

The Horseshoe Estimator for Sparse Signals

Carlos M. Carvalho

Nicholas G. Polson

University of Chicago Booth School of Business

James G. Scott

Duke University Department of Statistical Science

October 2008

Abstract

This paper proposes a new approach to sparse-signal detection called the horseshoe estimator. The horseshoe is a close cousin of other widely used Bayes rules arising from, for example, double-exponential and Cauchy priors, in that it is a member of the same family of multivariate scale mixtures of normals. Its advantage, however, is its robustness at handling unknown sparsity and large outlying signals. We give analytical results showing why the horseshoe is a good default for robust estimation of sparse normal means, along with a new representation theorem for the posterior mean under normal scale mixtures. This theorem is related to classic results of Stein and Masreliez, and gives qualitative insight into some aspects of robust Bayesian analysis in sparse settings. Most importantly, we show that the horseshoe estimator corresponds quite closely to the answers one would get by pursuing a full Bayesian model-averaging approach using a traditional “spike and slab” prior to model signals and noise. This correspondence holds both for the posterior mean itself and for the classification rule induced by a simple thresholding scheme, meaning that the resulting “thresholded horseshoe” can also be viewed as a novel Bayes multiple-testing procedure.

Keywords: double-exponential prior; normal scale mixtures; multiple testing; empirical Bayes.

1 Introduction

1.1 The proposed estimator

Suppose we observe a p -dimensional vector $(y|\theta) \sim N(\theta, \sigma^2 I)$ and wish to estimate θ under quadratic loss. Suppose further that θ is believed to be sparse, in the sense that many of its entries are zero, or nearly so.

Our proposed estimator arises as the posterior mean under an exchangeable model $\pi_H(\theta_i)$ that we call the horseshoe prior. This model has an interpretation as a scale mixture of normals:

$$\begin{aligned}(\theta_i | \lambda_i, \tau) &\sim N(0, \lambda_i^2 \tau^2) \\ \lambda_i &\sim C^+(0, 1),\end{aligned}$$

where $C^+(0, 1)$ is a standard half-Cauchy distribution on the positive reals.

Observe the difference from a typical model involving a common variance component τ : each θ_i is mixed over its own λ_i , and each λ_i has an independent half-Cauchy prior. This places the horseshoe in the widely studied class of multivariate scale mixtures of normals (West, 1987; Angers and Berger, 1991; Denison and George, 2000; Griffin and Brown, 2005). We call these “local shrinkage rules,” to distinguish them from global shrinkage rules with only a shared scale parameter τ .

The name “horseshoe prior” arises from the observation that

$$E(\theta_i | y) = \int_0^1 (1 - \kappa_i) y_i \pi(\kappa_i | y) d\kappa_i = \{1 - E(\kappa_i | y)\} y_i,$$

where $\kappa_i = 1/(1 + \lambda_i^2)$, assuming $\sigma^2 = \tau^2 = 1$. After the half-Cauchy prior on λ_i is transformed, κ_i is seen to have a $\text{Be}(1/2, 1/2)$ prior. By virtue of being symmetric and unbounded at both 0 and 1, this density function resembles a horseshoe. Its utility for discriminating signal from noise is readily apparent. The left side of the horseshoe, $\kappa_i \approx 0$, yields virtually no shrinkage; the right side of the horseshoe, $\kappa_i \approx 1$, yields near-total shrinkage.

Unlike most models for sparsity, our approach does not involve a mixture of a point mass for noise and a density for signals. It nonetheless proves impressively accurate at handling a wide variety of sparse situations. Indeed, many of the desirable properties of heavy-tailed discrete-mixture models (Johnstone and Silverman, 2004) will be shown to obtain for the horseshoe prior, as well. These “two-group” models, following the language of Efron (2008), are the recognized gold standard for handling sparsity. We will show that the horseshoe, despite being a one-group model, essentially matches the conclusions of the two-group model, with significantly less computational effort.

We will describe three main strengths of the horseshoe estimator. First, it is highly adaptive, both to unknown sparsity and to unknown signal-to-noise ratio. Second, it is robust to large, outlying signals, as we will demonstrate analytically using the

representation theorem of Pericchi and Smith (1992). Finally, it exhibits a strong form of multiplicity control by limiting the number of spurious signals. In this respect, the horseshoe shares one of the most appealing features of Bayesian and empirical-Bayes model-selection techniques: after a simple thresholding rule is applied, the horseshoe exhibits an automatic penalty for multiple hypothesis testing. The nature of this multiple-testing penalty is well understood in discrete mixture models (Berry, 1988; Scott and Berger, 2006), and we will clarify how a similar effect occurs when the thresholded horseshoe is used instead.

We focus primarily on the posterior mean, the Bayes estimator under quadratic loss, for which many interesting analytical results are available. We also give substantial numerical evidence to show that the horseshoe performs well under other common loss functions.

1.2 The horseshoe density function

The density $\pi_H(\theta_i)$ is not expressible in closed form, but very tight upper and lower bounds in terms of elementary functions are available.

Theorem 1. *The horseshoe prior satisfies the following:*

(a) $\lim_{\theta \rightarrow 0} \pi_H(\theta) = \infty$

(b) For $\theta \neq 0$,

$$\frac{K}{2} \log \left(1 + \frac{4}{\theta^2} \right) < \pi_H(\theta) < K \log \left(1 + \frac{2}{\theta^2} \right), \quad (1)$$

where $K = 1/\sqrt{2\pi^3}$.

Proof. See Appendix A. □

Plots of π_H compared to the standard double-exponential and standard Cauchy densities can be seen in Figure 1. The next several sections will study this prior in detail, but its most interesting features are the following:

- It is symmetric about zero.
- It has heavy, Cauchy-like tails that decay like θ_i^{-2} .
- It has an infinitely tall spike at 0, in the sense that the density approaches ∞ logarithmically fast as $\theta_i \rightarrow 0$ from either side.

This paper is primarily about why $\pi_H(\theta)$ is a good default shrinkage prior for sparse signals. Our argument is foreshadowed by the list above: the prior’s flat tails allow each θ_i to be large if the data warrant such a conclusion, and yet its infinitely tall spike at zero means that the estimate can also be quite severely shrunk. These properties—heavy tails, but also a heavy spike—parallel the $\text{Be}(1/2, 1/2)$ density for the shrinkage coefficient κ_i , which is unbounded both at 0 and 1.

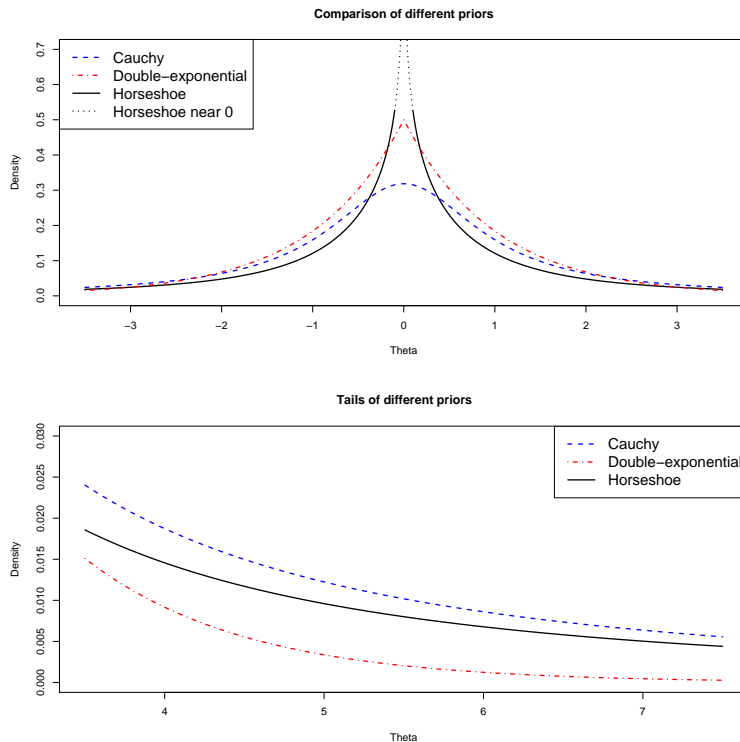


Figure 1: A comparison of π_H versus standard Cauchy and double-exponential densities; the dotted lines indicate that π_H approaches ∞ near 0.

2 Robust Shrinkage of Sparse Signals

2.1 A representation of the posterior mean

Recall the following well-known result. If $p(y - \theta)$ is a normal likelihood of known variance, $\pi(\theta)$ is the prior for the mean θ , and $m(y) = \int p(y - \theta)\pi(\theta) d\theta$ is the marginal density for y , then, for one sample of y :

$$E(\theta | y) = y + \frac{d}{dy} \ln m(y), \quad (2)$$

where we assume that $\sigma^2 = 1$ without loss of generality, unless otherwise noted. Versions of this result appear in Masreliez (1975), Polson (1991) and Pericchi and Smith (1992); analogous representation theorems are discussed by Brown (1971) and Stein (1981).

The theorem is useful for the insight it gives about an estimator's behavior in situations where y is very different from the prior mean. In particular, it shows that "Bayesian robustness" may be achieved by choosing a prior for θ such that the

derivative of the log predictive density is bounded as a function of y . Ideally, of course, this bound should converge to 0, which from (2) will lead to $E(\theta | y) \approx y$ for large $|y|$. This means that one will never miss too badly in the tails of the prior.

The representation in (2) does not directly apply for the horseshoe prior, as $\pi(\theta)$ fails to satisfy the condition that $\pi(\theta)$ be bounded. The following theorem, however, provides an alternative representation for all $\pi(\theta)$ in the scale mixture of normals family. This family's key feature is that $\pi(\theta | \lambda)$ is bounded, thereby allowing the required interchange of integration and differentiation. The theorem applies to likelihoods that may be non-normal, including the normal case with σ^2 unknown.

