

The Horseshoe Estimator for Sparse Signals

Carlos M. Carvalho

Nicholas G. Polson

Booth School of Business

University of Chicago

James G. Scott

McCombs School of Business

University of Texas at Austin

Submitted: October 2008

Revised: June 2009

Abstract

This paper proposes a new approach to sparsity 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. But the horseshoe enjoys a number of advantages over existing approaches, including its robustness, its adaptivity to different sparsity patterns, and its analytical tractability. We prove two theorems that formally characterize both the horseshoe's adeptness at large outlying signals, and its super-efficient rate of convergence to the correct estimate of the sampling density in sparse situations. Finally, using a combination of real and simulated data, 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 discrete mixture prior to model signals and noise.

Keywords: normal scale mixtures; ridge regression; shrinkage; sparsity.

1 Introduction

1.1 The proposed estimator

Suppose we observe a p -dimensional vector $(y|\theta) \sim N(\theta, \sigma^2 I)$, and that we wish both to estimate θ and to predict future realizations of y . Suppose further that θ is believed to be sparse, in the sense that many of its entries are zero, or nearly so. This setup, while simple, serves as a proving ground for methodology aimed at solving many of the challenges in modern statistics involving regression, classification, function estimation, covariance regularization, and others still.

We propose to estimate θ using the posterior mean under the following prior:

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

where $C^+(0, a)$ is a standard half-Cauchy distribution on the positive reals with scale parameter a . Crucially, each θ_i is mixed over its own λ_i , and each λ_i has an independent half-Cauchy prior with common, global scale τ .

We call this model the horseshoe prior. This name arises from the observation that

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

where $\kappa_i = 1/(1 + \lambda_i^2)$, assuming fixed values $\sigma^2 = \tau^2 = 1$. The half-Cauchy prior on λ_i yields a horseshoe-shaped $\text{Be}(1/2, 1/2)$ prior for the shrinkage coefficient κ_i . The left side of the horseshoe, $\kappa_i \approx 0$, yields virtually no shrinkage, and is meant to describe signals. The right side of the horseshoe, $\kappa_i \approx 1$, yields near-total shrinkage, and is meant to describe noise.

Unlike other similar procedures, the horseshoe prior is free of user-chosen hyperparameters. Nonetheless, it is both robust and highly adaptive; its excellent performance across a wide variety of situations suggests that the horseshoe is an attractive default choice among shrinkage priors.

This paper's goal, aside from introducing the horseshoe estimator, is to propose a theoretical framework under which the model can be compared with other similar shrinkage priors. This framework comprises two major issues:

Robustness to large signals: We will prove a new representation theorem that characterizes a prior's tail robustness in terms of the score function. This emphasizes the role of heavy-tailed priors in constructing robust, default estimators.

Shrinkage of noise: We will formally compare various estimators' asymptotic rates of convergence under the assumption that the true answer is sparse. This will

highlight the importance of the prior’s behavior near the origin.

Our procedure performs very strongly in light of both of these criteria. In sparse situations, the horseshoe prior will ensure that the Bayes estimator for the sampling density converges to the right answer at a super-efficient rate. Other common local shrinkage rules do not share this property. Yet when the true answer is far from zero, the horseshoe estimator exhibits a strong form of Bayesian robustness due to a redescending score function, and will leave the data unshrunk. This unique combination—super-efficiency when the real answer is sparse, robustness when the real answer is not sparse—proves to be quite powerful in discriminating signals from noise and in forming low-risk estimators.

1.2 The horseshoe density function

Assume fixed values of $\sigma^2 = \tau^2 = 1$. The univariate horseshoe density function lacks an analytic form, but very tight bounds are available.

Theorem 1. *When $\tau = 1$, the horseshoe density $p(\theta)$ satisfies the following:*

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

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

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

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

Proof. See the appendix. □

Upon integrating over τ , the marginal density for λ_i is

$$p(\lambda_i) = \frac{2}{\pi^2} \frac{\log \lambda_i^2}{\lambda_i^2 - 1},$$

though of course the terms are not independent once τ has been marginalized away. Indeed, the dependence structure induced by this marginalization will be difficult to visualize, making it easier to think in terms of p univariate conditional priors $p(\lambda_i | \tau)$ rather than a complex joint prior $p(\lambda_1, \dots, \lambda_p)$ over \mathbb{R}^p .

Figure 1 plots the density in (2) against the standard double-exponential and standard Cauchy densities. The horseshoe prior has heavy, Cauchy-like tails decaying like θ^{-2} , along with an infinitely tall spike at $\theta = 0$. These key features allow the prior to perform well in handling sparse vectors.

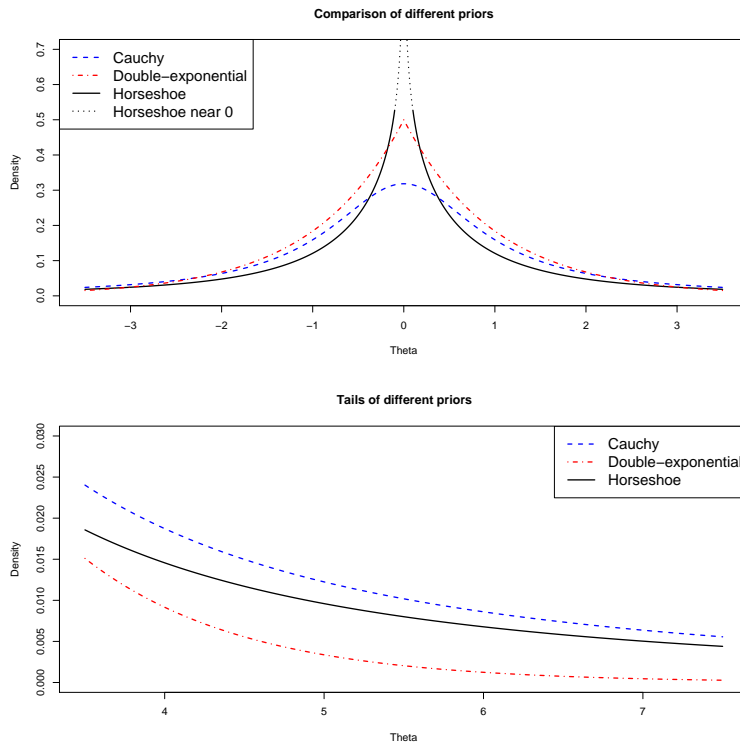


Figure 1: A comparison of p versus standard Cauchy and double-exponential densities; the dotted lines indicate that p approaches ∞ near 0.

