Bayesian Analysis (2017) **12**, Number 4, pp. 1105–1131

The Horseshoe+ Estimator of Ultra-Sparse Signals

Anindya Bhadra*, Jyotishka Datta†, Nicholas G. Polson‡, and Brandon Willard§

Abstract. We propose a new prior for ultra-sparse signal detection that we term the "horseshoe+ prior." The horseshoe+ prior is a natural extension of the horseshoe prior that has achieved success in the estimation and detection of sparse signals and has been shown to possess a number of desirable theoretical properties while enjoying computational feasibility in high dimensions. The horseshoe+ prior builds upon these advantages. Our work proves that the horseshoe+ posterior concentrates at a rate faster than that of the horseshoe in the Kullback–Leibler (K-L) sense. We also establish theoretically that the proposed estimator has lower posterior mean squared error in estimating signals compared to the horseshoe and achieves the optimal Bayes risk in testing up to a constant. For one-group global–local scale mixture priors, we develop a new technique for analyzing the marginal sparse prior densities using the class of Meijer-G functions. In simulations, the horseshoe+ estimator demonstrates superior performance in a standard design setting against competing methods, including the horseshoe and Dirichlet–Laplace estimators. We conclude with an illustration on a prostate cancer data set and by pointing out some directions for future research.

MSC 2010 subject classifications: primary 62F15; secondary 62F12, 62C10.

Keywords: Bayesian, global–local shrinkage, horseshoe, horseshoe+, normal means, sparsity.

1 Introduction

Ultra-sparse signal detection provides a challenge for developing statistical estimators. In the classical normal means inference problem, we observe data from the probability model $(y_i|\theta_i) \sim \mathcal{N}(\theta_i, 1)$ for $i = 1, \ldots, n$. We wish to provide an estimator for the vector of normal means $\theta = (\theta_1, \ldots, \theta_n)$. Sparsity occurs when a large portion of the parameter vector contains zeros. The "ultra-sparse" or "nearly black" vector case occurs when the parameter vector θ lies in the set $l_0[p_n] \equiv \{\theta : \#(\theta_i \neq 0) \leq p_n\}$ with the upper bound on the number of non-zero parameter values $p_n = o(n)$ as $n \to \infty$.

To motivate the need for developing new prior distributions, consider the classic James–Stein "global" shrinkage rule, $\hat{\theta}_{JS}(y)$. This estimator uniformly dominates the

*Department of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN 47907, bhadra@purdue.edu

†Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, jd033@uark.edu

‡Booth School of Business, The University of Chicago, 5807 S. Woodlawn Ave., Chicago, IL 60637, ngp@chicagobooth.edu

§Booth School of Business, The University of Chicago, 5807 S. Woodlawn Ave., Chicago, IL 60637, brandonwillard@gmail.com

 DOI: 10.1214/16-BA1028

traditional sample mean estimator, $\hat{\theta}$. For all values of the true parameter θ and for $n > 2$, we have the classical mean squared error (MSE) risk bound:

$$R(\hat{\theta}_{JS}, \theta) := \mathbb{E}_{y|\theta}\|\hat{\theta}_{JS}(y) - \theta\|^2 < n = \mathbb{E}_{y|\theta}\|y - \theta\|^2, \quad \forall\theta.$$

However, for a sparse signal, $\hat{\theta}_{JS}(y)$ performs poorly. Suppose that the true parameter θ is an "r-spike" with r coordinates of magnitude $\sqrt{n/r}$ and the rest set at zero, giving $\|\theta\|^2 = n$. Then Johnstone and Silverman (2004) showed that the classical risk satisfies $R(\hat{\theta}_{JS}, \theta) \geq n/2$ whereas simple thresholding at $\sqrt{2\log n}$ performs with risk $\sqrt{\log n}$.

To address this issue, a "global–local" shrinkage estimator called the horseshoe estimator was proposed by Carvalho et al. (2010). The horseshoe estimator, $\hat{\theta}_{HS}(y)$, provides a Bayes rule that inherits good MSE properties of global shrinkage estimators and simultaneously provides asymptotic minimax risk for estimating sparse signals. For example, Polson and Scott (2012) showed that $\hat{\theta}_{HS}(y)$ uniformly dominates the traditional sample mean estimator in terms of MSE and van der Pas et al. (2014) showed that the horseshoe estimator has good posterior concentration properties. Specifically, the horseshoe estimator achieves

$$\sup_{\theta \in l_0[p_n]} \mathbb{E}_{y|\theta}\|\hat{\theta}_{HS}(y) - \theta\|^2 \asymp p_n \log(n/p_n),$$

which is the asymptotically minimax risk rate in ℓ_2 for nearly black objects (Donoho et al., 1992). Here $a_n \asymp b_n$ means $\lim_{n\to\infty} a_n/b_n = 1$. The "worst" $\theta \in l_0[p_n]$ is obtained at the maximum absolute difference $|\hat{\theta}_{HS}(y) - y|$ where $\hat{\theta}_{HS}(y) = \mathbb{E}_{HS}(\theta|y)$ can be interpreted as a Bayes posterior mean which is optimal under the Bayes MSE.

Though the horseshoe prior was originally designed to provide an accurate and efficient estimator of a sparse normal mean vector, it turns out that the multiple testing rule induced by the horseshoe prior also enjoys the "oracle property" in testing under the 0–1 loss (Datta and Ghosh, 2013). For the multiple testing problem in the classical two-groups model, many approaches involve explicitly modeling the ultra-sparse mean as a mixture of a point mass at zero and a heavy-tailed alternative, also known as the "spike-and-slab" approach (Mitchell and Beauchamp, 1988). This results in a posterior distribution over a high-dimensional discrete space, exploring which often leads to extreme computational cost, although the posterior mean or some other point estimates might still be found using polynomial time algorithms (Castillo and van der Vaart, 2012). The one-group global–local model, inspired by the widespread popularity of the lasso for variable selection in regression (Tibshirani, 1996), is computationally more tractable, and can be used to select a model through concentration of measure in a space of pseudo-probabilities, rather than in the n-dimensional Euclidean space (Carvalho et al., 2010; Polson and Scott, 2010; Datta and Ghosh, 2013). In particular, the horseshoe prior leads to "pseudo-posterior" probabilities that mimic the true posterior inclusion probabilities from a two-groups mixture model, and induces a multiple testing rule with attractive properties. Specifically, Datta and Ghosh (2013) proved that the Bayes risk for the horseshoe estimator attains the Bayes risk of the oracle if the global shrinkage parameter is of the same order as the proportion of sparsity using the asymptotic framework introduced by Bogdan et al. (2011). Thus, it seems natural to require

that any new sparse signal recovery prior should attain the oracle risk up to a multiplicative constant, and improve upon the error rates in theory as well as in practice. The generality of the Bayes risk results was conjectured by Datta and Ghosh (2013) and proved by Ghosh et al. (2016) in a recent unpublished manuscript. Ghosh et al. (2016) proved that asymptotic Bayes optimality holds true for a general class of shrinkage priors where the local shrinkage parameter follows a distribution with a slowly-varying component bounded away from 0 and ∞. This class of shrinkage priors includes many of the recently introduced priors such as the horseshoe, the normal–exponential–gamma (Griffin and Brown, 2010), the three-parameter beta (Armagan et al., 2011), and the generalized double Pareto (Armagan et al., 2013), among others, but this class excludes the horseshoe+ prior, since its heavier tail is slowly varying but is not bounded above.

In the light of the previous works, the purpose of our article, then, is to provide an estimator that sharpens the ability of the Bayes estimator to extract signals from sparsity while maintaining the optimal properties of the induced decision rule. We provide theoretical justifications by demonstrating that the proposed estimator has sharper information theoretic bounds and better MSE bounds compared to the horseshoe estimator. We illustrate that the horseshoe+ estimator achieves greater separation of signals and noise in a standard simulation setting and we provide a comprehensive MSE comparison with existing sparse estimators. We develop a hierarchical model which is a natural extension of the horseshoe model of Carvalho et al. (2010) and hence our terminology for the horseshoe+ hierarchical model.

The rest of the paper is outlined as follows. Section 2 motivates the class of one-group global–local shrinkage priors for sparse signal estimation as a suitable alternative to the commonly used two-groups models. Section 3 describes the horseshoe+ estimator with a particular reference to global–local shrinkage estimators. Section 4 provides theoretical properties of our proposed estimator. Our major findings can be summarized as follows:

1. The decision rule induced by the horseshoe+ prior attains the risk of Bayes oracle under 0–1 loss up to a multiplicative constant, with the constant in Bayes risk close to the constant in oracle. We also obtain a sharper bound on the probability of type-I error compared to the horseshoe prior.

2. The posterior mean squared error for the horseshoe+ estimator is always smaller than the posterior mean squared error of the horseshoe estimator in estimating a large signal.

3. The estimated sampling density using the horseshoe+ prior converges to the true density at a super-efficient rate when the true parameter value is zero, when the efficiency is calculated using the Kullback–Leibler (K-L) distance between the true density and the estimated sampling density. The upper bound of the risk for horseshoe+ is shown to be smaller than that of the horseshoe estimator using asymptotic properties of the prior utilizing Meijer-G functions (Mathai et al., 2009).

Section 5 provides comparisons of our proposed approach with other shrinkage rules using a standard design setting. We compare horseshoe+ with the Dirichlet–Laplace es-

timator (Bhattacharya et al., 2015) and the horseshoe estimator (Carvalho et al., 2010), illustrating superior performance of the horseshoe+ estimator in both estimation (under squared error loss) and testing (under 0–1 loss). Section 6 discusses the application of the proposed prior on a high-dimensional prostate cancer data set. Section 7 concludes with some directions for future research.