Theorem 2. *Given likelihood $p(y - \theta)$, suppose that $\pi(\theta)$ is a mean-zero scale mixture of normals: $(\theta | \lambda) \sim N(0, \lambda^2)$, with λ having proper prior $\pi(\lambda)$ such that $m(y)$ is finite. Define*

$$\begin{aligned} m^*(y) &= \int p(y - \theta) \pi^*(\theta) d\theta \\ \pi^*(\theta) &= \int_0^\infty \pi(\theta | \lambda) \pi^*(\lambda) d\lambda \\ \pi^*(\lambda) &= \lambda^2 \pi(\lambda). \end{aligned}$$

Then

$$\begin{aligned} E(\theta | y) &= \frac{m^*(y)}{m(y)} \frac{d}{dy} \ln m^*(y) \\ &= \frac{1}{m(y)} \frac{d}{dy} m^*(y). \end{aligned} \tag{3}$$

Proof. Notice first that $m^*(y)$ exists by the case $\pi(\lambda^2) \equiv 1$, which leads to the harmonic estimator in the case of a normal likelihood. This is sufficient to allow the interchange of integration and differentiation. Also note the following identities:

$$\frac{d}{dy} p(y - \theta) = -\frac{d}{d\theta} p(y - \theta) \quad \text{and} \quad \lambda^2 \frac{d}{d\theta} \{N(\theta | 0, \lambda^2)\} = \theta N(\theta | 0, \lambda^2).$$

Clearly,

$$E(\theta|y) = \frac{1}{m(y)} \int \theta p(y - \theta) N(\theta | 0, \lambda^2) \pi(\lambda) d\theta d\lambda.$$

Using integration by parts and the above identities, we obtain

$$E(\theta|y) = \frac{1}{m(y)} \int \frac{d}{dy} p(y - \theta) N(\theta | 0, \lambda^2) \pi^*(\lambda) d\theta d\lambda,$$

from which the result follows directly. □

After some algebra, (3) can be shown to reduce to (2) where $p(y - \theta)$ is a normal

likelihood. This extends the results from Masreliez (1975), Polson (1991), and Pericchi and Smith (1992) to a more general family of scale mixtures of normals. Hence we will freely use the more familiar (2) in the rest of the paper.

This result also provides a complementary insight about an estimator's behavior for y near zero—precisely the case for sparse data. The key element is $\pi^*(\theta)$, the unnormalized prior density for θ under $\pi^*(\lambda) \equiv \lambda^2\pi(\lambda)$. If $\pi^*(\theta)$ ensures that $d/dy m^*(y)$ is small, then $E(\theta | y)$ will be strongly shrunk to 0. The region where $m^*(y)$ is flat thus corresponds to the region where the estimator declares y to be noise.

To understand at an intuitive level why the horseshoe prior yields both kinds of robustness, note two facts. First, its Cauchy-like tails ensure a redescending score function, in the venerable tradition of robust Bayes estimators involving heavy-tailed priors. Second, since $\pi(\lambda) \propto 1/(1 + \lambda^2)$, then $\pi^*(\lambda) \propto \lambda^2/(1 + \lambda^2)$. This is essentially uniform, leading to $m^*(y)$ having small derivative in a larger neighborhood near the origin than other priors. By Theorem 2, this will yield strong shrinkage of noise.

We now formalize these intuitions through a detailed study of the behavior of the horseshoe estimator.

2.2 The horseshoe score function

The following results speak to the horseshoe's robustness to large, outlying signals.

Theorem 3. *Suppose $y \sim N(\theta, 1)$. Let $m_H(y)$ denote the predictive density under the horseshoe prior for known scale parameter τ , i.e. where $(\theta | \lambda) \sim N(0, \tau^2\lambda^2)$ and $\lambda \sim C^+(0, 1)$. Then*

$$\lim_{|y| \rightarrow \infty} \frac{d}{dy} \ln m_H(y) = 0.$$

Proof. Clearly,

$$m_H(y) = \frac{1}{\sqrt{2\pi^3}} \int_0^\infty \exp\left(-\frac{y^2/2}{1 + \tau^2\lambda^2}\right) \frac{1}{\sqrt{1 + \tau^2\lambda^2}} \frac{1}{1 + \lambda^2} d\lambda.$$

Make a change of variables to $z = 1/(1 + \tau^2\lambda^2)$. Then

$$\begin{aligned} m_H(y) &= \frac{1}{\sqrt{2\pi^3}} \int_0^1 \exp(-zy^2/2) (1-z)^{-1/2} \left\{ \frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right) z \right\}^{-1} dz \\ &= \frac{2}{\tau\sqrt{2\pi^3}} \exp\left(-\frac{y^2}{2}\right) \Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}, \frac{y^2}{2}, 1 - \frac{1}{\tau^2}\right), \end{aligned} \tag{4}$$

where $\Phi_1(\alpha, \beta, \gamma, x, y)$ is the degenerate hypergeometric function of two variables (Gradshteyn and Ryzhik, 1965, 9.261).

By a similar transformation, it is easy to show that

$$\frac{d}{dy} m_H(y) = -\frac{4y}{3\tau\sqrt{2\pi^3}} \Phi_1\left(\frac{1}{2}, 1, \frac{5}{2}, \frac{y^2}{2}, 1 - \frac{1}{\tau^2}\right).$$

Hence

$$\frac{d}{dy} \ln m_H(y) = -\frac{2y \Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}, \frac{y^2}{2}, 1 - \frac{1}{\tau^2}\right)}{3 \Phi_1\left(\frac{1}{2}, 1, \frac{5}{2}, \frac{y^2}{2}, 1 - \frac{1}{\tau^2}\right)} \quad (5)$$

Next, note the following identity from Gordy (1998):

$$\Phi_1(\alpha, \beta, \gamma, x, y) = \exp(x) \sum_{n=0}^{\infty} \frac{(\alpha)_n (\beta)_n y^n}{(\gamma)_n n!} {}_1F_1(\gamma - \alpha, \gamma + n, -x)$$

for $0 \leq y < 1$, $0 < \alpha < \gamma$, where ${}_1F_1(a, b, x)$ is Kummer's function of the first kind, and $(a)_n$ is the rising factorial. Also note that for $y < 0$, $0 < \alpha < \gamma$,

$$\Phi_1(\alpha, \beta, \gamma, x, y) = \exp(x)(1-y)^{-\beta} \Phi_1\left(\gamma - \alpha, \beta, \gamma, -x, \frac{y}{y-1}\right).$$

The final identities necessary are from Chapter 4 of Slater (1960):

$$\begin{aligned} {}_1F_1(a, b, x) &= \frac{\Gamma(a)}{\Gamma(b)} e^x x^{a-b} \{1 + O(x^{-1})\}, \quad x > 0 \\ {}_1F_1(a, b, x) &= \frac{\Gamma(a)}{\Gamma(b-a)} (-x)^{-a} \{1 + O(x^{-1})\}, \quad x < 0 \end{aligned}$$

for real-valued x .

Hence regardless of the sign of $1 - 1/\tau^2$, expansion of (5) by combining these identities yields a polynomial of order y^2 or greater remaining in the denominator, from which the result follows. \square

Using the previously quoted identities, it is easy to show that, for fixed τ , the difference between y and $E(\theta | y)$ is bounded for all y . The horseshoe prior is therefore of bounded influence, with the bound decaying to 0 independently of τ for large $|y|$. Denoting this bound by b_τ and the horseshoe posterior mean by $\hat{\theta}^H$ for a vector θ of fixed dimension p , we have the following corollary.

Corollary 4. $E_{y|\theta}(\|\theta - \hat{\theta}^H\|^2)$ is uniformly bounded for all θ .

Proof.

$$\begin{aligned} E \left\{ \sum_{i=1}^p (\theta_i - \hat{\theta}_i^H)^2 \right\} &< E \left\{ \sum_{i=1}^p (|\theta_i - y| + b_\tau)^2 \right\} \\ &= p + pb_\tau^2 \end{aligned}$$

\square

Finally, by applying the result from (2) to (5), we get an explicit form for the posterior mean:

$$E_H(\theta_i | y_i) = y_i \left\{ 1 - \frac{2\Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}, \frac{y_i^2}{2}, 1 - \frac{1}{\tau^2}\right)}{3\Phi_1\left(\frac{1}{2}, 1, \frac{5}{2}, \frac{y_i^2}{2}, 1 - \frac{1}{\tau^2}\right)} \right\}. \quad (6)$$

Combining (6) with the closed-form marginal in (4) allows an empirical-Bayes estimate $E(\theta | y, \hat{\tau})$ to be computed extremely fast, even in very high dimensions.

2.3 Joint distribution for τ and the λ_i 's

With the exception of Corollary 4, the above results describe the behavior of the horseshoe estimator for a single θ_i when τ is known. Usually, however, one faces p different θ_i 's along with unknown τ , leading to a joint distribution for $(\tau, \lambda_1, \dots, \lambda_p)$. Inspection of this joint distribution yields an understanding of how sparsity is handled under the horseshoe model.

Let $y = (y_1, \dots, y_p)$. Recall that $\kappa_i = 1/(1 + \tau^2 \lambda_i^2)$, and let $\kappa = (\kappa_1, \dots, \kappa_p)$. For the horseshoe prior, $\pi(\lambda_i) \propto 1/(1 + \lambda_i^2)$, and so

$$\pi(\kappa_i | \tau) \propto \kappa_i^{-1/2} (1 - \kappa_i)^{-1/2} \frac{1}{1 + (\tau^2 - 1)\kappa_i}.$$

Some straightforward algebra leads to

$$p(y, \kappa, \tau^2) \propto \pi(\tau^2) \tau^p \prod_{i=1}^p \frac{e^{-\kappa_i y_i^2/2}}{\sqrt{1 - \kappa_i}} \prod_{i=1}^p \frac{1}{\tau^2 \kappa_i + 1 - \kappa_i}. \quad (7)$$

Form (7) yields several insights. As in other common multivariate scale mixtures, the global shrinkage parameter τ is conditionally independent of y , given κ . Similarly, the κ_i 's are conditionally independent of each other, given τ .