1.3 Relationship with similar methods

The horseshoe prior assumes independent mixing densities upon p idiosyncratic scale terms λ_i , and is thus in the well known family of multivariate scale mixtures of normals. We call these “local shrinkage rules,” to distinguish them from global shrinkage rules that have only a shared global scale parameter τ .

The following list, though far from exhaustive, summarizes some other popular local shrinkage rules that have been considered in the literature.

1. The discrete mixture prior, $\theta_i \sim w \cdot g(\theta_i) + (1 - w) \cdot \delta_0$, can also be represented as a variance mixture, with

$$\lambda_i \sim w \cdot h(\lambda_i) + (1 - w) \cdot \delta_0$$

The choice of h will induce the form of the non-null density g . If, for example, h is a point mass at τ , then g is a $N(0, \tau^2)$ distribution. Scott and Berger (2006) study this prior extensively.

2. The Student- t prior, $\theta_i \sim t_\xi(0, \tau)$, is defined by an inverse-gamma mixing den-

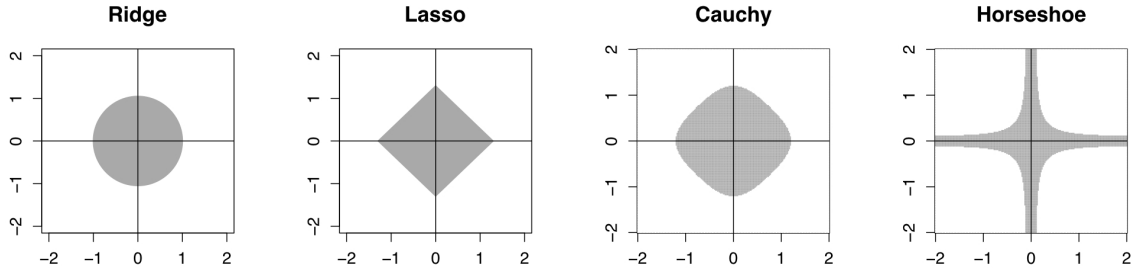


Figure 2: The implied constraint region of the horseshoe when viewed as a negative prior utility.

sity, $\lambda_i^2 \sim \text{IG}(\xi/2, \xi\tau^2/2)$. Tipping (2001) uses this model for sparsity by finding posterior modes under the assumption that $\xi \rightarrow 0$.

3. The double-exponential prior has undergone a recent surge in popularity:

$$p(\lambda_i^2 | \tau^2) = \frac{1}{2\tau^2} \exp\{\lambda_i^2/2\tau^2\}$$

$$\tau^2 \sim \text{IG}(\xi/2, \xi d^2/2).$$

The standard Markov-chain Monte Carlo algorithm for working with the double-exponential model is due to Carlin and Polson (1991), and the use of this model in robust Bayesian inference dates at least to Pericchi and Smith (1992). Theory for the wider class of powered-exponential priors appears in West (1987). More recently, Park and Casella (2008) and Hans (2008) have revitalized interest in this prior as a Bayesian alternative to the LASSO (Tibshirani, 1996).

4. The normal–Jeffreys prior has been studied by Figueiredo (2003) and Bae and Mallick (2004). This improper prior is induced by placing Jeffreys’ prior upon each variance term, $p(\lambda_i^2) \propto 1/\lambda_i^2$, leading to $p(\theta_i) \propto |\theta_i|^{-1}$ independently. This choice is commonly used in the absence of a global scale parameter, an issue that is carefully discussed in §3.1.
5. The Strawderman–Berger prior (Strawderman, 1971; Berger, 1980) lacks an analytic form, but arises from assuming $(\theta_i | \kappa_i) \sim N(0, \kappa_i^{-1} - 1)$, with $\kappa_i \sim \text{Be}(1/2, 1)$. Johnstone and Silverman (2004) call this the “quasi-Cauchy” density, and study it as a possible choice of g in the discrete mixture model. Denison and George (2000) also consider variations on this prior.
6. The normal–exponential–gamma family of priors proposed by Griffin and Brown (2005) is also based upon the exponential mixing density, but uses a $\text{Ga}(c, d^2)$

Table 1: Priors for λ_i and κ_i associated with some common local shrinkage rules. For the normal–exponential–gamma prior, it is assumed that $d = 1$. Densities are given up to constant terms.

Prior for θ_i	Density for λ_i	Density for κ_i
Double-exponential	$\lambda_i \exp\{\lambda_i^2/2\}$	$\kappa_i^{-2} e^{-\frac{1}{2\kappa_i}}$
Cauchy	$\lambda_i^{-2} \exp(-1/2\lambda_i^2)$	$\kappa_i^{-\frac{1}{2}} (1 - \kappa_i)^{-\frac{3}{2}} e^{-\frac{\kappa_i}{2(1-\kappa_i)}}$
Strawderman–Berger	$\lambda_i (1 + \lambda_i^2)^{-3/2}$	$\kappa_i^{-\frac{1}{2}}$
Normal–exponential–gamma	$\lambda_i (1 + \lambda_i^2)^{-(c+1)}$	κ_i^{c-1}
Normal–Jeffreys	$1/\lambda_i$	$\kappa_i^{-1} (1 - \kappa_i)^{-1}$
Horseshoe	$(1 + \lambda_i^2)^{-1}$	$\kappa_i^{-1/2} (1 - \kappa_i)^{-1/2}$

density rather than an inverse-gamma for the global scale term τ . The two hyperparameters allow control over tail weight (c) and scale (d). This leads to

$$p(\lambda_i^2) = \frac{c}{d^2} \left(1 + \frac{\lambda_i^2}{d^2}\right)^{-(c-1)}.$$

It is also common to view the negative log prior density as a penalty function, or equivalently as a constraint region of the form $\{\theta : -\log p(\theta) \leq C\}$ for some regularization parameter C . Figure 2 plots the horseshoe constraint region for $p = 2$, along with those of the ridge, lasso, and Cauchy estimators, to provide additional intuition into the prior’s behavior.

1.4 An intuitive basis for comparing shrinkage rules

Priors on shrinkage coefficients $\kappa_i = 1/(1 + \lambda_i^2)$ provide an intuitive way of understanding their associated Bayes rules, since $E(\theta_i | y_i) = \{1 - E(\kappa_i | y)\} y_i$. The behavior of $p(\kappa_i)$ near $\kappa_i = 0$ will control the tail robustness of the prior, and the behavior near $\kappa_i = 1$ will control the shrinkage of noise.