2 The one and two groups models

Consider the model of Section 1, i.e., $(y_i|\theta_i) \sim \mathcal{N}(\theta_i, 1)$, for $i = 1, \ldots, n$, where θ is ultra-sparse or nearly-black, in the sense that $\theta \in l_0[p_n]$. Our interest might lie in testing whether each θ_i is zero or non-zero, based on a suitably normalized test statistic or in proposing a suitable estimate $\hat{\theta}_i$, that has attractive properties, e.g., low mean squared error. The large number of parameters together with sparsity require further modeling of the data to facilitate learning via empirical Bayes or full Bayes methods. The two-groups or the spike-and-slab model (see, e.g., Mitchell and Beauchamp, 1988; Efron, 2008), provides a natural Bayesian hierarchical framework for the sparse multiple testing problem where conditionally i.i.d. θ_i are modeled as

$$\theta_i|\pi = (1-\pi)\delta_{\{0\}} + \pi\mathcal{N}(0, \psi^2), \tag{1}$$

where $\delta_{\{0\}}$ denotes a point mass at zero and the parameter $\psi^2 > 0$ is the non-centrality parameter that determines the separation between the two groups. Under this setting, the marginal distribution of $y_i|\pi$ is given by

$$y_i|\pi \sim (1-\pi)\mathcal{N}(0, 1) + \pi\mathcal{N}(0, 1+\psi^2). \tag{2}$$

As can be seen from Equation (2), the two-groups model leads to a sparse estimate, i.e., it puts exact zeros in the model. The two-groups model enjoys a number of attractive theoretical properties, detailed as follows:

1. Johnstone and Silverman (2004) showed that a thresholding-based estimator for θ under the two-groups model with an empirical Bayes estimate for π is minimax in ℓ_2 sense.
2. Castillo and van der Vaart (2012) treated a full Bayes version of the problem and again found an estimate that is minimax in ℓ_2.
3. Bogdan et al. (2011) found that the estimator under the two-groups model provides asymptotically optimal performance in testing, in the sense that its performance matches the Bayes oracle up to a constant.

Thus, while the two-groups approach is a recognized gold-standard for Bayesian sparse signal detection and estimation, a number of arguments favor an alternative approach via the one-group global–local shrinkage priors. First, in many real life applications, such as studies involving "high-dimensional, low sample size" gene expression data, the majority of the effect sizes are negligible, but not exactly zero, leading to an argument

against exact sparsity induced by the model in Equations (1)–(2) (Stephens and Balding, 2009; Guan and Stephens, 2008; Marchini and Howie, 2010; Stranger et al., 2011). From a more pragmatic point of view, the one-group global–local model leads to much faster computation, owing to the simple batch updating in the Gibbs sampler for the latent local shrinkage parameters (see, e.g., Section S.3 of the supplement of Bhadra et al., 2016a). We refer the readers to Carvalho et al. (2010) for further arguments and insights.

A useful outcome of the two-groups model is that the posterior mean $\mathbb{E}(\theta_i|y_i)$ can be written as follows:

$$\mathbb{E}(\theta_i|y_i) = \omega_i \frac{\psi^2}{1+\psi^2} y_i \approx \omega_i y_i (1 + o(1)) \quad \text{as} \quad \psi^2 \to \infty, \tag{3}$$

where $\omega_i = P(\theta_i \neq 0|y_i)$ is the posterior inclusion probability. Looking at the form of the posterior mean, one can see that it involves a global component $\psi^2/(1+\psi^2)$ that provides shrinkage towards zero for all the parameters. However, the local component ω_i allows the signal terms to escape from being too close to zero. The lack of a local shrinkage term explains why Stein-type global shrinkage estimators perform poorly in a nearly-black setting.

The key to success in a one-group model is to design a global–local shrinkage term that gives the same form of the posterior mean as in the two-groups model. The horseshoe prior of Carvalho et al. (2010) is one such one-group global–local shrinkage prior that has been shown to possess a number of theoretically attractive properties along with a considerably easier computational implementation compared to the two-groups model.

1. Carvalho et al. (2010) showed the horseshoe estimator has good information theoretic properties when the true parameter vector is sparse, in the sense that the K-L distance between the estimated and the true densities decreases at a super-efficient rate.

2. Datta and Ghosh (2013) proved that the decision rule induced by the horseshoe estimator is asymptotically Bayes optimal for multiple testing under 0–1 loss up to a multiplicative constant.

3. van der Pas et al. (2014) showed the horseshoe estimator is minimax in ℓ_2 in a nearly-black case up to a constant. The constant they have been able to achieve is at least twice as large as the minimax constant of Donoho et al. (1992).

These theoretical properties, coupled with the ease of computational implementation suggests the one-group global–local model holds considerable promise. Some other important examples of the one-group global–local model include the three-parameter beta prior (Armagan et al., 2011), the normal–exponential–gamma prior (Griffin and Brown, 2010), the generalized double Pareto prior (Armagan et al., 2013), the generalized shrinkage prior (Denison and George, 2012) and the Dirichlet–Laplace prior (Bhattacharya et al., 2015). Below we describe the one-group horseshoe hierarchical model and then proceed to propose the horseshoe+ model that leads to considerable improvements upon the horseshoe.

3 The horseshoe+ estimator

Given normally distributed data $(y_i|\theta_i) \sim \mathcal{N}(\theta_i, 1)$, the horseshoe hierarchical model is defined by the set of conditional distributions

$$\begin{aligned}(\theta_i|\lambda_i, \tau) &\sim \mathcal{N}\left(0, \lambda_i^2\right), \\ (\lambda_i|\tau) &\sim C^+(0, \tau),\end{aligned} \tag{4}$$

where C^+ denotes a half-Cauchy distributed scale parameter λ_i with density

$$p(\lambda_i|\tau) = \frac{2}{\pi\tau\{1 + (\lambda_i/\tau)^2\}}, \tag{5}$$

as discussed by Gelman (2006). The horseshoe+ hierarchical model is defined similarly by the set of conditionals

$$\begin{aligned}(\theta_i|\lambda_i, \eta_i, \tau) &\sim \mathcal{N}\left(0, \lambda_i^2\right), \\ (\lambda_i|\eta_i, \tau) &\sim C^+(0, \tau\eta_i), \\ \eta_i &\sim C^+(0, 1),\end{aligned} \tag{6}$$

where we have introduced a further half-Cauchy mixing variable η_i. In both models, the local shrinkage random effects λ_i's are not marginally independent after mixing over the global shrinkage parameter τ. The horseshoe+ model builds on the horseshoe by assuming that the λ_i's are conditionally independent given another level of local shrinkage parameters η_i's, in addition to τ. Integrating over η_i gives the density of λ_i as

$$p(\lambda_i|\tau) = \frac{4}{\pi^2\tau}\frac{\log(\lambda_i/\tau)}{(\lambda_i/\tau)^2 - 1}. \tag{7}$$

Although conceptually a natural extension, we will see that the additional $\log(\lambda_i/\tau)$ term in the numerator leads to very different properties of the proposed estimator compared to the horseshoe. There are a number of ways of dealing with the global shrinkage parameter τ. In a full Bayesian approach one can put a standard half-Cauchy prior or a Uniform$(0, 1)$ prior on τ. Another approach is to appeal to an asymptotic argument that suggests that the empirical Bayes estimator of τ to be set to $\hat{\tau} = p_n/n$, where p_n is the number of non-zero entries in θ (van der Pas et al., 2014).

To further develop the distributional properties of the horseshoe+ prior we write this as a member of the class of one-group global–local shrinkage priors with marginal prior density

$$p(\theta_i|\tau) = \int_0^\infty p(\theta_i|\lambda_i, \tau)p(\lambda_i|\tau)d\lambda_i.$$

Transforming to a shrinkage scale with $\kappa_i = 1/(1 + \lambda_i^2\tau^2)$ yields

$$p(\theta_i|\tau) = \int_0^1 p(\theta_i|\kappa_i, \tau)p(\kappa_i|\tau)d\kappa_i, \quad \text{with} \quad p(\theta_i|\kappa_i, \tau) \sim \mathcal{N}\left(0, \frac{1 - \kappa_i}{\kappa_i}\right),$$

where $\kappa_i \in [0,1]$ is a shrinkage weight. The corresponding ultra-sparse Bayes estimator is

$$\hat{\theta}_i = \mathbb{E}(\theta_i|y_i,\tau) = (1 - \mathbb{E}(\kappa_i|y_i,\tau))y_i, \tag{8}$$

where we need to compute $\mathbb{E}(\kappa_i|y_i,\tau)$. By comparing the expression for the posterior mean for θ_i for the one-group global–local model given by Equation (8) to the two-groups model given by Equation (3), it is apparent that the quantity $\hat{\omega}_i = 1-\mathbb{E}(\kappa_i|y_i,\tau)$ behaves as the posterior inclusion probability $P(\theta_i \neq 0|y_i)$. This results in a natural threshold for simultaneously testing $H_{0i} : \theta_i = 0$ vs. $H_{1i} : \theta_i \neq 0$ for $i = 1,\ldots,n$. We will consider the following multiple testing procedure proposed by Carvalho et al. (2010), and later shown to be optimal under 0–1 loss by Datta and Ghosh (2013), for the horseshoe prior:

$$\text{Reject } H_{0i} : \text{ if } 1 - \mathbb{E}(\kappa_i|y_i,\tau) > \frac{1}{2}. \tag{9}$$

3.1 Shrinkage profile

Note that the marginal data likelihood is $p(y_i|\kappa_i,\tau) = \kappa_i^{1/2}\exp(-\kappa_i y_i^2/2)$. Signals are identified when $\kappa_i \to 0$ and sparsity occurs when $\kappa_i \to 1$ in the posterior. We see that there are no shrinkage factors in the marginal likelihood to "help" identify signals in the normal model as $p(y_i|\kappa_i,\tau) \to 0$ as $\kappa_i \to 0$. This is precisely why the normal prior performs poorly for sparse settings. The horseshoe prior was designed to cancel the factor $\kappa_i^{1/2}$ and to simultaneously place prior mass at $\kappa_i = 1$ to introduce shrinkage (see Carvalho et al. (2010) for further discussion). The priors on the local shrinkage factor λ_i and the induced prior on κ_i for the horseshoe, the horseshoe+ and the generalized double Pareto prior are summarized in Table 1.

Prior for θ_i	Prior for λ_i	Prior for κ_i
GDP	$\frac{\sqrt{2}}{(\lambda_i^2)}\int_0^\infty \exp\left(\sqrt{\frac{2u}{\lambda_i^2}} - u\right)\sqrt{u}\mathrm{d}u$	$\frac{1}{2(1-\kappa_i)^2}\left[\frac{\sqrt{\pi}\exp\left\{\frac{\kappa_i}{2(1-\kappa_i)}\right\}Erfc\left\{\sqrt{\frac{\kappa_i}{2(1-\kappa_i)}}\right\}}{\sqrt{2\kappa_i(1-\kappa_i)}} - 1\right]$
Horseshoe	$2/\left\{\pi\tau(1+(\lambda_i/\tau)^2)\right\}$	$\frac{\tau}{\sqrt{\kappa_i(1-\kappa_i)}}\frac{1}{(1+\kappa_i(\tau^2-1))}$
Horseshoe+	$4\log(\lambda_i/\tau)/\left\{\pi^2\tau((\lambda_i/\tau)^2-1)\right\}$	$\frac{\tau}{\sqrt{\kappa_i(1-\kappa_i)}}\frac{\log\left\{(1-\kappa_i)/\kappa_i\tau^2\right\}}{(1-\kappa_i(\tau^2+1))}$

Table 1: Priors for λ_i and κ_i for some one-group global–local shrinkage rules.