More interestingly, the marginal posterior density for κ_i is always unbounded as $\kappa_i \rightarrow 1$, regardless of τ . The density, of course, remains integrable on the unit interval, with τ strongly influencing the amount of probability given to any set of positive Lebesgue measure containing $\kappa_i = 1$. Indeed, despite its unboundedness near 1, the posterior density of κ_i can quite heavily favor $\kappa_i \approx 0$, implying very little shrinkage in the posterior mean.

Finally, (7) clarifies that the global shrinkage parameter τ is estimated by the average ‘‘signal density.’’ To see this, observe that if p is large, the conditional posterior distribution for τ^2 , given κ , is well approximated by substituting $\bar{\kappa} = p^{-1} \sum_{i=1}^p \kappa_i$ for

each κ_i . Ignoring the contribution of the prior for τ^2 , this gives, up to a constant,

$$\begin{aligned} p(\tau^2 \mid \kappa) &\approx (\tau^2)^{-p/2} \left(1 + \frac{1 - \bar{\kappa}}{\tau^2 \bar{\kappa}}\right)^{-p} \\ &\approx (\tau^2)^{-p/2} \exp\left\{-\frac{1}{\tau^2} \frac{p(1 - \bar{\kappa})}{\bar{\kappa}}\right\}, \end{aligned}$$

or approximately a $\text{Ga}(\frac{p+2}{2}, \frac{p-p\bar{\kappa}}{\bar{\kappa}})$ distribution for $1/\tau^2$. If $\bar{\kappa}$ is close to 1, implying that most observations are shrunk near 0, then τ^2 will be very small with high probability, with an approximate mean $\mu = 2(1 - \bar{\kappa})/\bar{\kappa}$ and standard deviation of $\mu/\sqrt{p-2}$.

3 Comparison with other Bayes estimators

3.1 Comparison with multivariate scale mixtures

The advantage of the horseshoe estimator can now be stated directly. In sparse situations, posterior learning about τ allows most noise observations to be shrunk very near zero. Yet this small value of τ will not inhibit the estimation of large, obvious signals, due to a redescending score function.

This set of features is not shared by common Bayes estimators based upon other multivariate scale mixtures of normals. Take, for example, the commonly used double-exponential prior, where each λ_i has an independent $\text{Ex}(2)$ distribution. Results from Pericchi and Smith (1992) and Mitchell (1994) show that

$$\begin{aligned} E_{DE}(\theta_i \mid y_i) &= w_i(y_i + b) + (1 - w_i)(y_i - b) \\ w_i &= F(y_i)/\{F(y_i) + G(y_i)\} \\ F(y_i) &= e^{c_i} \Phi(-y - b) \\ G(y_i) &= e^{-c_i} \Phi(-y + b) \\ b &= \frac{\sqrt{2}}{\tau}, \quad c_i = \frac{\sqrt{2}(y - \mu)}{\tau}, \end{aligned}$$

where Φ is the normal cumulative distribution function. The double-exponential posterior mean thus has an interpretation as a data-based average of $y - b$ and $y + b$, with w_i becoming arbitrarily close to 0 for large positive y , or to 1 for large negative y . This can be seen in the score function, plotted in Figure 2.

Under the double-exponential prior, posterior learning about τ can also cause noise observations to be squelched. This phenomenon is well understood (Park and Casella, 2008; Hans, 2008); small values of τ lead to strong shrinkage near the origin, just as under the horseshoe.

Less appreciated, however, is the effect this has upon estimation in the tails. Notice that small values of τ yield large values of b . If the overall level of sparsity

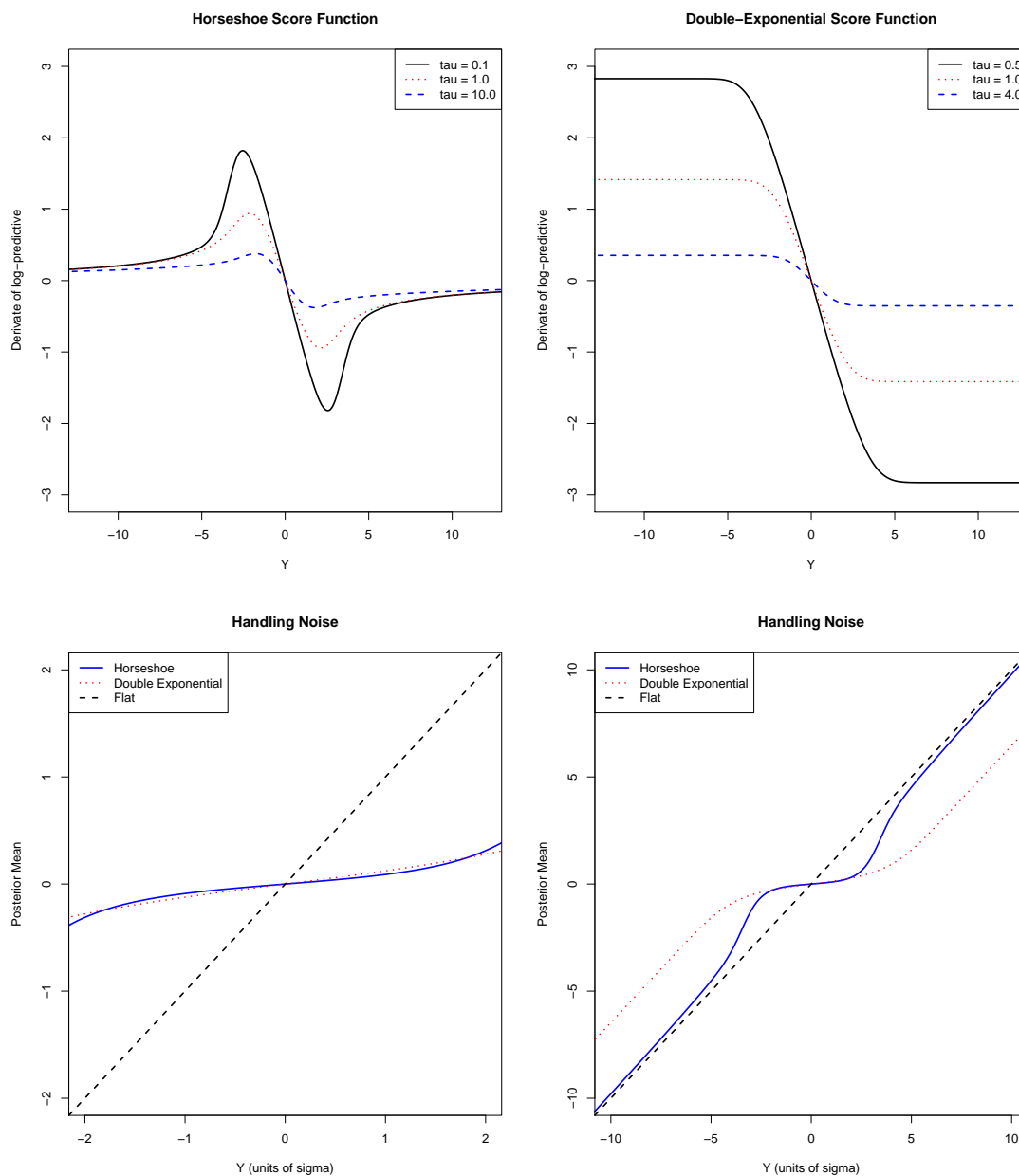


Figure 2: Top: A comparison of the score function for horseshoe and double-exponential priors for different values of τ . Bottom: a comparison of the posterior mean versus y for horseshoe and double-exponential priors for different values of τ , with the left pane zoomed in near the origin, and the right pane giving a broad view to encompass large signals.

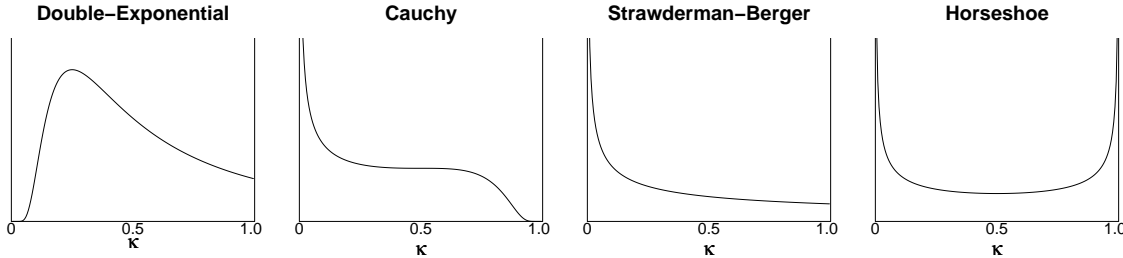


Figure 3: A comparison of the implied density for the shrinkage weights $\kappa_i \in [0, 1]$ for four different priors, where $\kappa_i = 0$ means no shrinkage and $\kappa_i = 1$ means total shrinkage to zero.

in θ is increased, τ must shrink to reduce risk at the origin. Yet the risk must be simultaneously increased in the tails, since $|\mathbb{E}_{DE}(\theta_i | y_i) - y_i| \approx b$ for large $|y_i|$. When θ is sparse, estimation of τ under the double-exponential model must strike a balance between reducing risk in estimating noise, and reducing risk in estimating large signals. This is demonstrated by the bottom two panels of Figure 2, which also show that no such tradeoff exists under the horseshoe prior.

These properties, and the properties of other common Bayes rules, can be understood intuitively in terms of the shrinkage coefficient κ_i . Table 1 gives the priors for κ_i implied by four different multivariate-scale-mixture priors: the double-exponential, the Strawderman–Berger density (Strawderman, 1971; Berger, 1980), the Cauchy, and the horseshoe. Figure 3 also plots these four densities.