Table 1 lists the priors for λ_i and κ_i implied by six different local shrinkage rules. Figure 3 also plots these six priors on the κ scale, which helps to frame the more formal developments of the rest of the paper:

- The normal–Jeffreys and horseshoe priors both yield $p(\kappa_i)$ unbounded near 1,

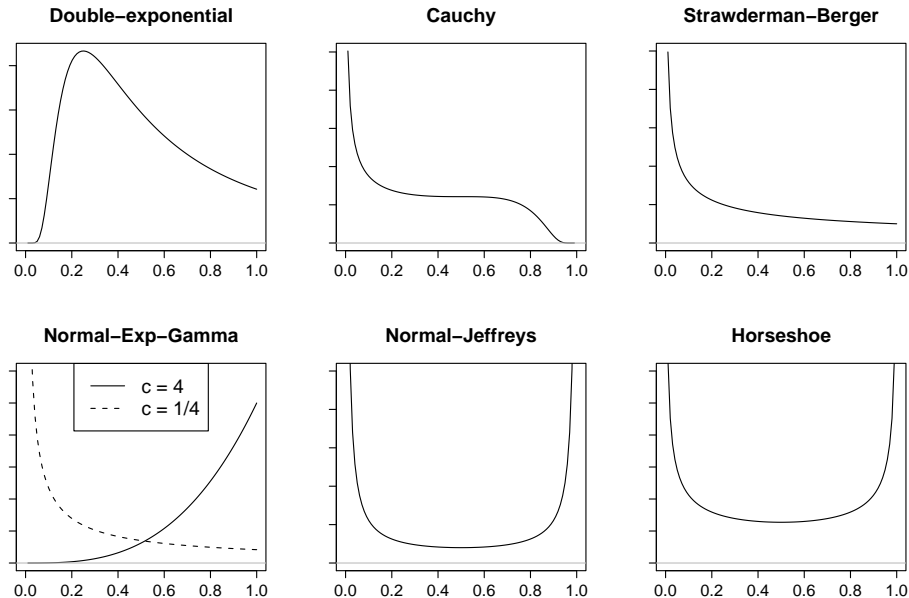


Figure 3: The implied densities $p(\kappa_i)$ for six different priors.

reflecting their infinite spikes at $\theta_i = 0$. The double-exponential, Strawderman–Berger, Cauchy, and normal–exponential–gamma priors all tend to fixed constants at $\kappa_i = 1$. These differences are highly significant for the behavior of the posterior mean when the true vector is sparse.

- The heavy-tailed priors—Cauchy, Strawderman–Berger, normal–Jeffreys, horseshoe, and normal–exponential–gamma with $c < 1$ —all yield $p(\kappa_i)$ unbounded near 0. The lighter-tailed priors, including the double-exponential and normal–exponential–gamma with $c \geq 1$, all cause $p(\kappa_i)$ to vanish at $\kappa_i = 0$. These differences affect the treatment of large, obvious signals.

The horseshoe prior is the only member of this group that exhibits tail robustness, unboundedness at the origin, and global adaptivity to different sparsity patterns. The first two of these features are a consequence of the prior behavior near $\kappa_i \approx 0$ and $\kappa_i \approx 1$, while the third is due to the presence of a global scale parameter.

Observe, however, that it is not enough for $p(\kappa_i)$ merely to diverge at 0 and 1; it must diverge sufficiently fast. Among priors of the form $\kappa_i \sim \text{Be}(a, b)$, the prior for λ_i is

$$p(\lambda_i) \propto \lambda_i^{2b-1} (1 + \lambda_i^2)^{-(a+b)},$$

which behaves like λ_i^{2b-1} near the origin, and like $\lambda_i^{-(2a+1)}$ in the upper tail. The horseshoe prior, which evaluates to $2/\pi$ at $\lambda_i = 0$, marks a phase transition between

two extremes. If $b < 1/2$, $p(\lambda_i)$ is unbounded at zero; if $b > 1/2$, $p(\lambda_i)$ vanishes at zero, and consequently $p(\theta_i)$ will be bounded.

One could also choose $\kappa_i \sim \text{Be}(\epsilon, \epsilon)$ for some $\epsilon < 1/2$. Compared to the horseshoe prior, this will lead to tails that are even heavier, and an infinite spike at $\theta = 0$ that is even more pronounced. Indeed, the normal–Jeffreys is the improper limiting case of this specification as $\epsilon \rightarrow 0$. Unfortunately, this makes for a very informative prior, since the prior for κ becomes concentrated ever closer near the extremes of 0 and 1. The horseshoe remains as noninformative as possible on the κ scale, placing 1/3 of its mass on $0.25 \leq \kappa_i \leq 0.75$.

2 Robustness to large signals

2.1 A representation of the posterior mean

Understanding tail-robustness is important in sparse settings, where one would like to shrink observations near zero much more forcefully than those far from zero. The following theorem can be used to characterize an estimator’s behavior in situations where y is very different from the prior mean.

Theorem 2. *Let $p(|y - \theta|)$ be the likelihood, and suppose that $p(\theta)$ is a mean-zero scale mixture of normals: $(\theta | \lambda) \sim N(0, \lambda^2)$, with λ having proper prior $p(\lambda)$. Assume further that the likelihood and $p(\theta)$ are such that the marginal density $m(y) < \infty$ for all y . Define the following three pseudo-densities, which may be improper:*

$$\begin{aligned} m^*(y) &= \int_{\mathbb{R}} p(|y - \theta|) p^*(\theta) d\theta \\ p^*(\theta) &= \int_{\mathbb{R}^+} p(\theta | \lambda) p^*(\lambda) d\lambda \\ p^*(\lambda) &= \lambda^2 p(\lambda). \end{aligned}$$

Then

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

Proof. See Appendix. □

If $p(|y - \theta|)$ is a normal likelihood, then (3) reduces to

$$E(\theta | y) = y + \frac{d}{dy} \log m(y). \tag{4}$$

Versions of (4) appear in Masreliez (1975), Polson (1991), and Pericchi and Smith (1992). But these results do not apply for the horseshoe prior, which fails to satisfy the common regularity condition that the density $p(\theta)$ is bounded. Theorem 2 relaxes this boundedness condition and extends the result to situations where $p(\theta)$ is a scale mixture of normals with proper mixing density and finite marginal $m(y)$.

