridge, and elastic net correspond to posterior modes under GL priors with appropriate choices of f and g. For example, one obtains a double-exponential prior corresponding to the popular L_1 or lasso penalty if f is an exponential distribution. However, many aspects of shrinkage priors are poorly understood, with the lack of exact zeroes compounding the difficulty in studying basic properties, such as prior expectation, tail bounds for the number of large signals, and prior concentration around sparse vectors. Hence, subjective Bayesians face difficulties in incorporating prior information regarding sparsity, and frequentists tend to be skeptical due to the lack of theoretical justification.

This skepticism is warranted, as it is clearly the case that reasonable seeming priors can have poor performance in high-dimensional settings. For example, choosing $\pi = 1/2$ in prior (3) leads to an exponentially small prior probability of 2^{-n} assigned to the null model, so that it becomes literally impossible to override that prior informativeness with the information in the data to pick the null model. However, with a beta prior on π, this problem can be avoided (Scott and Berger 2010). In the same vein, if one places iid $\mathrm{N}(0,1)$ priors on the entries of θ, then the induced prior on $\|\theta\|$ is highly concentrated around $\sqrt{n}$ leading to misleading inferences on θ almost everywhere. Although these are simple examples, similar *multiplicity problems* (Scott and Berger 2010) can transpire more subtly in cases where complicated models/priors are involved and hence it is

fundamentally important to understand properties of the prior and the posterior in the setting of (1).

There has been a recent awareness of these issues, motivating a basic assessment of the marginal properties of shrinkage priors for a single θ_j. Recent priors such as the horseshoe (Carvalho, Polson, and Scott 2010) and generalized double Pareto (Armagan, Dunson, and Lee 2013) are carefully formulated to obtain marginals having a high concentration around zero with heavy tails. This is well justified, but as we will see below, such marginal behavior alone is not sufficient; it is necessary to study the joint distribution of θ on $\mathbb{R}^n$. With such motivation, we propose a class of Dirichlet-kernel priors in the next subsection.

2.3 Dirichlet-Kernel Priors

Let ϕ_0 denote the standard normal density on $\mathbb{R}$. Also, let $\mathrm{DE}(\tau)$ denote a zero mean double-exponential or Laplace distribution with density $f(y) = (2\tau)^{-1}e^{-|y|/\tau}$ for $y \in \mathbb{R}$. Integrating out the local scales ψ_j's, (5) can be equivalently represented as a global scale mixture of a kernel $\mathcal{K}(\cdot)$,

$$\theta_j \overset{\text{iid}}{\sim} \mathcal{K}(\cdot\,, \tau), \quad \tau \sim g, \tag{6}$$

where $\mathcal{K}(x) = \int \psi^{-1/2}\phi_0(x/\sqrt{\psi})g(\psi)d\psi$ is a symmetric unimodal density on $\mathbb{R}$ and $\mathcal{K}(x, \tau) := \tau^{-1/2}\mathcal{K}(x/\sqrt{\tau})$. For example, $\psi_j \sim \mathrm{Exp}(1/2)$ corresponds to a double-exponential kernel $\mathcal{K} \equiv \mathrm{DE}(1)$, while $\psi_j \sim \mathrm{IG}(1/2, 1/2)$ results in a standard Cauchy kernel $\mathcal{K} \equiv \mathrm{Ca}(0, 1)$.

These choices lead to a kernel that is *bounded* in a neighborhood of zero. However, if one instead uses a half Cauchy prior $\psi_j^{1/2} \sim \mathrm{Ca}_+(0, 1)$, then the resulting horseshoe kernel (Carvalho, Polson, and Scott 2009, 2010) is unbounded with a singularity at zero. This phenomenon coupled with tail robustness properties leads to excellent empirical performance of the horseshoe. However, the joint distribution of θ under a horseshoe prior is understudied and further theoretical investigation is required to understand its operating characteristics. One can imagine that it concentrates more along sparse regions of the parameter space compared to common shrinkage priors since the singularity at zero potentially allows most of the entries to be concentrated around zero with the heavy tails ensuring concentration around the relatively small number of signals.

The above class of priors rely on obtaining a suitable kernel $\mathcal{K}$ through appropriate normal scale mixtures. In this article, we offer a fundamentally different class of shrinkage priors that alleviate the requirements on the kernel, while having attractive theoretical properties. In particular, our proposed class of Dirichlet-kernel (Dk) priors replaces the single global scale τ in (6) by a vector of scales $(\phi_1\tau, \ldots, \phi_n\tau)$, where $\phi = (\phi_1, \ldots, \phi_n)$ is constrained to lie in the $(n-1)$-dimensional simplex $\mathcal{S}^{n-1} = \{x = (x_1, \ldots, x_n)^{\mathrm{T}} : x_j \geq 0, \sum_{j=1}^n x_j = 1\}$ and is assigned a $\mathrm{Dir}(a, \ldots, a)$ prior:

$$\theta_j \mid \phi_j, \tau \sim \mathcal{K}(\cdot\,, \phi_j\tau), \quad \phi \sim \mathrm{Dir}(a, \ldots, a). \tag{7}$$

In (7), $\mathcal{K}$ is any symmetric (about zero) unimodal density with exponential or heavier tails; for computational purposes, we restrict attention to the class of kernels that can be represented as scale mixture of normals (West 1987). While previous shrinkage priors obtain marginal behavior similar to the point mass mixture priors (3), our construction aims at resembling the *joint distribution* of θ under a two-component mixture prior.

We focus on the Laplace kernel from now on for concreteness, noting that all the results stated below can be generalized to other choices. The corresponding hierarchical prior given τ,

$$\theta_j \mid \phi, \tau \sim \mathrm{DE}(\phi_j\tau), \quad \phi \sim \mathrm{Dir}(a, \ldots, a), \tag{8}$$

is referred to as a Dirichlet–Laplace prior, denoted $\theta \mid \tau \sim \mathrm{DL}_a(\tau)$.

To understand the role of ϕ, we undertake a study of the marginal properties of θ_j conditional on τ, integrating out ϕ_j. The results are summarized in Proposition 2.1.

Proposition 2.1. If $\theta \mid \tau \sim \mathrm{DL}_a(\tau)$, then the marginal distribution of θ_j given τ is unbounded with a singularity at zero for any $a < 1$. Further, in the special case $a = 1/n$, the marginal distribution is a wrapped Gamma distribution $\mathrm{WG}(\tau^{-1}, 1/n)$, where $\mathrm{WG}(\lambda, \alpha)$ has a density $f(x; \lambda, \alpha) \propto |x|^{\alpha-1}\, e^{-\lambda|x|}$ on $\mathbb{R}$.

Thus, marginalizing over ϕ, we obtain an unbounded kernel $\mathcal{K}$, so that the marginal density of $\theta_j \mid \tau$ has a singularity at 0 while retaining exponential tails. A proof of Proposition 2.1 can be found in the Appendix.

The parameter τ plays a critical role in determining the tails of the marginal distribution of θ_j's. We consider a fully Bayesian framework where τ is assigned a prior g on the positive real line and learnt from the data through the posterior. Specifically, we assume a $\mathrm{gamma}(\lambda, 1/2)$ prior on τ with $\lambda = na$. We continue to refer to the induced prior on θ implied by the hierarchical structure,

$$\theta_j \mid \phi, \tau \sim \mathrm{DE}(\phi_j\tau), \quad \phi \sim \mathrm{Dir}(a, \ldots, a), \quad \tau \sim \mathrm{gamma}(na, 1/2), \tag{9}$$

as a Dirichlet–Laplace prior, denoted $\theta \sim \mathrm{DL}_a$.

There is a frequentist literature on including a local penalty specific to each coefficient. The adaptive lasso (Zou 2006; Wang and Leng 2007) relies on empirically estimated weights that are plugged in. Leng (2010) instead sampled the penalty parameters from a posterior, with a sparse point estimate obtained for each draw. These approaches do not produce a full posterior distribution but focus on sparse point estimates.

2.4 Posterior Computation

The proposed class of DL priors leads to straightforward posterior computation via an efficient data augmented Gibbs sampler. The DL_a prior (9) can be equivalently represented as

$$\theta_j \sim \mathrm{N}(0, \psi_j\phi_j^2\tau^2),\ \psi_j \sim \mathrm{Exp}(1/2),\ \phi \sim \mathrm{Dir}(a, \ldots, a),\ \tau \sim \mathrm{gamma}(na, 1/2).$$

We detail the steps in the normal means setting noting that the algorithm is trivially modified to accommodate normal linear regression, robust regression with heavy tailed residuals, probit models, logistic regression, factor models, and other hierarchical Gaussian cases. To reduce auto-correlation, we rely on marginalization and blocking as much as possible. Our sampler cycles through (i) $\theta \mid \psi, \phi, \tau, y$, (ii) $\psi \mid \phi, \tau, \theta$, (iii) $\tau \mid \phi, \theta$, and (iv) $\phi \mid \theta$. We use the fact that the joint posterior of (ψ, ϕ, τ) is conditionally independent of y given θ. Steps (ii)–(iv) together give us a draw from the conditional distribution of $(\psi, \phi, \tau) \mid \theta$,

since

$$[\psi, \phi, \tau \mid \theta] = [\psi \mid \phi, \tau, \theta][\tau \mid \phi, \theta][\phi \mid \theta].$$

Steps (i)–(iii) are standard and hence not derived. Step (iv) is nontrivial and we develop an efficient sampling algorithm for jointly sampling ϕ. Usual one at a time updates of a Dirichlet vector lead to tremendously slow mixing and convergence, and hence the joint update in Theorem 2.1 is an important feature of our proposed prior; a proof can be found in the Appendix. Consider the following parameterization for the three-parameter generalized inverse Gaussian (giG) distribution: $Y \sim \text{giG}(\lambda, \rho, \chi)$ if $f(y) \propto y^{\lambda-1}e^{-0.5(\rho y+\chi/y)}$ for $y > 0$.