Notice that at $\kappa_i = 0$, both $\pi_C(\kappa_i)$ and $\pi_{SB}(\kappa_i)$ are both unbounded, while $\pi_{DE}(\kappa_i)$ vanishes. This suggests that Cauchy and Strawderman–Berger estimators will be good, and double-exponential estimators will be mediocre, at leaving large signals unshrunk. Meanwhile, at $\kappa_i = 1$, π_C tends to zero, while both π_{DE} and π_{SB} tend to fixed constants. This suggests that Cauchy estimators will be bad, while double-exponential and Strawderman–Berger estimators will be mediocre, at correctly shrinking noise all the way to 0.

The horseshoe prior, on the other hand, implies a $\text{Beta}(1/2, 1/2)$ distribution for κ_i . This is clearly unbounded both at 0 and 1, allowing both signals and noise to be accommodated by a single prior.

3.2 An illustrative example

An example will help illustrate these ideas. Two standard normal observations were simulated for each of 1000 means: 10 signals of mean 10, 90 signals of mean 2, and 900 noise of mean 0. Two models were then fit to this data: one that used independent horseshoe priors for each θ_i , and one that used independent double-exponential priors. For both models, a half-Cauchy prior was used for the global scale parameter τ

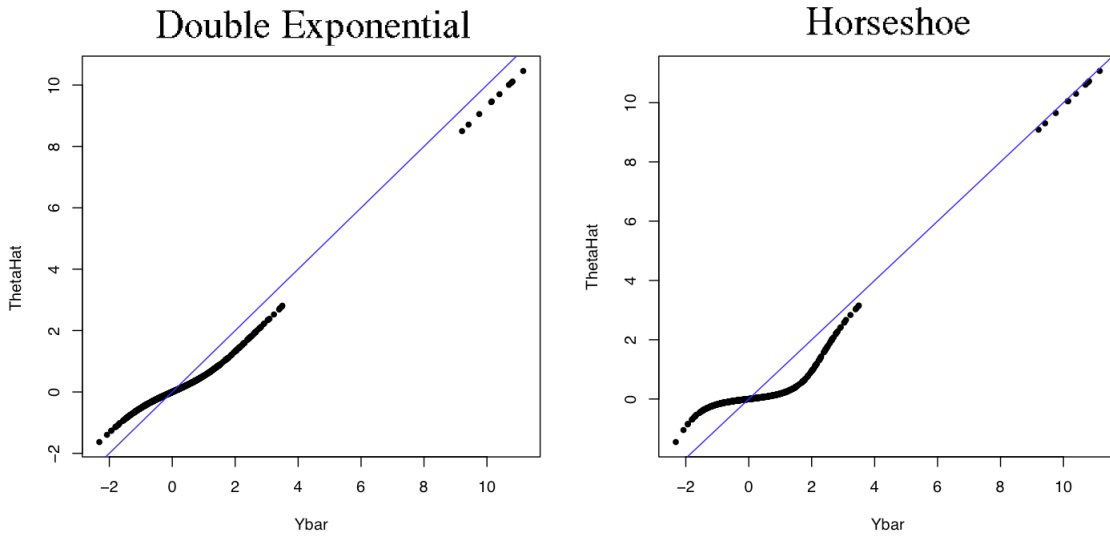


Figure 4: Plots of \bar{y}_i versus $\hat{\theta}_i$ for double-exponential (left) and horseshoe (right) priors on data where most of the means are zero. The diagonal lines are where $\theta_i = \bar{y}_i$.

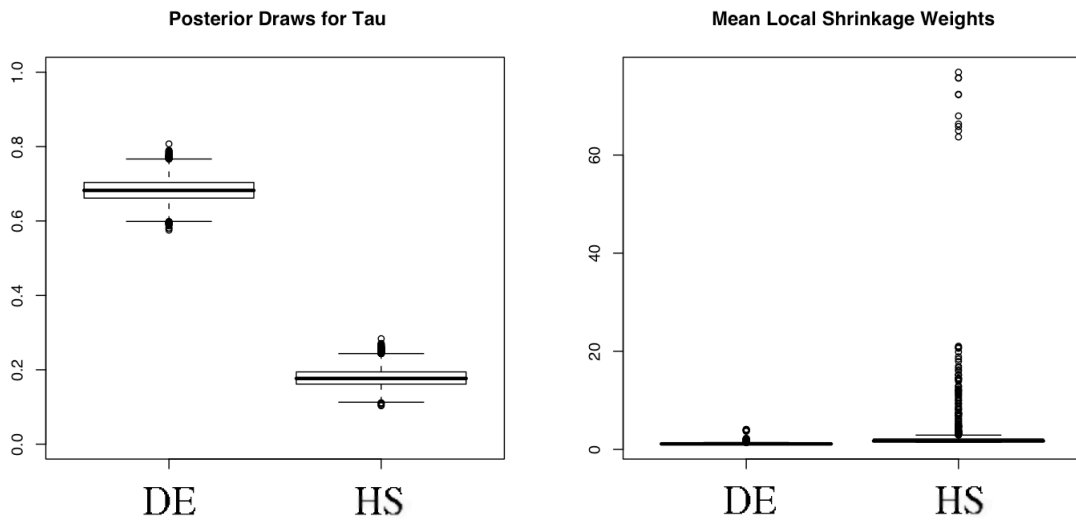


Figure 5: Left: posterior draws for the global shrinkage parameter τ under both double-exponential and horseshoe for the toy example. Right: boxplot of $\hat{\lambda}_i$'s, the posterior means of the local shrinkage parameters λ_i .

Prior for θ_i	Prior for λ_i	Prior for κ_i
Double-exponential	$\lambda_i^2 \sim \text{Ex}(2)$	$\pi_{DE}(\kappa_i) \propto \kappa_i^{-2} e^{-\frac{1}{2\kappa_i}}$
Cauchy	$\lambda_i \sim \text{IG}(1/2, 1/2)$	$\pi_C(\kappa_i) \propto \kappa_i^{-\frac{1}{2}} (1 - \kappa_i)^{-\frac{3}{2}} e^{-\frac{\kappa_i}{2(1-\kappa_i)}}$
Strawderman–Berger	$\pi(\lambda_i) \propto \lambda_i(1 + \lambda_i^2)^{-3/2}$	$\pi_{SB}(\kappa_i) \propto \kappa_i^{-\frac{1}{2}}$
Horseshoe	$\lambda_i \sim \text{C}^+(0, 1)$	$\pi_H(\kappa_i) \propto \kappa_i^{-1/2} (1 - \kappa_i)^{-1/2}$

Table 1: The implied priors for κ_i and λ_i associated with some common priors for shrinkage and sparsity.

following the advice of Gelman (2006) and Scott and Berger (2006), and Jeffreys’ prior $\pi(\sigma) \propto 1/\sigma$ was used for the error variance.

The shrinkage characteristics of these two fits are summarized in Figure 4. These plots show the posterior mean $\hat{\theta}_i^{DE}$ and $\hat{\theta}_i^H$ as a function of the observed data \bar{y}_i , with the diagonal lines showing where $\hat{\theta}_i = \bar{y}_i$. Key differences occur near $\bar{y}_i \approx 0$ and $\bar{y}_i \approx 10$. Compared with the horseshoe prior, the double-exponential prior tends to shrink small observations not enough, and the large observations too much.

These differences can also be seen in Figure 5. The left panel shows that the global shrinkage parameter τ is estimated to be much smaller under the horseshoe model than under the double-exponential model, roughly 0.2 versus 0.7. But under the horseshoe model, the local shrinkage parameters can take on quite large values and hence overrule this global shrinkage; this handful of large λ_i ’s under the horseshoe prior, corresponding to the observations near 10, can be seen in the right panel.

The horseshoe prior clearly does better at handling both aspects of the problem: leaving large signals unshrunk while squelching most of the noise. The double-exponential prior requires a delicate balancing act in estimating τ , which affects error near 0 and error in the tails. That this balance is often hard to strike is reflected in the mean-squared error, which was about 25% lower under the horseshoe model for this example.

3.3 Comparison with discrete-mixture rules

As an alternative procedure, consider a discrete mixture model of the form

$$\theta_i \sim (1 - w)\delta_0 + w g(\theta_i). \tag{8}$$

with the mixing probability w , often called the prior inclusion probability, being unknown. Sparse mixture models of this sort have become quite the workhorse for a wide variety of sparse problems; see Berry (1988) and Mitchell and Beauchamp (1988) for discussion and other references on their early development.

The crucial choice here is that of g , which must allow convolution with a normal

likelihood in order to evaluate the predictive density under the alternative model g . One common choice is a normal prior, the properties of which are well understood in a multiple-testing context (Scott and Berger, 2006; Bogdan et al., 2008b). Also see Johnstone and Silverman (2004) for an empirical-Bayes treatment of a heavy-tailed version of this model.

For these and other conjugate choices of g , it is straightforward to compute the posterior inclusion probabilities, $w_i = \Pr(\theta_i \neq 0 \mid y)$. These quantities will adapt to the level of sparsity in the data through shared dependence upon the unknown mixing probability w . This effect can most easily be seen if one imagines testing a small number of signals in the presence of an increasingly large number of noise observations. As the noise comes to predominate, the posterior for w concentrates near 0, making it increasingly more difficult for any w_i to be large.

The discrete mixture can be thought of as adding a point mass at $\kappa_i = 1$, allowing total shrinkage. In broad terms, this is much like the unbounded density under the horseshoe prior as $\kappa_i \rightarrow 1$, suggesting that these models may be similar in the degree to which they shrink noise variables to 0.