The theorem provides a key insight about an estimator’s behavior in the presence of large signals: “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 (4) will lead to $E(\theta | y) \approx y$ for large $|y|$. This is precisely what happens under the horseshoe prior and others with sufficiently heavy tails, ensuring that large signals will not be overshrunk.

2.2 The horseshoe score function

Due to its heavy tails, the horseshoe prior is of bounded influence, leading to an estimator that is tail-robust.

Theorem 3. *Suppose $y \sim N(\theta, 1)$. Let $m(y)$ denote the predictive density under the horseshoe prior for known scale parameter $\tau < \infty$, i.e. where $(\theta | \lambda) \sim N(0, \tau^2 \lambda^2)$ and $\lambda \sim C^+(0, 1)$. Let $E(\theta | y)$ denote the posterior mean. Then $|y - E(\theta | y)| \leq b_\tau$ for some $b_\tau < \infty$ that depends upon τ , and*

$$\lim_{|y| \rightarrow \infty} \frac{d}{dy} \log m(y) = 0$$

Proof. See Appendix. □

The following corollary is immediate, and shows that the risk of the horseshoe estimator is bounded for all possible configurations of the true mean vector, whether sparse or not.

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

Proof.

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

□

Finally, although the horseshoe prior itself has no analytic form, it does yield an

expression for the posterior mean:

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

where $\Phi_1(\alpha, \beta, \gamma, x, y)$ is the degenerate hypergeometric function of two variables (Gradshteyn and Ryzhik, 1965, 9.261). Combining (5) with the marginal density in (9) allows an empirical-Bayes estimate $\mathbb{E}(\theta | y, \hat{\tau})$ to be computed very rapidly.

3 Efficiency in handling sparsity

3.1 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 each θ_i when τ is known. Usually, however, τ is unknown, leading to a joint distribution $p(y, \tau, \lambda_1, \dots, \lambda_p)$. Inspecting this joint distribution yields an understanding of how sparsity is handled under the global-local framework of 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, $p(\lambda_i) \propto 1/(1 + \lambda_i^2)$, and so

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

Some straightforward algebra leads to

$$p(y, \kappa, \tau^2) \propto p(\tau^2) \tau^p \prod_{i=1}^p \frac{e^{-\kappa_i y_i^2/2}}{\sqrt{1 - \kappa_i}} \prod_{i=1}^p \frac{1}{\tau^2 \kappa_i + 1 - \kappa_i}. \quad (6)$$

From (6) 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, (6) 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

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

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 $\mu/\sqrt{p-2}$.

Shared global parameters are of fundamental importance in high-dimensional inferences. This is the insight of Stein, and it applies regardless of whether sparsity is present. This fact is also central to the work of Johnstone and Silverman (2004) in the context of discrete mixtures, where a global parameter that characterizes sparsity in a data-adaptive way is crucial in bounding the risk of the resulting procedure.

Models that lack global parameters, or do not estimate them from the data, will not enjoy the benefits of this adaptivity. This issue is intimately related to the notion of multiplicity control in Bayesian hypothesis testing (Berry, 1988; Scott and Berger, 2006), where global parameters plays a central role in controlling the rate of Type-I errors. In fact, one way of viewing our procedure is that we are asking τ to play the role of w , the so-called ‘‘prior inclusion probability’’ in the discrete-mixture model. This highlights the importance of $p(\kappa_i)$: if κ_i is constrained by the prior from being very close to either 0 or 1, then the interpretation of $\bar{\kappa}$ as an average signal density breaks down, and τ will not be a faithful measure of underlying sparsity even if it is learned from the data.

3.2 Comparison with other Bayes rules

The advantages of the horseshoe prior are not shared by other common scale-mixture rules. Under the double-exponential prior, for example, small values of τ can also lead to strong shrinkage near the origin. This shrinkage, however, can severely compromise performance in the tails. Results from Pericchi and Smith (1992) and Mitchell (1994) show that the posterior mean $E(\theta_i | y_i) = w_i(y_i + b) + (1 - w_i)(y_i - b)$, where

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

and 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 $y + b$. This can be seen in the score function, plotted in Figure 4. Small values of τ may help to reduce risk at the origin, but do so at the expense of increased risk in the tails, since $|E(\theta_i | y_i) - y_i| \approx \sqrt{2}/\tau$ for large $|y_i|$.

Therefore, when θ is sparse, estimation of τ under the double-exponential model must balance two competing forces: risk due to under-shrinking noise, and risk due to over-shrinking large signals. This compromise is forced by the structure of the prior, and will be required under any model without tails sufficiently heavy to ensure

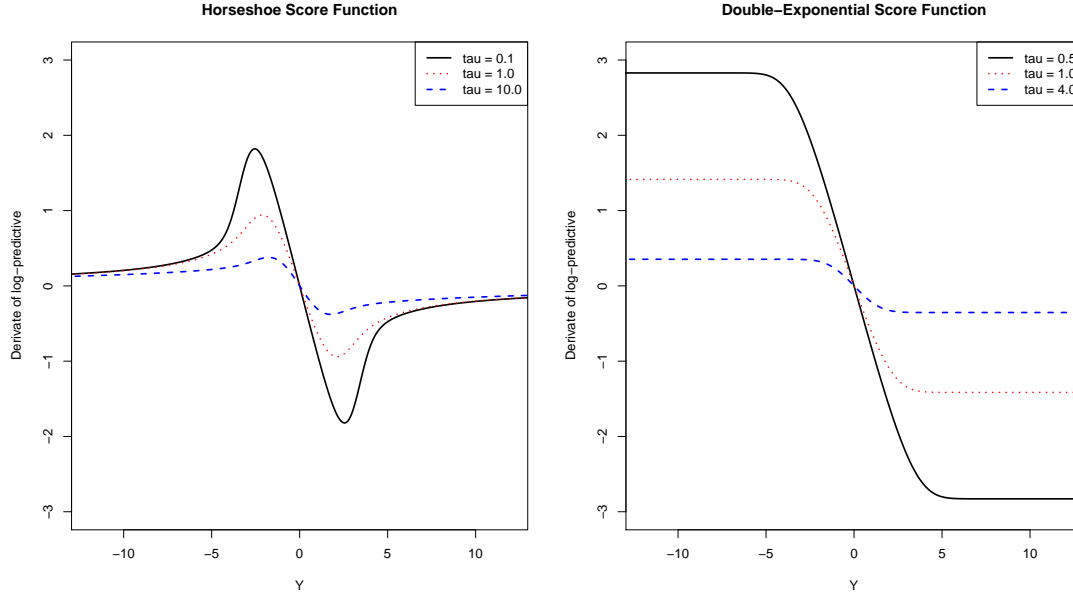


Figure 4: A comparison of the score function for horseshoe and double-exponential priors for different values of τ .

a redescending score function. As Figure 4 shows, the horseshoe prior requires no compromise of this sort.