Theorem 2.1. The joint posterior of $\phi \mid \theta$ has the same distribution as $(T_1/T, \ldots, T_n/T)$, where T_j are independently distributed according to a $\text{giG}(a-1, 1, 2|\theta_j|)$ distribution, and $T = \sum_{j=1}^n T_j$.

Summaries of each step are provided below.

1. To sample $\theta \mid \psi, \phi, \tau, y$, draw θ_j independently from an $\text{N}(\mu_j, \sigma_j^2)$ distribution with
$$\sigma_j^2 = \{1 + 1/(\psi_j\phi_j^2\tau^2)\}^{-1}, \quad \mu_j = \{1 + 1/(\psi_j\phi_j^2\tau^2)\}^{-1}y.$$
2. The conditional posterior of $\psi \mid \phi, \tau, \theta$ can be sampled efficiently in a block by independently sampling $\tilde{\psi}_j \mid \phi, \theta$ from an inverse-Gaussian distribution $\text{iG}(\mu_j, \lambda)$ with $\mu_j = \phi_j\tau/|\theta_j|$, $\lambda = 1$ and setting $\psi_j = 1/\tilde{\psi}_j$.
3. Sample the conditional posterior of $\tau \mid \phi, \theta$ from a $\text{giG}(\lambda - n, 1, 2\sum_{j=1}^n |\theta_j|/\phi_j)$ distribution.
4. To sample $\phi \mid \theta$, draw $T_1, \ldots, T_n$ independently with $T_j \sim \text{giG}(a-1, 1, 2|\theta_j|)$ and set $\phi_j = T_j/T$ with $T = \sum_{j=1}^n T_j$.

3. CONCENTRATION PROPERTIES OF DIRICHLET–LAPLACE PRIORS

In this section, we study a number of properties of the joint density of the Dirichlet–Laplace prior DL_a on $\mathbb{R}^n$ and investigate the implied rate of posterior contraction (4) in the normal means setting (1). Recall the hierarchical specification of DL_a from (9). Letting $\psi_j = \phi_j\tau$ for $j = 1, \ldots, n$, a standard result (see, e.g., Lemma IV.3 of Zhou 2012) implies that $\psi_j \sim \text{gamma}(a, 1/2)$ independently for $j = 1, \ldots, n$. Therefore, (9) can be alternatively represented as (this formulation only holds when $\tau \sim \text{gamma}(na, 1/2)$ and is not true for the general $\text{DL}_a(\tau)$ class with $\tau \sim g$)

$$\theta_j \mid \psi_j \sim \text{DE}(\psi_j), \quad \psi_j \sim \text{Ga}(a, 1/2). \tag{10}$$

The formulation (10) is analytically convenient since the joint distribution factors as a product of marginals and the marginal density can be obtained analytically in Proposition 3.1. The proof follows from standard properties of the modified Bessel function (Gradshteyn and Ryzhik 1980); a proof is sketched in the Appendix.

Proposition 3.1. The marginal density Π of θ_j for any $1 \le j \le n$ is given by

$$\Pi(\theta_j) = \frac{1}{2^{(1+a)/2}\Gamma(a)}\,|\theta_j|^{(a-1)/2}\,K_{1-a}\big(\sqrt{2|\theta_j|}\big), \tag{11}$$

where

$$K_\nu(x) = \frac{\Gamma(\nu + 1/2)(2x)^\nu}{\sqrt{\pi}}\int_0^\infty \frac{\cos t}{(t^2 + x^2)^{\nu+1/2}}dt$$

is the modified Bessel function of the second kind.

Figure 1 plots the marginal density (11) to compare with common shrinkage priors.

We continue to denote the joint density of θ on $\mathbb{R}^n$ by Π, so that $\Pi(\theta) = \prod_{j=1}^n \Pi(\theta_j)$. For a subset $S \subset \{1, \ldots, n\}$, let Π_S denote the marginal distribution of $\theta_S = \{\theta_j : j \in S\} \in \mathbb{R}^{|S|}$. For a Borel set $A \subset \mathbb{R}^n$, let $\mathbb{P}(A) = \int_A \Pi(\theta)d\theta$ denote the prior probability of A, and $\mathbb{P}(A \mid y^{(n)})$ the posterior probability given data $y^{(n)} = (y_1, \ldots, y_n)$ and the model (1). Finally, let $E_{\theta_0}/P_{\theta_0}$, respectively, indicate an expectation/probability w.r.t. the $\text{N}_n(\theta_0, \text{I}_n)$ density. We now establish that under mild restrictions on $\|\theta_0\|$, the posterior arising from the DL_a prior (9) contracts at the minimax rate of convergence for an appropriate choice of the Dirichlet concentration parameter a.

Theorem 3.1. Consider model (1) with $\theta \sim \text{DL}_{a_n}$ as in (9), where $a_n = n^{-(1+\beta)}$ for some $\beta > 0$ small. Assume $\theta_0 \in l_0[q_n; n]$ with $q_n = o(n)$ and $\|\theta_0\|_2^2 \le q_n \log^4 n$. Then, with $s_n^2 = q_n \log(n/q_n)$ and for some constant $M > 0$,

$$\lim_{n\to\infty} E_{\theta_0}\mathbb{P}(\|\theta - \theta_0\|_2 < Ms_n \mid y^{(n)}) = 1. \tag{12}$$

If $a_n = 1/n$ instead, then (12) holds when $q_n \gtrsim \log n$.

A proof of Theorem 3.1 can be found in Section 6. Theorem 3.1 is the first result obtaining posterior contraction rates for a continuous shrinkage prior in the normal means setting or the closely related high-dimensional regression problem. Theorem 3.1 posits that when the parameter a in the Dirichlet–Laplace prior is chosen, depending on the sample size, to be $n^{-(1+\beta)}$ for any $\beta > 0$ small, the resulting posterior contracts at the minimax rate (2), provided $\|\theta_0\|_2^2 \le q_n \log^4 n$. Using the Cauchy–Schwartz inequality, $\|\theta_0\|_1^2 \le q_n \|\theta_0\|_2^2$ for $\theta_0 \in l_0[q_n; n]$ and the bound on $\|\theta_0\|_2$ implies that $\|\theta_0\|_1 \le q_n(\log n)^2$. Hence, the condition in Theorem 3.1 permits each nonzero signal to grow at a $(\log n)^2$ rate, which is a fairly mild assumption. Moreover, in a recent technical report, the authors showed that a large subclass of global-local priors (5) including the Bayesian lasso lead to a suboptimal rate of posterior convergence, that is, the expression in (12) converges to 0 whenever $\|\theta_0\|_2^2/q_n \to \infty$. Therefore, Theorem 3.1 indeed provides a substantial improvement over a large class of GL priors.

The choice $a_n = n^{-(1+\beta)}$ will be evident from the various auxiliary results in Section 3.1, specifically Lemma 3.2 and Theorem 3.2. The conclusion of Theorem 3.1 continues to hold when $a_n = 1/n$ under an additional mild assumption on the sparsity q_n. In Tables 1 and 2, detailed empirical results are provided with $a_n = 1/n$ as a default choice. Based on empirical experience, we find that $a = 1/n$ may have a tendency to oversshrink the signals in cases where there are several relatively small signals. In such settings, depending on the practitioner's utility function, the singularity at zero can be softened using a DL_a prior for a larger value of a. We report the results for $a = 1/2$ in Section 4, whence computational gains arise as the distribution of T_j in (iv) turns out to be inverse-Gaussian (iG), for which exact samplers are available. A fully Bayesian procedure

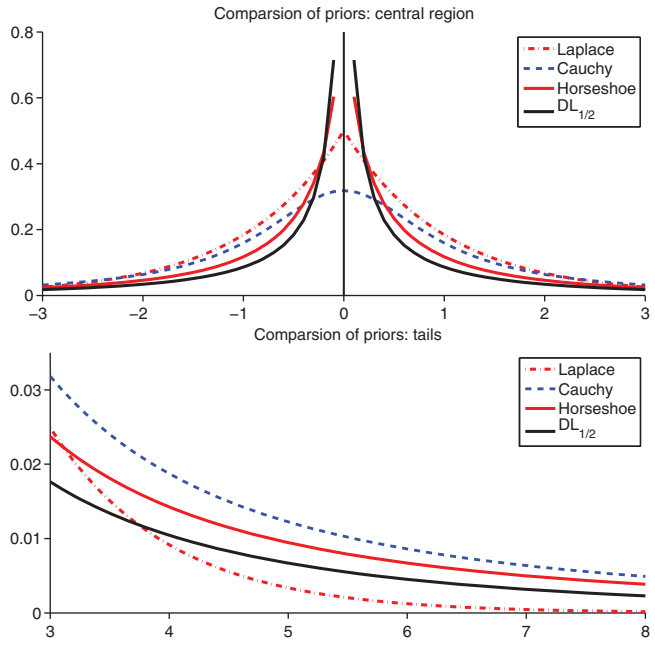

Figure 1. Marginal density of the DL prior with $a = 1/2$ in comparison to other shrinkage priors.

with a discrete hyperprior on a is considered in the real data application.