It is therefore interesting to consider the differences between shrinkage profiles of the horseshoe (π_H) and the discrete mixture using Strawderman–Berger priors (π_{DM}):

$$\pi_{DM} : (\theta_i \mid \kappa_i) \sim (1 - w) \delta_0 + w \text{N} \left(0, \frac{\tau^2 + \sigma^2}{2\kappa_i} - \sigma^2 \right),$$

with $\kappa_i \sim \text{Be}(1/2, 1)$ and $\tau > \sigma$. The Strawderman–Berger prior has a number of desirable properties for describing the nonzero θ_i 's, since it is both heavy-tailed and yet still allows closed-form convolution with the normal likelihood.

Both the horseshoe and the discrete mixture have global scale parameters τ and local shrinkage weights κ_i . Yet τ plays a very different role in each model. Under the horseshoe prior,

$$E_H(\theta_i \mid \kappa_i, \tau, y_i) = \left(1 - \frac{\kappa_i}{\kappa_i + \tau^2(1 - \kappa_i)} \right) y_i, \quad (9)$$

recalling that σ^2 is assumed to be 1. And in the mixture model,

$$E_{DM}(\theta_i \mid \kappa_i, w_i, y_i) = w_i \left(1 - \frac{2\kappa_i}{1 + \tau^2} \right) y_i,$$

where w_i is the posterior inclusion probability. Let $G^*(y_i \mid \tau)$ denote the predictive density, evaluated at y_i , under the Strawderman–Berger prior. Then these probabilities are

$$\frac{w_i}{1 - w_i} = \frac{w G^*(y_i \mid \tau)}{(1 - w) \text{N}(y_i \mid 0, 1)}. \quad (10)$$

Several differences between the approaches are apparent:

- In the discrete model, local shrinkage is controlled by the Bayes factor in (10),

which is a function of the signal-to-noise ratio τ/σ . In the scale-mixture model, local shrinkage is determined entirely by the κ_i 's, or equivalently the λ_i 's.

- In the discrete model, global shrinkage is primarily controlled through the prior inclusion probability w . In the scale-mixture model, global shrinkage is controlled by τ , since if τ is small in (9), then κ_i must be very close to 0 for extreme shrinkage to be avoided. Hence in the former model, w adapts to the overall sparsity of θ , while in the latter model, τ performs this role.
- There will be a strong interaction between w and τ in the discrete model which is not present in the scale-mixture model. Intuitively, as w changes, τ must adapt to the scale of the observations that are reclassified as signals or noise.

Nonetheless, these structural differences between the procedures are small compared to their operational similarities, as Sections 4 and 5 will show. Both models imply priors for κ_i that are unbounded at 0 and at 1. They have similarly favorable risk properties for estimation under squared-error or absolute-error loss. And perhaps most remarkably, they yield thresholding rules that are nearly indistinguishable in practice despite originating from very different goals.

4 Estimation Risk

4.1 Overview

In this section, we describe the results of a large bank of simulation studies meant to assess the risk properties of the horseshoe prior under both squared-error and absolute-error loss. We benchmark its performance against three alternatives: the double-exponential model, along with fully Bayesian and empirical-Bayes versions of the discrete-mixture model with Strawderman–Berger priors.

Since any estimator is necessarily optimal with respect to its assumed prior, we do not use any of these as the true model in our simulation study; this answers only an uninteresting question. But neither do we fix a single, supposedly representative θ and merely simulate different noise configurations, since this is tantamount to asking the equally uninteresting question of which prior best describes the arbitrarily chosen θ . Instead, we describe two studies in which we simulate repeatedly from models that correspond to archetypes of “strong” sparsity in Experiment 1 and “weak” sparsity in Experiment 2, but that match none of the priors we are evaluating.

For both the double-exponential and the horseshoe, the following default hyper-priors were used in all studies:

$$\begin{aligned}\pi(\sigma) &\propto 1/\sigma \\ (\tau \mid \sigma) &\sim C^+(0, \sigma).\end{aligned}$$

A slight modification is necessary in the fully Bayesian discrete-mixture model, since under the Strawderman–Berger prior of (3.3), τ must be larger than σ :

$$\begin{aligned} w &\sim \text{Unif}(0, 1) \\ \pi(\sigma) &\propto 1/\sigma \\ (\tau \mid \sigma) &\sim \text{C}(\sigma, \sigma) 1_{\tau \geq \sigma}. \end{aligned}$$

These are similar to the recommendations of Scott and Berger (2006), with the half-Cauchy prior on τ being appropriately scaled by σ and yet decaying only polynomially fast. In the empirical-Bayes approach, w , σ , and τ were estimated by marginal maximum likelihood, subject to the constraint that $\tau \geq \sigma$.

Experiment 1: Strongly sparse signals

Strongly sparse signals are vectors in which some of the components are identically zero. Since we are also interested in the question of robustness in detecting large, outlying signals, we simulate strongly sparse data sets from the following model:

$$\begin{aligned} y_i &\sim \text{N}(\theta_i, \sigma^2) \\ \theta_i &\sim w t_3(0, \tau) + (1 - w) \delta_0 \\ w &\sim \text{Be}(1, 4), \end{aligned}$$

where the nonzero θ_i 's follow a t distribution with 3 degrees of freedom, and where θ has 20% nonzero entries on average.

We simulated from this model under many different configurations of the signal-to-noise ratio. In all cases τ was fixed at 3, and we report the results on 1000 simulated data sets for each of $\sigma^2 = 1$ and $\sigma^2 = 9$. Each simulated vector was of length 250.

Experiment 2: Weakly sparse signals

A vector θ is considered weakly sparse if none of its components are identically zero, but its component nonetheless follow some kind of power-law or l^α decay; see Johnstone and Silverman (2004) for a more formal description. Weakly sparse vectors have most of their total “energy” concentrated on a relatively few number of elements.

We simulate 1000 weakly sparse data sets of 250 means each, where

$$\begin{aligned} y_i &\sim \text{N}(\theta_i, \sigma^2) \\ (\theta_i \mid \eta, \alpha) &\sim \text{Unif}(-\eta c_i, \eta c_i) \\ \eta &\sim \text{Ex}(2) \\ \alpha &\sim \text{Unif}(a, b), \end{aligned}$$

for $c_i = n^{1/\alpha} i^{-1/\alpha}$ for $i = 1, \dots, n$. These θ 's correspond to a weak- l^α bound on

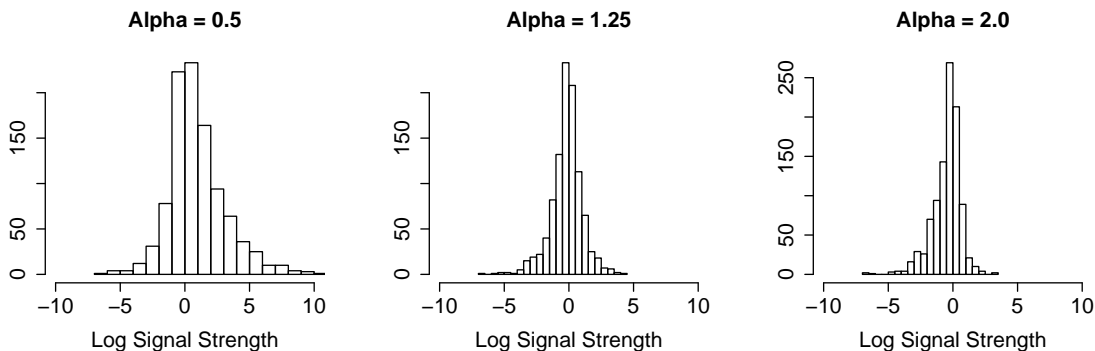


Figure 6: Log signal strength for three weak- l^α vectors of 1000 means, with α at 0.5, 1.25, and 2.0.

the coefficients, as described by Johnstone and Silverman (2004): the ordered θ_i 's follow a power-law decay, with the size of the largest coefficients controlled by the exponentially distributed random bound η , and the speed of the decay controlled by the random norm α .

Two experiments were conducted: one where α was drawn uniformly on $(0.5, 1)$, and another where α was uniform on $(1, 2)$. Small values of α give vectors where the cumulative signal strength is concentrated on a few very large elements. Larger values of α , on the other hand, yield vectors where the signal strength is more uniformly distributed among the components. For illustration, see Figure 6, which shows the log signal-strength of three weak- l^α vectors of 1000 means with α fixed at each of 0.5, 1.25, and 2.0.

4.2 Results

The results of Experiments 1 and 2 are summarized in Tables 2 and 3

Two conclusions are readily apparent from the tables. First, the double-exponential systematically loses out to the horseshoe, and to both versions of the discrete mixture rule, under both squared-error and absolute-error loss. The difference in performance is substantial. In experiment 1, the double-exponential averaged between 50% and 75% more risk regardless of the specific value of σ^2 and regardless of which loss function is used. In experiment 2, the double-exponential typically had ≈ 25 –40% more risk.

Close inspection of some of these simulated data sets led us to conclude that the double-exponential prior loses on both ends here, much as it did in the toy example summarized in Figures 4 and 5. It lacks tails that are heavy enough to robustly estimate the large t_3 signals, and it also lacks sufficient mass near 0 to adequately squelch the substantial noise in θ .

		$\sigma^2 = 1$				$\sigma^2 = 9$			
		DE	HS	DMF	DME	DE	HS	DMF	DME
SE Loss	DE	209	1.62	1.62	1.71	850	1.47	1.51	1.50
	HS		77	0.95	1.04		416	0.99	0.99
	DMF			93	1.18			440	1.00
	DME				74				437
AE Loss	DE	178	1.50	1.60	1.73	341	1.56	1.75	1.76
	HS		80	1.02	1.13		142	1.10	1.10
	DMF			83	1.20			123	1.00
	DME				60				122

Table 2: Risk under squared-error loss and absolute-error loss in experiment 1. Bold diagonal entries in the top and bottom halves are median sum of squared-errors and absolute errors, respectively, in 1000 simulated data sets. Off-diagonal entries are average risk ratios, risk of row divided by risk of column, in units of σ . DE: double exponential. HS: horseshoe. DMF: discrete mixture, fully Bayes. DME: discrete mixture, empirical Bayes.