The main difference between horseshoe+ and the others is in the extra Jacobian term introduced in the representation on the shrinkage scale. This term has a fundamentally different behavior for separating signals ($\kappa_i = 0$) from the noise terms ($\kappa_i = 1$). The horseshoe+ prior introduces another horseshoe U-shaped Jacobian factor that pushes posterior mass to the places of most interest, $\kappa_i = 0, 1$. This provides horseshoe+ prior with an additional power to detect signals in the ultra sparse signal case. Figure 1 plots

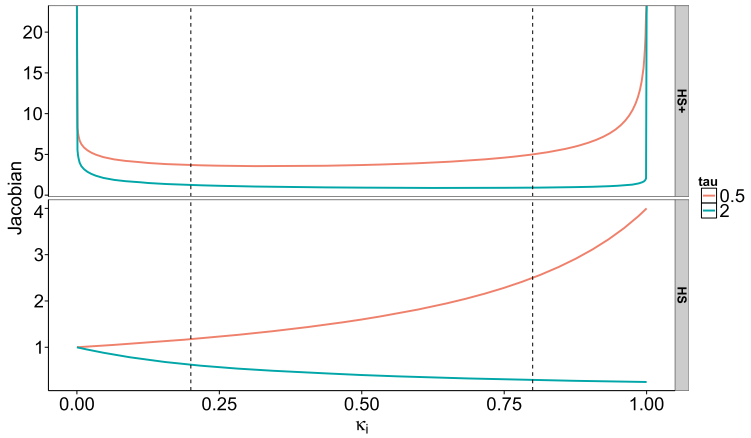

Figure 1: The horseshoe+ (top) and horseshoe (bottom) prior Jacobian terms against κ_i for $\tau = 0.5$ and 2. The vertical lines are at $\kappa = 1/(1+\tau^2)$.

the Jacobians of the horseshoe and horseshoe+ priors with τ set to 0.5 and 2 to make the difference explicit. The horseshoe Jacobian displays unequal shrinkage behavior near the two extremities of κ_i.

This extra shrinkage through the Jacobian is an unique property of the horseshoe+ prior not shared by any of the other univariate shrinkage priors. To see this, note that the 'sensible' priors can be expressed in terms of a slowly varying function following Theorem 1 of Polson and Scott (2010):

$$p(\lambda_i^2) \propto (\lambda_i^2)^{(-a-1)} L(\lambda_i^2) \text{ for } \tau^2 = 1, \tag{10}$$

$$p(\kappa_i) \propto (1-\kappa_i)^{(-a-1)} \kappa_i^{(a-1)} L(1/\kappa_i - 1), \tag{11}$$

where $L(\cdot)$ is a slowly varying function with the property $L(ty)/L(y) \to 1$ as $y \to \infty$. In a recent unpublished manuscript, Ghosh et al. (2016) showed that the popular shrinkage priors like the three-parameter beta (TPB, Armagan et al. (2011)), which includes the popular Strawderman–Berger prior, the horseshoe prior and the normal–exponential–gamma prior, as well as the generalized double Pareto (GDP, Armagan et al. (2011)) prior fall into this class. Furthermore, the authors proved that the slowly varying component of (10) is bounded as $\lambda_i^2 \to \infty$ for these popular shrinkage rules, i.e. $\lim_{\lambda_i^2 \to \infty} L(\lambda_i^2) \in (0, \infty)$ for priors such as TPB and GDP. This is where the horseshoe+ prior stands out from the rest, as the slowly-varying component for the prior density $p_{HS+}(\lambda_i^2)$ is unbounded as $\lambda \to \infty$, i.e.

$$\lim_{\lambda_i^2 \to \infty} L_{HS+}(\lambda_i^2) = \lim_{\lambda_i^2 \to \infty} \log(\lambda_i^2) \left(1 - \frac{1}{\lambda_i^2}\right)^{-1} \to \infty.$$

Since $\kappa_i \to 0$ as $\lambda_i^2 \to \infty$, the unboundedness of $L_{HS+}(\lambda_i^2) \equiv L(1/\kappa_i - 1)$ together with (11) implies that the extra shrinkage at $\lim_{\kappa_i \to 0} p(\kappa_i) \to \infty$ only holds for the

Horseshoe+ prior among all shrinkage priors expressible as heavy-tailed Gaussian scale mixtures. The Jacobian term can also be interpreted on the shrinkage scale. Specifically, for $\kappa = 1/(1+\tau^2)$, we have

$$p(\kappa_1, \ldots, \kappa_p | \kappa, y) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{1-\kappa_i}} \exp\left\{-\kappa_i \frac{y_i^2}{2}\right\} \frac{|\log\left((1-\kappa_i^{-1})/(1-\kappa^{-1})\right)|}{|\kappa - \kappa_i|}.$$

This representation shows that the horseshoe+ prior allows differential shrinkage for κ_i around κ (and is continuous at $\kappa_i = \kappa$), and suggests that the global shrinkage parameter τ^2 can also be interpreted as a scaling factor for the shrinkage weights κ_i.

4 Theoretical properties of the horseshoe+ estimator

In this section we establish a few theoretical properties for the proposed prior and the resulting posterior, from both a decision theoretic and information theoretic viewpoint. We present our main results in the form of seven theorems. Proofs and technical details are given in Supplementary Sections S.1–S.7 (Bhadra et al., 2016b).

4.1 Marginal density for the horseshoe+ prior

We start by formally establishing that the marginal prior density for horseshoe+ is unbounded at the origin.

Theorem 1. *Assume $\tau^2 = 1$. Then the marginal density of the horseshoe+ prior, $p_{HS+}(\theta)$, satisfies the following properties:*

1. $$\frac{1}{\pi^2\sqrt{2\pi}} \log\left(1 + \frac{4}{\theta^2}\right) < p_{HS+}(\theta) \leq \frac{1}{\pi^2|\theta|},$$

2. $$\lim_{|\theta| \to 0} p_{HS+}(\theta) = \infty.$$

A proof is given in Supplementary Section S.1. Figures 2 and 3 show the behavior of several one-group global–local shrinkage priors near the origin and at the tails. The priors considered here are: horseshoe+, horseshoe (Carvalho et al., 2010), Dirichlet–Laplace (Bhattacharya et al., 2015), generalized double Pareto (Armagan et al., 2013), standard Cauchy, and standard Laplace (double-exponential). Note that the horseshoe+, horseshoe and Dirichlet–Laplace densities are unbounded near the origin. Perhaps more importantly, horseshoe+ puts more mass compared to the horseshoe in a small neighborhood of the origin and has heavier tails compared to both horseshoe and Dirichlet–Laplace. Carvalho et al. (2010) established that a prior with unbounded density near the origin leads to super-efficiency in density estimation in a sparse signal setting. Due to Theorem 1, the horseshoe+ estimator enjoys the resultant advantages, as we shall show in Section 4.3.

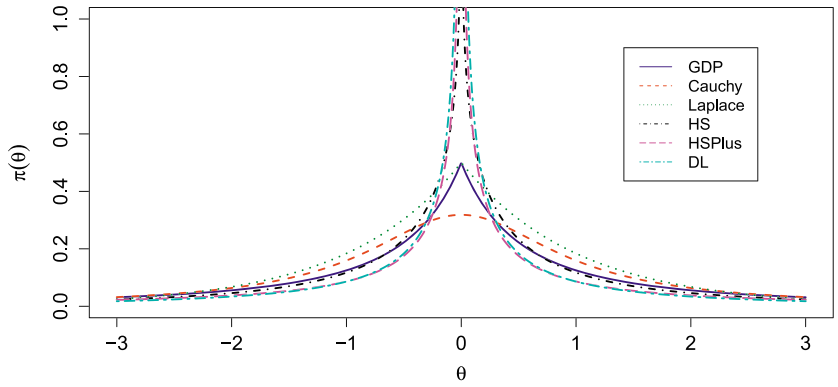

Figure 2: Marginal prior densities near the origin. The legends denote the horseshoe+ (HSPlus), horseshoe (HS), Dirichlet–Laplace (DL), generalized double Pareto (GDP), Cauchy and Laplace priors.

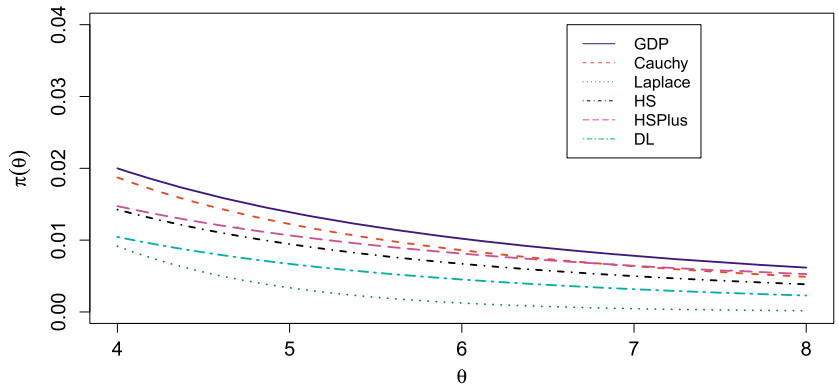

Figure 3: Marginal prior densities in the tail regions. The legends denote the horseshoe+ (HSPlus), horseshoe (HS), Dirichlet–Laplace (DL), generalized double Pareto (GDP), Cauchy and Laplace priors.

4.2 Asymptotic Bayes optimality under sparsity

Datta and Ghosh (2013) proved that the Bayes risk optimality for the horseshoe prior leverages the fact that the shrinkage weight $1-\hat{\kappa}_i$ concentrates near one (uniformly in y_i) if the global shrinkage parameter $\tau \to 0$, and concentrates near zero if $|y_i| \to \infty$ for any fixed τ in $(0,1)$. To attain the Bayes risk of the oracle, one additionally needs the global shrinkage parameter τ to adapt to the underlying proportion of non-zero effects μ_n, i.e. $\lim_{n\to\infty} \tau\mu_n^{-1} \in (0,\infty)$, where $\mu_n = \#\{\theta_i \neq 0\}/n$. It turns out that similar concentration inequalities, but with sharper bounds, hold for the posterior distribution of κ_i under the new horseshoe+ prior. At an intuitive level, this suggests that the decision rule induced by the horseshoe+ prior will also inherit the same, if not better,

optimality properties. In this section, we state two posterior concentration inequalities along with the asymptotic type-I and type-II error probabilities to establish the oracle property for horseshoe+.