In the regression setting $y \sim \mathrm{N}_n(X\beta, \mathrm{I}_n)$ with the number of covariates $p_n \leq n$, it is typically assumed that the eigenvalues of $(X^T X)/n$ are bounded between two absolute constants (Johnson and Rossell 2012; Armagan et al. 2013). In such cases, Theorem 3.1 can be used to derive posterior convergence rates for the vector of regression coefficients β in l_2 norm. Armagan et al. (2013) studied posterior consistency in this setting for a class of shrinkage priors but did not provide contraction rates. While our methods can be extended trivially to the ultra high-dimensional regression setting, substantial work will be required to establish theoretical properties under standard restricted isometry type conditions on the design matrix commonly used to prove oracle inequalities for frequentist penalized methods (Van De Geer et al. 2009). Even with point mass mixture priors, such contraction results are unknown with the exception of a recent technical report (Castillo, Schmidt-Hieber, and van der Vaart 2014). While we propose a heuristic variable selection method, which seems to work well in practice, obtaining model selection consistency results similar to Johnson and Rossell (2012) and Narisetty et al. (2014) is challenging and will be addressed elsewhere.

3.1 Auxiliary Results

In this section, we state a number of properties of the DL prior, which provide a better understanding of the joint prior structure and also crucially help us in proving Theorem 3.1. We first provide useful bounds on the joint density of the DL prior in Lemma 3.1; a proof can be found in the Appendix.

Lemma 3.1. Consider the DL_a prior on $\mathbb{R}^n$ for a small. Let $S \subset \{1, \ldots, n\}$ and $\eta \in \mathbb{R}^{|S|}$.

If $\min_{1 \leq j \leq |S|} |\eta_j| > \delta$ for δ small, then

$$\log \Pi_S(\eta) \leq C|S| \log(1/\delta), \tag{13}$$

where $C > 0$ is an absolute constant.

Table 1. Squared error comparison over 100 replicates. Average squared error across replicates reported for DL (Dirichlet–Laplace) with $a = 1/n$ and $a = 1/2$, LS (lasso), EBMed (empirical Bayes median), PM (point mass prior), BL (Bayesian lasso), and HS (horseshoe). Signal strength $A = 5, 6$

n	100						200					
$\frac{q_n}{n}$%	5		10		20		5		10		20	
A	5	6	5	6	5	6	5	6	5	6	5	6
$\mathrm{DL}_{1/n}$	21.90	14.09	32.43	21.81	55.26	38.71	44.86	31.57	50.24	35.39	95.14	85.54
$\mathrm{DL}_{1/2}$	14.77	12.54	21.79	18.26	37.13	31.06	28.52	25.46	43.74	37.45	75.37	65.29
LS	22.21	20.67	36.41	36.54	65.91	66.60	43.37	43.51	74.05	75.98	139.20	137.94
EBMed	15.42	13.68	28.26	27.86	57.93	58.00	28.94	27.50	56.03	57.65	120.67	119.85
PM	14.50	11.99	24.97	23.66	49.92	49.16	26.66	24.86	49.96	50.89	103.40	101.69
BL	30.84	31.61	44.32	46.65	61.23	63.46	59.70	62.30	88.10	94.02	125.89	128.72
HS	12.50	9.31	21.63	17.63	39.60	35.89	23.41	18.75	42.95	38.21	81.07	74.31

Table 2. Squared error comparison over 100 replicates. Average squared error across replicates reported for DL (Dirichlet–Laplace) with $a = 1/n$ and $a = 1/2$, LS (lasso), EBMed (empirical Bayes median), PM (point mass prior), BL (Bayesian lasso), and HS (horseshoe). Signal strength $A = 7, 8$

n	100						200					
$\frac{q_n}{n}$%	5		10		20		5		10		20	
A	7	8	7	8	7	8	7	8	7	8	7	8
$DL_{1/n}$	8.20	7.19	17.29	15.35	32.00	29.40	16.07	14.28	33.00	30.80	65.53	59.61
$DL_{1/2}$	11.77	11.87	17.57	18.36	30.24	30.42	20.30	22.70	36.51	35.30	61.43	65.79
LS	21.25	19.09	38.68	37.25	68.97	69.05	41.82	41.18	75.55	75.12	137.21	136.25
EBMed	13.64	12.47	29.73	27.96	60.52	60.22	26.10	25.52	57.19	56.05	119.41	119.35
PM	12.15	10.98	25.99	24.59	51.36	50.98	22.99	22.26	49.42	48.42	101.54	101.62
BL	33.05	33.63	49.85	50.04	68.35	68.54	64.78	69.34	99.50	103.15	133.17	136.83
HS	8.30	7.93	18.39	16.27	37.25	35.18	15.80	15.09	35.61	33.58	72.15	70.23

If $\|\eta\|_2 \leq m$ for m large, then

$$-\log \Pi_S(\eta) \leq C\{|S| \log(1/a) + |S|^{3/4} m^{1/2}\}, \qquad (14)$$

where $C > 0$ is an absolute constant.

It is evident from Figure 1 that the univariate marginal density Π has an infinite spike near zero. We quantify the probability assigned to a small δ-neighborhood of the origin in Lemma 3.2.

Lemma 3.2. Assume $\theta_1 \in \mathbb{R}$ has a probability density Π as in (11). Then, for $\delta > 0$ small,

$$\mathbb{P}(|\theta_1| > \delta) \leq C \log(1/\delta)/\Gamma(a),$$

where $C > 0$ is an absolute constant.

A proof of Lemma 3.2 can be found in the Appendix.

In case of point mass mixture priors (3), the induced prior on the model size $|\text{supp}(\theta)|$ follows a Binomial(n, π) prior (given π), facilitating study of the multiplicity phenomenon (Scott and Berger 2010). However, $\mathbb{P}(\theta = 0) = 0$ for any continuous shrinkage prior, which compounds the difficulty in studying the degree of shrinkage for these classes of priors. Letting $\text{supp}_\delta(\theta) = \{j : |\theta_j| > \delta\}$ to be the entries in θ larger than δ in magnitude, we propose $|\text{supp}_\delta(\theta)|$ as an approximate measure of model size for continuous shrinkage priors. We show in Theorem 3.2 that for an appropriate choice of δ, $|\text{supp}_\delta(\theta)|$ does not exceed a constant multiple of the true sparsity level q_n with *posterior probability* tending to one, a property that we refer to as *posterior compressibility*.

Theorem 3.2. Consider model (1) with $\theta \sim DL_{a_n}$ as in (9), where $a_n = n^{-(1+\beta)}$ for some $\beta > 0$ small. Assume $\theta_0 \in l_0[q_n; n]$ with $q_n = o(n)$. Let $\delta_n = q_n/n$. Then,

$$\lim_{n\to\infty} E_{\theta_0} \mathbb{P}(\left|\text{supp}_{\delta_n}(\theta)\right| > A q_n \mid y^{(n)}) = 0, \qquad (15)$$

for some constant $A > 0$. If $a_n = 1/n$ instead, then (15) holds when $q_n \gtrsim \log n$.

The choice of δ_n in Theorem 3.2 guarantees that the entries in θ smaller than δ_n in magnitude produce a negligible contribution to $\|\theta\|$. Observe that the prior distribution of $|\text{supp}_{\delta_n}(\theta)|$ is Binomial(n, ζ_n), where $\zeta_n = \mathbb{P}(|\theta_1| > \delta_n)$. When $a_n = n^{-(1+\beta)}$, ζ_n can be bounded above by $\log n/n^{1+\beta}$ in view of Lemma 3.2 and the fact that $\Gamma(x) \geq 1/x$ for x small. Therefore, the prior expectation $\mathbb{E}\left|\text{supp}_{\delta_n}(\theta)\right| \leq \log n/n^\beta$. This implies an exponential tail bound for $\mathbb{P}(\left|\text{supp}_{\delta_n}(\theta)\right| > A q_n)$ by Chernoff's method, which is instrumental in deriving Theorem 3.2. A proof of Theorem 3.2 along these lines can be found in the Appendix.

The posterior compressibility property in Theorem 3.2 ensures that the dimensionality of the posterior distribution of θ (in an approximate sense) does not substantially overshoot the true dimensionality of θ_0, which together with the bounds on the joint prior density near zero and infinity in Lemma 3.1 delivers the minimax rate in Theorem 3.1.

4. SIMULATION STUDY

To illustrate the finite-sample performance of the proposed DL prior, we tabulate the results from a replicated simulation study with various dimensionality n and sparsity level q_n in Tables 1 and 2. In each setting, we sample 100 replicates of an n-dimensional vector y from an $N_n(\theta_0, I_n)$ distribution with θ_0 having q_n nonzero entries that are all set to be a constant $A > 0$. We chose two values of n, namely, $n = 100, 200$. For each n, we let the model size q_n to be 5%, 10%, 20% of n and vary A over 5, 6, 7, 8. This results in 24 simulation settings in total; for each setting the squared error loss corresponding to the posterior median averaged across simulation replicates is tabulated. The simulations were designed to mimic the setting in Section 3 where θ_0 is sparse with a few moderate-sized coefficients.