Like any shrinkage estimator, the double-exponential model is fine when its prior describes reality, but suffers in problems that correspond to very reasonable, commonly held notions of sparsity. The horseshoe, on the other hand, allows each shrinkage coefficient κ_i to be arbitrarily close to 0 or 1, and hence can accurately estimate both signal and noise.

Second, it is equally interesting that no meaningful systematic edges could be found for any of the other three approaches: the horseshoe, the Bayes discrete mixture, or the empirical-Bayes discrete mixture. All three models have heavy tails, and all three models can shrink y_i arbitrarily close to 0. Though we have summarized only a limited set of results here, this trend held for a wide variety of other data sets that we investigated.

By and large, the horseshoe estimator, despite providing a one-group answer, acts just like a model-averaged Bayes estimator arising from a two-group mixture.

5 Classification Risk

5.1 Overview

We now describe a simple thresholding rule for the horseshoe estimator that can yield accurate decisions about whether each θ_i is signal or noise. As we have hinted before, these classifications turn out to be nearly indistinguishable from those of the Bayesian discrete-mixture model under a simple 0–1 loss function, suggesting an even deeper correspondence between the two procedures than was shown in the previous section.

Recall that under the discrete mixture model of (8), the Bayes estimator for each θ_i is $\hat{\theta}_i^{DM} = w_i E_g(\theta_i | y_i)$, where w_i is the posterior inclusion probability for

		$\alpha \in (0.5, 1.0)$				$\alpha \in (1.0, 2.0)$			
		DE	HS	DMF	DME	DE	HS	DMF	DME
SE Loss	DE	231	1.36	1.42	1.38	139	1.34	1.34	1.32
	HS		170	1.05	1.01		69	0.97	0.96
	DMF			194	0.95			73	0.99
	DME				227				73
AE Loss	DE	189	1.23	1.31	1.30	144	1.23	1.24	1.23
	HS		148	1.07	1.05		91	1.00	0.99
	DMF			142	0.98			92	1.00
	DME				150				92

Table 3: Risk under squared-error loss and absolute-error loss in experiment 2. Bold diagonal entries in the top and bottom halves are median sum of squared-errors and absolute errors, respectively, in 1000 simulated data sets. Off-diagonal entries are average risk ratios, risk of row divided by risk of column, in units of σ . DE: double exponential. HS: horseshoe. DMF: discrete mixture, fully Bayes. DME: discrete mixture, empirical Bayes.

θ_i . For appropriately heavy-tailed g such as the Strawderman–Berger prior of Section 3.3, this expression is approximately $w_i y_i$, meaning that w_i can be construed in two different ways:

- as a posterior probability, which forms the basis for a classification rule that is optimal in both Bayesian and frequentist senses.
- as an indicator of how much shrinkage should be performed on y_i , thereby giving rise to an estimator $\hat{\theta}_i^{DM} \approx w_i y_i$ with excellent risk properties under squared-error and absolute-error loss.

The horseshoe estimator also yields “significance weights” $w_i = 1 - \kappa_i$, with $\hat{\theta}_i^H = w_i y_i$. Though these lack an interpretation as posterior probabilities, the previous section showed that these weights behave similarly to those arising from the discrete mixture. Hence by analogy with the decision rule one would apply to the discrete-mixture w_i ’s under a symmetric 0–1 loss function, one possible threshold is to call θ_i a signal if the horseshoe yields $w_i \geq 0.5$, and to call it noise otherwise.

The following simulations will demonstrate the surprising fact that, even though the horseshoe w_i ’s are not posterior probabilities, and even though the horseshoe model itself makes no allowance for two different groups, this simple thresholding rule nonetheless displays very strong control over the number of false-positive classifications. Indeed, in problem after problem, it is hard to tell the difference between the w_i ’s from the two-group model and those from the horseshoe.

We will study two different asymptotic scenarios: fixed- k asymptotics, and what we call “ideal signal-recovery” asymptotics. As before, we will benchmark the horseshoe against the double-exponential.

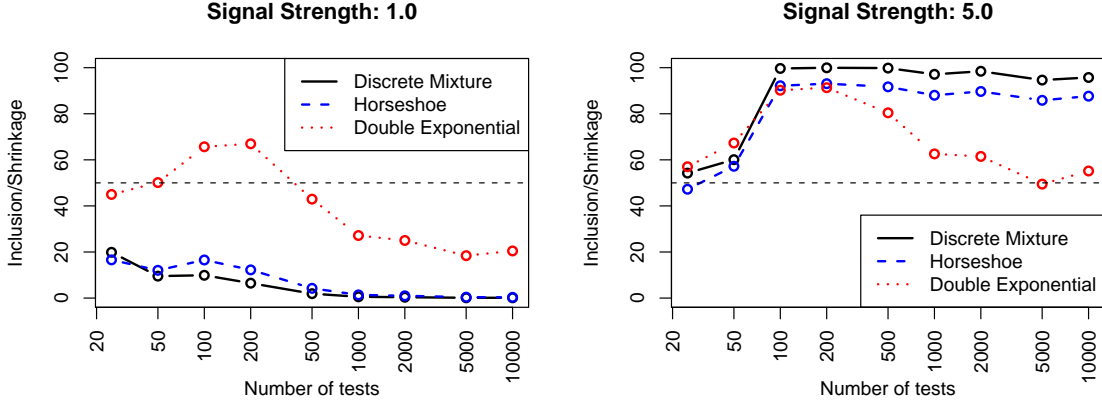


Figure 7: Inclusion probabilities/significance weights w_i versus number of tests n in Experiment 3.

Experiment 3: Fixed- k asymptotics

Under fixed- k asymptotics, the number of true signals remains fixed, while the number of noise observations grows without bound. We study this asymptotic regime by fixing 10 true signals that are repeatedly tested in the face of an increasingly large number of noise observation. The error variance σ^2 remains fixed at 1 throughout. The 10 signals were the half-integers between 0.5 and 5.0 with random signs.

Experiment 4: Ideal signal-recovery asymptotics

Unfortunately, fixed- k asymptotics are in some sense hopeless for signal recovery: as $n \rightarrow \infty$, every signal must eventually be classified as noise under the discrete mixture model, since each Bayes factor remains bounded while the prior odds ratio comes to favor the null hypothesis arbitrarily strongly.

A set of asymptotic conditions does exist, however, that makes near-perfect signal recovery possible. Suppose that the nonzero θ_i 's follow a $N(0, \tau^2)$ distribution. Define $s = \tau^2/\sigma^2$, and recall that p is the fraction of signals among all observations. Bogdan et al. (2008a) show that if

$$s \rightarrow \infty \tag{11}$$

$$w \rightarrow 0 \tag{12}$$

$$s^{-1} \log \left(\frac{1-w}{w} \right) \rightarrow C \text{ for some } 0 < C < \infty, \tag{13}$$

then as the number of tests n grows without bound, the probability of a false positive ($w_i \geq 0.5, \theta_i = 0$) converges to 0, while the probability of a false negative ($w_i <$

0.5, $\theta_i \neq 0$) converges to a fixed constant less than one, the exact value of which will depend upon C .

Essentially, the situation is one in which the fraction of signal observations can approach 0 as long as the signal-to-noise ratio s is growing larger at a sufficiently rapid rate. The Bayesian mixture model can then recover all but a fixed fraction of the true signals without committing any Type-I errors.

To study ideal signal-recovery asymptotics, we used the same 10 signal observations as under the fixed- k asymptotics. The variance of the noise observations, however, decayed to 0 as n grew, instead of remaining fixed at 1:

$$\sigma_n^2 = \frac{D}{\log\left(\frac{n-k}{k}\right)},$$

where D is any constant, n is the total number of tests being performed, and k is the fixed number of true signals. We chose $D = 0.4$ and $k = 10$. It is easy to show that as $n \rightarrow \infty$, the conditions of (11), (12), and (13) will hold, and the results of Bogdan et al. (2008a) will obtain, if for each sample size the noise variance is σ_n^2 .

5.2 Results

The general character of the results can be understood from Figure 7. These two plots shows the inclusion probabilities/significance weights for two signals—a weak signal of 1σ , and a strong signal of 5σ —as the number of noise observations increases gradually from 10 to 10,000 under fixed- k asymptotics.

Intuitively, the weak signal should rapidly be overwhelmed by noise, while the strong signal should remain significant. This is precisely what happens under both the discrete mixture and the horseshoe prior, whose significance weights coincide almost perfectly as n grows. This is not, however, what happens under the double-exponential prior, which shrinks the strong signal almost as much as it shrinks the weak signal at all levels.

Comprehensive results for Experiments 3 and 4 are given in Tables 4–9. A close inspection of these numbers bears out the story of Figure 7: regardless of the asymptotic regime and regardless of the signal strength, the horseshoe significance weights are a good stand-in for the posterior inclusion probabilities under the discrete mixture. They lead to nearly identical numerical summaries of the strength of evidence in the data, and nearly identical classifications of signal versus noise. One anomaly worth mentioning occurs in Table 5.2, which shows 11 false positives on 50 tests for the discrete mixture, compared to none for the horseshoe. This difference seems large, but further inspection showed that of these 11 discrepancies, all but one had their w_i 's within 5% for the horseshoe and the discrete mixture.

The horseshoe thresholding rule is, of course, quite accurate in its own right, aside from its correspondence with the discrete mixture. As the tables show, it exhibits

# Tests	Signal Strength										FP
	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	
25	18	20	23	27	31	35	39	44	49	54	0
50	8	10	12	15	19	25	32	41	50	60	0
100	8	10	15	27	46	69	86	95	99	100	0
200	5	6	11	21	42	70	90	98	100	100	2
500	1	2	3	6	14	35	67	91	98	100	1
1000	0	1	1	2	3	9	24	55	85	97	0
2000	0	0	1	1	3	8	24	59	89	98	0
5000	0	0	0	0	1	3	9	32	72	95	0
10000	0	0	0	0	1	3	9	32	74	96	3

Table 4: Posterior probabilities as n grows for 10 fixed signals; discrete mixture model, fixed- k asymptotics.