To further illustrate the importance of this point, consider the following simple example. Two repeated standard normal observations y_{i1} and y_{i2} were simulated for each of 1000 means: 10 signals with $\theta_i = 10$, 90 signals $\theta_i = 2$ and 900 noise components where $\theta_i = 0$.

Figure 5 plots \bar{y}_i against $\hat{\theta}_i = E(\theta_i|y)$ under the horseshoe and double-exponential priors. The important differences occur when $\bar{y}_i \approx 0$ and when \bar{y}_i is large. Compared to the horseshoe prior, the double-exponential tends to over-shrink the large signals and yet under-shrink the noise observations. This is a direct effect of the prior on κ_i , which in the double-exponential case is bounded both at 0 and 1. These differences can also be seen in the bottom-left panel. The global shrinkage parameter τ is estimated to be much smaller under the horseshoe 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 bottom-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

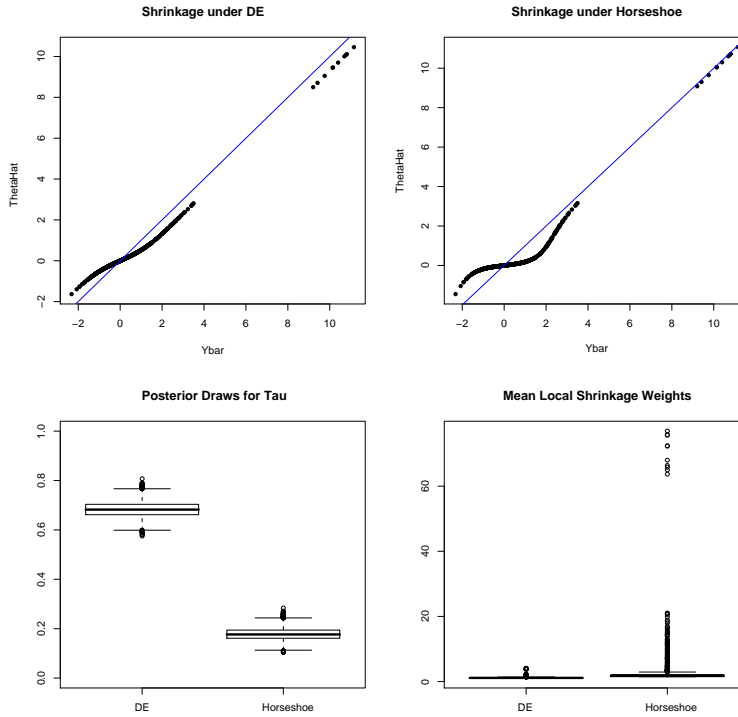


Figure 5: Plots of \bar{y}_i versus $\hat{\theta}_i$ for double-exponential (top left) and horseshoe (top right) priors. The diagonal lines are where $\hat{\theta}_i = \bar{y}_i$.

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.

Note that other local shrinkage priors with tails at least as heavy as the Cauchy will be similarly robust. This includes the Strawderman–Berger, the normal–Jeffreys, the normal–exponential–gamma with $c \leq 1/2$, and of course the Cauchy itself. Tails lighter than Cauchy but heavier than exponential may also be sufficient in practice, though we have not investigated this fully.

3.3 Kullback–Leibler risk bounds

We have argued at an intuitive level that the horseshoe is better at suppressing noise than many other scale-mixture priors. This intuition can be formalized by relating the behavior of the prior near the origin to its efficiency in handling sparsity.

The following theorem demonstrates that, when the true mean is zero, the horseshoe Bayes estimator for the sampling density converges to the right answer at a super-efficient rate compared to that of other common estimators. This efficiency is

measured using the Kullback–Leibler divergence between the true sampling model and the Bayes estimator of the density function. The theorem is proved for the univariate case, with convergence in the multivariate case following from a component-wise application of the results for a fixed value of τ .

A preliminary lemma is required. To avoid notational confusion between priors and sampling models, we use θ_0 to denote the true value of the parameter, $p_\theta = p(y \mid \theta)$ to denote a sampling model with parameter θ , and $\mu(A)$ to denote the prior or posterior measure of some set A . We also let $L(p_1, p_2) = \mathbb{E}_{p_1} \{\log(p_1/p_2)\}$ denote the Kullback–Leibler divergence of p_2 from p_1 .

Lemma 5. *Let $A_\epsilon = \{\theta : L(p_{\theta_0}, p_\theta) \leq \epsilon\} \subset \mathbb{R}$ denote the Kullback–Leibler information neighborhood of size ϵ , centered at θ_0 . Let $\mu_n(d\theta)$ be the posterior distribution under some prior measure $\mu(d\theta)$ after observing data $y_{(n)} = (y_1, \dots, y_n)$, and let $\hat{p}_n = \int p_\theta \mu_n(d\theta)$ be the posterior mean estimator of the density function.*

Suppose that the prior $\mu(d\theta)$ is information dense at p_{θ_0} , in the sense that $\mu(A_\epsilon) > 0$ for all $\epsilon > 0$. Then the following bound for R_n , the Cesàro-average risk of the Bayes estimator \hat{p}_n , holds for all $\epsilon > 0$:

$$R_n = \frac{1}{n} \sum_{j=1}^n L(p_{\theta_0}, \hat{p}_j) \leq \epsilon - \frac{1}{n} \mu(A_\epsilon).$$

The full proof of this lemma is given in Barron (1988). Intuitively, it follows from the fact that, for any θ , $\{n^{-1} \log(p_{\theta_0}/p_\theta)\} \rightarrow L(p_{\theta_0}, p_\theta)$ almost surely under p_{θ_0} , which allows the approximation

$$\frac{1}{n} \mathbb{E}_{p_{\theta_0}} \{\log(p_{\theta_0}/\hat{p}_n)\} \approx \frac{1}{n} \log \int \exp\{nL(p_{\theta_0}, p_\theta)\} \mu(d\theta).$$

This lemma can then be used to characterize the Kullback–Leibler risk in terms of $\mu(A_\epsilon)$, the amount of prior mass in a neighborhood of θ_0 . The horseshoe prior’s infinite spike at zero produces a super-efficient rate of convergence.