Based on the discussion in Section 3, we present the results for the DL prior with $a = 1/n$ and $a = 1/2$. To offer grounds for comparison, we have tabulated the results for lasso (LS), empirical Bayes median (EBMed) (the EBMed procedure was implemented using the package; Johnstone and Silverman 2005) (Johnstone and Silverman 2004), posterior median with a point mass prior (PM) (Castillo and van der Vaart 2012), posterior median corresponding to the Bayesian lasso Park and Casella (2008), and the horseshoe (Carvalho, Polson, and Scott 2010). For the fully Bayesian analysis using point mass mixture priors, we use a complexity prior on the subset-size, $\pi_n(s) \propto \exp\{-\kappa s \log(2n/s)\}$ with $\kappa = 0.1$ and independent standard Laplace priors for the nonzero entries as in Castillo and van der Vaart (2012). (Given a draw for s, a subset S of size s is drawn uniformly. Set $\theta_j = 0$ for all $j \notin S$ and draw θ_j, $j \in S$ iid from standard Laplace. The beta-Bernoulli

Table 3. Squared error comparison over 100 replicates. Average squared error for the posterior median reported for BL (Bayesian lasso), HS (horseshoe), and DL (Dirichlet–Laplace) with $a = 1/n$ and $a = 1/2$, respectively

n	1000					
A	2	3	4	5	6	7
BL	299.30	385.68	424.09	450.20	474.28	493.03
HS	306.94	353.79	270.90	205.43	182.99	168.83
$DL_{1/n}$	368.45	679.17	671.34	374.01	213.66	160.14
$DL_{1/2}$	267.83	315.70	266.80	213.23	192.98	177.20

priors in (3) induce a similar prior on the subset size.) A wide difference between most of the competitors and the proposed $DL_{1/n}$ is observed in Table 2. As evident from Tables 1 and 2, $DL_{1/2}$ is robust to smaller signal strength as is the horseshoe.

We also illustrate a high-dimensional simulation setting akin to an example in Carvalho, Polson, and Scott (2010), where one has a single observation y from an $n = 1000$ dimensional $N_n(\theta_0, I_n)$ distribution, with $\theta_0[1:10] = 10$, $\theta_0[11:100] = A$, and $\theta_0[101:1000] = 0$. We then vary A from 2 to 7 and summarize the squared error averaged across 100 replicates in Table 3. We only compare the Bayesian shrinkage priors here; the squared error for the posterior median is tabulated. Table 3 clearly illustrates the need for prior elicitation in high dimensions.

For visual illustration and comparison, we present the results from a single replicate in the first simulation setting with $n = 200$, $q_n = 10$, and $A = 7$ in Figures 2 and 3. The blue circles indicate the entries of y, while the red circles correspond to the posterior median of θ. The shaded region corresponds to a 95% pointwise credible interval for θ.

5. PROSTATE DATA APPLICATION

We consider a popular dataset (Efron 2008, 2010) from a microarray experiment consisting of expression levels for 6033 genes for 50 normal control subjects and 52 patients diagnosed with prostate cancer. The data take the form of a 6033×102 matrix with the (i, j)th entry corresponding to the expression level for gene i on patient j; the first 50 columns correspond to the normal control subjects with the remaining 52 for the cancer patients. The goal of the study is to discover genes whose expression levels *differ* between the prostate cancer patients (treatment) and normal subjects (control). A two sample t-test with 100 degrees of freedom was implemented for each gene and the resulting t-statistic t_i was converted to a z-statistic $z_i = \Phi^{-1}(T_{100}(t_i))$. Under the null hypothesis H_{0i} of no difference in expression levels between the treatment and control group for the ith gene, the null distribution of z_i is N(0, 1). Figure 4 shows a histogram of the z-values, comparing it to an N(0, 1) density with a multiplier chosen to make the curve integrate to the same area as the histogram. The shape of the histogram suggests the presence of certain interesting genes (Efron 2008).

The classical Bonferroni correction for multiple testing flags only 6 genes as significant, while the two-group empirical Bayes' method of Johnstone and Silverman (2004) found 139 significant genes, being much less conservative. The local Bayes' false discovery rate (fdr) (Benjamini and Hochberg 1995) control method identified 54 genes as nonnull. For detailed analysis of this dataset using existing methods, refer to Efron (2008, 2010).

To apply our method, we set up a normal means model $z_i = \theta_i + \epsilon_i$, $i = 1, \ldots, 6033$ and assign θ a DL_a prior. Instead of fixing a, we place a discrete uniform prior on a supported on the interval $[1/6000, 1/2]$, with the support points of the form $10(k+1)/6000$, $k = 0, 1, \ldots, K$. Such a fully Bayesian approach allows the data to dictate the choice of the tuning parameter a, which is only specified up to a constant by the theory and also avoids potential numerical issues arising from fixing $a = 1/n$ when n is large. Updating a is straightforward since the full conditional distribution of a is again a discrete distribution on the chosen support points.

A referee pointed out that the z-values for 6033 genes obtained using $n = 102$ observations are unlikely to be independent and recommended investigating robustness of our approach under correlated errors. We conducted a small simulation study to this effect to evaluate the performance of our method when the error distribution is misspecified. We generated data from the model $y_i = \theta_i + \epsilon_i$ with $(\epsilon_1, \ldots, \epsilon_n)^T \sim N_n(0, \Omega)$, where Ω corresponds to the covariance matrix of an auto-regression sequence with pure error variance σ^2 and auto-regressive coefficient ρ. We placed a discrete uniform prior on a supported on the interval $[1/1000, 1/2]$, with the support points of the form $10(k+1)/1000$, $k = 0, 1, \ldots, K$. We observe that the mean squared error of the posterior median for 50 replicated datasets with $n = 1000$, $q_n = 50$, $A = 4, 5, 6$, $\sigma^2 = 1$ increases by only 3% on average when ρ increases from 0 to 0.1. Therefore, the proposed method is robust to mild misspecification in the covariance structure. However, if further prior information is available regarding the covariance structure, that should be incorporated in the modeling framework.

For the real data application, we implemented the Gibbs sampler in Section 2.4 for 10,000 draws discarding a burn-in of 5000. Mixing and convergence of the Gibbs sampler was satisfactory based on examination of trace plots, with the 5000 retained samples having an effective sample size of 2369.2 averaged across the θ_i's. The computational time per iteration scaled approximately linearly with the dimension. The posterior mode of a was at $1/20$.

In this application, we expect there to be two clusters of $|\theta_i|$'s, with one concentrated closely near zero corresponding to genes that are effectively not differentially expressed and another away from zero corresponding to interesting genes for further study. As a simple automated approach, we cluster $|\theta_i|$'s at each Markov chain Monte Carlo (MCMC) iteration using k means with two clusters. For each iteration, the number of nonzero signals is then estimated by the smaller cluster size out of the two clusters. A final estimate (M) of the number of nonzero signals is obtained by taking the mode over all the MCMC iterations. The M largest (in absolute magnitude) entries of the posterior median are identified as the nonzero signals.

Using the above selection scheme, our method declared 128 genes as nonnull. Interestingly, out of the 128 genes, 100 are common with the ones selected by EBMed. Also all the 54 genes obtained using FDR control form a subset of the selected 128 genes. Horseshoe is overly conservative; it selected only 1 gene