# Tests	Signal Strength										FP
	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	
25	15	17	18	21	24	28	32	37	42	47	0
50	11	12	14	17	20	26	33	40	49	57	0
100	14	17	22	31	46	62	75	85	89	92	0
200	11	12	17	26	43	63	79	87	91	93	1
500	4	4	6	10	18	36	61	80	89	92	1
1000	1	1	2	3	5	10	25	52	76	88	0
2000	1	1	1	2	4	9	24	54	80	90	0
5000	0	0	0	1	1	3	10	33	67	86	0
10000	0	0	0	1	1	3	10	30	68	88	2

Table 5: Significance weights as n grows for 10 fixed signals; horseshoe prior, fixed- k asymptotics.

# Tests	Signal Strength										FP
	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	
25	44	45	46	48	50	51	53	55	56	57	0
50	47	50	52	55	58	60	62	64	66	67	5
100	60	66	72	77	81	84	86	88	89	90	8
200	59	67	73	79	83	86	88	89	90	91	29
500	39	43	49	56	62	68	72	76	78	80	15
1000	26	27	30	34	39	44	50	54	59	63	0
2000	23	25	27	31	36	41	47	53	58	61	0
5000	18	18	20	22	25	30	35	40	45	49	0
10000	19	20	22	25	29	34	39	45	50	55	2

Table 6: Significance weights as n grows for 10 fixed signals; double-exponential prior, fixed- k asymptotics.

# Tests	Signal Strength										FP
	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	
25	12	13	15	18	20	24	27	31	36	42	0
50	45	63	82	94	98	100	100	100	100	100	11
100	21	37	68	92	99	100	100	100	100	100	3
200	8	23	69	97	100	100	100	100	100	100	3
500	3	13	67	99	100	100	100	100	100	100	2
1000	2	11	76	100	100	100	100	100	100	100	1
2000	1	9	79	100	100	100	100	100	100	100	2
5000	1	5	76	100	100	100	100	100	100	100	0
10000	0	4	82	100	100	100	100	100	100	100	1

Table 7: Posterior probabilities as n grows for 10 fixed signals; discrete mixture model, ideal signal-recovery asymptotics.

# Tests	Signal Strength										FP
	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	
25	12	13	14	16	18	21	24	28	32	38	0
50	38	54	71	84	90	94	95	97	97	98	0
100	25	40	64	82	91	94	96	97	97	98	1
200	16	30	64	86	92	95	96	97	98	98	2
500	7	19	62	89	94	96	97	98	98	99	1
1000	5	15	69	91	95	96	97	98	98	99	0
2000	3	11	70	92	95	97	98	98	99	99	0
5000	1	6	70	93	96	97	98	98	99	99	0
10000	1	5	74	94	96	97	98	99	99	99	0

Table 8: Significance weights as n grows for 10 fixed signals; horseshoe prior, ideal signal-recovery asymptotics.

# Tests	Signal Strength										FP
	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	
25	37	39	40	41	43	44	45	47	48	50	0
50	89	93	95	96	97	97	98	98	98	98	0
100	95	97	98	98	99	99	99	99	99	99	4
200	81	90	93	95	96	97	97	97	98	98	3
500	71	84	89	92	94	95	95	96	96	97	6
1000	69	83	89	92	93	94	95	96	96	97	7
2000	57	74	83	87	90	91	93	94	94	95	12
5000	44	63	75	81	85	87	89	91	92	92	11
10000	36	54	69	77	81	85	87	88	90	91	9

Table 9: Significance weights as n grows for 10 fixed signals; double-exponential prior, ideal signal-recovery asymptotics.

very strong control over the number of false positive declarations while retaining a reasonable amount of power, even under fixed- k asymptotics.

On the other hand, there is no sense in which the significance weights from the double-exponential can be trusted to sort signal from noise. These weights are inappropriately uniform as a function of signal strength, buttressing our analytic results from earlier that the underlying joint model for τ and κ_i cannot adapt sufficiently to the degree of sparsity in the data.

6 Estimation of Hyperparameters

6.1 Bayes, empirical Bayes, and cross-validation

This section discusses the estimation of key model hyperparameters.

One possibility is to proceed with a fully Bayesian solution by placing priors upon model hyperparameters, just as we have done throughout the rest of the paper. An excellent reference on hyperpriors for variance components can be found in Gelman (2006). A second possibility is to estimate σ and τ , along with w if the discrete mixture model is being used, by empirical Bayes. Marginal maximum-likelihood solutions along these lines are explored in, for example, George and Foster (2000) and Johnstone and Silverman (2004). A third possibility is cross-validation, a common approach when fitting double-exponential priors to regression models (Tibshirani, 1996).

Empirical Bayes solutions under the horseshoe prior, as we have already hinted, are extremely fast to compute, since once a two-dimensional optimization over τ and σ is undertaken, all posterior expectations are known. Caution, however, is in order. When few signals are present, it is quite common for the posterior mass of τ to concentrate near 0 and for the signals to be flagged via large values of the local shrinkage parameters λ_i . The marginal maximum-likelihood solution is therefore always in danger of collapsing to the degenerate $\hat{\tau} = 0$. See, for example, Tiao and Tan (1965).

Cross-validation will not suffer from this difficulty, but still involves plugging in a point estimate for the signal-to-noise ratio. This can be misleading, given that σ and τ will typically have an unknown correlation structure that should ideally be averaged over. Indeed, as the following example serves to illustrate, careful handling of uncertainty in the joint distribution for τ and σ can be crucial.

Example: Suppose the true model is $\theta = 20$ and $\sigma^2 = 1$. Two observations are available: $y_1 = 19.6$ and $y_2 = 20.4$. Two different Bayesian versions of the horseshoe model in (1) are entertained. In both cases, σ is unknown and assigned the noninformative prior $1/\sigma$. But in the first fit, τ is assigned a $C^+(0, 1)$ prior, while in the second fit, τ is assigned a $C^+(0, \sigma)$ distribution, allowing it to scale with the uncertain error variance.

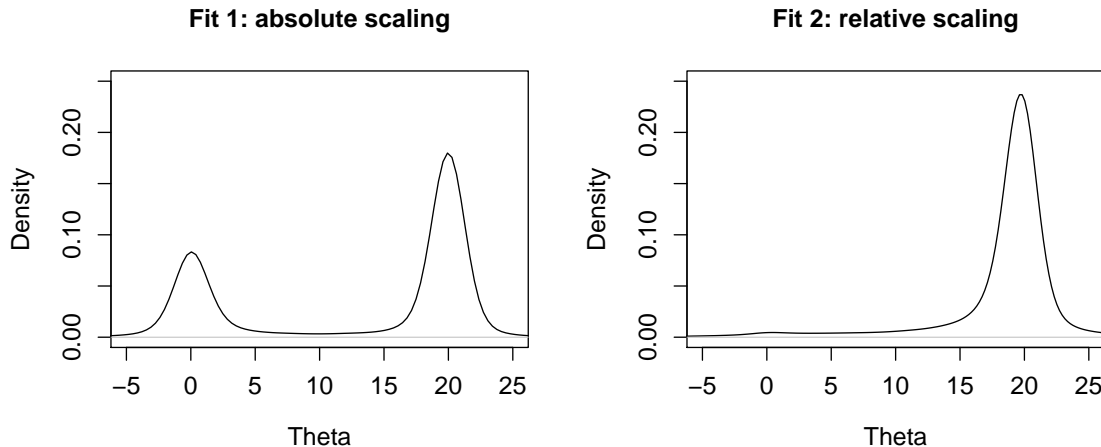


Figure 8: Example 1. Left: the posterior for θ when $\tau \sim C^+(0, 1)$. Right: the posterior when $\tau \sim C^+(0, \sigma)$.

The two posterior distributions for θ under these fits are shown in Figure 8. In the first fit using absolute scaling for τ , the posterior is bimodal, with one mode around 20 and the other around 0. This bimodality is absent in the second fit, where τ was allowed to scale relative to σ .

A situation with only two observations is highly stylized, to be sure, and yet the differences between the two fits are still striking. Note that the issue is not one of failing to condition on σ in the prior for τ ; indeed, the first fit involved plugging the true value of σ into the prior for τ , which is exactly what an empirical-Bayes analysis aims to accomplish asymptotically. Rather, the issue is one of averaging over uncertainty about σ in estimating the signal-to-noise ratio. Similar phenomena can be observed with other scale mixtures; see Fan and Berger (1992) for a general discussion of the issue.

6.2 A notable discrepancy under the double-exponential prior

Scott and Berger (2008) give a detailed comparison of Bayes and empirical-Bayes approaches for handling w , the prior inclusion probability, in the context of discrete mixture models for variable selection. Since in the horseshoe model, τ plays the role of w in exercising multiplicity control, an analogous set of issues surely arises here.

A full theoretical discussion of these issues is beyond the scope of this paper. Nonetheless, we include the following example as a warning of the fact that marginalizing over uncertainty in hyperparameters can drastically change the implied regularization penalty. Surprisingly, this difference between Bayesian and plug-in analyses may not disappear even in the limit.

Suppose that in the basic normal-means problem, $\theta_i = \mu + \tau\eta_i$, where $\eta_i \sim \text{DE}(2)$ has a double-exponential distribution. Hence

$$\pi(\theta \mid \mu, \tau) \propto \tau^{-p} \exp\left(-\frac{1}{\tau} \sum_{i=1}^p |\theta_i - \mu|\right),$$

leading to the joint distribution

$$p(\theta, y \mid \mu, \nu) \propto \nu^{-p} \exp\left\{-\frac{1}{2\sigma^2} \left(\sum_{i=1}^p (y_i - \theta_i)^2 + \nu^{-1} \sum_{i=1}^p |\theta_i - \mu|\right)\right\},$$

where ν is the regularization penalty (for known σ).