Below, we briefly describe the notion of Bayes oracle in the context of multiple testing following the asymptotic framework of Bogdan et al. (2011). Assume the two-groups model of Equations (1)–(2). The optimal Bayes rule under a 0–1 additive loss for testing $H_{0i} : \theta_i = 0$ vs. $H_{1i} : \theta_i \neq 0$ is given by:

$$\text{Reject } H_{0i} \text{ if } |y_i| > C,$$

where,

$$C^2 = C^2_{\psi,f} = \frac{1+\psi^2}{\psi^2}\left(\log(\psi^2+1) + 2\log f\right) \text{ where } f = \frac{1-\mu}{\mu}. \tag{12}$$

We call this rule the *Bayes oracle* as the risk for this is the lower bound of $(1/n)$ times the risk for any multiple testing procedure under the two-groups model. Bogdan et al. (2011) further re-parametrized this by $u = \psi^2$ and $v = uf^2$, to obtain the following simpler form for the threshold in the oracle:

$$C^2 = \left(1 + \frac{1}{u}\right)\left(\log v + \log\left(1 + \frac{1}{u}\right)\right). \tag{13}$$

For maintaining clarity of notations and preserving correspondence with the original work of Bogdan et al. (2011), we use the same asymptotic framework in form of the following assumption:

Assumption 1. *The sequence of vectors $\gamma_n = (\psi_n, \mu_n)$ satisfies the following conditions:*

$$\mu_n \to 0; u_n \doteq \psi_n^2 \to \infty; v_n \doteq u_n f_n^2 \doteq \psi_n^2 \left(\frac{1-\mu_n}{\mu_n}\right)^2 \to \infty;$$
$$\frac{\log v_n}{u_n} \to C \in (0, \infty) \text{ as } n \to \infty.$$

Remark 1. *The asymptotic framework provides a natural way to study the properties of the Bayes risk as the parameter vector $\gamma = (\psi, \mu)$ defining the Bayes oracle in Equation (12) varies through an infinite sequence indexed by the number of tests n increasing to infinity. To reduce notational complexity, we will suppress the index n from $\gamma_n, \mu_n, \tau_n, \psi_n$ throughout the remainder of this section. The statements such as $\mu \to 0$ should imply that $\mu_n \to 0$ as $n \to \infty$.*

Remark 2. *It should be pointed out that the conditions are not restrictive, and are in fact minimal conditions for optimality in some sense. On one hand, the Bayes Oracle is no better than a coin toss if the limit $\psi^2/2\log(1/\mu) \to \infty$, making the test powerless, and has zero type-II error when the limit goes to zero, which could happen if one has an infinite number of replicates. The interesting cases are obtained for a finite, non-zero limit which Bogdan et al. (2011) term as "verge of detectability" and our results pertain to this situation.*

Under Assumption 1, the type-I and type-II error probabilities of the Bayes oracle are given by Bogdan et al. (2011):

$$\begin{aligned} t_1^{\mathrm{BO}} &= \mathrm{e}^{-C/2}\sqrt{\frac{2}{\pi v \log v}}(1+o_n), \\ t_2^{\mathrm{BO}} &= (2\Phi(\sqrt{C})-1)(1+o_n), \\ R_{opt} &= n\left((1-\mu)t_1^{BO}+\mu t_2^{BO}\right) = n\mu(2\Phi(\sqrt{C})-1)(1+o_n), \end{aligned} \tag{14}$$

where o_n denotes an infinite sequence of terms, indexed by n (the number of tests), converging to zero as $n \to \infty$. The last expression follows from the fact that the Bayes risk for a fixed-threshold multiple testing rule is given by $R = n((1-\mu)t_1 + \mu t_2)$ for an additive 0–1 loss, when t_1, t_2 denote the type-I and type-II error probabilities respectively. A decision rule is said to attain the asymptotic Bayes optimality under sparsity (or, ABOS) if the ratio of the Bayes risk of the decision rule to the risk of the Bayes oracle (Equation (14)) goes to 1 as multiplicity $n \to \infty$. Now, we present the first concentration inequality on the posterior distribution of κ_i providing the conditions under which the posterior mass of κ_i concentrates near one. We show that an upper bound to the he posterior mass of $\kappa_i \in (0, \epsilon)$, decays as τ^2.

Theorem 2. *Suppose we have observations $y_1, \ldots, y_n$ where $y_i \sim \mathcal{N}(\theta_i, 1)$, for $i = 1, \ldots, n$, and the prior on θ_i is distributed as horseshoe+ with the hierarchical model given by (6). Then the posterior distribution of $\kappa_i = (1+\lambda_i^2\tau^2)^{-1}$ given y_i and τ satisfies the following:*

$$\mathbb{P}(\kappa_i < \epsilon | y_i, \tau) \leq e^{\frac{y_i^2}{2}}\tau^2\epsilon(1-\epsilon)^{-2}, \tag{15}$$

for any fixed $\epsilon \in (0,1)$, and any $\tau \in (0,1)$.

The proof is given in Supplementary Section S.2. Theorem 2 implies that the posterior distribution of κ_i given τ and the observation y_i would converge to a point mass one if $\tau \to 0$. This leads to the following bound on the probability of type-I error rate for horseshoe+ prior, with proof given in Supplementary Section S.3.

Theorem 3. *Suppose we have observations $y_1, \ldots, y_n$ from the 'two-groups' model in Equation (2), and we want to test $H_{0i} : \theta_i = 0$ vs. $H_{1i} : \theta_i \neq 0$, using the decision rule of Equation (9) induced by the horseshoe+ prior. Suppose furthermore that Assumption 1 holds for the parameter vector (ψ, μ), then the probability of type-I error for horseshoe+ decision rule is given by:*

$$t_1 \leq \sqrt{\frac{2}{\pi}}\frac{\tau^2}{\sqrt{\log(1/2\tau)}}(1+o(1)).$$

Remark 3. *It should be noted that one of the bounds (and the type-I error rate) obtained for the horseshoe+ prior are sharper than that obtained for the horseshoe prior. Theorem 2 shows $\mathbb{P}_{HS+}(\kappa_i < \epsilon | y_i, \tau) = O(\tau^2)$ whereas Datta and Ghosh (2013) obtained $\mathbb{P}_{HS}(\kappa_i < \epsilon | y_i, \tau) = O(\tau)$. This relative gain will not affect the asymptotic order of the total Bayes risk derived here, but this result has interesting implications (e.g. lower false positives) nonetheless.*

We now present the second concentration inequality in the other direction, with a proof in Supplementary Section S.4.

Theorem 4. *Suppose we have observations $y_1, \ldots, y_n$ where $y_i \sim \mathcal{N}(\theta_i, 1)$, for $i = 1, \ldots, n$, and the prior on θ_i is distributed as horseshoe+ with the hierarchical model given by Equation (6). Then the posterior distribution of $\kappa_i = (1 + \lambda_i^2\tau^2)^{-1}$ given y_i and τ satisfies the following:*

$$\mathbb{P}(\kappa_i > \eta | y_i, \tau) \leq e^{-\eta(1-\delta)\frac{y_i^2}{2}} \frac{1}{\tau^2} C(\eta, \delta), \tag{16}$$

for any fixed $\eta \in (0,1)$, any fixed $\delta \in (0, 1/\eta(1+\tau^2))$ and uniformly in $y_i \in \mathbb{R}$, where $C(\eta, \delta)$ is a constant independent of y_i.

A corollary of Theorem 4 is that the posterior distribution of κ_i given τ and y_i would converge to a point mass at zero if $|y_i| \to \infty$.

A crucial step for proving the optimality for the horseshoe prior is the choice of the global shrinkage parameter τ. Datta and Ghosh (2013) chose τ to be of the same order as the proportion of signals μ, i.e. $\tau = \tau_n = O(\mu_n)$. They also argued that the optimality of the decision rule induced by the horseshoe prior depends on how well the sparsity is captured in the hyper-parameter τ. This was further supported by van der Pas et al. (2014) who showed that the condition $\tau = O(\mu)$ is a sufficient condition for the minimaxity properties of the horseshoe estimator. Since the role of τ as a global scale parameter for the prior on local shrinkage parameters λ_i does not change with the horseshoe+ prior, intuitively the same choice on τ would lead to the optimal type-II error rates. Under this choice of τ, it follows that the type-II error for horseshoe+ decision rule has the same asymptotic order as that of the type-II error rate for the Bayes oracle. Let C denote the constant in the expression for the risk of the Bayes oracle as appears in Equation (13). Then it follows from Theorem 4 that the type-II error rate has the following upper bound:

Theorem 5. *Suppose we have observations $y_1, \ldots, y_n$ from the 'two-groups' model in Equation (2), and wish to test $H_{0i} : \theta_i = 0$ vs. $H_{1i} : \theta_i \neq 0$, using the decision rule of Equation (9). Suppose furthermore that Assumption 1 holds for the parameter vector (ψ, μ), and the global shrinkage parameter τ decreases to zero such that $\tau = O(\mu)$. Then for all $\eta \in (0,1)$ and $\delta \in (0, 1/\eta(1+\tau^2))$, the probability of type-II error of the decision rules induced by the horseshoe+ prior is bounded above by:*

$$t_2 \leq \left(2\Phi\left(\sqrt{\frac{2}{\eta(1-\delta)}}\sqrt{C}\right) - 1 \right)(1 + o(1)).$$

The proof is given in Supplementary Section S.5. The proof of this theorem follows similar steps as the proof of type-II error rate for horseshoe prior in Datta and Ghosh (2013), where a fixed $\eta = 1/4$ and $\delta = 1/9$ were used for deriving an explicit expression. Then it follows from Theorems 3 and 5 that the risk of the horseshoe+ decision rule is given by

$$R_{\text{HS}+} = n \left\{ \mu\left(2\Phi\left(\sqrt{\frac{2}{\eta(1-\delta)}}\sqrt{C}\right) - 1\right) + (1-\mu)\frac{\sqrt{2}\tau^2}{\sqrt{\pi \log(1/2\tau)}} \right\}(1 + o(1))$$

$$= n\left\{\mu(2\Phi(\sqrt{\frac{2}{\eta(1-\delta)}}\sqrt{C})-1)\right\}(1+o(1)) \quad \text{as} \quad \tau \to 0.$$

Since the risk of the Bayes oracle is $R_{\text{BO}} = n\{\mu(2\Phi(\sqrt{C})-1)\}(1+o(1))$, it follows that the horseshoe+ decision rule attains the Bayes oracle up to a multiplicative constant.