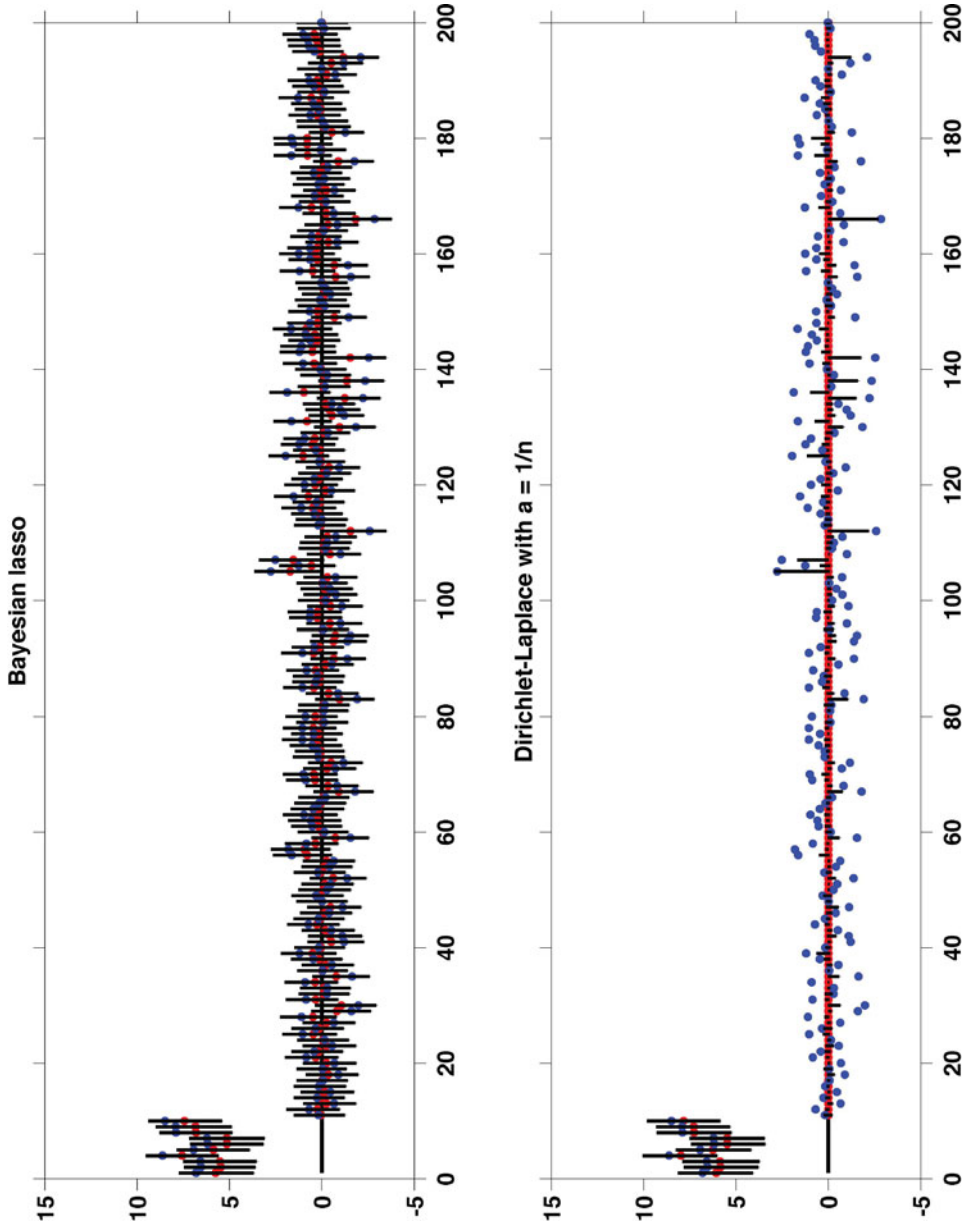

Figure 2. Simulation results from a single replicate with $n = 200$, $q_n = 10$, $A = 7$. Blue circles = entries of y, red circles = posterior median of θ, shaded region: 95% pointwise credible interval for θ. Left panel: Bayesian lasso, right panel: $\mathrm{DL}_{1/n}$ prior.

(index: 610) using the same clustering procedure; the selected gene was the one with the largest *effect size* (refer to Table 11.2 in Efron 2010).

6. PROOF OF THEOREM 3.1

We prove Theorem 3.1 for $a_n = 1/n$ in details and note the places where the proof differs in case of $a_n = n^{-(1+\beta)}$. Recall $\theta_S := \{\theta_j, j \in S\}$ and for $\delta \geq 0$, $\mathrm{supp}_\delta(\theta) := \{j : |\theta_j| > \delta\}$. Let $E_{\theta_0}/P_{\theta_0}$, respectively, indicate an expectation/probability w.r.t. the $\mathrm{N}_n(\theta_0, \mathrm{I}_n)$ density.

For a sequence of positive real numbers r_n to be chosen later, let $\delta_n = r_n/n$. Define $\mathcal{D}_n = \int \prod_{i=1}^n f_{\theta_i}(y_i)/f_{\theta_{0i}}(y_i)\, d\Pi(\theta)$. Let

$$\mathcal{A}_n = \{\mathcal{D}_n \geq e^{-4r_n^2}\mathbb{P}(\|\theta - \theta_0\|_2 \leq 2r_n)\}$$

be a subset of $\sigma(y^{(n)})$, the sigma-field generated by $y^{(n)}$, as in Lemma 5.2 of Castillo and van der Vaart (2012) such that $P_{\theta_0}(\mathcal{A}_n^c) \leq e^{-r_n^2}$. Let $\mathcal{S}_n$ be the collection of subsets $S \subset \{1, 2, \ldots, n\}$ such that $|S| \leq Aq_n$. For each such S and a positive integer j, let $\{\theta^{S,j,i} : i = 1, \ldots, N_{S,j}\}$ be a $2jr_n$ net of $\Theta_{S,j,n} = \{\theta \in \mathbb{R}^n : \mathrm{supp}_{\delta_n}(\theta) = S, 2jr_n \leq \|\theta - \theta_0\|_2 \leq 2(j+1)r_n\}$ created as follows. Let $\{\phi^{S,j,i} : i = 1, \ldots, N_{S,j}\}$ be a jr_n net of the $|S|$-dimensional ball $\{\|\phi - \theta_{0S}\| \leq 2(j+1)r_n\}$; we can choose this net in a way that $N_{S,j} \leq C^{|S|}$ for some constant C (see, e.g., Lemma 5.2 of Vershynin 2010). Letting $\theta_S^{S,j,i} = \phi^{S,j,i}$ and $\theta_k^{S,j,i} = 0$ for $k \in S^c$, we show this collection indeed forms a $2jr_n$ net of $\Theta_{S,j,n}$. To that end, fix $\theta \in \Theta_{S,j,n}$. Clearly, $\|\theta_S - \theta_{0S}\| \leq 2(j+1)r_n$. Find $1 \leq i \leq N_{S,j}$ such that $\|\theta_S^{S,j,i} - \theta_S\| \leq jr_n$. Then,

$$\left\|\theta^{S,j,i} - \theta\right\|_2^2 = \left\|\theta_S^{S,j,i} - \theta_S\right\|_2^2 + \|\theta_{S^c}\|_2^2 \leq (jr_n)^2 + (n - q_n)r_n^2/n^2 \leq 4j^2r_n^2,$$

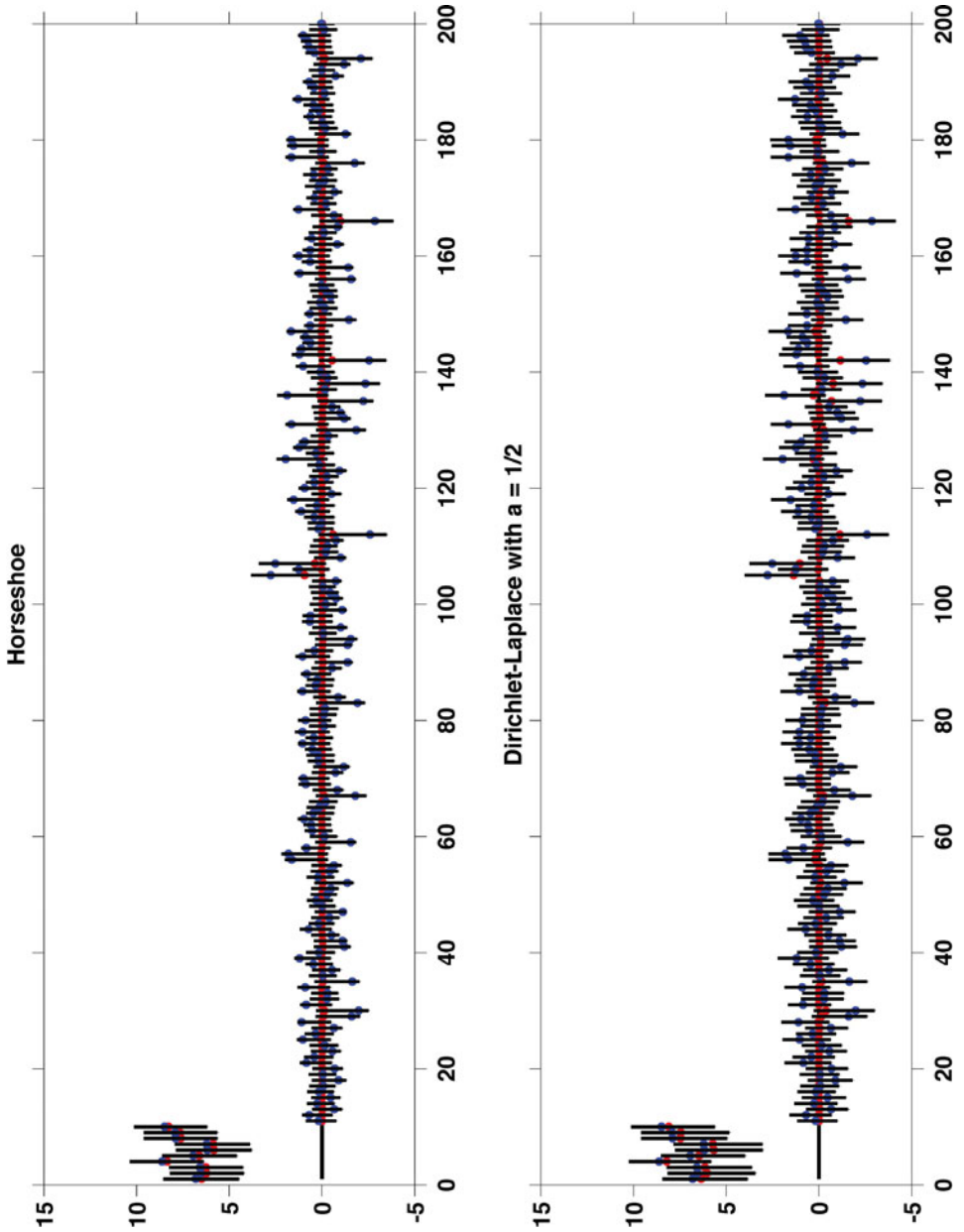

Figure 3. Simulation results from a single replicate with $n = 200$, $q_n = 10$, $A = 7$. Blue circles = entries of y, red circles = posterior median of θ, shaded region: 95% pointwise credible interval for θ. Left panel: Horseshoe, right panel: $\text{DL}_{1/2}$ prior.