Theorem 6. *Suppose the true sampling model p_{θ_0} is $y_j \sim N(\theta_0, \sigma^2)$. Then:*

1. *For \hat{p}_n under the horseshoe prior, the optimal rate of convergence of R_n when $\theta_0 = 0$ is*

$$R_n = O\left(\frac{\log n - b \log \log n}{n}\right)$$

where b is a constant. When $\theta_0 \neq 0$, the optimal rate is

$$R_n = O\left(\frac{\log n}{n}\right).$$

2. *Suppose $p(\theta)$ is any other density that is continuous, bounded above, and strictly positive on a neighborhood of the true value θ_0 . For \hat{p}_n under $p(\theta)$, the optimal*

rate of convergence of R_n , regardless of θ_0 , is

$$R_n = O\left(\frac{\log n}{n}\right).$$

Proof. See appendix. □

Two further remarks are in order. First, the horseshoe estimator’s super-efficient rate occurs only on a set of prior measure zero. But this set is of special importance in sparse situations, since the hypothesis that some components of θ are zero has been explicitly flagged as an interesting possibility. And if $\theta_0 \neq 0$, then the horseshoe has no worse a convergence rate than any other common prior.

Second, this super-efficient rate of Kullback–Leibler convergence cannot be shared by any prior whose density function is bounded at the origin. Of course, priors with bounded density functions may exhibit large differences in the constant that multiplies the basic $O\left(\frac{\log n}{n}\right)$ rate, which can lead to substantial differences in performance on real problems.

3.4 Thresholding

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. 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 interesting correspondence between the two procedures.

Recall that under the discrete mixture model described in the introduction, the Bayes estimator for each θ_i is $\hat{\theta}_i = w_i E_g(\theta_i | y_i)$, where w_i is the posterior inclusion probability for θ_i , and g is the distribution of the non-zero means. For appropriately heavy-tailed g , 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 $\hat{\theta}_i \approx w_i y_i$ with excellent risk properties under squared-error and absolute-error loss.

The horseshoe estimator also yields weights $w_i = 1 - \kappa_i$, with $\hat{\theta}_i = w_i y_i$. Though these weights lack an interpretation as posterior probabilities, they turn out to behave similarly to those w_i 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.

Table 2: Posterior probabilities for 10 fixed signals as the number of noise observations grows, discrete-mixture model. FP: false positive declarations among noise observations.

Number of Noise	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

To test this thresholding rule, we fixed ten true signals at the half-integers between 0.5 and 5.0, and repeatedly applied the horseshoe thresholding rule in the face of an increasingly large number of standard-normal noise observations. We compared these results to those of the discrete-mixture rule using Strawderman–Berger priors. Results are shown in Tables 2 and 3.

These simulations 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 all situations we have investigated, it is hard to tell the difference between the weights w_i from the two-group model and those from the horseshoe. This can be seen from the Table, in which the horseshoe w_i are quite close to the corresponding posterior probabilities under the discrete-mixture prior across a wide-variety of sparsity configurations. Though the weights w_i under the double-exponential prior are not shown, we note that they do not behave at all like those from the discrete mixture model. These results, and many more simulations along these lines, can be found in Scott (2009).

We emphasize that this thresholding procedure is merely a heuristic, and lacks a firm probabilistic interpretation. Nonetheless, it seems to perform very well. Other possible thresholding rules, such as those based on posterior credible intervals, will be more applicable to non-orthogonal situations, but we do not investigate these here. Rather, our focus is on the posterior mean without thresholding, to which applies all of the previous theory developed in this section.

Table 3: Significance weights for 10 fixed signals as the number of noise observations grows, horseshoe prior. FP: false positive declarations among noise observations.

Number of Noise	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

4 Examples

4.1 Simulated data

Table 4 shows the results of a simulation study to assess the risk properties of the horseshoe prior. In this study, we benchmarked our model’s performance against four alternatives: the maximum-likelihood estimator, the double-exponential model, the normal–exponential–gamma model, and the empirical-Bayes mixture model described in Johnstone and Silverman (2004). This last approach uses a mixture of a point mass at zero with a double-exponential prior to differentiate signals from noise, and estimates θ_0 using the posterior median. This last comparison in particular is an important benchmark, as it is widely recognized as the gold standard in handling sparsity.

Our study involved simulating from the following sparse model:

$$\begin{aligned} (y_i | \theta_i) &\sim N(\theta_i, 1) \\ \theta_i &\sim w t_\xi(0, \tau) + (1 - w)\delta_0, \end{aligned}$$

where δ_0 is a point mass at zero, and where $t_\xi(0, \tau)$ is a Student- t density centered at zero, with ξ degrees of freedom and scale parameter τ .

In all our simulations, we set $\tau = 3$, and investigated six configurations of tail weight and sparsity by choosing $\xi \in \{2, 10\}$ and $w \in \{0.05, 0.2, 0.5\}$. These combinations span a wide range of behavior, from very sparse signals with very heavy tails, to mildly sparse signals with much lighter tails. For each combination we simulated 500 fake data sets.

When fitting the scale-mixture priors, we used Jeffreys’ prior for the variance, $p(\sigma^2) \propto 1/\sigma^2$. In the empirical-Bayes approach, σ and τ were estimated by marginal

Table 4: Results on simulated data, with w reflecting the degree of sparsity and ξ the tail weight of the density from which signals were drawn.

	$w = 0.05$		$w = 0.2$		$w = 0.5$	
	$\xi = 2$	$\xi = 10$	$\xi = 2$	$\xi = 10$	$\xi = 2$	$\xi = 10$
MLE	250	248	249	251	252	251
Double-exponential	171	127	237	217	247	235
NEG ($c = 4.0, d = 3$)	121	121	134	134	186	183
NEG ($c = 2.0, d = 3$)	165	164	170	171	187	187
NEG ($c = 1.0, d = 3$)	199	197	201	202	208	208
NEG ($c = 0.5, d = 3$)	219	217	220	222	227	225
Empirical-Bayes	32	38	111	129	417	442
NEG (best fixed c, d)	33	39	96	98	179	178
Horseshoe	32	33	94	95	178	244

maximum likelihood, along with the mixing weight w describing the probability of θ_i being a signal *a priori*.