The plug-in solution is to estimate μ and ν by cross-validation or marginal maximum likelihood. Meanwhile, a reasonable fully Bayesian solution is to use the non-informative prior $\pi(\mu, \tau) \propto 1/\tau$. This yields a marginal prior distribution for θ of

$$\begin{aligned} \pi(\theta) &= \int \pi(\theta \mid \mu, \tau) \pi(\mu, \tau) \, d\mu \, d\tau \\ &\propto \exp\left\{-\frac{1}{2}Q(\theta)\right\}, \end{aligned}$$

where $Q(\theta)$ is piecewise linear and depends upon the order statistics $\theta_{(j)}$ (Uthoff, 1973). Specifically, define $v_j(\theta) \equiv v_j = \sum_{i=1}^p |\theta_{(i)} - \theta_{(j)}|$. Then

$$\pi(\theta) = (p-2)! 2^{-p+1} \sum_{j=1}^p w_j^{-1}, \quad (14)$$

where

$$w_j = \begin{cases} 4v_j^{p-1} (j - \frac{p}{2}) (\frac{p}{2} + 1 - j), & j \neq \frac{p}{2}, \frac{p}{2} + 1 \\ 4v_j^{p-1} [1 + (p-1) (\theta_{(p/2+1)} - \theta_{(p/2)}) v_j^{-1}], & j = \frac{p}{2}, \frac{p}{2} + 1 \end{cases}.$$

Hence the non-Bayesian estimates θ using

$$\pi_{EB}(\theta \mid y, \hat{\nu}, \hat{\mu}) \propto \exp\left\{-\frac{1}{2\sigma^2} \left(\sum_{i=1}^p (y_i - \theta_i)^2 + \hat{\nu}^{-1} \sum_{i=1}^p |\theta_i - \hat{\mu}|\right)\right\}, \quad (15)$$

while the Bayesian estimates θ using

$$\pi_{FB}(\theta \mid y) \propto \exp\left\{-\frac{1}{2\sigma^2} \left(\sum_{i=1}^p (y_i - \theta_i)^2\right) + \frac{(p-2)!}{2^{p-1}} \log\left(\sum_{i=1}^p [w_i(\theta)]^{-1}\right)\right\}. \quad (16)$$

The former is the traditional double-exponential prior, while the latter prior exhibits a rather complicated dependence upon the order statistics of the θ_i 's (which do not appear in the plug-in expression). It is by no means certain that the two procedures will reach similar answers asymptotically, since this difference in functional form persists for all p .

The double-exponential prior coupled with the noninformative prior on μ and τ is just one example where the marginalization in (14) is analytically tractable. But it serves to convey the essence of the problem, which is quite general. The Bayes and plug-in approaches for estimating τ imply fundamentally different regularization penalties for θ , regardless of whether θ is estimated by the mean or the mode, and regardless of whether marginal maximum likelihood or cross-validation is used.

Of course, neither prior is wrong *per se*, but the stark difference between (15) and (16) is interesting in its own right, and also calls into question the extent to which the plug-in analysis can approximate the fully Bayesian one. While some practitioners may have different goals for empirical Bayes or cross-validation, such comparison is at least reasonable. Many Bayesians use empirical-Bayes as a computational simplification, and many non-Bayesians appeal to complete-class theorems that rely upon an empirical-Bayes procedure's asymptotic correspondence with a fully Bayesian procedure. Hence questions about where the two approaches agree, and where they disagree, is of interest both to Bayesians and non-Bayesians.

While we do not wish to argue that plugging in point estimates for these parameters is necessarily wrong, we feel that in light of these issues, it should be done only with caution, or as an approximate analysis.

7 Final Remarks

We do not claim, of course, that the horseshoe is a panacea for sparse problems, merely a good default option. It is both surprising and interesting that its answers coincide so closely with the answers from the two-group gold standard of a Bayesian mixture model. Indeed, our results show an interesting duality between the two procedures. While the discrete mixture arrives at a good shrinkage rule by way of a multiple-testing procedure, the horseshoe estimator goes in the opposite direction, arriving at a good multiple-testing procedure by way of a shrinkage rule. Its combination of strong global shrinkage through τ , along with robust local adaptation to signals through the λ_i 's, is unmatched by other common Bayes rules using scale mixtures.

This paper studies sparsity in the simplified context where θ is a vector of normal means: $(y|\theta) \sim N(\theta, \sigma^2 I)$, where σ^2 may be unknown. It is here that the lessons drawn from a comparison of different approaches for modeling sparsity are most readily understood, but these lessons generalize straightforwardly to more difficult problems—regression, covariance regularization, function estimation—where many of the challenges of modern statistics lie. Multivariate scale mixtures of normals are often used as shrinkage priors in regression and basis expansion; see, for example, Park and

Casella (2008), Hans (2008), and Dunson (2008). In these situations, the behavioral similarities of the horseshoe and the Bayesian two-group model may be exploited even more profitably, especially in light of the massive computational savings offered by a one-group model.

References

- J. Angers and J. Berger. Robust hierarchical bayes estimation of exchangeable means. *The Canadian Journal of Statistics*, 19(1):39–56, 1991.
- J. O. Berger. A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics*, 8(4):716–761, 1980.
- D. Berry. Multiple comparisons, multiple tests, and data dredging: A Bayesian perspective. In J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, editors, *Bayesian Statistics 3*, pages 79–94. Oxford University Press, 1988.
- M. Bogdan, A. Chakrabarti, and J. K. Ghosh. Optimal rules for multiple testing and sparse multiple regression. Technical report, Purdue University, 2008a.
- M. Bogdan, J. K. Ghosh, and S. T. Tokdar. 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, pages 211–30. Institute of Mathematical Statistics, 2008b.
- L. Brown. Admissible estimators, recurrent diffusions and insoluble boundary problems. *The Annals of Mathematical Statistics*, 42:855–903, 1971.
- D. Denison and E. George. Bayesian prediction using adaptive ridge estimators. Technical report, Imperial College, London, 2000.
- D. B. Dunson. Kernel local partition processes for functional data. Technical report, Duke University Department of Statistical Science, 2008.
- B. Efron. Microarrays, empirical Bayes and the two-groups model (with discussion). *Statistical Science*, 1(23):1–22, 2008.
- T. Fan and J. O. Berger. Behaviour of the posterior distribution and inferences for a normal mean with t prior distributions. *Stat. Decisions*, 10:99–120, 1992.
- A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.*, 1(3):515–33, 2006.
- E. I. George and D. P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, 2000.

- M. B. Gordy. A generalization of generalized beta distributions. Technical report, Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series, 1998.
- I. Gradshteyn and I. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, 1965.
- J. Griffin and P. Brown. Alternative prior distributions for variable selection with very many more variables than observations. Technical report, University of Warwick, 2005.
- C. M. Hans. Bayesian lasso regression. Technical report, Ohio State University, 2008.
- I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: Empirical-Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.
- C. Masreliez. Approximate non-Gaussian filtering with linear state and observation relations. *IEEE. Trans. Autom. Control*, 1975.
- A. F. Mitchell. A note on posterior moments for a normal mean with double-exponential prior. *Journal of the Royal Statistical Society, Series B*, 56(4):605–10, 1994.
- T. Mitchell and J. Beauchamp. Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association*, 83:1023–36, 1988.
- T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–6, 2008.
- L. R. Pericchi and A. Smith. Exact and approximate posterior moments for a normal location parameter. *Journal of the Royal Statistical Society (Series B)*, 54(3):793–804, 1992.
- N. G. Polson. A representation of the posterior mean for a location model. *Biometrika*, 78:426–30, 1991.
- J. G. Scott and J. O. Berger. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7):2144–2162, 2006.
- J. G. Scott and J. O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Discussion Paper 2008-10, Duke University Department of Statistical Science, 2008.
- L. J. Slater. *Confluent Hypergeometric Functions*. Cambridge University Press, 1960.

- C. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–51, 1981.
- W. Strawderman. Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Statistics*, 42:385–8, 1971.
- G. C. Tiao and W. Tan. Bayesian analysis of random-effect models in the analysis of variance. i. Posterior distribution of variance components. *Biometrika*, 51:37–53, 1965.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–88, 1996.
- V. Uthoff. The most powerful scale and location invariant test of the normal versus the double exponential. *The Annals of Statistics*, 1(1):170–4, 1973.
- M. West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–8, 1987.

A Proofs

Proof of Theorem 1. Clearly,

$$\pi_H(\theta) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left\{\frac{-\theta^2}{2\lambda^2}\right\} \frac{2}{\pi(1+\lambda^2)} d\lambda.$$

Let $u = 1/\lambda^2$. Then

$$\pi_H(\theta) = K \int_0^\infty \frac{1}{1+u} \exp\left\{-\frac{\theta^2 u}{2}\right\} du,$$

or equivalently, for $z = 1 + u$:

$$\pi_H(\theta) = K e^{\theta^2/2} \int_1^\infty \frac{1}{z} e^{-z\theta^2/2} dz \tag{17}$$

$$= K e^{\theta^2/2} E_1(\theta^2/2), \tag{18}$$

where $E_1(\cdot)$ is the exponential integral function (closely related to the upper incomplete gamma function). This function satisfies very tight upper and lower bounds:

$$\frac{e^{-t}}{2} \log\left(1 + \frac{2}{t}\right) < E_1(t) < e^{-t} \log\left(1 + \frac{1}{t}\right)$$

for all $t > 0$, which proves Part (b). Part (a) then follows from the lower bound in Equation (1), which approaches ∞ as $\theta \rightarrow 0$. \square