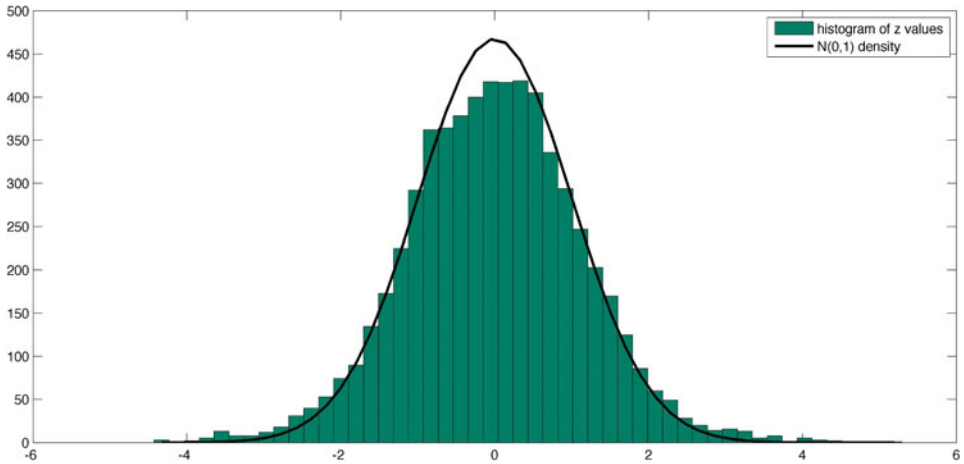

Figure 4. Histogram of z-values.

proving our claim. Therefore, the union of balls $B_{S,j,i}$ of radius $2jr_n$ centered at $\theta^{S,j,i}$ for $1 \le i \le N_{S,j}$ cover $\Theta_{S,j,n}$. Since $E_{\theta_0}\mathbb{P}(|\text{supp}_{\delta_n}(\theta)| > Aq_n \mid y^{(n)}) \to 0$ by Theorem 3.2, it is enough to work with $E_{\theta_0}\mathbb{P}(\theta : \|\theta - \theta_0\|_2 > 2Mr_n, \text{supp}_{\delta_n}(\theta) \in \mathcal{S}_n \mid y^{(n)})$. Using the standard testing argument for establishing posterior convergence rates (see, e.g., the proof of Proposition 5.1 in Castillo and van der Vaart 2012), we arrive at

$$E_{\theta_0}\mathbb{P}(\theta : \|\theta - \theta_0\|_2 > 2Mr_n, \text{supp}_{\delta_n}(\theta) \in \mathcal{S}_n \mid y^{(n)})$$
$$I_{\mathcal{A}_n} \le \sum_{S \in \mathcal{S}_n} \sum_{j \ge M} \sum_{i=1}^{N_{S,j}} 2\sqrt{\beta_{j,S,i}} e^{-Cj^2 r_n^2}, \tag{16}$$

where

$$\beta_{S,j,i} = \frac{\mathbb{P}(B_{S,j,i})}{e^{-4r_n^2}\mathbb{P}(\theta : \|\theta - \theta_0\|_2 < 2r_n)}.$$

Let $r_n^2 = q_n \log n$. The proof of Theorem 3.1 is completed by deriving an upper bound to $\beta_{S,j,i}$ in the following Lemma 6.1 akin to Lemma 5.4 in Castillo and van der Vaart (2012).

Lemma 6.1. $\log \beta_{S,j,i} \le |S| \log(2j) + C(|S| + |S_0|)\log n + C' r_n^2$.

Proof.

$$\begin{aligned}
\beta_{S,j,i} &\le \frac{\mathbb{P}\big(\theta \in \mathbb{R}^n : \text{supp}_{\delta_n}(\theta) = S, \big\|\theta_S - \tilde{\theta}_S^{S,j,i}\big\|_2 < 2jr_n\big)}{e^{-4r_n^2}\mathbb{P}(\theta \in \mathbb{R}^n : \|\theta - \theta_0\|_2 < 2r_n)} \\
&\le \frac{\mathbb{P}\big(\theta \in \mathbb{R}^n : |\theta|_j \le \delta_n \,\forall\, j \in S^c, |\theta_j| > \delta_n \,\forall\, j \in S, \big\|\theta_S - \tilde{\theta}_S^{S,j,i}\big\|_2 < 2jr_n\big)}{e^{-4r_n^2}\mathbb{P}(\theta \in \mathbb{R}^n : \|\theta_{S_0} - \theta_{0S_0}\|_2 < r_n, \|\theta_{S_0^c}\| < r_n)} \\
&\le \frac{e^{4r_n^2}\mathbb{P}(|\theta_1| < \delta_n)^{n-|S|}\mathbb{P}\big(|\theta_j| > \delta_n \,\forall\, j \in S, \big\|\theta_S - \tilde{\theta}_S^{S,j,i}\big\|_2 < 2jr_n\big)}{\mathbb{P}(|\theta_1| < \delta_n)^{n-q_n}\mathbb{P}(\|\theta_{S_0} - \theta_{0S_0}\|_2 < r_n)}.
\end{aligned} \tag{17}$$

Next, we find an upper bound to

$$R_{S,j,i} = \frac{\mathbb{P}\big(|\theta_j| > \delta_n \,\forall\, j \in S, \big\|\theta_S - \tilde{\theta}_S^{S,j,i}\big\|_2 < 2jr_n\big)}{\mathbb{P}(\|\theta_{S_0} - \theta_{0S_0}\|_2 < r_n)}. \tag{18}$$

Let $v_q(r)$ denote the q-dimensional Euclidean ball of radius r centered at zero and $|v_q(r)|$ denote its volume. For the sake of brevity, denote $v_q = |v_q(1)|$, so that $|v_q(r)| = r^q v_q$. The numerator of (18) can be clearly bounded above by $|v_{|S|}(2jr_n)| \sup_{|\theta_j| > \delta_n \,\forall\, j \in S} \Pi_S(\theta_S)$. Since the set $\{\|\theta_{S_0} - \theta_{0S_0}\|_2 < r_n\}$ is contained in the ball $v_{|S_0|}(\|\theta_{0S_0}\|_2 + r_n) = \{\|\theta_{S_0}\|_2 \le \|\theta_{0S_0}\|_2 + r_n\}$ and $\|\theta_{0S_0}\|_2 = \|\theta_0\|_2$, the denominator of (18) can be bounded below by $|v_{|S_0|}(r_n)| \inf_{v_{|S_0|}(t_n)} \Pi_{S_0}(\theta_{S_0})$, where $t_n = \|\theta_0\|_2 + r_n$. Putting together these inequalities and invoking Lemma 3.1, we have

$$R_{S,j,i} \le \frac{(2jr_n)^{|S|}\, v_{|S|} \exp\big\{C|S|\log(1/\delta_n)\big\}}{r_n^{|S_0|}\, v_{|S_0|} \exp\big\{-C(|S_0|\log n + |S_0|^{3/4} t_n^{1/2})\big\}}. \tag{19}$$

Using $v_q \asymp (2\pi e)^{q/2} q^{-q/2-1/2}$ (see Lemma 5.3 in Castillo and van der Vaart 2012) and $r_n^2 \ge q_n = |S_0|$, we can bound $\log\{r_n^{|S|} v_{|S|}/(r_n^{|S_0|} v_{|S_0|})\}$ from above by $C(|S|\log n + r_n^2)$. Therefore, we have

$$\begin{aligned}
\log R_{S,j,i} \le{}& |S|\log(2j) + C\{|S|\log n + r_n^2 + |S_0| \\
&\log n + |S|\log(1/\delta_n) + |S_0|^{3/4} t_n^{1/2}\}.
\end{aligned} \tag{20}$$

Now, since $\|\theta_0\|_2^2 \le q_n \log^4 n$ and $r_n^2 = q_n \log n$, we have $t_n \lesssim q_n^{1/2}\log^2 n$ and hence $|S_0|^{3/4} t_n^{1/2} \lesssim q_n \log n = r_n^2$. Substituting in (20), we have

$$\log R_{S,j,i} \le |S|\log(2j) + C(|S| + |S_0|)\log n + C' r_n^2.$$

Finally, $\mathbb{P}(|\theta_1| < \delta_n)^{n-|S|}/\mathbb{P}(|\theta_1| < \delta_n)^{n-q_n} \le \mathbb{P}(|\theta_1| < \delta_n)^{-|S|}$. Using Lemma 3.2, $\mathbb{P}(|\theta_1| < \delta_n) \ge (1 - \log n/n)$, which implies that $\mathbb{P}(|\theta_1| < \delta_n)^{-|S|} \le e^{\log n}$. □

Substituting the upper bound for $\beta_{S,j,i}$ obtained in Lemma 6.1, and noting that $|S| \le Aq_n$ and $|N_{S,j}| \le e^{C|S|}$, the expression in the left-hand side of (16) can be bounded above by

$$\begin{aligned}
&2\sum_{S \in \mathcal{S}_n} \sum_{j \ge M} \sum_{i=1}^{N_{S,j}} \exp\{Aq_n \log(2j)/2 + C/2(A+1)q_n \\
&\quad \log n + C' 2r_n^2/2\} e^{-Cj^2 r_n^2} \\
&\le 2\sum_{S \in \mathcal{S}_n} \sum_{j \ge M} \exp\{Cq_n + Aq_n \log(2j)/2 + C/2(A+1)q_n \\
&\quad \log n + C' r_n^2/2\} e^{-Cj^2 r_n^2}.
\end{aligned}$$

Since $|\mathcal{S}_n| \le Aq_n \binom{n}{Aq_n} \le Aq_n e^{Aq_n \log(ne/Aq_n)}$, it follows that $E_{\theta_0}\mathbb{P}(\theta : \|\theta - \theta_0\|_2 > 2Mr_n, \text{supp}_{\delta_n}(\theta) \in \mathcal{S}_n \mid y^{(n)}) \to 0$ for large $M > 0$.