4.3 Kullback–Leibler risk bounds

Carvalho et al. (2010) proved that for horseshoe the Bayes estimate for the sampling density, measured using the Kullback–Leibler distance between the true model and the estimator of the density function, converges to the truth at a super-efficient rate. Let θ_0 be the true parameter value and $f(y|\theta)$ be the sampling model. Further, let $K(q_1, q_2) = \mathbb{E}_{q_1} log(q_1/q_2)$ denote the K-L divergence of a density q_2 from q_1. The proof utilizes the following result by Clarke and Barron (1990).

Proposition 1. *(Clarke and Barron, 1990). Let $\nu_n(d\theta|y_1,\ldots,y_n)$ be the posterior distribution corresponding to some prior $\nu(d\theta)$ after observing data $y_{(n)} = (y_1,\ldots, y_n)$ according to the sampling model $f(y|\theta)$. Define the posterior predictive density $\hat{q}_n(y) = \int f(y|\theta)\nu_n(d\theta|y_1,\ldots,y_n)$. Assume further that $\nu(A_\epsilon) > 0$ for all $\epsilon > 0$. Then the Cesàro-average risk of the Bayes estimator, defined as $R_n \equiv n^{-1}\sum_{j=1}^{n} K(q_{\theta_0}, \hat{q}_j)$, satisfies*

$$R_n \le \epsilon - \frac{1}{n}\log \nu(A_\epsilon),$$

where $\nu(A_\epsilon)$ denotes the measure of the set $\{\theta : K(q_{\theta_0}, q_\theta) \le \epsilon\}$.

Using the above proposition, Theorem 4 of Carvalho et al. (2010) proves that for the horseshoe estimator the Cesàro-average risk satisfies

$$R_n = O\left(\frac{1}{n}\log\left(\frac{n}{(\log n)^b}\right)\right), \tag{17}$$

when the true parameter $\theta_0 = 0$. This rate is faster than any prior without a pole at zero. It is super-efficient, in the sense that the risk is lower than that of the MLE, which has the rate $O(\log n/n)$. The same result holds for the horseshoe+ estimator due to its infinite mass near zero (by Theorem 1). However, we demonstrate that the horseshoe+ prior in fact has a better rate of convergence than the horseshoe prior. Our result is based on the following theorem.

Theorem 6. *Let $p^0_{HS+}(\theta)$ and $p^0_{HS}(\theta)$ denote the marginal densities of the horseshoe+ and horseshoe priors at the origin when $\tau = 1$. Then we have*

$$\int_0^{\frac{1}{\sqrt{n}}} p^0_{HS+}(\theta)d\theta = \frac{1}{\sqrt{2}\pi^{5/2}\sqrt{n}}\left(\frac{\log^2(n)}{4} + \left(1 - \frac{\gamma}{2} + \frac{\log(4)}{4}\right)\log(n) + O(1)\right),$$

where γ is the Euler–Mascheroni constant and

$$\int_0^{\frac{1}{\sqrt{n}}} p^0_{HS}(\theta)d\theta = \frac{1}{\sqrt{2}\pi^{3/2}\sqrt{n}}\left(\frac{\log(n)}{2} + O(1)\right).$$

The proof is given in Supplementary Section S.6. Due to the extra $\log(n)$ factor, the horseshoe+ prior places more mass around a neighborhood of the origin compared to the horseshoe prior. Thus, when $\theta_0 = 0$, setting $\epsilon = 1/n$ gives after some algebra from Proposition 1 that

$$R_n(HS) \leq \frac{\log n}{2n} + \frac{1}{n} - \frac{\log\log n}{n} + \text{const},$$

and

$$R_n(HS+) \leq \frac{\log n}{2n} + \frac{1}{n} - \frac{2\log\log n}{n} + \text{const}.$$

Therefore, the multiplier of the $\log\log n$ term improves for horseshoe+.

4.4 Mean squared error

It is well known that if $p(|y_i - \theta_i|)$ is the standard normal density and $p(\theta_i)$ is a zero mean scale mixture of normals, with the scale parameter λ^2 following a proper prior law, the posterior moments of θ_i admits the following representations, also known as "Tweedie's formula" (Efron, 2011):

$$\mathbb{E}(\theta_i|y_i) = y_i + \frac{d}{dy_i}\log m(y_i), \tag{18}$$

$$\mathbb{V}(\theta|y_i) = 1 + \frac{d^2}{dy_i^2}\log m(y_i), \tag{19}$$

where $m(y_i)$ is the marginal for y_i (see for example Pericchi and Smith (1992) and Carvalho et al. (2010)). Furthermore, we can use properties of slowly varying functions to show that if the prior on θ_i can be written as a normal scale mixture with a "slowly-varying" prior on the scale parameter, the marginal inherits the slowly varying property. For priors with a polynomially heavy tail it can also be shown that the resulting posterior mean is asymptotically robust, in that the difference $|\mathbb{E}(\theta_i|y_i, \tau) - y_i|$ vanishes for large $|y_i|$ while τ is fixed.

Heavy-tailed distributions are often characterized by the notion of regular variation. The following definition is due to Karamata (see Mikosch (1999) or Bingham et al. (1989) for a detailed discussion).

Definition 1. *A positive, measurable function $L(\cdot)$ is said to be regularly varying at infinity with index α if it is defined on the interval $[x_0, \infty)$ for some x_0 and*

$$\lim_{x\to+\infty} \frac{L(tx)}{L(x)} = t^\alpha \quad \textit{for all} \quad t > 0.$$

$L(\cdot)$ is said to be slowly varying at infinity if $\alpha = 0$.

Using the above definition, we state the following result from Theorem 6.1 of Barndorff-Nielsen et al. (1982).

Proposition 2. *(Barndorff-Nielsen et al., 1982). Consider the Gaussian scale mixture* $y|\lambda^2 \sim \mathcal{N}(0, \lambda^2)$ *and suppose the prior density of* λ^2 *is given by* $f(\lambda^2) = (\lambda^2)^{\alpha-1}L(\lambda^2)$ *as* $\lambda^2 \to \infty$*, where* $L(\cdot)$ *is a slowly varying function. Then the marginal* $m(y)$ *after integrating out* λ^2 *has the property that* $m(y) \propto |y|^{2\alpha-1}L(y^2)$ *as* $|y| \to \infty$*.*

Let $m_{HS+}(y_i)$ and $m_{HS}(y_i)$ denote the marginals under the horseshoe+ and horseshoe priors respectively. Proposition 2 immediately shows that we have $m_{HS+}(y_i) = m_{HS}(y_i)\log(|y_i|)(1+o(1))$ as $|y_i| \to \infty$, since the only difference between the horseshoe and horseshoe+ mixing densities is the additional slowly varying $(\log \lambda_i)$ term in the scale mixing density for the horseshoe+ prior. In particular, as $|y_i| \to \infty$, we have

$$m_{HS+}(y_i) = m_{HS}(y_i) \times \log(|y_i|) \times \frac{y_i^2 - 1}{y_i^2 + 1} \times \text{constant},$$

where "constant" denotes the collection of all terms that does not involve y_i. Thus,

$$\frac{d}{dy_i}\log m_{HS+}(y_i) = \frac{d}{dy_i}\log m_{HS}(y_i) + \frac{1}{|y_i|\log|y_i|} - \underbrace{\frac{4y_i^2}{y_i^4 - 1}}_{O(1/y_i^2)}, \tag{20}$$

and,

$$\frac{d^2}{dy_i^2}\log m_{HS+}(y_i) = \frac{d^2}{dy_i^2}\log m_{HS}(y_i) - \frac{1 + \log y_i}{(y_i \log y_i)^2} + O\left(1/y_i^3\right). \tag{21}$$

Using Equations (18) and (19), in combination with Equations (20) and (21), allows one to relate the bias and variance, and hence the MSE, for the horseshoe and the horseshoe+ estimators. We have the following result:

Theorem 7. *Suppose* $p(|y_i - \theta_i|)$ *is the standard normal density, and* $p_{HS}(\theta_i)$ *and* $p_{HS+}(\theta_i)$ *denote the horseshoe and horseshoe+ prior densities on* θ_i *when* $\tau = 1$*, leading to the posterior mean squared errors* $\text{MSE}_{HS}(\theta_i|y_i)$ *and* $\text{MSE}_{HS+}(\theta_i|y_i)$ *respectively. Then, for large values of* $|y_i|$*, we have,*

$$\text{MSE}_{HS+}(\theta_i|y_i) = \text{MSE}_{HS}(\theta_i|y_i) - \frac{1}{y_i^2 \log|y_i|} + O\left(\frac{1}{y_i^3}\right).$$

The proof is given in Supplementary Section S.7. This theorem establishes that the horseshoe+ estimator has asymptotically lower MSE compared to the horseshoe estimator when $|y_i|$ is large, due to the extra $(\log|y_i|)$ factor in the marginal, which in turn is due to the extra $(\log \lambda_i)$ term in the prior mixing density.

5 Numerical examples

5.1 Sum of squared error about the posterior median

We follow the simulation setting described in Bhattacharya et al. (2015). We simulate data $y_i|\theta_i \sim \mathcal{N}(\theta_i, 1)$ for $i = 1, \ldots, n$, where $\theta_i = A$ in fraction q of its components

with the magnitude of $A = 7, 8$ and $\theta_i = 0$ in the remaining components. We report simulation results for $n = 200$ in Table 2. Each configuration is replicated 100 times and the average sum of squared error about the posterior median is reported.

q	A	D-L	HS Cauchy	HS+ Cauchy	HS Unif	HS+ Unif
0.05	7	26.86	**15.95**	18.58	17.11	18.08
	8	22.49	**14.47**	15.97	15.26	17.42
0.1	7	43.76	33.92	**31.65**	35.13	33.51
	8	43.81	32.28	**29.77**	33.67	32.23
0.2	7	78.11	69.29	**59.26**	83.61	59.92
	8	82.64	70.72	**62.64**	118.52	63.69
0.3	7	103.46	104.33	**86.77**	322.93	100.26
	8	121.04	108.12	**93.21**	373.71	220.16

Table 2: Average SSE about the posterior median for $n = 200$ for the competing priors. The averages are computed over 100 replicates. The lowest SSE for each setting (in rows) is in bold.

We compare the proposed horseshoe+ prior with two competitors: the horseshoe prior of Carvalho et al. (2010) and the Dirichlet–Laplace (D-L) prior of Bhattacharya et al. (2015). To deal with the global shrinkage parameter τ for the horseshoe and the horseshoe+ priors, we try two scenarios: (a) $\tau \sim \mathrm{C}^+(0, 1/n)$ and (b) $\tau \sim \mathrm{Uniform}(0, 1)$. For posterior sampling, we use the Stan software package (Stan Development Team, 2014) to draw 10,000 samples in each case, half of which are treated as burn-in and discarded. We monitored Markov chain Monte Carlo (MCMC) convergence and found no evidence of mixing problems. The D-L prior is implemented in its hierarchal normal–exponential form, and the horseshoe and horseshoe+ priors by the hierarchical model in Equations (4) and (6) respectively.