The normal–exponential–gamma prior requires specifying two hyperparameters: c for tail weight, and d^2 for scale. To study the effect of these choices, we computed posterior means using a grid of values spanning $0.1 \leq d \leq 10$ and $1/2 \leq c \leq 8$. We report results for five of these choices in Table 4. Four of these choices involve fixing the scale hyperparameter d at 3 to reflect the known, true scale of the coefficients. The fifth result reported is the single best performer for each configuration of ξ and w , which could only be judged after the fact.

To summarize our results:

- The double-exponential prior systematically, and substantially, loses out to the horseshoe. Clearly this prior lacks tails that are heavy enough to estimate the largest signals in a robust fashion. It also lacks sufficient mass near 0 to adequately squelch the substantial amount of noise.
- The horseshoe prior systematically beats the default normal–exponential–gamma priors, and has a slight edge over the best fixed choice of c and d hyperparameters. Given the difficulty of eliciting these hyperparameters, we judge this to be a major advantage of the horseshoe prior as a default choice.
- Empirical-Bayes thresholding can do quite poorly in the signal-rich configurations, when $w = 0.5$.
- The horseshoe prior was beaten only in the situation when the signal was neither

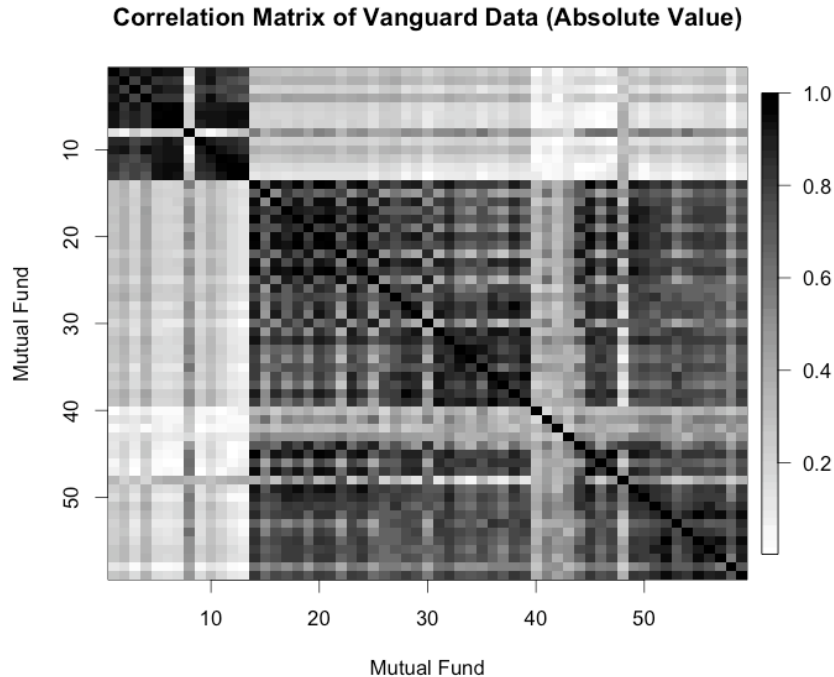


Figure 6: Empirical correlation matrix of Vanguard mutual-fund data.

sparse nor heavy-tailed, which was when $w = 0.5$ and $\xi = 10$. This is unsurprising, since the normal–exponential–gamma priors yield admissible estimators that seem especially well-suited to signals fitting this description.

The above results strongly support our claims that the horseshoe prior is indeed a good default choice for the estimation of sparse vectors.

4.2 Vanguard mutual-fund data

We now describe the use of the horseshoe prior in linear regression, with an example intended to show how the horseshoe can provide a regularized estimate of a large covariance matrix whose inverse may be sparse. As a test problem, we will use the data set on Vanguard mutual funds that appears in Carvalho and Scott (2009), which contains $n = 86$ weekly returns for $p = 59$ Vanguard mutual funds. Figure 6 shows the absolute value of the empirical correlation matrix.

The connection with regression is as follows. Suppose we observe a matrix of samples $Y' = (y^1 \cdots y^n)$, with each p -dimensional vector y^i drawn from a mean-zero normal distribution with unknown covariance matrix Σ . When p is large relative to n , traditional estimators of Σ are known to perform poorly, and some form of

Table 5: Covariance-estimation example. The table entries are risk ratios versus Bayesian model averaging in the out-of-sample prediction exercise. SE: squared-error loss. AE: absolute-error loss.

	MLE	Lasso AND	Lasso OR	Horseshoe	BMA
Risk ratio (SE)	10.63	1.25	2.12	1.07	1.00
Risk ratio (AE)	3.51	1.22	1.47	1.04	1.00

regularization is necessary to reduce their variance. We choose to model the Cholesky decomposition of Σ^{-1} and estimate the ensemble of regression models in the implied triangular system $\{Y_j \mid Y_1, \dots, Y_{j-1}\}_{j=2, \dots, p}$, where Y_j is the j^{th} column of the matrix of samples. Horseshoe priors are assumed for the vector of coefficients in each of these regressions, and posterior means were computed using Markov-chain Monte Carlo.

The intuition here is that some of these conditional regressions may be sparse, reflecting a joint distribution with a conditional independence, or Markov, structure. Such joint distributions are often called Gaussian graphical models.

We will compare the out-of-sample predictive performance of the horseshoe model against four different approaches for estimating Σ :

- the maximum-likelihood estimate $\hat{\Sigma} = Y'Y$.
- the AND and the OR versions of the LASSO, described by Meinshausen and Bühlmann (2006).
- Bayesian model-averaging over different Gaussian graphical models, using fractional Bayes factors for computing marginal likelihoods and fitted with the feature-inclusion stochastic search algorithm (Scott and Carvalho, 2008).

To assess out-of-sample performance, we used each of the above procedures to estimate Σ after observing the first 60 samples. We then attempted to impute random subsets of “missing” values among the remaining 26 samples, using the “non-missing” values as regressors. The full details of this exercise are given in Carvalho and Scott (2009). Both the data and relevant Matlab code are available from the authors upon request.

The results are in Table 5, and are expressed in terms of the error relative to the Bayesian model-averaging solution. It is clear that the horseshoe performs very closely to this benchmark, which is much more computationally intensive than any procedure based on local shrinkage rules. At the same time, the horseshoe significantly outperforms the classical LASSO solution, regardless of which version is used.

5 Final Remarks

The goal of this paper has not been to show that the horseshoe is a panacea for sparse problems—merely that it is a good default option. It is both surprising and interesting that its answers coincide so closely with the answers from the gold standard of a Bayesian discrete-mixture model. This lesson came through quite strongly in results both on simulated and real data.