When $a_n = n^{-(1+\beta)}$, the conclusion of Lemma 6.1 remains unchanged and the proof of Theorem 3.2 does not require $q_n \gtrsim \log n$. The rest of the proof remains exactly the same.

APPENDIX: Proofs of Other Results

A.1 Proof of Proposition 2.1

When $a = 1/n$, $\phi_j \sim \text{Beta}(1/n, 1 - 1/n)$ marginally. Hence, the marginal distribution of θ_j given τ is proportional to

$$\int_{\phi_j=0}^{1} e^{-|\theta_j|/(\phi_j \tau)} \left(\frac{\phi_j}{1-\phi_j}\right)^{1/n} \phi_j^{-2} d\phi_j.$$

Substituting $z = \phi_j/(1-\phi_j)$ so that $\phi_j = z/(1+z)$, the above integral reduces to

$$e^{-|\theta_j|/\tau} \int_{z=0}^{\infty} e^{-|\theta_j|/(\tau z)} z^{-(2-1/n)} dz \propto e^{-|\theta_j|/\tau} |\theta_j|^{1/n-1}.$$

In the general case, $\phi_j \sim \text{Beta}(a, (n-1)a)$ marginally. Substituting $z = \phi_j/(1-\phi_j)$ as before, the marginal density of θ_j is proportional to

$$e^{-|\theta_j|/\tau} \int_{z=0}^{\infty} e^{-|\theta_j|/(\tau z)} z^{-(2-a)} \left(\frac{1}{1+z}\right)^{na-1} dz.$$

The above integral can clearly be bounded below by a constant multiple of

$$e^{-|\theta_j|/\tau} \int_{z=0}^{1} e^{-|\theta_j|/(\tau z)} z^{-(2-a)} dz.$$

The above expression clearly diverges to infinity as $|\theta_j| \to 0$ by the monotone convergence theorem.

A.2 Proof of Theorem 2.1

Integrating out τ, the joint posterior of $\phi \mid \theta$ has the form

$$\pi(\phi_1, \ldots, \phi_{n-1} \mid \theta) \propto \prod_{j=1}^{n} \left[\phi_j^{a-1} \frac{1}{\phi_j} \right] \int_{\tau=0}^{\infty} e^{-\tau/2} \tau^{\lambda - n - 1} e^{-\sum_{j=1}^{n} |\theta_j|/(\phi_j \tau)} d\tau. \tag{A.1}$$

We now state a result from the theory of normalized random measures (see, e.g., (36) in Kruijer, Rousseau, and van der Vaart 2010). Suppose $T_1, \ldots, T_n$ are independent random variables with T_j having a density f_j on $(0, \infty)$. Let $\phi_j = T_j/T$ with $T = \sum_{j=1}^{n} T_j$. Then, the joint density f of $(\phi_1, \ldots, \phi_{n-1})$ supported on the simplex $\mathcal{S}^{n-1}$ has the form

$$f(\phi_1, \ldots, \phi_{n-1}) = \int_{t=0}^{\infty} t^{n-1} \prod_{j=1}^{n} f_j(\phi_j t) dt, \tag{A.2}$$

where $\phi_n = 1 - \sum_{j=1}^{n-1} \phi_j$. Setting $f_j(x) \propto \frac{1}{x^{\delta}} e^{-|\theta_j|/x} e^{-x/2}$ in (22), we get

$$f(\phi_1, \ldots, \phi_{n-1}) = \left[\prod_{j=1}^{n} \frac{1}{\phi_j^{\delta}} \right] \int_{t=0}^{\infty} e^{-t/2} t^{n-1-n\delta} e^{-\sum_{j=1}^{n} |\theta_j|/(\phi_j t)} dt. \tag{A.3}$$

We aim to equate the expression in (23) with the expression in (21). Comparing the exponent of ϕ_j gives us $\delta = 2 - a$. The other requirement $n - 1 - n\delta = \lambda - n - 1$ is also satisfied, since $\lambda = na$. The proof is completed by observing that f_j corresponds to a giG$(a - 1, 1, 2|\theta_j|)$ when $\delta = 2 - a$.

A.3 Proof of Proposition 3.1

By (10),

$$\begin{aligned} \Pi(\theta_j) &= \frac{(1/2)^a}{2\Gamma(a)} \int_{\psi_j=0}^{\infty} e^{-|\theta_j|/\psi_j} \psi_j^{a-2} e^{-\psi_j/2} d\psi_j \\ &= \frac{(1/2)^a}{2\Gamma(a)} \int_{z=0}^{\infty} e^{-z|\theta_j|} z^{-a} e^{-2/z} dz. \end{aligned}$$

The result follows from 8.432.7 in Gradshteyn and Ryzhik (1980).

A.4 Proof of Lemma 3.1

Letting $h(x) = \log \Pi(x)$, we have $\log \Pi_S(\eta) = \sum_{1 \le j \le |S|} h(\eta_j)$.

We first prove (13). Since $\Pi(x)$, and hence $h(x)$, is monotonically decreasing in $|x|$, and $|\eta_j| > \delta$ for all j, we have $\log \Pi_S(\eta) \le |S| h(\delta)$. Using $K_\alpha(z) \asymp z^{-\alpha}$ for $|z|$ small and $\Gamma(a) \asymp a^{-1}$ for a small, we have from (11) that $\Pi(\delta) \asymp a^{-1} |\delta|^{(a-1)}$ and hence $h(\delta) \asymp (1 - a) \log(\delta^{-1}) - \log a^{-1} + C \le C \log(\delta^{-1})$.

We next prove (14). Noting that $K_\alpha(z) \gtrsim e^{-z}/z$ for $|z|$ large (Section 9.7 of Abramowitz and Stegun 1965), we have from (11) that $-h(x) \le \log a^{-1} + 3/2 \log |x| + \sqrt{2}\sqrt{|x|}$ for $|x|$ large. Using Cauchy–Schwartz inequality twice, we have $(\sum_{j=1}^{|S|} \sqrt{|\eta_j|})^4 \le |S|^3 \|\eta\|_2^2$, which implies $\sum_{j=1}^{|S|} \sqrt{|\eta_j|} \le |S|^{3/4} \|\eta\|_2^{1/2} \le |S|^{3/4} m^{1/2}$.

A.5 Proof of Lemma 3.2

Using the representation (9), we have $\mathbb{P}(|\theta_1| > \delta \mid \psi_1) = e^{-\delta/\psi_1}$, so that,

$$\begin{aligned} &\mathbb{P}(|\theta_1| > \delta) \\ &= \frac{(1/2)^a}{\Gamma(a)} \int_0^{\infty} e^{-\delta/x} x^{a-1} e^{-x/2} dx \\ &= \frac{(1/2)^a}{\Gamma(a)} \left\{ \int_0^{4\delta} e^{-\delta/x} x^{a-1} e^{-x/2} dx + \int_{4\delta}^{\infty} e^{-\delta/x} x^{a-1} e^{-x/2} dx \right\} \\ &\le \frac{(1/2)^a}{\Gamma(a)} \left\{ C + \int_{4\delta}^{\infty} \frac{e^{-x/2}}{x} dx \right\} \\ &\le \frac{(1/2)^a}{\Gamma(a)} \left\{ C + \int_{2\delta}^{\infty} \frac{e^{-t}}{t} dt \right\}, \end{aligned} \tag{A.4}$$

where $C > 0$ is a constant independent of δ. Using a bound for the incomplete gamma function from Theorem 2 of Alzer (1997),

$$\int_{2\delta}^{\infty} \frac{e^{-t}}{t} dt \le -\log(1 - e^{-2\delta}) \le -\log(\delta), \tag{A.5}$$

for δ small. The proof is completed by noting that $(1/2)^a$ is bounded above by a constant and $C + \log(1/\delta) \le 2 \log(1/\delta)$ for δ small enough.

A.6 Proof of Theorem 3.2

For $\theta \in \mathbb{R}^n$, let $f_\theta(\cdot)$ denote the probability density function of an $\mathrm{N}_n(\theta, \mathrm{I}_n)$ distribution and f_{θ_i} denote the univariate marginal $\mathrm{N}(\theta_i, 1)$ distribution. Let $S_0 = \mathrm{supp}(\theta_0)$. Since $|S_0| = q_n$, it suffices to prove that

$$\lim_{n \to \infty} E_{\theta_0} \mathbb{P}(\left|\mathrm{supp}_{\delta_n}(\theta) \cap S_0^c\right| > A q_n \mid y^{(n)}) \to 0.$$