In Table 2, the estimator with the lowest average SSE is in bold in each simulation setting (in rows). The horseshoe+ prior with the half-Cauchy prior on τ has the lowest SSE in all but two cases, in which the horseshoe prior performs the best. The $\mathrm{C}^+(0, 1/n)$ prior on τ results in better performance over a $\mathrm{Uniform}(0, 1)$ prior for both horseshoe and horseshoe+ since the former puts more mass in a neighborhood close to zero, helping τ adapt to the sparsity level of the data. Additional simulation results for different values of n and A are presented in Table S.1 in the Supplementary Material. Horseshoe+ outperforms the competing approaches in most cases.

To make the difference between the horseshoe and the horseshoe+ estimates clear, we plot $\mathbb{E}(\kappa_i|y_i)$ and $\mathbb{E}(\theta_i|y_i)$ for $i = 1, \ldots, n$, for horseshoe in Figure 4 and for horseshoe+ in Figure 5. In both cases, the prior on τ is $\mathrm{Uniform}(0, 1)$. We used $n = 200$ and simulated y_i with 10 components with a mean equal to 7 and the rest with mean 0. Without loss of generality, the components (true values and estimates) with true non-zero means are plotted as the first 10 data points and those with true zero means are plotted afterwards. The posterior means are shown as dots and the middle 95% posterior credible intervals by solid lines. By comparing the estimates, it is clear that horseshoe+ does a much better job compared to horseshoe in terms of shrinking the noise terms to zero (estimated $\hat{\kappa}_i$ closer to 1 or equivalently, estimated $\hat{\theta}_i$ closer to zero).

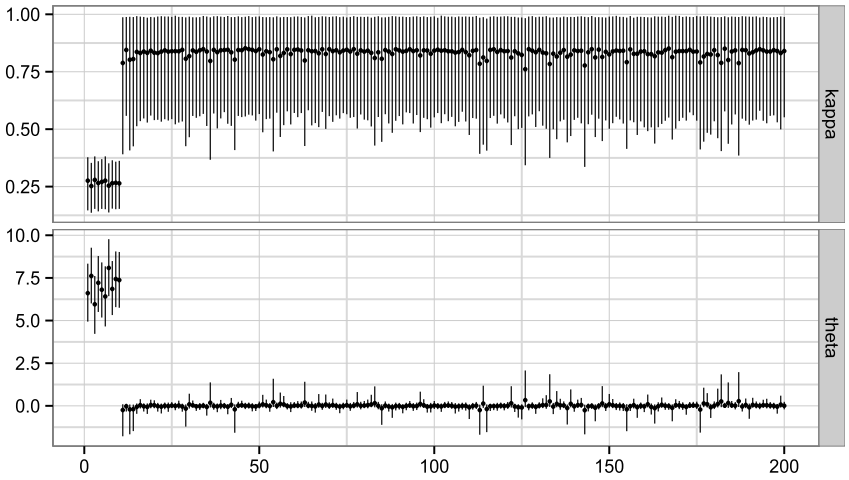

Figure 4: Estimated κ_i and θ_i for horseshoe for $n = 200$ with first 10 true θ_i equal to 7 and rest true values set to 0. Dots are posterior means and solid lines are the middle 95% posterior credible intervals. We used $\tau \sim \text{Uniform}(0, 1)$.

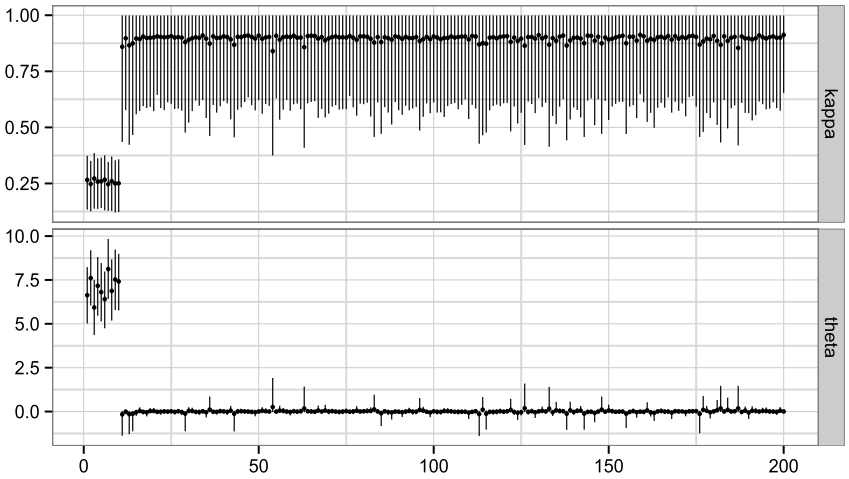

Figure 5: Estimated κ_i and θ_i for horseshoe+ for $n = 200$ with first 10 true θ_i equal to 7 and rest true values set to 0. Dots are posterior means and solid lines are the middle 95% posterior credible intervals. We used $\tau \sim \text{Uniform}(0, 1)$.

5.2 Misclassification probabilities

We compared the performance of the multiple testing rule induced by the horseshoe+ prior with two other one-group global–local shrinkage priors: the horseshoe prior of Carvalho et al. (2010) and the Dirichlet–Laplace prior of Bhattacharya et al. (2015) in

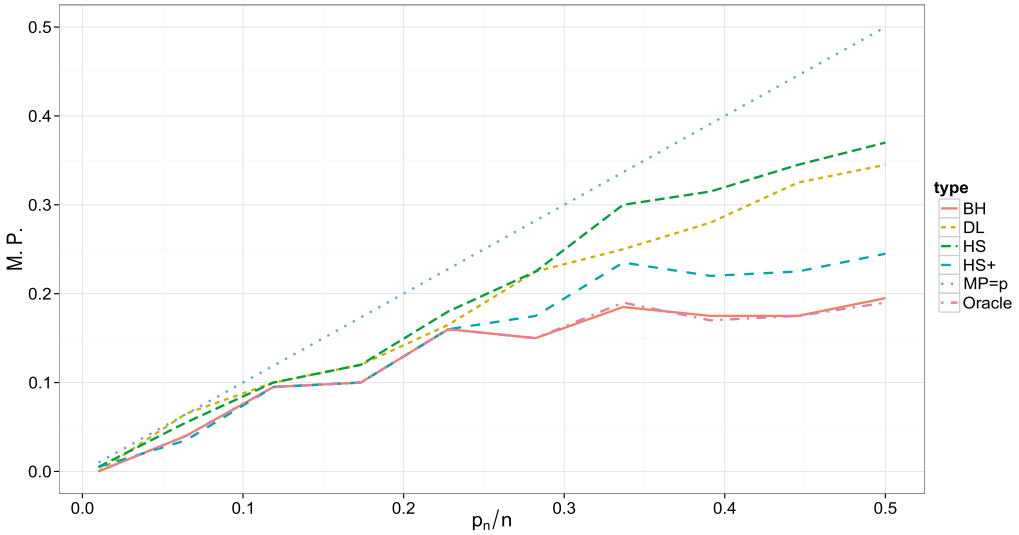

Figure 6: Misclassification probability plots for the horseshoe+, horseshoe, and the Dirichlet–Laplace ($DL_{1/n}$) shrinkage priors, Benjamini–Hochberg and the Bayes oracle for $\mu \in (0.1, 0.5)$.

terms of the misclassification probability (MP). We use the misclassification probability as a criteria for our experiment as it is equal to the Bayes risk under a 0–1 additive loss for data generated by a two-groups model. We follow the same experimental set up in Bogdan et al. (2008), replicated in Datta and Ghosh (2013), where the Bayes oracle (BO) acts as the lower bound and the $MP = \mu$ line as the upper bound, where μ is the proportion of signals. We simulated data of size $n = 200, \psi_n = \sqrt{2 \log n} = 3.26$. Our data generation scheme follows the conditions provided by Bogdan et al. (2011), which guarantees the optimality of the Benjamini–Hochberg procedure to use it as another practical lower bound along with the Bayes Oracle.

Figure 6 shows the misclassification probabilities (henceforth abbreviated as MP) for different shrinkage priors considered for ten equispaced values of $\mu \in [0.01, 0.5]$ along with the oracle and the straight line (MP $= \mu$). Figure 6 shows that the misclassification probability for the horseshoe+ prior is very close to that of the Bayes oracle for a wide range of values of μ, and departs a little for values higher than 0.2. Furthermore, the horseshoe+ decision rule leads to a superior performance compared both the horseshoe and the Dirichlet–Laplace prior. We have also plotted the MP for the Benjamini–Hochberg rule, for $\alpha = 1/\log n = 0.1887$, along with the one-group global–local shrinkage priors. Under this setting, the Benjamini–Hochberg rule achieves the same MP as the oracle. This is in concordance with the theoretical results for optimality of BH in Bogdan et al. (2011).

We used the full Bayes estimates for the hyperparameters for both the horseshoe prior and the double exponential prior. For estimating τ, we assumed standard half-

Cauchy prior on τ for deriving the full conditionals using a Gibbs sampler. As pointed out by Carvalho et al. (2009) and Scott and Berger (2006), the fully Bayesian approach for estimating τ has a few advantages over its alternatives, viz. empirical Bayes and cross-validation. In the extremely sparse case, the empirical Bayes estimate of τ might collapse to 0 (Scott and Berger, 2010; Bogdan et al., 2008). Cross-validation, though free of this problem, uses plug-in estimates for the signal-to-noise ratio. Carvalho et al. (2009) argue that the plug-in estimates are not necessarily wrong, but caution should be exercised while using them for extremely sparse problems.