Indeed, these results show an interesting duality between the two procedures. While the discrete mixture arrives at a good shrinkage rule by way of a procedure for sparsity, the horseshoe estimator goes in the opposite direction, arriving at a good procedure for sparsity 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.

Finally, a word on sparsity. Many similar procedures—most notably the LASSO—estimate θ using the posterior mode. This can cause some components of the estimated vector to be identically zero. Nonetheless, we prefer the posterior mean, and have chosen to study this rather than the mode. For one thing, the posterior mean is the Bayes estimator under quadratic loss, while the mode is the Bayes estimator under so-called “0–1” loss. In situations where estimation and prediction are the goals, the mean therefore embodies a loss function that is more likely to be closer to the true loss function—even though the mean itself is not sparse. Moreover, the insight of Bayesian model averaging is that different configurations of zeros in θ can always be treated as a nuisance parameter to be averaged over, and that averaging over models typically produces better results than selecting a single model. This marginalization over different sparsity patterns will produce an estimator for θ just like ours, in the sense that it will contain no entries that are exactly zero.

Under normal scale-mixture priors, using the mode is akin to selecting a model, while using the mean is akin to averaging over models—in particular, to averaging over the two peaks at 0 and 1 in the posterior distribution for each local shrinkage parameter κ_i . While the mean will lack zeros, the example of Bayesian model averaging demonstrates quite clearly that estimators of sparse objects need not be sparse themselves in order to yield excellent performance.

A Proofs

Proof of Theorem 1. Clearly,

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

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

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

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

$$p(\theta) = K e^{\theta^2/2} \int_1^\infty \frac{1}{z} e^{-z\theta^2/2} dz \tag{7}$$

$$= K e^{\theta^2/2} E_1(\theta^2/2), \tag{8}$$

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 (2), which approaches ∞ as $\theta \rightarrow 0$. \square

Proof of Theorem 2. Notice first that $m^*(y)$ exists by the case $p(\lambda^2) \equiv 1$, 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

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

from which the result follows directly. \square

Proof of Theorem 3. Clearly,

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

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

$$\begin{aligned} m(y) &= \frac{1}{\sqrt{2\pi^3}} \int_0^1 \exp(-zy^2/2) (1-z)^{-1/2} \left\{ \frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right) z \right\}^{-1} dz \\ &= \frac{2}{\tau\sqrt{2\pi^3}} \exp\left(-\frac{y^2}{2}\right) \Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}, \frac{y^2}{2}, 1 - \frac{1}{\tau^2}\right). \end{aligned} \quad (9)$$

By a similar transformation, it is easy to show that

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

Hence

$$\frac{d}{dy} \log m(y) = -\frac{2y \Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}, \frac{y^2}{2}, 1 - \frac{1}{\tau^2}\right)}{3 \Phi_1\left(\frac{1}{2}, 1, \frac{5}{2}, \frac{y^2}{2}, 1 - \frac{1}{\tau^2}\right)} \quad (10)$$

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 (10) by combining these identities yields a polynomial of order y^2 or greater remaining in the denominator, from which the redescending score function follows.

The bound $|y - \mathbb{E}(\theta \mid y)| \leq b_\tau$ then follows from the continuity of (9), which evaluates to 0 at $y = 0$. \square

Proof of Theorem 6. The optimal rate of convergence, following Barron (1988), comes from choosing $\epsilon_n = 1/n$, which reflects the ideal case of independent samples y_1, \dots, y_n .

First, note that for any prior $p(\theta)$ satisfying the stated regularity conditions in Part 2 of the theorem,

$$\mu(A_\epsilon) = \int_{A_\epsilon} p(\theta) d\theta \leq \int_{-\sqrt{\epsilon}}^{\sqrt{\epsilon}} p(\theta) d\theta = O(n^{-1/2})$$

since the density is bounded above. Applying Lemma 5, the optimal rate is given by

$$R_n \leq \frac{1}{n} - \frac{1}{n} \log\{Cn^{-1/2}\} = O\left(\frac{\log n}{n}\right),$$

proving Part 2.

Under the horseshoe prior, this same bound holds when $\theta_0 \neq 0$, since the horseshoe density function is bounded by a constant on a sufficiently small neighborhood near θ_0 . When $\theta_0 = 0$, we can use the bound on the density given previously, $2\sqrt{2\pi^3}p(\theta) \geq \log\left(1 + \frac{4}{\theta^2}\right)$. Ignoring constant factors not depending upon n , this leads to

$$\mu(A_\epsilon) \geq \int_0^{\sqrt{\epsilon}} \log\left(1 + \frac{4}{\theta^2}\right) d\theta.$$

Let $u = 1/\theta^2$. This yields

$$\mu(A_\epsilon) \geq \int_{4/\epsilon}^{\infty} \frac{\log(1+u)}{u^{3/2}} du.$$

Upon integrating by parts, we then have

$$\mu(A_\epsilon) \geq \epsilon^{1/2} \log\left(1 + \frac{4}{\epsilon}\right) + 2 \int_{4/\epsilon}^{\infty} \frac{1}{u^{1/2}(1+u)} du.$$

This last integral is easily computed and of order $\epsilon^{1/2}$. Setting $\epsilon = 1/n$ and applying Lemma 5 then gives the optimal rate bound

$$R_n = O\left(\frac{\log n - b \log \log n}{n}\right),$$

proving Part 1. \square

References

- K. Bae and B. Mallick. Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, 20(18):3423–30, 2004.
- A. R. Barron. The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Technical Report 7, University of Illinois at Urbana–Champaign, 1988.
- 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.
- B. P. Carlin and N. G. Polson. Inference for nonconjugate bayesian models using the gibbs sampler. *The Canadian Journal of Statistics*, 19(4):399–405, 1991.
- C. M. Carvalho and J. G. Scott. Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, 2009. to appear.
- D. Denison and E. George. Bayesian prediction using adaptive ridge estimators. Technical report, Imperial College, London, 2000.
- M. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1150–9, 2003.
- 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. 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.

- N. Meinshausen and P. Buhlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.
- 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. 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. *Bayesian Adjustment for Multiplicity*. PhD thesis, Duke University, 2009.
- 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 C. M. Carvalho. Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17(790–808), 2008.
- L. J. Slater. *Confluent Hypergeometric Functions*. Cambridge University Press, 1960.
- W. Strawderman. Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Statistics*, 42:385–8, 1971.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–88, 1996.
- M. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–44, 2001.
- M. West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–8, 1987.