Let $\mathcal{B}_n = \{\left|\mathrm{supp}_{\delta_n}(\theta) \cap S_0^c\right| > A q_n\}$. By (10), $\{\theta_i, i \in S_0^c\}$ is independent of $\{\theta_i, i \in S_0\}$ conditionally on $y^{(n)}$. Hence

$$\mathbb{P}(\mathcal{B}_n \mid y^{(n)}) = \frac{\int_{\mathcal{B}_n} \prod_{i \in S_0^c} \frac{f_{\theta_i}(y_i)}{f_0(y_i)} d\Pi(\theta_i)}{\int \prod_{i \in S_0^c} \frac{f_{\theta_i}(y_i)}{f_0(y_i)} d\Pi(\theta_i)} := \frac{\mathcal{N}_n'}{\mathcal{D}_n'}, \tag{A.6}$$

where $\mathcal{N}_n'$ and $\mathcal{D}_n'$, respectively, denote the numerator and denominator of the expression in (26). Observe that

$$E_{\theta_0} \mathbb{P}(\mathcal{B}_n \mid y^{(n)}) \le E_{\theta_0} \mathbb{P}(\mathcal{B}_n \mid y^{(n)}) 1_{\mathcal{A}_n'} + P_{\theta_0}(\mathcal{A}_n'^c), \tag{A.7}$$

where $\mathcal{A}_n'$ is a subset of $\sigma(y^{(n)})$ as in Lemma 5.2 of Castillo and van der Vaart (2012) (replacing θ by $\theta_{S_0^c}$ and θ_0 by 0) defined as

$$\mathcal{A}_n' = \{\mathcal{D}_n' \ge e^{-r_n^2} \mathbb{P}(\left\|\theta_{S_0^c}\right\|_2 \le r_n)\},$$

with $P_{\theta_0}(\mathcal{A}_n'^c) \le e^{-r_n^2}$ for some sequence of positive real numbers r_n. We set $r_n^2 = q_n$ here. With this choice, from (27),

$$E_{\theta_0} \mathbb{P}(\mathcal{B}_n \mid y^{(n)}) \le \frac{\mathbb{P}(\mathcal{B}_n)}{e^{-r_n^2} \mathbb{P}(\left\|\theta_{S_0^c}\right\|_2 \le r_n)} + e^{-r_n^2}. \tag{A.8}$$

We have $\mathbb{P}(\left\|\theta_{S_0^c}\right\|_2 \le r_n) \ge \mathbb{P}(|\theta_j| < r_n/\sqrt{n} \, \forall \, j \in S_0^c) = \mathbb{P}(|\theta_1| < r_n/\sqrt{n})^{n - q_n}$, with the equality following from the representation in (10). Using Lemma 3.2, $\mathbb{P}(|\theta_1| < r_n/\sqrt{n}) \ge 1 - \log n/n$, implying $\mathbb{P}(\left\|\theta_{S_0^c}\right\|_2 \le r_n) \ge e^{-C \log n}$. Next, clearly $\mathbb{P}(\mathcal{B}_n) \le \mathbb{P}(\left|\mathrm{supp}_{\delta_n}(\theta)\right| > A q_n)$. As indicated in Section 3.1, $|\mathrm{supp}_{\delta_n}(\theta)| \sim \mathrm{Binomial}(n, \zeta_n)$, with $\zeta_n = \mathbb{P}(|\theta_1| > \delta_n) \le \log n/n$ in view of Lemma 3.2. A version of Chernoff's inequality for the binomial distribution (Hagerup and Rüb 1990) states that for $B \sim \mathrm{Binomial}(n, \zeta)$ and $\zeta \le a < 1$,

$$\mathbb{P}(B > an) \le \left\{ \left(\frac{\zeta}{a} \right)^a e^{a - \zeta} \right\}^n. \tag{A.9}$$

When $a_n = 1/n$, $q_n \ge C_0 \log n$ for some constant $C_0 > 0$. Setting $a_n = A q_n/n$, clearly $\zeta_n < a_n$ for some $A > 1/C_0$. Substituting in (29), we have $\mathbb{P}(\left|\mathrm{supp}_{\delta_n}(\theta)\right| > A q_n) \le e^{A q_n \log(e \log n) - A q_n \log A q_n}$. Choosing

$A \geq 2e/C_0$ and using the fact that $q_n \geq C_0 \log n$, we obtain $\mathbb{P}(\left|\text{supp}_{\delta_n}(\theta)\right| > Aq_n) \leq e^{-Aq_n \log 2}$. Substituting the bounds for $\mathbb{P}(\mathcal{B}_n)$ and $\mathbb{P}(\left\|\theta_{S_0^c}\right\|_2 \leq r_n)$ in (28) and choosing larger A if necessary, the expression in (27) goes to zero.

If $a_n = n^{-(1+\beta)}$, $\zeta_n \leq \log n / n^{1+\beta}$ in view of Lemma 3.2. In (29), set $a_n = Aq_n/n$ as before. Clearly $\zeta_n < a_n$. Substituting in (29), we have $\mathbb{P}(\left|\text{supp}_{\delta_n}(\theta)\right| > Aq_n) \leq e^{-CAq_n \log n}$. Substituting the bounds for $\mathbb{P}(\mathcal{B}_n)$ and $\mathbb{P}(\left\|\theta_{S_0^c}\right\|_2 \leq r_n)$ in (28), the expression in (27) goes to zero.

[Received December 2013. Revised July 2014.]

REFERENCES

Abramowitz, M., and Stegun, I. A. (1965), *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables* (Vol. 55), USA: Dover Publications. [1489]

Alzer, H. (1997), "On Some Inequalities for the Incomplete Gamma Function," *Mathematics of Computation*, 66, 771–778. [1489]

Armagan, A., Dunson, D., and Lee, J. (2013), "Generalized Double Pareto Shrinkage," *Statistica Sinica*, 23, 119–143. [1479,1480,1481]

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., and Strawn, N. (2013), "Posterior Consistency in Linear Models Under Shrinkage Priors," *Biometrika*, 100, 1011–1018. [1483]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling The False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300. [1485]

Bontemps, D. (2011), "Bernstein–Von Mises Theorems for Gaussian Regression With Increasing Number of Regressors," *The Annals of Statistics*, 39, 2557–2584. [1479]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009), "Handling Sparsity via the Horseshoe," *Journal of Machine Learning Research W&CP*, 5, 73–80. [1481]

——— (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [1480,1481,1484,1485]

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. W. (2014), "Bayesian Linear Regression With Sparse Priors." arXiv:1403.0735. [1483]

Castillo, I., and van der Vaart, A. (2012), "Needles and Straws in a Haystack: Posterior Concentration for Possibly Sparse Sequences," *The Annals of Statistics*, 40, 2069–2101. [1480,1484,1486,1488,1489]

Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992), "Maximum Entropy and the Nearly Black Object," *Journal of the Royal Statistical Society*, Series B, 54, 41–81. [1480]

Efron, B. (2008), "Microarrays, Empirical Bayes and the Two-Groups Model," *Statistical Science*, 23, 1–22. [1485]

——— (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Vol. 1), Cambridge, UK: Cambridge University Press. [1485]

Ghosal, S. (1999), "Asymptotic Normality of Posterior Distributions in High-Dimensional Linear Models," *Bernoulli*, 5, 315–331. [1479]

Gradshteyn, I. S., and Ryzhik, I. M. (1980), *Corrected and Enlarged Edition. Tables of Integrals, Series and Products,* New York: Academic Press. [1482,1489]

Griffin, J. E., and Brown, P. J. (2010), "Inference With Normal-Gamma Prior Distributions in Regression Problems," *Bayesian Analysis*, 5, 171–188. [1480]

Hagerup, T., and Rüb, C. (1990), "A Guided Tour of Chernoff Bounds," *Information Processing Letters*, 33, 305–308. [1489]

Hans, C. (2011), "Elastic Net Regression Modeling With the Orthant Normal Prior," *Journal of the American Statistical Association*, 106, 1383–1393. [1480]

Johnson, V. E., and Rossell, D. (2012), "Bayesian Model Selection in High-Dimensional Settings," *Journal of the American Statistical Association*, 107, 649–660. [1483]

Johnstone, I. M., and Silverman, B. W. (2004), "Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences," *The Annals of Statistics*, 32, 1594–1649. [1480,1484,1485]

Johnstone, I. M., and Silverman, B. W. (2005), "Ebayesthresh: R and s-Plus Programs for Empirical Bayes Thresholding," *Journal of Statistical Software*, 12, 1–38. [1484]

Kruijer, W., Rousseau, J., and van der Vaart, A. (2010), "Adaptive Bayesian Density Estimation With Location-Scale Mixtures," *Electronic Journal of Statistics*, 4, 1225–1257. [1489]

Leng, C. (2010), "Variable Selection and Coefficient Estimation via Regularized Rank Regression," *Statistica Sinica*, 20, 167–181. [1481]

Narisetty, N. N., and He, X. (2014), "Bayesian Variable Selection With Shrinking and Diffusing Priors," *The Annals of Statistics*, 42, 789–817. [1480,1483]

Park, T., and Casella, G. (2008), "The Bayesian Lasso," *Journal of the American Statistical Association*, 103, 681–686. [1480,1484]

Polson, N. G., and Scott, J. G. (2010), "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction," in *Bayesian Statistics 9*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, New York: Oxford University Press, pp. 501–538. [1480]

Scott, J. G., and Berger, J. O. (2010), "Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem," *The Annals of Statistics*, 38, 2587–2619. [1480,1484]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1479]

Van De Geer, S. A., Bühlmann, P. (2009), "On the Conditions Used to Prove Oracle Results for the Lasso," *Electronic Journal of Statistics*, 3, 1360–1392. [1483]

Vershynin, R. (2010), "Introduction to the Non-Asymptotic Analysis of Random Matrices." arxiv:1011.3027. [1486]

Wang, H., and Leng, C. (2007), "Unified Lasso Estimation by Least Squares Approximation," *Journal of the American Statistical Association*, 102, 1039–1048. [1481]

West, M. (1987), "On Scale Mixtures of Normal Distributions," *Biometrika*, 74, 646–648. [1481]

Zhou, M., and Carin, L. (2012), "Negative Binomial Process Count and Mixture Modeling." arXiv:1209.3442, accepted. [1482]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [1481]