6 Application on a prostate cancer data set

We illustrate the performance of the horseshoe+ prior for the benchmark *prostate cancer data*, introduced by Singh et al. (2002) and made popular by Efron (2008, 2010b,a), among others. The *prostate cancer data* has gene expression values for $n = 6{,}033$ genes for $m = 102$ subjects, with $m_1 = 50$ normal controls and $m_2 = 52$ prostate cancer patients. The goal is to identify genes that are differentially expressed between controls and the cancer patients. To analyze this data further, the test statistic values are calculated for each of the 6,033 genes by first calculating a two-sample t-statistic, say $t_i, i = 1, 2, \ldots, n = 6{,}033$ for each of the genes and then applying the inverse normal cumulative distribution function (CDF) transformation to obtain $y_i = \Phi^{-1}(F_{t_{100}}(t_i))$, where $F_{t_{100}}$ is the CDF of a t-distribution with 100 degrees of freedom. The y_i-values can be modeled as independent Gaussian variables with mean θ_i's, i.e. $y_i \sim \theta_i + \epsilon_i$ to cast this problem as a high-dimensional normal means inference problem. The corresponding multiple testing problem would be to simultaneously test the hypotheses $H_{0i} : \theta_i = 0$, for $i = 1, \ldots, n$. Under the global null hypothesis of no 'differentially expressed' genes, one should expect the histogram of the test statistics to follow a $\mathcal{N}(0, 1)$ density curve but the histogram shows a heavier tail, suggesting the presence of a few regulatory genes.

For a proper appraisal of the extra shrinkage by the horseshoe+ prior at the tails compared to the horseshoe prior, we do the following experiment: We consider the top 10 genes selected by Efron (2010a) and their effect sizes estimated by a two-groups normal hierarchical model. We apply both the horseshoe and the horseshoe+ prior to the 6,033 test statistics, and compare the 'effect-size' estimates $\hat{\theta}_i$ for these genes. One would expect that the horseshoe+ prior would shrink these "top" genes even less than the horseshoe prior and as a result the posterior mean $\hat{\theta}_i = (1 - \mathbb{E}(\kappa_i | y_i, \tau))y_i$ would be closer to the observed test statistics y_i.

Table 3 shows the top 10 genes selected by Efron (2010a), and the effect size estimates by the horseshoe and the horseshoe+ priors. For both the horseshoe and horseshoe+ prior, we implemented a Gibbs sampler with 15,000 draws with a burn-in period of 3,000 draws. The benefits of a heavier tail become apparent from this table as in 9 out of the top 10 genes, the horseshoe+ estimates are closer to the observed test statistics compared to the horseshoe estimates. One might naturally wonder about the performance of the two competing Bayesian models for the "uninteresting" genes, and it turns out that both the priors have equal strength in squelching the noisy test statistics to zero. Figure 7 shows the posterior mean for the two priors against the observed test statistics.

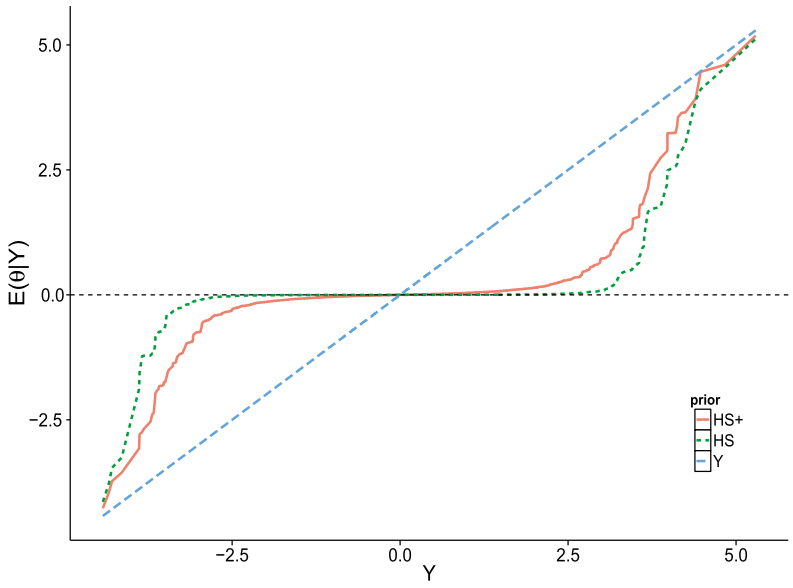

Figure 7: Posterior mean $\mathbb{E}(\theta_i|y_i)$ against y_i for 6,033 genes for the horseshoe and horseshoe+ priors applied to the prostate cancer example.

Gene	y-value	$\hat{\theta}_i^{HS+}$	$\hat{\theta}_i^{HS}$	$\hat{\theta}_i^{\text{Efron}}$
610	5.29	5.20	5.12	4.11
1720	4.83	4.77	4.54	3.65
332	4.47	3.24	4.11	3.24
364	-4.42	-4.43	-4.14	-3.57
914	4.40	4.40	3.89	3.16
3940	-4.33	-3.78	-3.77	-3.52
4546	-4.29	-3.88	-3.46	-3.47
1068	4.25	3.71	3.03	2.99
579	4.19	3.99	2.88	2.92
4331	-4.14	-3.48	-3.26	-3.30

Table 3: The test statistics (y-values) and the effect-size estimates for the top 10 genes selected by Efron (2010a) by the horseshoe, horseshoe+ models, and Efron's two-groups model estimates.

It can be clearly seen that all the procedures show good shrinkage properties near zero, and the only difference comes from the performance near tails, or *robustness to large signals*. This is also reflected in the value of the estimated mean squared prediction error calculated as $MSE = (1/n)\sum_{i=1}^{n}(\hat{\theta}_i - y_i)^2$. The values of the mean squared prediction error for the horseshoe+ and the horseshoe prior are 1.189 and 1.045 illustrating the superiority of horseshoe+ prior over the horseshoe prior.

7 Discussion

We have provided a default Bayesian shrinkage estimator for extracting signals from a sparse parameter vector. The proposed prior is called horseshoe+ prior as it renders itself to a natural extension of the horseshoe prior and provides substantial improvement for the "nearly-black" or "ultra-sparse" situation. In particular, the heavier tails of the horseshoe+ prior leads to an increasing ability of separating the signals, and the larger mass near origin leads to better handling of sparsity and a higher order of super-efficiency for the risk in density estimation. We have examined this new prior both theoretically and empirically by considering the estimation accuracy for a high-dimensional parameter vector as well as the error rates for the multiple testing rule induced by applying a threshold rule to the pseudo posterior inclusion probabilities.

Our asymptotic results demonstrate that the horseshoe+ estimator achieves a lower MSE and the horseshoe+ decision rule attains the Bayes oracle in testing up to $O(1)$ with a sharper bound on the type-I error rate compared to horseshoe. While we have not discussed the asymptotic minimaxity properties of the horseshoe+ estimator in the ℓ_2 sense, we conjecture that the asymptotic minimaxity will continue to hold, likely with a sharper bound on the constant term compared to van der Pas et al. (2014). The sharpening effect of the horseshoe+ prior can be attributed to the extra shrinkage gained by having a U-shaped Jacobian over a lopsided one, in addition to the U-shaped prior induced on κ_i. It is also worth noting that asymptotic minimaxity results for a class of priors with $p(\lambda_i^2) \propto (\lambda_i^2)^{-a-1} L(\lambda_i^2)$ where $L(\cdot)$ is a bounded, slowly varying function has been established by Ghosh and Chakrabarti (2014). For horseshoe+, $L(\lambda_i) = \lambda_i^2 \log \lambda_i / (\lambda_i^2 - 1)$ is slowly varying but not bounded. At the time of preparing the current manuscript, we became aware of the work by van der Pas et al. (2016), who use an alternative technique to establish the asymptotic minimaxity of a class of estimators, including the horseshoe+, up to a constant (see Section 3.3 of their paper). It must be noted, however, while asymptotic minimaxity results hold for a class of priors, their finite sample performances can often be quite different, as evidenced by our simulations. It is also an open question how close the minimax constant can possibly get to the optimal value of two for ultra-sparse objects, as established by Donoho et al. (1992).

In the recent past, there have been a few shrinkage priors that we collectively call the 'global–local' shrinkage priors following Polson and Scott (2010). These priors include the generalized double Pareto (Armagan et al., 2013), the normal–exponential–gamma (Griffin and Brown, 2010), the three parameter beta (Armagan et al., 2011), and the Dirichlet–Laplace (Bhattacharya et al., 2015), among others. These priors exhibit similar shrinkage properties as the horseshoe prior in that they simultaneously squelch the noise to zero and recover the signals. Though these priors lead to competitive performances in the sparse signal recovery problem, they also have unique, distinguishable characteristics. For example, the generalized double Pareto prior leads to a closed form prior density of θ and induces a sparsity favoring penalty in regularized least squares, while the Dirichlet–Laplace prior models the joint distribution of θ under the two-groups model via the joint distribution of the shrinkage parameters. The behavior of

the marginal prior densities on θ can be seen from Figures 2 and 3, and our simulation results suggest improvements for both estimation and testing, but it is an open question whether a set of necessary conditions can be imposed on the class of one-group global–local shrinkage priors that guarantees certain desirable properties.

A key insight we gain from the success of the family of the one-group global–local shrinkage priors is that the global shrinkage parameter plays a vital role in controlling the behavior of the posterior. Specifically, the global shrinkage parameter in the horseshoe prior needs to be of the order of the proportion of non-null effects to ensure asymptotic minimaxity in estimation (van der Pas et al., 2014) as well as the optimality of the induced decision rule in testing (Datta and Ghosh, 2013). We have proved that the same condition also guarantees the optimal performance for the horseshoe+ prior in testing.

Finally, the horseshoe+ prior can be further extended by modeling the local shrinkage parameter λ_i as a higher order product of independent half-Cauchy random variables, leading to an even heavier tail and larger spike at zero. The moments and densities of the Cauchy product $C_1C_2\ldots C_k$ are given in Bourgade et al. (2007). The density $\Psi_k(\cdot)$ of the k-product $C_1C_2\ldots C_k$ for the even and the odd cases are as follows:

$$\Psi_{2i+1}(x) = \frac{2^{2i}}{\pi(2i)!}\left(\prod_{j=1}^{i}((j-\frac{1}{2})^2+\frac{(\log|x|)^2}{\pi^2})\right)\frac{1}{1+x^2},$$

$$\Psi_{2i}(x) = \frac{2^{2i-1}}{\pi(2i-1)!}\left(\prod_{j=1}^{i-1}(j^2+\frac{(\log|x|)^2}{\pi^2})\right)\frac{\log|x|}{x^2-1}.$$

Furthermore, one might use the "universal prior" due to Rissanen (1983) over the number of terms k in the product density. The "universal prior" is defined with the mass function:

$$Q(i) = 2^{-L^0(i)},\ \text{for } i = 1,2,\ldots; \quad L^0(i) = \log^*(i) + \log c,$$

where, $\log^*(x) = \log x + \log\log x + \ldots$, where the sum involves only non-negative terms and $c = \sum 2^{-\log^* i} \approx 2.865064$.

The family of Cauchy-product densities can be used in conjunction with Rissanen's universal prior described above to define an adaptive shrinkage estimator such as the *Polyshrink* estimator due to Foster and Stine (2005), where the amount of shrinkage varies adaptively with the estimation task. For an n-dimensional parameter θ, the Polyshrink estimator uses a collection of discrete mixture models $\mathcal{G}_p = \{g_{\epsilon_k}(y) = (1-\epsilon_k)\phi(y) + \epsilon_k\psi(z),\ \epsilon_k = 2^{k-(K+1)}\}$, for $k = 1,\ldots,K$ with $K = 1 + \lfloor\log_2(p)\rfloor$ and $\phi(\cdot)$ and $\psi(\cdot)$ denote the standard normal density and Cauchy density with scale $\sqrt{2}$ respectively. We conjecture the advantages of the one-group global–local model over the two-groups model would naturally carry over to this case if we use a collection of one-group priors defined by Cauchy products of different orders to achieve different amounts of shrinkage. The possibility of such extensions was first discussed in Polson and Scott (2012), and it would be interesting to settle this issue theoretically.

Supplementary Material

Supplementary Material to "The Horseshoe+ Estimator of Ultra-Sparse Signals" (DOI: 10.1214/16-BA1028SUPP; .pdf). Contains the proofs of theorems and an additional simulation example.

References

Armagan, A., Clyde, M., and Dunson, D. B. (2011). "Generalized beta mixtures of Gaussians." In *Advances in Neural Information Processing Systems*, 523–531. 1107, 1109, 1112, 1126

Armagan, A., Dunson, D. B., and Lee, J. (2013). "Generalized double Pareto shrinkage." *Statistica Sinica*, 23(1): 119–143. MR3076161. 1107, 1109, 1113, 1126

Barndorff-Nielsen, O., Kent, J., and Sørensen, M. (1982). "Normal variance–mean mixtures and z distributions." *International Statistical Review/Revue Internationale de Statistique*, 50: 145–159. MR0678296. doi: http://dx.doi.org/10.2307/1402598. 1119, 1120

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. T. (2016a). "Default Bayesian analysis with global–local shrinkage Priors." *Biometrika*, to appear. arXiv:1510.03516 1109

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016b). "Supplementary material to "The horseshoe+ estimator of ultra-sparse signals"." *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/16-BA1028SUPP. 1113

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). "Dirichlet–Laplace priors for optimal shrinkage." *Journal of the American Statistical Association*, 110: 1479–1490. MR3449048. doi: http://dx.doi.org/10.1080/01621459.2014.960967. 1108, 1109, 1113, 1120, 1121, 1122, 1126

Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1989). *Regular variation*, volume 27 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press. MR1015093. 1119

Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). "Asymptotic Bayes-optimality under sparsity of some multiple testing procedures." *The Annals of Statistics*, 39(3): 1551–1579. MR2850212. doi: http://dx.doi.org/10.1214/10-AOS869. 1106, 1108, 1115, 1116, 1123

Bogdan, M., Ghosh, J. K., and Tokdar, S. T. (2008). "A comparison of the Benjamini–Hochberg procedure with some Bayesian rules for multiple testing." In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*, volume 1 of *Inst. Math. Stat. Collect.*, 211–230. Inst. Math. Statist., Beachwood, Ohio, USA. MR2462208. doi: http://dx.doi.org/10.1214/193940307000000158. 1123, 1124

Bourgade, P., Fujita, T., and Yor, M. (2007). "Euler's formulae for ζ (2n) and products of Cauchy variables." *Electronic Communications in Probability*, 12: 73–80. MR2300217. doi: http://dx.doi.org/10.1214/ECP.v12-1244. 1127

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. 1124

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97: 465–480. MR2650751. doi: http://dx.doi.org/10.1093/biomet/asq017. 1106, 1107, 1108, 1109, 1111, 1113, 1118, 1119, 1121, 1122

Castillo, I. and van der Vaart, A. (2012). "Needles and straw in a haystack: Posterior concentration for possibly sparse sequences." *The Annals of Statistics*, 40(4): 2069–2101. MR3059077. doi: http://dx.doi.org/10.1214/12-AOS1029. 1106, 1108

Clarke, B. and Barron, A. R. (1990). "Information-theoretic asymptotics of Bayes methods." *IEEE Transactions on Information Theory*, 36(3): 453–471. MR1053841. doi: http://dx.doi.org/10.1109/18.54897. 1118

Datta, J. and Ghosh, J. K. (2013). "Asymptotic properties of Bayes risk for the horseshoe prior." *Bayesian Analysis*, 8(1): 111–132. MR3036256. doi: http://dx.doi.org/10.1214/13-BA805. 1106, 1107, 1109, 1111, 1114, 1116, 1117, 1123, 1127

Denison, D. G. and George, E. I. (2012). *Bayesian prediction with adaptive ridge estimators*, volume 8 of *IMS Collections*, 215–234. Beachwood, Ohio, USA: Institute of Mathematical Statistics. MR3202513. doi: http://dx.doi.org/10.1214/11-IMSCOLL815. 1109

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 (Methodological)*, 54: 41–81. MR1157714. 1106, 1109, 1126

Efron, B. (2008). "Microarrays, empirical Bayes and the two-groups model." *Statistical Science*, 23(1): 1–22. MR2431866. doi: http://dx.doi.org/10.1214/07-STS236. 1108, 1124

Efron, B. (2010a). "The future of indirect evidence." *Statistical Science*, 25(2): 145–157. MR2789983. doi: http://dx.doi.org/10.1214/09-STS308. 1124, 1125

Efron, B. (2010b). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press. MR2724758. doi: http://dx.doi.org/10.1017/CBO9780511761362. 1124

Efron, B. (2011). "Tweedie's formula and selection bias." *Journal of the American Statistical Association*, 106(496): 1602–1614. MR2896860. doi: http://dx.doi.org/10.1198/jasa.2011.tm11181. 1119

Foster, D. P. and Stine, R. A. (2005). "Polyshrink: An adaptive variable selection procedure that is competitive with Bayes experts." Technical report, Univ. of Penn. 1127

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515–534. MR2221284. 1110

Ghosh, P. and Chakrabarti, A. (2014). "Posterior Concentration Properties of a General Class of Shrinkage Estimators around Nearly Black Vectors." arXiv:1412.8161. 1126

Ghosh, P., Tang, X., Ghosh, M., and Chakrabarti, A. (2016). "Asymptotic properties of Bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity." *Bayesian Analysis*, 11: 753–796. MR3498045. doi: http://dx.doi.org/10.1214/15-BA973. 1107, 1112

Griffin, J. E. and Brown, P. J. (2010). "Inference with normal–gamma prior distributions in regression problems." *Bayesian Analysis*, 5(1): 171–188. MR2596440. doi: http://dx.doi.org/10.1214/10-BA507. 1107, 1109, 1126

Guan, Y. and Stephens, M. (2008). "Practical issues in imputation-based association mapping." *PLoS Genet*, 4(12): e1000279. 1109

Johnstone, I. M. and Silverman, B. W. (2004). "Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences." *Annals of Statistics*, 32: 1594–1649. MR2089135. doi: http://dx.doi.org/10.1214/009053604000000030. 1106, 1108

Marchini, J. and Howie, B. (2010). "Genotype imputation for genome-wide association studies." *Nature Reviews Genetics*, 11(7): 499–511. 1109

Mathai, A., Saxena, R. K., and Haubold, H. J. (2009). *The H-function*. New York, NY: Springer. 1107

Mikosch, T. (1999). *Regular variation, subexponentiality and their applications in probability theory*. Volume 99 of EURANDOM report. Eindhoven, The Netherlands: Eindhoven University of Technology. 1119

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression." *Journal of the American Statistical Association*, 83(404): 1023–1032. MR0997578. 1106, 1108

Pericchi, L. and Smith, A. (1992). "Exact and approximate posterior moments for a normal location parameter." *Journal of the Royal Statistical Society. Series B (Methodological)*, 54: 793–804. MR1185223. 1119

Polson, N. G. and Scott, J. G. (2010). "Shrink globally, act locally: Sparse Bayesian regularization and prediction." *Bayesian Statistics*, 9: 501–538. MR3204017. doi: http://dx.doi.org/10.1093/acprof:oso/9780199694587.003.0017. 1106, 1112, 1126

Polson, N. G. and Scott, J. G. (2012). "On the half-Cauchy prior for a global scale parameter." *Bayesian Analysis*, 7(4): 887–902. MR3000018. doi: http://dx.doi.org/10.1214/12-BA730. 1106, 1127

Rissanen, J. (1983). "A universal prior for integers and estimation by minimum description length." *The Annals of Statistics*, 11: 416–431. MR0696056. doi: http://dx.doi.org/10.1214/aos/1176346150. 1127

Scott, J. G. and Berger, J. O. (2006). "An exploration of aspects of Bayesian multiple testing." *Journal of Statistical Planning and Inference*, 136(7): 2144–2162. MR2235051. doi: http://dx.doi.org/10.1016/j.jspi.2005.08.031. 1124

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 38(5): 2587–2619. MR2722450. doi: http://dx.doi.org/10.1214/10-AOS792. 1124

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., et al. (2002). "Gene expression correlates of clinical prostate cancer behavior." *Cancer Cell*, 1(2): 203–209. 1124

Stan Development Team (2014). "Stan: A C++ Library for Probability and Sampling, Version 2.2." http://mc-stan.org/. 1121

Stephens, M. and Balding, D. J. (2009). "Bayesian statistical methods for genetic association studies." *Nature Reviews Genetics*, 10(10): 681–690. 1109

Stranger, B. E., Stahl, E. A., and Raj, T. (2011). "Progress and promise of genome-wide association studies for human complex trait genetics." *Genetics*, 187(2): 367–383. 1109

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society (Series B)*, 58: 267–288. MR1379242. 1106

van der Pas, S., Kleijn, B., and van der Vaart, A. (2014). "The horseshoe estimator: Posterior concentration around nearly black vectors." *Electronic Journal of Statistics*, 8: 2585–2618. MR3285877. doi: http://dx.doi.org/10.1214/14-EJS962. 1106, 1109, 1110, 1117, 1126, 1127

van der Pas, S., Salomond, J.-B., and Schmidt-Hieber, J. (2016). "Conditions for posterior contraction in the sparse normal means problem." *Electronic Journal of Statistics*, 10: 976–1000. MR3486423. doi: http://dx.doi.org/10.1214/16-EJS1130. 1126

Acknowledgments

The authors thank an anonymous referee, the Associate Editor and the Editor for their constructive suggestions. Bhadra acknowledges a Research Fellowship from the Statistical and Applied Mathematical Sciences Institute (SAMSI), where part of this research was conducted. This material is based upon work supported by the National Science Foundation under Grant No. DMS-1613063.